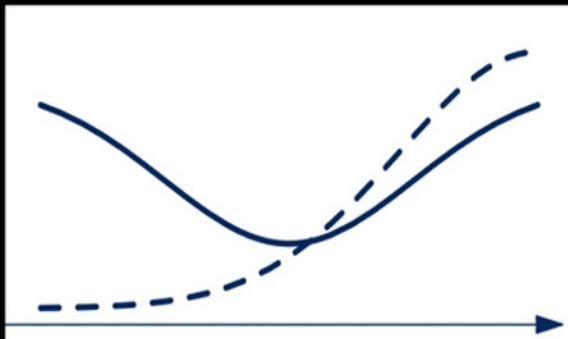
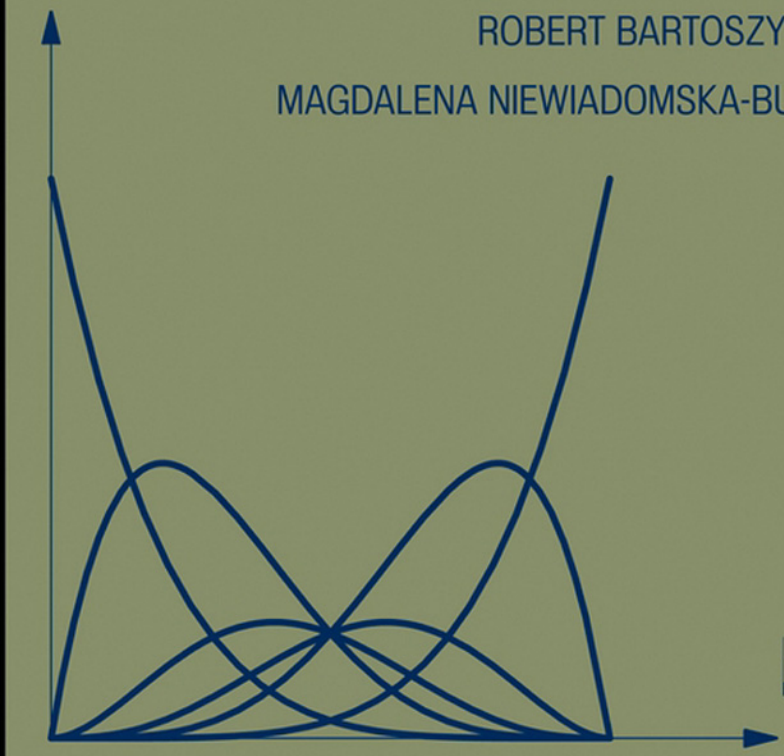


PROBABILITY AND STATISTICAL INFERENCE

Second Edition

ROBERT BARTOSZYŃSKI

MAGDALENA NIEWIADOMSKA-BUGAJ



PROBABILITY AND STATISTICAL INFERENCE

Second Edition

Robert Bartoszyński

Magdalena Niewiadomska-Bugaj

WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

This Page Intentionally Left Blank

PROBABILITY AND STATISTICAL INFERENCE

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

PROBABILITY AND STATISTICAL INFERENCE

Second Edition

Robert Bartoszyński

Magdalena Niewiadomska-Bugaj

WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 978-0-471-69693-3

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface

The first edition of this book was published in 1996. Since then, powerful computers have come into wide use, and it became clear that our text should be revised and material on computer-intensive methods of statistical inference should be added. To my delight, Steve Quigley, Executive Editor of John Wiley and Sons, agreed with the idea, and work on the second edition began.

Unfortunately, Robert Bartoszyński passed away in 1998, so I was left to carry out this revision by myself. I revised the content by creating a new chapter on random samples, adding sections on Monte Carlo methods, bootstrap estimators and tests, and permutation tests. More problems were added, and existing ones were reorganized. Hopefully nothing was lost of the "spirit" of the book which Robert liked so much and of which he was very proud.

This book is intended for seniors or first-year graduate students in statistics, mathematics, natural sciences, engineering, and any other major where an intensive exposure to statistics is necessary. The prerequisite is a calculus sequence that includes multivariate calculus. We provide the material for a two-semester course that starts with the necessary background in probability theory, followed by the theory of statistics.

What distinguishes our book from other texts is the way the material is presented and the aspects that are stressed. To put it succinctly, understanding "why" is prioritized over the skill of "how to." Today, in an era of undreamed-of computational facilities, a reflection in an attempt to understand is not a luxury but a necessity.

Probability theory and statistics are presented as self-contained conceptual structures. Their value as a means of description and inference about real-life situations

lies precisely in their level of abstraction—the more abstract a concept is, the wider is its applicability. The methodology of statistics comes out most clearly if it is introduced as an abstract system illustrated by a variety of real-life applications, not confined to any single domain.

Depending on the level of the course, the instructor can select topics and examples, both in the theory and in applications. These can range from simple illustrations of concepts, to introductions of whole theories typically not included in comparable textbooks (e.g., prediction, extrapolation, and filtration in time series as examples of use of the concepts of covariance and correlation). Such additional, more advanced, material (e.g., Chapter 5 on Markov Chains) is marked with asterisks. Other examples are: the proof of the extension theorem (Theorem 6.2.4), showing that the cumulative distribution function determines the measure on the line; the construction of Lebesgue, Riemann-Stieltjes and Lebesgue-Stieltjes integrals; and the explanation of the difference between double integral and iterated integrals (Section 8.3).

In the material that is seldom included in other textbooks on mathematical statistics, we stress the consequences of nonuniqueness of a sample space and illustrate, by examples, how the choice of a sample space can facilitate the formulation of some problems (e.g., issues of selection or randomized response). We introduce the concept of conditioning with respect to partition (Section 4.4); we explain the Borel-Kolmogorov paradox by way of the underlying measurement process that provides information on the occurrence of the condition (Example 7.22); we present the Neyman-Scott theory of outliers (Example 10.4); we give a new version of the proof of the relation between mean, median, and standard deviation (Theorem 8.7.3); we show another way of conditioning in the secretary problem (Example 4.10). Among examples of applications, we discuss the strategies of serves in tennis (Problem 4.2.12), and a series of problems (3.2.14-3.2.20) concerning combinatorial analysis of voting power. In Chapter 11 we discuss the renewal paradox, the effects of importance sampling (Example 11.6), and the relevance of measurement theory for statistics (Section 11.6). Chapter 14 provides a discussion of true regression versus linear regression and concentrates mostly on explaining why certain procedures (in regression analysis and ANOVA) work, rather than on computational details. In Chapter 15 we provide a taste of rank methods—one line of research starting with the Glivenko-Cantelli Theorem and leading to Kolmogorov-Smirnov tests, and the other line leading to Mann-Whitney and Wilcoxon tests. In this chapter we also show the traps associated with multiple tests of the same hypothesis (Example 15.3). Finally, Chapter 16 contains information on partitioning contingency tables—the method that provides insight into the dependence structure. We also introduce McNemar's test and various indices of association for tables with ordered categories.

The backbone of the book is the examples used to illustrate concepts, theorems, and methods. Some examples raise the possibilities of extensions and generalizations, and some simply point out the relevant subtleties.

Another feature that distinguishes our book from most other texts is the choice of problems. Our strategy was to integrate the knowledge students acquired thus far, rather than to train them in a single skill or concept. The solution to a problem in a given section may require using knowledge from some preceding sections, that is, reaching back into material already covered. Most of the problems are intended

to make the students aware of facts they might otherwise overlook. Many of the problems were inspired by our teaching experience and familiarity with students' typical errors and misconceptions.

Finally, we hope that our book will be "friendly" for students at all levels. We provide (hopefully) lucid and convincing explanations and motivations, pointing out the difficulties and pitfalls of arguments. We also do not want good students to be left alone. The material in starred chapters, sections, and examples can be skipped in the main part of the course, but used at will by interested students to complement and enhance their knowledge. The book can also be a useful reference, or source of examples and problems, for instructors who teach courses from other texts.

I am indebted to many people without whom this book would not have reached its current form. First, thank you to many colleagues who contributed to the first edition and whose names are listed there. Comments of many instructors and students who used the first edition were influential in this revision. I wish to express my gratitude to Samuel Kotz for referring me to Stigler's (1996) article about the "right and lawful rood," which we previously used in the book without reference (Example 8.40). My sincere thanks are due to Jung Chao Wang for his help in creating the R-code for computer-intensive procedures that, together with additional examples, can be found on the book's ftp site

ftp://ftp.wiley.com/public/sci_tech_med/probability_statistical/

Particular thanks are due to Katarzyna Bugaj for careful proofreading of the entire manuscript, Łukasz Bugaj for meticulously creating all figures with the Mathematica software, and Agata Bugaj for her help in compiling the index. Changing all those diapers has finally paid off.

I wish to express my appreciation to the anonymous reviewers for supporting the book and providing valuable suggestions, and to Steve Quigley, Executive Editor of John Wiley & Sons, for all his help and guidance in carrying out the revision.

Finally, I would like to thank my family, especially my husband Jerzy, for their encouragement and support during the years this book was being written.

Magdalena Niewiadomska-Bugaj

October 2007

CHAPTER 1

EXPERIMENTS, SAMPLE SPACES, AND EVENTS

1.1 INTRODUCTION

The consequences of making a decision today often depend on what will happen in the future, or at least on that limited part of the world and of the future that is relevant to the decision. The main purpose of using statistical methods is to help in making better decisions under uncertainty.

Judging from the failures of weather forecasts, to more spectacular prediction failures, such as bankruptcies of large companies and stock market crashes, it would appear that statistical methods do not perform very well. However, with a possible exception of weather forecasting, these examples are, at best, only partially statistical predictions. Moreover, failures tend to be better remembered than successes. Whatever the case, statistical methods are at present, and are likely to remain indefinitely, our best and most reliable prediction tools.

To analyze a given fragment of reality relevant for the specific purpose at hand, one usually needs to collect some *data*. Data may come from past experiences and observations, or may result from some controlled processes, such as laboratory or field experiments. The data are then used to hypothesize about the laws (often called *mechanisms*) that govern the fragment of reality of interest. In our book we are interested in laws expressed in probabilistic terms: They specify directly, or allow

us to compute, the chances of some events to occur. Knowledge of these chances is, in most cases, the best one can get with regard to prediction and decisions.

Probability theory is a domain of pure mathematics and as such, it has its own conceptual structure. To enable a variety of applications (typically comprising of all areas of human endeavor, ranging from biological, medical, social and physical sciences, to engineering, humanities, business, etc.), such structure must be kept on an abstract level. An application of probability to the particular situation analyzed requires a number of initial steps, in which the elements of the real situation are *interpreted* as abstract concepts of probability theory. Such interpretation is often referred to as building a *probabilistic model* of the situation at hand. How well this is done is crucial to the success of application.

One of the main concepts here is that of an *experiment*—a term used in a sense somewhat broader than usual. It means any process, possibly under partial control, that we may observe and whose behavior in the future is not totally determined because it is influenced, at least in part, by chance.

1.2 SAMPLE SPACE

In analyzing an experiment, one is primarily interested in its *outcome*—the concept that is not defined (i.e., *a primitive concept*) but has to be specified in every particular application. This specification may be done in different ways, with the only requirements being that (1) outcomes exclude one another and (2) they exhaust the set of all logical possibilities.

■ **EXAMPLE 1.1**

Consider an experiment consisting of two tosses of a regular die. An outcome is most naturally represented by a pair of numbers that turn up on the upper faces of the die so that they form a pair (x, y), with $x, y = 1, 2, \ldots, 6$ (see Table 1.1).

Table 1.1 Outcomes on a Pair of Dice

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

In the case of an experiment of tossing a die three times, the outcomes will be triplets (x, y, z), with x, y, and z being integers between 1 and 6.

Since the outcome of an experiment is not known in advance, it is important to determine the set of all possible outcomes. This set, called the *sample space*, forms the conceptual framework for all further considerations of probability.

Definition 1.2.1 The *sample space*, denoted by $\mathcal{S}$, is the set of all outcomes of an experiment. The elements of the sample space are called *elementary* outcomes, or *sample points*. □

■ **EXAMPLE 1.2**

In Example 1.1 the sample space $\mathcal{S}$ has $6^2 = 36$ sample points in the case of two tosses, and $6^3 = 216$ points in the case of three tosses of a die. The first statement can be verified by direct counting of the elements of the sample space. Similar verification of the second claim, although possible in principle, would be cumbersome. In Chapter 3 we will introduce some methods of determining the sizes of sets without actually counting sample points.

■ **EXAMPLE 1.3**

Suppose that the only available information about the numbers, those that turn up on the upper faces of the die, is their sum. In such a case as outcomes we take 11 possible values of the sum so that

$$\mathcal{S} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

For instance, all outcomes on the diagonal of Table 1.1—(6, 1), (5, 2), (4, 3), (3, 4), (2, 5), and (1, 6)—are represented by the same value 7.

■ **EXAMPLE 1.4**

If we are interested in the number of accidents that occur at a given intersection within a month, the sample space might be taken as the set $\mathcal{S} = \{0, 1, 2, \ldots\}$ consisting of all nonnegative integers. Realistically, there is a practical limit, say 1000, of the monthly numbers of accidents at this particular intersection. Although one may think that it is simpler to take the sample space $\mathcal{S} = \{0, 1, 2, \ldots, 1000\}$, it turns out that it is often much simpler to take the infinite sample space if the "practical bound" is not very precise.

Since outcomes can be specified in various ways (as illustrated by Examples 1.1 and 1.3), it follows that the same experiment can be described in terms of different sample spaces $\mathcal{S}$. The choice of a sample space depends on the goal of description. Moreover, certain sample spaces for the same experiment lead to easier and simpler analysis. The choice of a "better" sample space requires some skill, which is usually gained through experience. The following two examples illustrate this point.

■ **EXAMPLE 1.5**

Let the experiment consist of recording the lifetime of a piece of equipment, say a light bulb. An outcome here is the time until the bulb burns out. An

outcome typically will be represented by a number $t \geq 0$ ($t = 0$ if the bulb is not working at the start), and therefore $\mathcal{S}$ is the nonnegative part of the real axis. In practice, t is measured with some precision (in hours, days, etc.), so one might instead take $\mathcal{S} = \{0, 1, 2, \ldots\}$. Which of these choices is better depends on the type of subsequent analysis.

■ **EXAMPLE 1.6**

Two persons enter a cafeteria and sit at a square table, with one chair on each of its sides. Suppose we are interested in the event "they sit at a corner" (as opposed to sitting across from one another). To construct the sample space, we let A and B denote the two persons, and then take as $\mathcal{S}$ the set of outcomes represented by 12 ideograms in Figure 1.1.

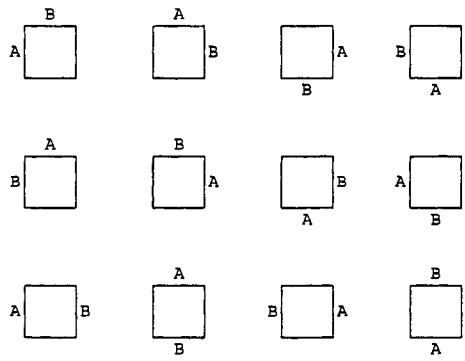


Figure 1.1 Possible seatings of persons A and B at a square table

One could argue, however, that such a sample space is unnecessarily large. If we are interested only in the event "they sit at a corner," then there is no need to label the persons as A and B. Accordingly the sample space $\mathcal{S}$ may be reduced to the set of six outcomes depicted in Figure 1.2.

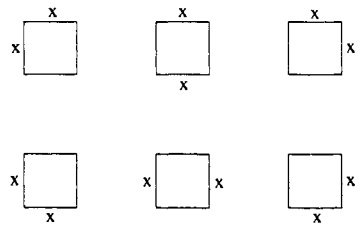


Figure 1.2 Possible seatings of any two persons at a square table

But even this sample space can be simplified. Indeed, one could use the rotational symmetry of the table and argue that once the first person selects

a chair (it does not matter which one), then the sample space consists of just three chairs remaining for the second person (see Figure 1.3).

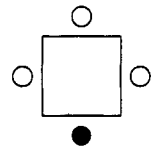


Figure 1.3 Possible seatings of one person if the place of the other person is fixed

Sample spaces can be classified according to the number of sample points they contain. *Finite* sample spaces contain finitely many outcomes, and elements of *infinitely countable* sample spaces can be arranged into an infinite sequence; other sample spaces are called *uncountable.*

The next concept to be introduced is that of an *event.* Intuitively, an event is anything about which we can tell whether or not it has occurred, as soon as we know the outcome of the experiment. This leads to the following definition:

Definition 1.2.2 An *event* is a subset of the sample space S. □

■ **EXAMPLE 1.7**

In Example 1.1, concerning two tosses of a die, an event such as "the sum equals 7" containing six outcomes $(1,6), (2,5), (3,4), (4,3), (5,2)$, and $(6,1)$ is a subset of the sample space S. In Example 1.3, the same event consists of one outcome, 7.

When an experiment is performed, we observe its outcome. In the interpretation developed in this chapter, this means that we observe a point chosen randomly from the sample space. If this point belongs to the subset representing the event A, we say that *the event A has occurred.*

We will let events be denoted either by letters $A, B, C, \ldots$, possibly with identifiers, such as $A_1, B_k, \ldots$, or by more descriptive means, such as $\{X = 1\}$ and $\{a < Z < b\}$, where X and Z are some numerical attributes of the sample points (formally: random variables, to be discussed in Chapter 6). Events can also be described through verbal phrases, such as "two heads in a row occur before the third tail" in the experiment of repeated tosses of a coin.

In all cases considered thus far, we assumed that an outcome (a point in the sample space) can be observed. To put it more precisely, all sample spaces S considered so far were constructed in such a way that their points were observable. Thus, for any event A, we were always able to tell whether it occurred or not.

The following examples show experiments and corresponding sample spaces with sample points that are only partially observable:

■ **EXAMPLE 1.8 Selection**

Candidates for a certain job are characterized by their level z of skills required for the job. The actual value of z is not observable, though; what we observe is the candidate's score x on a certain test. Thus, the sample point in $\mathcal{S}$ is a pair $s = (z, x)$, and only one coordinate of s, x, is observable.

The objective might be to find selection thresholds z_0 and x_0, such that the rule: "accept all candidates whose score x exceeds x_0" would lead to maximizing the (unobservable) number of persons accepted whose true level of skill z exceeds z_0. Naturally, to find such a solution, one needs to understand statistical relation between observable x and unobservable z.

Another example when the points in the sample space are only partially observable concerns studies of incidence of activities about which one may not wish to respond truthfully, or even to respond at all. These are typically studies related to sexual habits or preferences, abortion, law and tax violation, drug use, and so on.

■ **EXAMPLE 1.9 Randomized Response**

Let Q generally be the activity analyzed, and assume that the researcher is interested in the frequency of persons who ever participated in activity Q (for simplicity, we will call them Q-persons). It ought to be stressed that the objective is *not* to identify the Q-persons, but only to find the proportion of such persons in the population.

The direct question reduced to something like "Are you a Q-person?" is not likely to be answered truthfully, if at all. It is therefore necessary to make the respondent safe, guaranteeing that their responses will reveal nothing about them as regards Q. This can be accomplished as follows: The respondent is given a pair of distinguishable dice, for example, one green and one white. She throws them both at the same time, in such a way that the experimenter does not know the results of the toss (e.g., the dice are in a box and only the respondent looks into the box after it is shaken). The instruction is: If the green die shows an odd face (1, 3, or 5), then respond to the question "Are you a Q-person?" If the green die shows an even face (2, 4, or 6), then respond to the question "Does the white die show an ace?" The scheme of this response is summarized by the flowchart in Figure 1.4.

The interviewer knows the answer "yes" or "no" but does not know whether it is the answer to the question about Q or the question about the white die. Here a natural sample space consists of points $s = (i, x, y)$, where x and y are outcomes on green and white die, respectively, while i is 1 or 0 depending on whether or not the respondent is a Q-person. We have $\phi(s) = \phi(i, x, y)$= "yes" if $i = 1$ and $x = 1, 3$, or 5 for any y, or if $x = 2, 4, 6$, and $y = 1$ for any i. In all other cases $\phi(s) =$ "no."

One could wonder what is a possible advantage, if any, of not knowing the question asked and observing only the answer. This does not make sense if we need to know the truth about each individual respondent. However, one should remember that we are only after the overall frequency of Q-persons.

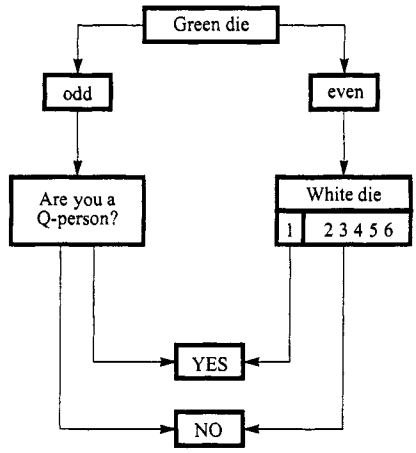


Figure 1.4 Scheme of a randomized response

We are in fact "contaminating" the question by making the respondent answer either a Q-question or some other auxiliary question. But this is a "controlled contamination": we know how often (on average) the respondents answer the auxiliary question, and how often their answer is "yes." Consequently, as we will see in Chapter 12, we can still make an inference about the proportion of Q-persons from the observed responses.

PROBLEMS

1.2.1 List all sample points in sample spaces for the following experiments: **(i)** We toss a coin. If heads come up, we toss a die. Otherwise, we toss the coin two more times. **(ii)** A coin is tossed until the total of two tails occurs, but no more than four times (i.e., a coin is tossed until the second tail or fourth toss, whichever comes first).

1.2.2 Alice, Bob, Carl, and Diana enter the elevator on the first floor of a four-story building. Each of them leaves the elevator on either the second, third, or fourth floor. **(i)** Find a simple way of describing the sample space (do not list all sample points). **(ii)** List all sample points such that Carl and Diana leave the elevator on the third floor. **(iii)** List all sample points if Carl and Diana leave the elevator at the same floor.

1.2.3 An urn contains five chips, labeled $1, \ldots, 5$. Three chips are drawn. List all outcomes included in the event "the second largest number drawn was 3."

1.2.4 In a game of craps, the player rolls a pair of dice. If he gets a total of 7 or 11, he wins at once; if the total is 2, 3, or 12, he loses at once. Otherwise, the sum, say x, is his "point," and he keeps rolling dice until either he rolls another x (in which case he wins) or he rolls a 7 in which case he loses. Describe the event "the player wins with a point of 5."

1.2.5 The experiment consists of placing six balls in three boxes. List all outcomes in the sample space if: **(i)** The balls are indistinguishable, but the boxes are distinguishable. (*Hint:* There are 28 different placements.) **(ii)** Neither the balls nor the boxes are distinguishable. **(iii)** Two balls are white and four are red; the boxes are distinguishable.

1.2.6 John and Mary plan to meet each other. Each of them is to arrive at the meeting place at some time between 5 p.m. and 6 p.m. John is to wait 20 minutes (or until 6 p.m., whichever comes first), and then leave if Mary does not show up. Mary will wait only 5 minutes (or until 6 p.m., whichever comes first), and then leave if John does not show up. Letting x and y denote the arrival times of John and Mary, determine the sample space and describe events (i)–(viii) by drawing pictures, or by appropriate inequalities for x and y. If you think that the description is impossible, say so. **(i)** John arrives before Mary does. **(ii)** John and Mary meet. **(iii)** Either Mary comes first or they do not meet. **(iv)** Mary comes first but they do not meet. **(v)** John comes very late. **(vi)** They arrive less than 15 minutes apart and they do not meet. **(vii)** Mary arrives at 5:15 p.m. and meets John, who is already there. **(viii)** They almost miss one another.

Problems 1.2.7–1.2.8 concern the possibility of expressing some events, depending on the choice of the sample space.

1.2.7 Let $\mathcal{E}$ be the experiment consisting of tossing a coin three times, with H and T standing for heads and tails, respectively.
(i) The following set of outcomes is an incomplete list of the points of the sample space $\mathcal{S}$ of the experiment $\mathcal{E}$: {HHH, HTT, TTT, HHT, TTH, HTH, THH}. Find the missing outcome.
(ii) An alternative sample space $\mathcal{S}'$ for the same experiment $\mathcal{E}$ consists of the following four outcomes: no heads (0), one head (1), two heads (2), and three heads (3). Which of the following events can be described as subsets of $\mathcal{S}$ but not as subsets of $\mathcal{S}' = \{0, 1, 2, 3\}$?

A_1 = More than two heads.
A_2 = Head on the second toss.
A_3 = More tails than heads.
A_4 = At least one tail, with head on the last toss.
A_5 = At least two faces the same.
A_6 = Head and tail alternate.

(iii) Still another sample space $\mathcal{S}''$ for the experiment $\mathcal{E}$ consists of the four outcomes $(0, 0), (0, 1), (1, 0)$, and $(1, 1)$. The first coordinate is 1 if the first two tosses show the same face and 0 otherwise; the second coordinate is 1 if the last two tosses show the same face, and 0 otherwise. For instance, if we observe HHT, the outcome is $(1, 0)$. List the outcomes of $\mathcal{S}$ that belong to the event $A = \{(1, 1), (0, 1)\}$ of $\mathcal{S}''$.
(iv) Which of the following events can be represented as subsets of $\mathcal{S}$, but cannot be represented as subsets of $\mathcal{S}''$?

B_1 = First and third tosses show the same face.

B_2 = Heads on all tosses.
B_3 = All faces the same.
B_4 = Each face appears at least once.
B_5 = More heads than tails.

1.2.8 Let $\mathcal{E}$ be the experiment consisting of tossing a die twice. Let $\mathcal{S}$ be the sample space with sample points $(i, j), i, j = 1, 2, \ldots 6$, with i and j being the numbers of dots that appear in the first and second toss, respectively.
(i) Let $\mathcal{S}'$ be the sample space for the experiment $\mathcal{E}$ consisting of all possible sums $i + j$ so that $\mathcal{S}' = \{2, 3, \ldots, 12\}$. Which of the following events can be defined as subsets of $\mathcal{S}$ but not of $\mathcal{S}'$?

A_1 = One face odd, the other even.
A_2 = Both faces even.
A_3 = Faces different.
A_4 = Result on the first toss less than the result on the second.
A_5 = Product greater than 10.
A_6 = Product greater than 30.

(ii) Let $\mathcal{S}''$ be the sample space for the experiment $\mathcal{E}$ consisting of all possible absolute values of the difference $|i - j|$ so that $\mathcal{S}'' = \{0, 1, 2, 3, 4, 5\}$. Which of the following events can be defined as subsets of $\mathcal{S}$ but not of $\mathcal{S}''$?

B_1 = One face shows twice as many dots as the other,
B_2 = Faces the same,
B_3 = One face shows six times as many dots as the other,
B_4 = One face odd, the other even,
B_5 = The ratio of the numbers of dots on the faces is different from 1.

1.2.9 Referring to Example 1.9, suppose that we modify it as follows: The respondent tosses a green die (with the outcome unknown to the interviewer). If the outcome is odd, he responds to the Q-question; otherwise, he responds to the question "Were you born in April?" Again, the interviewer observes only the answer "yes" or "no." Apart from the obvious difference in frequency of the answer "yes" to the auxiliary question (on the average one in 12 instead of one in 6), are there any essential differences between this scheme and the scheme in Example 1.9? Explain your answer.

1.3 ALGEBRA OF EVENTS

Next we introduce some concepts that will allow us to form composite events out of simpler ones. We begin with the relations of *inclusion* and *equality*.

Definition 1.3.1 The event A is *contained* in the event B, or B *contains* A, if every sample point of A is also a sample point of B. Whenever this is true, we will write $A \subset B$, or equivalently, $B \supset A$. □

An alternative terminology here is that A *implies* (or *entails*) B.

Definition 1.3.2 Two events A and B are said to be *equal*, $A = B$, if $A \subset B$ and $B \subset A$. □

It follows that two events are equal if they consist of exactly the same sample points.

■ **EXAMPLE 1.10**

Consider two tosses of a coin, and the corresponding sample space $\mathcal{S}$ consisting of four outcomes: HH, HT, TH, and TT. The event A = "heads in the first toss" = {HH, HT} is contained in the event B = "at least one head" = {HH, HT, TH}. The events "the results alternate" and "at least one head and one tail" imply one another, and hence are equal.

Definition 1.3.3 The set containing no elements is called the *empty* set and is denoted by $\emptyset$. The event corresponding to $\emptyset$ is called a *null* (*impossible*) event. □

■ **EXAMPLE 1.11** *

[1] The reader may wonder whether it is correct to use the definite article in the definition above and speak of "*the* empty set," since it would appear that there may be many different empty sets. For instance, the set of all kings of the United States and the set of all real numbers x such that $x^2 + 1 = 0$ are both empty, but one consists of people and the other of numbers, so they cannot be equal. This is not so, however, as is shown by the following formal argument (to appreciate this argument, one needs some training in logic). Suppose that $\emptyset_1$ and $\emptyset_2$ are two empty sets. To prove that they are equal, one needs to prove that $\emptyset_1 \subset \emptyset_2$ and $\emptyset_2 \subset \emptyset_1$. Formally, the first inclusion is the implication: "if s belongs to $\emptyset_1$, then s belongs to $\emptyset_2$." This implication is true, because its premise is false: there is no s that belongs to $\emptyset_1$. The same holds for the second implication, so $\emptyset_1 = \emptyset_2$.

We now give the definitions of three principal operations on events: *complementation*, *union*, and *intersection*.

Definition 1.3.4 The set that contains all sample points that are not in the event A will be called the *complement* of A and denoted A^c, to be read also as "not A." □

Definition 1.3.5 The set that contains all sample points belonging either to A or to B (so possibly to both of them) is called the *union* of A and B and denoted $A \cup B$, to be read as "A or B." □

Definition 1.3.6 The set that contains all sample points belonging to both A and B is called the *intersection* of A and B, and denoted $A \cap B$, to be read as "A and B." □

An alternative notation for a complement is A' or $\overline{A}$, whereas in the case of an intersection one often writes AB instead of $A \cap B$.

[1] Asterisks denote more advanced material, as explained in the Preface.

The operations above have the following interpretations in terms of occurrences of events:

1. Event A^c occurs if event A does not occur.

2. Event $A \cup B$ occurs when either A or B or both events occur.

3. Event $A \cap B$ occurs when both A and B occur.

■ **EXAMPLE 1.12**

Consider the experiment of tossing a coin three times, with the sample space consisting of outcomes described as HHH, HHT, and so on. Let A be the event "heads and tails alternate," and let B be "heads on the last toss." The event A^c occurs if either heads or tails occur at least twice in a row so that $A^c = \{\text{HHH, HHT, THH, HTT, TTT, TTH}\}$ while B^c is "tails on the last toss," hence $B^c = \{\text{HHT, THT, HTT, TTT}\}$. The union $A \cup B$ is the event "either the results alternate or it is heads on the last toss," meaning $A \cup B = \{\text{HTH, THT, HHH, THH, TTH}\}$. Observe that while A has two outcomes and B has four outcomes, their union has only five outcomes, since the outcome HTH appears in both events. This common part is the intersection $A \cap B$.

Some formulas can be simplified by introducing the operation of the *difference* of two events.

Definition 1.3.7 The *difference* $A \setminus B$ of events A and B contains all sample points that belong to A but not to B

$$A \setminus B = A \cap B^c.$$

The *symmetric difference*, $A \div B$, contains sample points that belong to A or to B, but not to both of them:

$$A \div B = (A \cap B^c) \cup (A^c \cap B) = (A \cup B) \setminus (A \cap B). \qquad \square$$

■ **EXAMPLE 1.13**

In Example 1.12, the difference $B^c \setminus A$ is described as "at least two identical outcomes in a row and tails on the last toss," which means the event {HHT, HTT, TTT}.

Next we have the following important concept:

Definition 1.3.8 If $A \cap B = \emptyset$, then the events A and B are called *disjoint*, or *mutually exclusive*. $\square$

■ **EXAMPLE 1.14**

Based on Example 1.12 we know that the following two events are disjoint: C = "more heads than tails" and the intersection $A \cap B^c$ = "the results alternate, ending with tails."

Example 1.14 shows that to determine whether or not events are disjoint, it is not necessary to list the outcomes in both events and check whether there exist common outcomes. Apart from the fact that such listing is not feasible when sample spaces are large, it is often simpler to employ some logical reasoning, for instance, that one of the events is contained in the complement of the other (i.e., if one of them occurs, the other does not). In the case above, if the results alternate and end with tails, then the outcome must be THT. Since there are more tails than heads, C does not occur.

The definitions of union and intersection can be extended to the case of a finite and even infinite number of events (discussed in the Section 1.4). Thus

$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_i \tag{1.1}$$

is the event that contains the sample points belonging to A_1 or A_2 or ...or A_n. Consequently, (1.1) is the event " at least one A_i occurs." Similarly

$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n} A_i \tag{1.2}$$

is the event that contains the sample points belonging to A_1 and A_2 and ...and A_n. Consequently, the event (1.2) is " all A_i's occur."

■ **EXAMPLE 1.15**

Suppose that n shots are fired at a target, and let A_i, $i = 1, 2, \ldots, n$ denote the event "the target is hit on the ith shot." Then the union $A_1 \cup \cdots \cup A_n$ is the event "the target is hit" (at least once). Its complement $(A_1 \cup \cdots \cup A_n)^c$ is the event "the target is missed" (on every shot), which is the same as the intersection $A_1^c \cap \cdots \cap A_n^c$.

A perceptive reader may note that the unions $A_1 \cup \cdots \cup A_n$ and intersections $A_1 \cap \cdots \cap A_n$ do not require an extension of the definition of union and intersection for two events. Indeed, we could consider unions such as

$$A_1 \cup (A_2 \cup (\cdots (A_{n-2} \cup (A_{n-1} \cup A_n)) \cdots)),$$

where the union of only two events is formed in each set of parentheses. The property of associativity (below) shows that parentheses can be omitted so that the expression $A_1 \cup \cdots \cup A_n$ is unambiguous. The same argument applies to intersections.

The operations on events defined in this section obey some laws. The most important ones are listed below.

Idempotence:

$$A \cup A = A, \quad A \cap A = A.$$

Double Complementation:

$$(A^c)^c = A.$$

Absorption:

$$A \cup B = B \text{ iff } A \cap B = A \text{ iff } A \subset B. \tag{1.3}$$

In particular,

$$A \cup \emptyset = A, \quad A \cup \mathcal{S} = \mathcal{S}, \quad A \cap \emptyset = \emptyset, \quad A \cap \mathcal{S} = A,$$

which in view of (1.3) means that $\emptyset \subset A \subset \mathcal{S}$.

Commutativity:

$$A \cup B = B \cup A, \qquad A \cap B = B \cap A.$$

Associativity:

$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C.$$

Distributivity:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

De Morgan's Laws:

$$\begin{aligned} (A_1 \cup \cdots \cup A_n)^c &= A_1^c \cap \cdots \cap A_n^c, \\ (A_1 \cap \cdots \cap A_n)^c &= A_1^c \cup \cdots \cup A_n^c. \end{aligned} \tag{1.4}$$

When studying mutual relations between composite events in the same sample space, it is often helpful to use Venn diagrams, where the sample space $\mathcal{S}$ is represented by a rectangle, while its subsets represent events (see Figure 1.5).

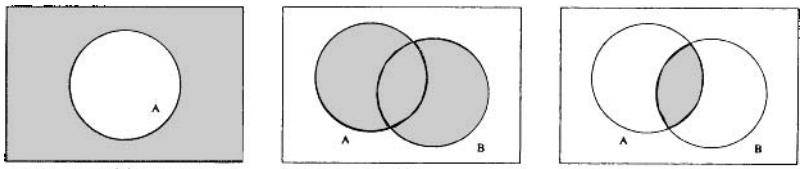


Figure 1.5 Complement, union and intersection

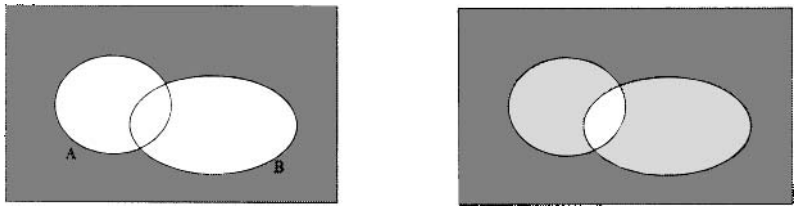


Figure 1.6 The first De Morgan's law

The complement of event A is represented in Figure 1.5(a), the union and intersection of the events A and B are represented in Figure 1.5(b) and (c), respectively.

Venn diagrams can also be used to check the validity of formulas. For example, consider the first De Morgan's law (1.4) for the case of two events:

$$(A \cup B)^c = A^c \cap B^c. \tag{1.5}$$

Venn diagrams made separately for the left-hand side and the right-hand side of (1.5) (see Figure 1.6) indicate that both regions are the same. Although a picture does not constitute a proof, it may provide convincing evidence that the statement is true, and sometimes may even suggest a method of proving the statement.

PROBLEMS

For the problems below, remember that a statement (expressed as a sentence or formula) is true if it is true under *all* circumstances, and it is false if there is at least one case where it does not hold.

1.3.1 Answer true or false. Justify your answer.
(i) If A and B are distinct events (i.e., $A \neq B$) such that A and B^c are disjoint, then A^c and B are also disjoint.
(ii) If A and B are disjoint, then A^c and B^c are also disjoint.
(iii) If A and B are disjoint, and also B and C are disjoint, then A and C are disjoint.
(iv) If A and B are both contained in C, then $C^c \subset A^c \cap B^c$.
(v) If A is contained in B, C is contained in D, and B is disjoint from D, then A is disjoint from C.
(vi) If $A \cup B^c = B^c$, then $B \subset A^c$.

1.3.2 In the statements below A, B, C and D, are events. Find those statements or formulas that are true.
(i) If $A \cap B = A \cap C$ then $B = C$.
(ii) $A \cup (A \cap B) \cup (A \cap B^c) = A$.
(iii) $A \cup (A \cap B) \cup (A \cap B^c) = B$.
(iv) If $A \setminus B = C$, then $A = B \cup C$.
(v) $(A \cup B) \cap (C \cup D) = (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D)$.
(vi) $(A \cap B) \cup (C \cap D) = (A \cup C) \cap (A \cup D) \cap (B \cup C) \cap (B \cup D)$.

(vii) $(A^c \cup B^c \cup C^c)^c = A^c \cap B^c \cap C^c$.
(viii) If $A \subset B$, and $B \cap C = \emptyset$, then $C^c \cap A \cap B^c = \emptyset$.
(ix) If $A \cap B$, $A \cap C$ and $B \cap C$ are not empty, then $A \cap B \cap C$ is not empty.
(x) Show that $(A \div B) \div C = A \div (B \div C)$.

1.3.3 Find X if: **(i)** $A \div X = \emptyset$. **(ii)** $A \div X = A$. **(iii)** $A \div X = S$. **(iv)** $A \div X = B$.

1.3.4 In a group of 1000 students of a certain college, 60 take French, 417 take calculus, and 509 take statistics. Moreover, 20 take French and calculus, 17 take French and statistics, and 147 take statistics and calculus. However, 196 students do not take any of these three subjects. Determine the number of students who take French, calculus, and statistics.

1.3.5 Let A, B, and C be three events. Match, where possible, events D_1 through D_{10} with events E_1 through E_{11}. Matching means that the events are exactly the same; that is, if one occurs, so must the other and conversely (see the Definition 1.3.2). (*Hint:* Draw a Venn diagram for each event $D_1, \ldots, D_{10}$, do the same for events $E_1, \ldots, E_{11}$, and then compare the diagrams.)

Among events A, B, C:
D_1 = two or more occur.
D_2 = exactly one occurs.
D_3 = only A occurs.
D_4 = all occur.
D_5 = none occurs.
D_6 = at most one occurs.
D_7 = at least one occurs.
D_8 = exactly two occur.
D_9 = no more than two occur.
$D_{10} = B$ occurs.

$E_1 = A \cup B \cup C$.
$E_2 = (A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C)$.
$E_3 = (A \cap B)^c \cap (A \cap C)^c \cap (B \cap C)^c$,
$E_4 = (A \cup B \cup C)^c$.
$E_5 = A^c \cap B^c \cap C^c$.
$E_6 = A \cap B \cap C$.
$E_7 = B$.
$E_8 = A \cap B^c \cap C^c$.
$E_9 = (A \cap B \cap C^c) \cup (A \cap B^c \cap C) \cup (A^c \cap B \cap C)$.
$E_{10} = (A \cap B \cap C)^c$.
$E_{11} = (A \cap B) \cup (A \cap C) \cup (B \cap C)$.

1.3.6 A standard deck of cards is dealt among players N, S, E, and W. Let N_k, $k = 1, 2, 3, 4$ be the event "N has at least k aces," and let S_k, E_k and W_k be defined similarly. For each of the events below, determine the number of aces that N has. **(i)** $N_1 \cap S_1 \cap E_1 \cap W_1$. **(ii)** $E_2 \cap (W_2 \cup S_2)$. **(iii)** $N_3 \setminus N_4$. **(iv)** $S_3 \cap W_1$. **(v)** $S_1^c \cap W_1^c \cap E_1^c$. **(vi)** $N_2 \cap E_2$.

1.3.7 Five burglars, A, B, C, D, and E, divide the loot, consisting of 5 identical gold bars and 4 identical diamonds. Let A_{jk} be the event that A got *at least* j gold bars and *at most* k diamonds. Let B_{jk}, C_{jk} denote analogous events for burglars B, C (e.g., B_{21} is the event that B got 2, 3, 4, or 5 gold bars and 0 or 1 diamond). Determine the number x of gold bars and the number y of diamonds received by E if the following events occur (if determination of x and/or y is impossible, give the range of values): **(i)** $(A_{20} \cup B_{20} \cup C_{20}) \cap D_{30}$. **(ii)** E_{12}^c. **(iii)** $A_{23} \cap B_{13} \cap C_{13} \cap D_{13}$. **(iv)** $A_{23} \cup B_{13} \cup C_{13} \cup D_{13}$.

1.3.8 Let A^{nc} be defined inductively by $A^{0c} = A$, $A^{(n+1)c} = (A^{nc})^c$. Find $A^{mc} \cap A^{nc}$ and $A^{mc} \cup A^{nc}$ for $m, n > 0$.

1.4 INFINITE OPERATIONS ON EVENTS

As already mentioned, the operations of union and intersection can be extended to infinitely many events. Let $A_1, A_2, \ldots$ be an infinite sequence of events. Then

$$A_1 \cup A_2 \cup \cdots = \bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad A_1 \cap A_2 \cap \cdots = \bigcap_{i=1}^{\infty} A_i$$

are events " at least one A_i occurs" and "all A_i's occur."

If at least one event A_i occurs, then there is one that occurs first. This remark leads to the following useful decomposition of a union of events into a union of *disjoint* events:

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_1^c \cap A_2) \cup (A_1^c \cap A_2^c \cap A_3) \cup \cdots, \tag{1.6}$$

where $A_1^c \cap \cdots \cap A_{k-1}^c \cap A_k$ is the event "A_k is the first event in the sequence that occurs."

For an infinite sequence $A_1, A_2, \ldots$ one can define two events:

$$\limsup A_n = \bigcap_{k=1}^{\infty} \bigcup_{i=k}^{\infty} A_i \tag{1.7}$$

and

$$\liminf A_n = \bigcup_{k=1}^{\infty} \bigcap_{i=k}^{\infty} A_i, \tag{1.8}$$

these being, respectively, the event that "infinitely many A_i's occur" and the event that "all except finitely many A_i's occur." Here the inner union in the event (1.7) is the event "at least one event A_i with $i \geq k$ will occur"; call this event B_k. The intersection over k means that the event B_k occurs for every k. No matter how large k we take, there will be at least one event A_i with $i \geq k$ that will occur. But this is possible only if infinitely many A_i's occur.

For the event $\liminf A_n$ the argument is similar. The intersection $A_k \cap A_{k+1} \cap \cdots = C_k$ occurs if all events A_i with $i \geq k$ occur. The union $C_1 \cup C_2 \cup \cdots$ means

that at least one of the events C_k will occur, and that means that all A_i will occur, except possibly finitely many.

If all events (except possibly finitely many) occur, then infinitely many of them must occur, so that $\limsup A_n \supset \liminf A_n$. If $\limsup A_n \subset \liminf A_n$, then (see the definition of equality of events) we say that the sequence $\{A_n\}$ *converges*, and $\limsup A_n = \liminf A_n$.

The most important class of convergent sequences of events consists of *monotone* sequences, when $A_1 \subset A_2 \subset \cdots$ (increasing sequence) or when $A_1 \supset A_2 \supset \cdots$ (decreasing sequence). We have the following theorem:

Theorem 1.4.1 *If the sequence $A_1, A_2, \ldots$ is increasing, then*

$$\lim A_n = \bigcup_{n=1}^{\infty} A_n,$$

and in case of a decreasing sequence, we have

$$\lim A_n = \bigcap_{n=1}^{\infty} A_n.$$

Proof. If the sequence is increasing, then the inner union $(\bigcup_{i=1}^{\infty} A_i)$ in $\limsup A_n$ remains the same independently of k so that $\limsup A_n = \bigcup_{i=1}^{\infty} A_i$. On the other hand, the inner intersection in $\liminf A_n$ equals A_k so that $\liminf A_n = \bigcup_{k=1}^{\infty} A_k$, which is the same as $\limsup A_n$, as was to be shown. A similar argument holds for decreasing sequences. □

The following two examples illustrate the concept of convergence of events.

■ **EXAMPLE 1.16**

Let $B(r)$ and $C(r)$ be the sets of points on the plane (x, y) satisfying the conditions $x^2+y^2 < r^2$ and $x^2+y^2 \leq r^2$, respectively. If $A_n = B(1+1/n)$, then $\{A_n\}$ is a decreasing sequence, and therefore $\lim A_n = \bigcap_{n=1}^{\infty} B(1 + 1/n)$. Since $x^2 + y^2 < (1 + 1/n)^2$ for all n if and only if $x^2 + y^2 \leq 1$, we have $\bigcap_{n=1}^{\infty} B(1+1/n) = C(1)$. On the other hand, if $A_n = C(1-1/n)$, then $\{A_n\}$ is an increasing sequence, and $\lim A_n = \bigcup_{n=1}^{\infty} C(1 - 1/n) = B(1)$. We leave a justification of the last equality to the reader.

■ **EXAMPLE 1.17**

Let $A_n = B(1 + 1/n)$ for n odd and $A_n = B(1/3 - 1/2n)$ for n even. The sequence $\{A_n\}$ is now $B(2), B(1/12), B(4/3), B(5/24), \ldots$, so it is not monotone. We have here $\limsup A_n = C(1)$, since every point (x, y) with $x^2 + y^2 \leq 1$ belongs to infinitely many A_n. On the other hand, $\liminf A_n = B(1/3)$. For $x^2 + y^2 < 1/9$, we have $x^2 + y^2 < (1/3 - 1/2n)^2$ if n is large enough (and also $x^2+y^2 < 1+1/n$ for all n). However, if $x^2+y^2 \geq 1/3$, then (x, y) does not belong to any A_n with even n. Thus $\limsup A_n \neq \liminf A_n$, and the sequence $\{A_n\}$ does not converge.

Infinite operations on events play a very important role in the development of the theory, especially in determining limiting probabilities.

The definitions below will prepare the ground for the considerations in the following chapters. In Chapter 2 we will introduce probability as a number assigned to an event. Formally, we will be considering numerical functions defined on events, that is, on subsets of the sample space $\mathcal{S}$. As long as $\mathcal{S}$ is finite or countably infinite, we can take the class of all subsets of $\mathcal{S}$ as the domain of definition of probability. In case of infinite but *not* countable $\mathcal{S}$ (e.g., where $\mathcal{S}$ is an interval, the real line, or a plane) it may not be possible to define probability on the class of *all* subsets of $\mathcal{S}$. Although the explanation lies beyond the scope of this book, we will show how the difficulties can be avoided by suitable restriction of the class of subsets of $\mathcal{S}$ that are taken as events. We begin with the concept of *closure* under some operation.

Definition 1.4.1 We say that the class $\mathcal{A}$ of subsets of $\mathcal{S}$ is *closed* under a given operation if the sets resulting from performing this operation on elements of $\mathcal{A}$ are also elements of $\mathcal{A}$. □

Complementation A^c, finite union $A_1 \cup \cdots \cup A_n$, infinite union $A_1 \cup A_2 \cup \cdots$, limits of sequences $\lim A_n$, are few examples of such operations.

■ **EXAMPLE 1.18**

Let $\mathcal{S} = \{0, 1, 2, \ldots\}$, and let $\mathcal{A}$ consist of all subsets of $\mathcal{S}$ that are finite. $\mathcal{A}$ is closed under finite unions and all intersections, finite or not. Indeed, if $A_1, \ldots, A_n$ are finite sets, then $A = A_1 \cup \cdots \cup A_n$ is also finite. Similarly, if $A_1, A_2, \ldots$ are finite, then $\cap_i A_i \subset A_1$, and hence $\cap_i A_i$ is also finite. However, $\mathcal{A}$ is *not closed* under complementation: if A is finite ($A \in \mathcal{A}$), then A^c is not finite, and hence $A^c \notin \mathcal{A}$. On the other hand, if $\mathcal{A}$ is the class of all subsets of $\mathcal{S}$ that contain some fixed element, say 0, then $\mathcal{A}$ is closed under all intersections and unions, but it is not closed under complementation.

■ **EXAMPLE 1.19**

Let $\mathcal{S}$ be the real line, and let $\mathcal{A}$ be the class of all intervals closed on the right and open on the left—meaning intervals of the form $(a, b] = \{x : a < x \leq b\}$. Assume that we allow here $b \leq a$, in which case $(a, b]$ is empty. Then $\mathcal{A}$ is closed under the operation of intersection with $(a, b] \cap (c, d] = (\max(a, c), \min(b, d)]$.

The following three concepts play a central role in the construction of the probability theory:

Definition 1.4.2 A nonempty class $\mathcal{A}$ of subsets of $\mathcal{S}$ that is closed under complementation and all finite operations (i.e., finite union, finite intersection) is called a *field*. If $\mathcal{A}$ is closed under complementation and all countable operations, it is called

a *σ-field*. Finally, if $\mathcal{A}$ is closed under monotone passage to the limit,[2] it is called a *monotone class*. □

Let us observe that Definition 1.4.2 can be formulated in a more efficient way. For $\mathcal{A}$ to be a field, it suffices to require that if $A, B \in \mathcal{A}$ then $A^c \in \mathcal{A}$ and $A \cap B \in \mathcal{A}$ (or $A^c \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$). Any of these two conditions implies (by induction and De Morgan's laws) the closure of $\mathcal{A}$ under all finite operations. Consequently, for $\mathcal{A}$ to be a σ-field, it suffices to require that whenever $A_1, A_2, \ldots \in \mathcal{A}$, then $A_i^c \in \mathcal{A}$ and $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$ (or $A_i^c \in \mathcal{A}$ and $\bigcup_{n=1}^{\infty} A_i \in \mathcal{A}$); this follows again from De Morgan's laws.[3]

It is important to realize that closure under countable operations is stronger than closure under any finite operations. This means that there exist classes of sets that are fields but not σ-fields. This is illustrated by the following example:

■ **EXAMPLE 1.20**

Let $\mathcal{S} = \{1, 2, 3, \ldots\}$, and let $\mathcal{A}$ be the class of all subsets A of $\mathcal{S}$ such that either A or A^c is finite. Then $\mathcal{A}$ is a field but not a σ-field. First, if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$ because the definition of $\mathcal{A}$ is symmetric with respect to complementation. Next, if A and B are both in $\mathcal{A}$, so is their union. If A and B are both finite, then $A \cup B$ is finite and hence belongs to $\mathcal{A}$. On the other hand, if either A^c or B^c (or both) are finite, then $(A \cup B)^c = A^c \cap B^c$ is also finite because it is contained in A^c and also in B^c.

Thus $\mathcal{A}$ is a field. However, $\mathcal{A}$ is not a σ-field. Let A_n be the set consisting only of the element n (i.e., $A_n = \{n\}$). Clearly, $A_n \in \mathcal{A}$. Take now $\bigcup_{n=1}^{\infty} A_{2n} = \{2, 4, 6, \ldots\}$. This is a countable union of sets in $\mathcal{A}$ that is not in $\mathcal{A}$ (since the set of all even numbers is not finite, nor does it have a finite complement).

Typically it is easy to determine that a class of sets is a field, while direct verification that it is a σ-field can be difficult. On the other hand, it is occasionally easy to verify that a class of sets is a monotone class. Thus the following theorem is sometimes useful:

Theorem 1.4.2 *A σ-field is a monotone class. Conversely, a field that is a monotone class is a σ-field.*

Proof. To prove this theorem, assume first that $\mathcal{A}$ is a σ-field, and let $A_1, A_2, \ldots$ be a monotone sequence of elements of $\mathcal{A}$. If $A_1 \subset A_2 \subset \cdots$, then $\lim A_n = \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$, whereas if $A_1 \supset A_2 \supset \cdots$, then $\lim A_n = \bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$. So $\mathcal{A}$ is a monotone class. On the other hand, let $\mathcal{A}$ be a monotone class and a field, and let $A_1, A_2, \ldots$ be an arbitrary sequence of elements of $\mathcal{A}$. Put $B_n = A_1 \cup \cdots \cup A_n$. Then since $\mathcal{A}$ is a field, and also $B_1 \subset B_2 \subset \cdots$, $B_n \in \mathcal{A}$ for every n. Further, since $\mathcal{A}$

[2] In view of the fact proved earlier that all monotone sequences converge, this condition means that (a) if $A_1 \subset A_2 \subset \cdots$ is an increasing sequence of sets in $\mathcal{A}$, then $\cup_i A_i \in \mathcal{A}$, and (b) if $A_1 \supset A_2 \supset \cdots$ is a decreasing sequence of sets in $\mathcal{A}$, then $\cap_i A_i \in \mathcal{A}$.

[3] For various relations among classes of sets defined through closure properties under operations, for example, see Chow and Teicher (1997) and Chung (2001).

is a monotone class, $\lim B_n \in \mathcal{A}$. However, $\lim B_n = \bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$, so $\mathcal{A}$ is a σ-field, as asserted. □

The last in this series of concepts is that of the minimal field (or σ-field, or monotone class) containing a given set or collection of sets. We begin with some examples.

■ **EXAMPLE 1.21**

Let $\mathcal{S}$ be any set. On one extreme, the class consisting of two sets, $\emptyset$ and $\mathcal{S}$, is closed under any operation so that $\mathcal{A} = \{\emptyset, \mathcal{S}\}$ is a field, a σ-field, and a monotone class. On the other extreme, the class of *all* subsets of $\mathcal{S}$ is also closed under any operations, finite or not, and hence is a field, a σ-field, and a monotone class. These two classes of subsets of $\mathcal{S}$ form the smallest and the largest fields (σ-field, monotone class).

For any event A it is easy to check that the class $\mathcal{A}$, consisting of the four events $\{\emptyset, A, A^c, \mathcal{S}\}$, is closed under any operations: unions, intersections, and complements of members of $\mathcal{A}$ are again members of $\mathcal{A}$. This class is an example of a field (σ-field, monotone class) that contains the events A and A^c, and it is the smallest such field (σ-field, monotone class).

On the other hand, the class $\mathcal{A}$, consisting of events $\{\emptyset, A, S\}$, is a monotone class, but neither a field nor σ-field. If A and B are two events, then the smallest field $\mathcal{A}$ containing A and B must contain also the sets A^c, B^c, the intersections $A \cap B, A \cap B^c, A^c \cap B, A^c \cap B^c$, as well as their unions $A \cup B, A \cup B^c, A^c \cup B$, and $A^c \cup B^c$. The closure property implies that unions such as $(A \cap B) \cup (A \cup B^c)$, must also belong to $\mathcal{A}$.

We are ready to present the final step in our construction.

Theorem 1.4.3 *For any nonempty class $\mathcal{K}$ of subsets of $\mathcal{S}$ there exists a unique smallest field (σ-field, monotone class) containing all sets in $\mathcal{K}$. It is called the field (σ-field, monotone class) generated by $\mathcal{K}$.*

Proof. We will prove the assertion for fields. Observe first that if $\mathcal{A}_1$ and $\mathcal{A}_2$ are fields, then their intersection $\mathcal{A}_1 \cap \mathcal{A}_2$ (i.e., the class of sets that belong to both $\mathcal{A}_1$ and $\mathcal{A}_2$) is also a field. For instance, if $A, B \in \mathcal{A}_i$ $(i = 1, 2)$, then $A \cup B \in \mathcal{A}_i$ because each $\mathcal{A}_i$ is a field, and consequently $A \cup B \in \mathcal{A}_1 \cap \mathcal{A}_2$. A similar argument holds for intersections and complements.

Note that if $\mathcal{A}_1$ and $\mathcal{A}_2$ contain the class $\mathcal{K}$, then the intersection $\mathcal{A}_1 \cap \mathcal{A}_2$ also contains $\mathcal{K}$. The foregoing property extends to any intersection of fields containing $\mathcal{K}$ (not only the intersections of two such fields).

Now let $\mathcal{C}$ be the intersection of *all* fields containing $\mathcal{K}$. We claim that $\mathcal{C}$ is the minimal unique field containing $\mathcal{K}$. We have to show that (1) $\mathcal{C}$ exists, (2) $\mathcal{C}$ is a field containing $\mathcal{K}$, (3) $\mathcal{C}$ is unique, and (4) $\mathcal{C}$ is minimal.

For property (1) it is enough to show that there exists at least one field containing $\mathcal{K}$. We may take here the class of all subsets of S: it is a field (as well as a σ-field and monotone class), and it contains all sets in $\mathcal{K}$. Property (2) follows from the fact that the intersection of fields containing $\mathcal{K}$ is a field containing $\mathcal{K}$. Property (3) (i.e.,

uniqueness of $\mathcal{C}$) follows from the fact that the result of the operation of intersection is unique.

Finally, suppose that there exists a field $\mathcal{C}'$ containing $\mathcal{K}$ such that $\mathcal{C}' \subset \mathcal{C}$. Then $\mathcal{C}'$ must appear as one of the factors in the intersection defining $\mathcal{C}$ so that $\mathcal{C} \subset \mathcal{C}'$. Consequently $\mathcal{C}' = \mathcal{C}$. This completes the proof for the case of fields. The proofs for σ-fields and monotone classes are exactly the same, since an intersection of σ-fields (or monotone classes) containing $\mathcal{K}$ is again a σ-field (monotone class) containing $\mathcal{K}$. □

One may find it disturbing that Theorem 1.4.3 asserts the existence and uniqueness of some objects without giving a clue as to how to find them in practical situations. In fact the nonconstructive character of the theorem, combined with its generality, is instead a great help. As we will see in Chapter 2, the natural objects of our interest (the domains of definition of probability) will be σ-fields of events. Beyond the trivial situations of finite or countably infinite sample spaces $\mathcal{S}$, where one can always consider the maximal σ-field consisting of all subsets of $\mathcal{S}$, one is forced to restrict consideration to classes of events that form σ-fields generated by some "simple" events. The events in these σ-fields are typically of a very rich structure, and one seldom has useful criteria for distinguishing events (elements of the σ-field in question) from "non-events," that is, subsets of $\mathcal{S}$ to which probabilities are not assigned. However, as shown by the two examples below, the smallest σ-field generated by some class is richer than the smallest field generated by the same class.

■ **EXAMPLE 1.22**

A point moves randomly on the plane, and its location is recorded at some time t. The outcome of this experiment is the pair (x, y) of coordinates of the observed location of the point (e.g., imagine here the location of a particle of dust in a liquid, tossed about by random hits from molecules of the medium, and performing Brownian motion; or imagine a location of a previously marked bird at the time of its capture in a bird migration study or the ages of both husband and wife at the time one of them dies).

In any study of this kind (regardless its ultimate purpose), the "natural" sample space $\mathcal{S}$ is a plane or part of the plane, (the positive quadrant, etc.). The "simple" events here are of the form $a < x \leq b, c < x \leq d$, that is, rectangles with sides parallel to the axes. The reason for distinguishing these events as "simple" is that, as will be explained in later chapters, it is often easy to assign probabilities to these events. The reason for the particular configuration of strict and non-strict inequalities (i.e., north and east side included, south and west side excluded) will also become apparent from the analysis below. To simplify the language, we will call such events Rectangles, and use a capital letter to signify the specific assumption about which sides are included and which are not. Naturally we will allow for infinite Rectangles, such as $\{a < x \leq \infty, -\infty < y \leq b\}$.

It is easy to determine the field generated by all Rectangles: These are events that result from finite operations on Rectangles. Clearly, the comple-

ment of a Rectangle is a union of at most eight disjoint (infinite) Rectangles (see Figure 1.7), whereas the intersection of Rectangles is again a Rectangle (or is empty).

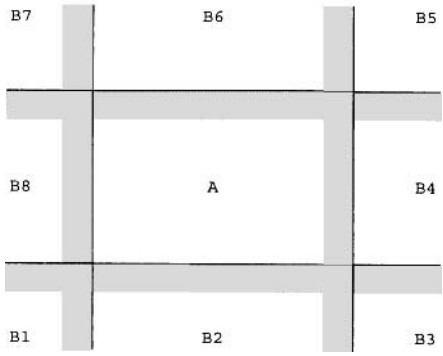


Figure 1.7 Complement of a Rectangle

Since unions are reduced to intersections of complements by De Morgan's laws, every element of the smallest field containing all Rectangles is the union of a finite number of disjoint Rectangles. On the other hand, there exist events that do not belong to this field of events. As a simple example, one might be interested in the event that the point (x, y) lies within distance r from some fixed point (from the initial location of the particle, the point of release of the bird, etc.). This event is a circle on the plane, and hence a subset of $\mathcal{S}$, which is not decomposable into a finite number of Rectangles. On the other hand, a circle does belong to the σ-field spanned by Rectangles: it is representable as a countable union of Rectangles, or equivalently, as an infinite intersection of sets built up of Rectangles.

Similarly in this example there are other events, which are not in the field generated by Rectangles and which could be considered, such as triangles, rectangles with sides not parallel to the axes, and ellipses.

■ **EXAMPLE 1.23**

Take an experiment consisting of tossing a coin infinitely many times. The "natural" sample space $\mathcal{S}$ is the space of all infinite sequences $x = (\xi_1, \xi_2, \ldots)$, where $\xi_i = 0$ or 1 (or any other two distinct symbols representing heads and tails). The "simple" events here are of the form "heads on the nth toss," that is, sets of all infinite sequences $x = (\xi_1, \xi_2, \ldots)$ with the nth coordinate ξ_n satisfying $\xi_n = 0$. The events in the field generated by the simple events are of the form "heads on tosses $k_1, \ldots, k_n$ and tails on tosses $r_1, \ldots, r_m$," with both m and n finite and the outcomes of all other tosses remaining unspecified.

An event that does not belong to this field, but does belong to the σ-field generated by the simple events, is the event that "as the number of tosses increases, the frequency of heads approaches a limit." Clearly, to determine

whether or not this event occurs, it does not suffice to know any finite number of coordinates ξ_n.

To generalize this example, replace the outcome of the coin tosses by the result of some experiment repeated infinitely many times. This way the coordinate ξ_n carries more information than it does for the outcome of nth coin toss. The "simple" events are now of the form of sets of sequences $x = (\xi_1, \xi_1, \ldots)$ with $\xi_i \in A_i$ for $i = 1, \ldots, n$, while the ξ_i's for $i > n$ are unconstrained. Here $A_1, \ldots, A_n$ are events that occur at the first n times of observations. The "simple" events described above, of an obvious interest and importance both in applications and in building the theory, are called "cylinder" events. The smallest σ-field containing all cylinder events comprises all events that may be of interest, including those that are obtained through limits of sequences of cylinder events.

PROBLEMS

1.4.1 Let $B_1, B_2, \ldots$ be a countable partition of $\mathcal{S}$; that is, $B_i \cap B_j = \emptyset$ for all $i \neq j$, and $\bigcup_i B_i = \mathcal{S}$. Let $A_n = B_n \cup B_{n+1} \cup \cdots$. Find $\lim A_n$.

1.4.2 Assume that John will live forever. He plays a certain game each day. Let A_i be the event that he wins the game on the ith day.
(i) Let B be the event that John will win every game starting on January 1, 2015. Label the following statements as true or false: (a) $B = \liminf A_n$. (b) $B \subset \liminf A_n$. (c) $B \supset \limsup A_n$. (d) $B = \limsup A_n$.
(ii) Assume now that John starts playing on a Monday. Match the following events C_1 through C_9 with events D_1 through D_{11}:

C_1 = John loses infinitely many games.
C_2 = When John loses on a Thursday, he wins on the following Sunday.
C_3 = John never wins on three consecutive days.
C_4 = John wins every Wednesday.
C_5 = John wins on infinitely many Wednesdays.
C_6 = John wins on a Wednesday.
C_7 = John never wins on a weekend.
C_8 = John wins infinitely many games and loses infinitely many games.
C_9 = If John wins on some day, he never loses on the next day.

$$D_1 = \bigcap_{j=0}^{\infty} \left[A_{7j+4} \cup A_{7(j+1)}\right].$$

$$D_2 = \bigcap_{j=0}^{\infty} A_{7j+3}.$$

$$D_3 = \bigcap_{j=0}^{\infty} \left[A_{7j+6}^c \cap A_{7(j+1)}^c\right].$$

$$D_4 = \bigcup_{n=0}^{\infty} \left\{ \left[\bigcap_{i=1}^{n} A_i^c\right] \cap \left[\bigcap_{k=n+1}^{\infty} A_k\right] \right\}.$$

$$D_5 = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_{7k+3}.$$

$$D_6 = \bigcap_{i=1}^{\infty} \left[A_i^c \cup A_{i+1}^c \cup A_{i+2}^c\right].$$

$$D_7 = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^c.$$

$$D_8 = \bigcup_{j=0}^{\infty} A_{7j+3}.$$

$$D_9 = \left\{\bigcup_{j=0}^{\infty} \left[A_{7j+6} \cup A_{7(j+1)}\right]\right\}^c.$$

$$D_{10} = \bigcap_{i=0}^{\infty} \left[A_i \cap A_{i+1} \cap A_{i+2}\right]^c.$$

$$D_{11} = \left[\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^c\right] \cap \left[\bigcap_{j=1}^{\infty} \bigcup_{m=j}^{\infty} A_m\right].$$

1.4.3 Let $A_1, \ldots, A_n$ be distinct subsets of $\mathcal{S}$. **(i)** Find the maximum number of sets (including $\mathcal{S}$ and $\emptyset$) of the smallest field containing $A_1, \ldots, A_n$. **(ii)** Find the maximum number of sets in this field if $A_{n-1} \subset A_n$. **(iii)** Answer (ii) if $A_1 \subset A_2 \subset \cdots \subset A_n$. **(iv)** Answer (ii) if $A_1 = \cdots = A_n = \emptyset$. **(v)** Answer (i)–(iv) for a σ-field.

1.4.4 For $0 < \alpha < 1$, let $I(\alpha) = \{x : 1 - \alpha < x < 1 + \alpha\}$. Consider a sequence $\alpha_1, \alpha_2, \ldots$ of numbers satisfying $0 < \alpha_n < 1$ for all n, and let $A_n = I(\alpha_n)$. **(i)** Find $\limsup A_n$ and $\liminf A_n$. **(ii)** Find conditions, expressed in terms of α_n, under which $\lim A_n$ exists, and find this limit. **(iii)** Define $J(\alpha) = \{x : 1-\alpha \leq x \leq 1+\alpha\}$ and $B_n = J(\alpha_n)$. Answer questions (i) and (ii) for sequence $\{B_n\}$.

1.4.5 Let $\mathcal{S} = \{0, 1, 2, \ldots\}$ be the set of all integers. For $A \subset \mathcal{S}$, let $f_n(A)$ be the number of elements in the intersection $A \cap \{0, 1, \ldots, n\}$. Let $\mathcal{A}$ be the class of all sets A for which the limit

$$q(A) = \lim_{n \to \infty} \frac{f_n(A)}{n}$$

exists. Show that $\mathcal{A}$ is not a field. [*Hint:* Let $A_1 = \{1, 3, 5, \ldots\}$ and $A_2 =$ { all odd integers between 2^{2n} and 2^{2n+1} and all even integers between 2^{2n+1} and 2^{2n+2} for $n = 0, 1, \ldots\}$. Show that both A_1 and A_2 are in $\mathcal{A}$ but $A_1 \cap A_2 \notin \mathcal{A}$.]

1.4.6 Let $S = (-\infty, +\infty)$. Show that the class of all finite unions of intervals of the form $[a, b]$, (a, b), $[a, b)$, and $(a, b]$, with possibly infinite a or b (intervals of the form $[a, \infty)$, etc.) forms a field.

CHAPTER 2

PROBABILITY

2.1 INTRODUCTION

The concept of probability has been an object of debate among philosophers, logicians, mathematicians, statisticians, physicists, and psychologists for the last couple of centuries, and this debate is not likely to be over in the foreseeable future. As advocated by Bertrand Russell in his essay on scepticism, when experts disagree, the layman would do best by refraining from forming a strong opinion. Accordingly, we will not enter into the discussion about the nature of probability; rather, we will start from the issues and principles that are commonly agreed upon.

Probability is a number associated with an event that is intended to represent its "likelihood," "chance of occurring," "degree of certainty," and so on. The phrases above have to be explicated so as to obtain workable principles. This can be done in several ways, the most common being (1) the *frequency* (or *objective*) interpretation of probability, (2) the *classical* (sometimes called *logical*) interpretation of probability, and (3) the *subjective* or *personal* interpretation of probability.

2.2 PROBABILITY AS A FREQUENCY

According to the common interpretation, probability is the "long-run" relative frequency of an event. Before attempting to provide a more formal explication, let us

observe that the idea connecting probability and frequency is (and had been for a long time) very well grounded in everyday intuition. For instance, loaded dice were on several occasions found in the graves of ancient Romans. That indicates that they were aware of the possibility of modifying long-run frequencies of outcomes, and perhaps making some profit in such a way.

Today the intuition regarding relationship between probabilities and frequencies is even more firmly established. For instance, the phrases "there is 3% chance that an orange picked at random from this shipment will be rotten" and "the fraction of rotten oranges in this shipment is a 3%" appear almost synonymous. But on closer reflection one realizes that the first phrase refers to the probability of an event "randomly selected orange will be rotten," while the second phrase refers to the population of oranges.

The precise nature of the relation between probability and frequency is hard to formulate. But the usual explanation is as follows: Consider an experiment that can be repeated under identical conditions, potentially an infinite number of times. In each of these repetitions, some event, say A, may occur or not. Let $N(A)$ be the number of occurrences of A in the first N repetitions. The frequency principle states that the ratio $N(A)/N$ approximates the probability $P(A)$ of event A, with the accuracy of the approximation increasing as N increases.

Let us observe that this principle serves as a basis for estimating probabilities of various events in the real world, especially those probabilities that might not be attainable by any other means (e.g., the probability of heads in tossing a biased coin).

We start this chapter by putting a formal framework (axiom system) on a probability regarded as a *function on the class of all events*. That is, we impose some general conditions on a set of individual probabilities. This axiom system, due to Kolmogorov (1933), will be followed by the derivation of some of its immediate consequences. The latter will allow us to compute probabilities of some composite events given the probabilities of some other ("simpler") events.

2.3 AXIOMS OF PROBABILITY

Let $\mathcal{S}$ be the sample space, namely the set of all outcomes of an experiment. Formally, probability, to be denoted by P, is a function defined on the class of all events[4], satisfying the following conditions (usually referred to as *axioms*):

Axiom 1 (Nonnegativity):

$$P(A) \geq 0 \quad \textit{for every event } A.$$

Axiom 2 (Norming):

$$P(\mathcal{S}) = 1.$$

Axiom 3 (Countable Additivity):

$$P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

[4]The nature of the class of all events will be (to a certain extent) explicated in Section 2.6. See also Section 1.4.

for every sequence of pairwise disjoint events $A_1, A_2, \ldots$, *so that* $A_i \cap A_j = \emptyset$ *for all* $i \neq j$. □

If the sample space $\mathcal{S}$ is finite or countable, one can define a probability function P as follows: Let f be a nonnegative function defined on $\mathcal{S}$, satisfying the condition $\sum_{s\in\mathcal{S}} f(s) = 1$. Then P may be defined for every subset A of $\mathcal{S}$ as $P(A) = \sum_{s\in A} f(s)$. One can easily check that P satisfies all three axioms.

Indeed, $P(A) \geq 0$ because f is nonnegative, and $P(S) = \sum_{s\in S} f(s) = 1$. Finally, let $A_1, A_2, \ldots$ be a sequence of disjoint subsets of S. Then

$$\begin{aligned} P(A_1) + P(A_2) + \cdots &= \sum_{s\in A_1} f(s) + \sum_{s\in A_2} f(s) + \cdots \\ &= \sum_{s\in A_1\cup A_2\cup\cdots} f(s) = P(\bigcup_{i=1}^{\infty} A_i). \end{aligned}$$

Passing from the first to the second line is allowed because $A_1, A_2, \ldots$ are disjoint, so each term appears only once. The sum in the second line is well defined (i.e., it does not depend on the order of summation because the terms are nonnegative).

However, if $\mathcal{S}$ is not countable, one usually needs to replace summation by integration, $P(A) = \int_A f(s)ds$. This imposes some conditions on functions f and on the class of events A. For a detailed discussion the reader is referred to more advanced probability texts (e.g., Chung, 2001).

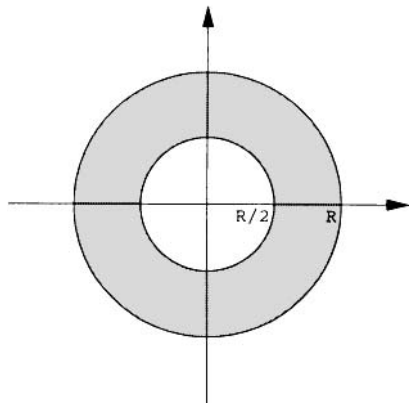


Figure 2.1 Hitting a target

■ EXAMPLE 2.1 Geometric Probability

One of the first examples of an uncountable sample space is associated with "the random choice of a point from a set." This phrase is usually taken to mean the following: a point is selected at random from a certain set $\mathcal{S}$ in a finite-dimensional space (line, plane, etc.), where $\mathcal{S}$ has finite measure (length, area, etc.), denoted generally by $|\mathcal{S}|$. The choice is such that if A ($A \subset \mathcal{S}$) has measure $|A|$, then the probability of the chosen point falling into A is

proportional to $|A|$. Identifying $\mathcal{S}$ with the sample space, we can then write $P(A) = |A|/|\mathcal{S}|$.

To better see this, suppose that in shooting at a circular target $\mathcal{S}$, one is certain to score a hit, and that the point where one hits $\mathcal{S}$ is assumed to be chosen at random in the way described above. What is the probability that the point of hit is farther from the center than half of the radius of the target?

From Figure 2.1 it is clear that the point of hit must lie somewhere in the shaded annulus A. Its area is $|A| = \pi R^2 - \pi(R/2)^2$ so that $P(A) = |A|/\pi R^2 = 3/4$. Of course, the assumption under which this solution is obtained is not very realistic: typically sets closer to the center are more likely to be hit than sets of the same area located closer to the perimeter.

The concept of "random choice" from an uncountable set is sometimes ambiguous. This is illustrated by the next example.

■ **EXAMPLE 2.2 Bertrand's Paradox**

A chord is chosen at random in a circle. What is the probability that the length of the chord will exceed the length of the side of an equilateral triangle inscribed in the circle?

This problem was originally posed by Joseph Bertrand, a French mathematician, who provided three solutions, all valid, but yielding inconsistent results.

SOLUTION 1. Choose point A as one of the ends of the chord. The chord is uniquely determined by the angle α (see Figure 2.2). These angles are chosen

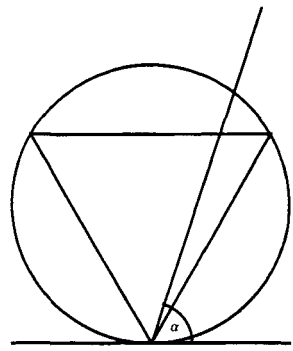


Figure 2.2 First solution of Bertrand's problem

at random from the interval $(0, \pi)$. It is clear that the length of the chord exceeds the side of the equilateral triangle if α lies between $\pi/3$ and $2\pi/3$, so the answer to the question is $1/3$.

SOLUTION 2. Let us draw a diameter QQ' (see Figure 2.3) perpendicular to the chord P. Then the length of the chord exceeds the side of the equilat-

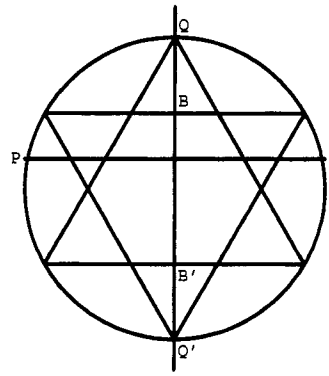


Figure 2.3 Second solution of Bertrand's problem

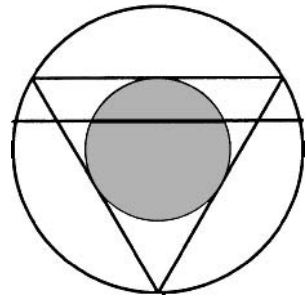


Figure 2.4 Third solution of Bertrand's problem

eral triangle if it intersects the line QQ' between points B and B'. Elementary calculations give $|BB'| = |QQ'|/2$, so the answer is $1/2$.

SOLUTION 3. The location of the chord is uniquely determined by the location of its center (except when the center coincides with the center of the circle, which is an event with probability zero). For the chord to be longer than the side of the equilateral triangle inscribed in the circle, its center must fall somewhere inside the shaded circle in Figure 2.4. Again, by elementary calculations we obtain probability $1/4$.

The discovery of Bertrand's paradox was one of the impulses that made researchers in probability and statistics acutely aware of the need to clarify the foundations of the theory, and ultimately led to the publication of Kolmogorov's book (1933). In the particular instance of the Bertrand "paradoxes," they are explained simply by the fact that each of the solutions refers to a different sample scheme:

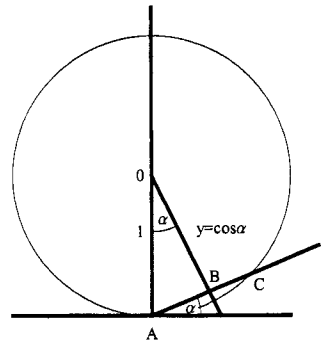


Figure 2.5 Explanation of Bertrand's Paradox

(1) choosing a point on the circumference and then choosing the angle between the chord and the tangent at the point selected, (2) choosing a diameter perpendicular to the chord and then selecting the point of intersection of the chord with this diameter, and (3) choosing a center of the chord. Random choice according to one of these schemes is not equivalent to a random choice according to the other two schemes.

To see why it is so, we will show that the first and second scheme are not equivalent. The analogous arguments for the other two possible pairs of schemes are left as an exercise.

■ **EXAMPLE 2.3**

Imagine different devices (physical mechanisms, computer programs, etc.) built for sampling random chords. One scheme chooses a point on the circumference, and then the angle α between the chord and the tangent to the circle at the point chosen (Figure 2.2). The second scheme chooses first the direction of the diameter and then the point B on the diameter, at which the chord perpendicular to this diameter intersects it (Figure 2.3). From Figure 2.5 it is seen that the angle AOB is α, and therefore $y = |OB| = \cos\alpha$. Thus $dy = (\sin\alpha)\, d\alpha$, which means that equal changes of α do not produce equal changes of y. In fact, these changes are smaller when α is small. Consequently, a device that chooses angles α at random will tend to produce more intersections of the diameter that are farther from the center (i.e., more chords will be shorter).

PROBLEMS

2.3.1 Label all statements below as true or false.
(i) If A is more likely to occur than A^c, then $P(A) > 0.5$.
(ii) If A occurs whenever B does, then $P(B) \leq P(A)$.
(iii) If $P(A) \leq P(B)$, then whenever B occurs, A does also.
(iv) If $P(A) = 0.75$, then A must occur three times out of every four.
(v) The sum of probabilities of disjoint events A and B cannot exceed 1.

(vi) If A and B are not disjoint, the sum of their probabilities exceeds 1.
(vii) If $P(A \cap B)$, $P(A \cap C)$, and $P(B \cap C)$ are all positive, then $P(A \cap B \cap C)$ is also positive.
(viii) If sample spaces for two experiments are identical, then the probability of the same event A must be the same for both experiments.

2.3.2 A bathroom floor is covered by square tiles with side length a. You drop a coin with diameter b, where $b < a$. Find:
(i) The probability that the coin will rest entirely within one tile.
(ii) The probability that the coin will partially cover four different tiles.

2.3.3 **(i)** A point (a, b) is selected at random from the square $[-1, 1]^2$. Find the probability that the equation $ax^2 + bx + 1 = 0$ has two distinct real solutions. **(ii)** Answer the same question if the point (a, b) is selected at random from the rectangle $A_1 \leq a \leq A_2, B_1 \leq b \leq B_2$. Discuss all possible choices of A_1, A_2, B_1, B_2.

2.3.4 Show that first and third, as well as second and third, schemes of sampling chords (see Bertrand's paradox) are not equivalent.

2.4 CONSEQUENCES OF THE AXIOMS

The simplest consequences of the axioms of probability are as follows:

1. *The probability of the impossible event is zero:*

$$P(\emptyset) = 0. \tag{2.1}$$

This follows from the fact that the sequence $\mathcal{S}, \emptyset, \emptyset, \ldots$ satisfies conditions of Axiom 3 so that $1 = P(\mathcal{S}) = P(\mathcal{S} \cup \emptyset \cup \emptyset \cup \cdots) = P(\mathcal{S}) + P(\emptyset) + P(\emptyset) + \cdots$, which is possible only if $P(\emptyset) = 0$. It is important to realize that the converse is not true: the condition $P(A) = 0$ does not imply that $A = \emptyset$. This is shown by the following example:

■ **EXAMPLE 2.4**

Consider an experiment consisting of tossing a coin infinitely many times. The outcomes may be represented as infinite sequences of the form HHTHTTHT ... so that the sample space $\mathcal{S}$ contains infinitely many of such sequences. The event "heads only," that is, the set consisting of just one sequence HHHH ..., is not empty. However, the chance of such an outcome is, at least intuitively, zero: tails should come up sooner or later.

2. *Probability is finitely additive:*

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + (A_n)$$

for any $n = 1, 2, \ldots$, if the events A_i are pairwise disjoint.

In an infinite sequence $A_1, \ldots, A_n, \emptyset, \emptyset, \ldots$, events are pairwise disjoint only if $A_1, \ldots, A_n$ are, so Axiom 3 applies. Therefore, using (2.1), we have

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n \cup \emptyset \cup \cdots) &= P(A_1) + \cdots + P(A_n) + P(\emptyset) + \cdots \\ &= P(A_1) + \cdots + P(A_n), \end{aligned}$$

while the left-hand side is $P(A_1 \cup \cdots \cup A_n)$.

3. *Monotonicity:* If $A \subset B$ then $P(A) \leq P(B)$. This follows from the fact that $B = A \cup (B \cap A^c)$. The events on the right-hand side are disjoint, so we have $P(B) = P(A) + P(B \cap A^c) \geq P(A)$ by Axiom 1. Since $B \cap A^c = B \setminus A$, the equality part gives a useful consequence: if $A \subset B$, then

$$P(B \setminus A) = P(B) - P(A). \tag{2.2}$$

Since $A \subset \mathcal{S}$ for every event A, we have $0 \leq P(A) \leq 1$.

4. *Probability is countably subadditive:*

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq P(A_1) + P(A_2) + \cdots, \tag{2.3}$$

for every sequence of events $A_1, A_2, \ldots$. This follows from representation (1.6) as a union of disjoint events, and then from monotonicity. We have

$$\begin{aligned} P\left(\bigcup_{n=1}^{\infty} A_n\right) &= P(A_1) + P(A_1^c \cap A_2) + P(A_1^c \cap A_2^c \cap A_3) + \cdots \\ &\leq P(A_1) + P(A_2) + P(A_3) + \cdots. \end{aligned}$$

5. *Complementation:*

$$P(A^c) = 1 - P(A). \tag{2.4}$$

This follows from Axiom 2, by the fact that A and A^c are disjoint and $A \cup A^c = \mathcal{S}$.

6. *Probability of a union of events:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.5}$$

Indeed, we can recast $A \cup B = A \cup (A^c \cap B)$ so that $P(A \cup B) = P(A) + P(A^c \cap B)$. On the other hand, $B = (A \cap B) \cup (A^c \cap B)$, and hence $P(B) = P(A \cap B) + P(A^c \cap B)$. Solving for $P(A^c \cap B)$ in one equation and substituting into the other, we obtain (2.5).

A more intuitive argument may be made by using Venn diagrams (see Figure 2.6). In the sum $P(A) + P(B)$ each sample point from the intersection $A \cap B$ is included twice, so to obtain the probability of the union $A \cup B$, we must subtract the probability $P(A \cap B)$.

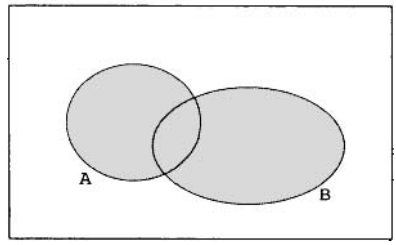


Figure 2.6 Union of two events

■ **EXAMPLE 2.5**

Suppose that $P(A) = 0.4, P(B^c) = 0.6, P(A^c \cap B) = 0.1$. Find $P(A \cup B^c)$.

SOLUTION. The best strategy for solving this kind of problem is usually to start by finding the probabilities of all the intersections (in this case, $A \cap B, A \cap B^c, A^c \cap B, A^c \cap B^c$). The probability $P(A^c \cap B) = 0.1$ is given. Next, $(A \cap B) \cup (A^c \cap B) = B$, and the events on the left are disjoint. So $P(A \cap B) + P(A^c \cap B) = P(B) = 1 - P(B^c)$, which means that $P(A \cap B) + 0.1 = 1 - 0.6$, and hence $P(A \cap B) = 0.3$. Then using $A = (A \cap B) \cup (A \cap B^c)$ we have $0.4 = P(A) + P(A \cap B) + P(A \cap B^c) = 0.3 + P(A \cap B^c)$; hence $P(A \cap B^c) = 0.1$. Finally, in the same way we obtain $P(A^c \cap B^c) = 0.5$. Applying formula (2.4), we have $P(A \cup B^c) = P(A) + P(B^c) - P(A \cap B^c) = 0.3 + 0.6 - 0.1 = 0.8$.

For the case of three events: A, B, and C, the same argument based on Venn diagrams gives the formula

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C). \end{aligned} \tag{2.6}$$

It can be checked at Figure 2.7 that the formula (2.6) includes each part of the union $A \cup B \cup C$ exactly once.

Formula (2.6) may be generalized to the case of the union of any finite number of events.

Theorem 2.4.1 *For any events* $A_1, \ldots, A_n$

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n) &= \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i_1 < i_2 \le n} P(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{1 \le i_1 < i_2 < i_3 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) + \cdots \\ &\quad + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n). \end{aligned} \tag{2.7}$$

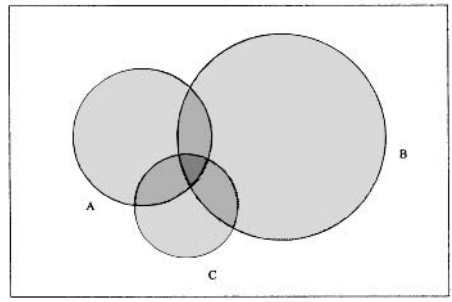


Figure 2.7 Union of three events

Proof. We will proceed by induction. The theorem is true for $n = 1$. Assume now that formula (2.7) holds and write

$$P\left(\bigcup_1^{n+1} A_i\right) = P\left(\bigcup_1^{n} A_i\right) + P\left(A_{n+1} \setminus \bigcup_1^{n} A_i\right).$$

Since

$$P(A_{n+1}) = P\left(A_{n+1} \cap \bigcup_1^{n} A_i\right) + P\left(A_{n+1} \setminus \bigcup_1^{n} A_i\right),$$

we have

$$P\left(\bigcup_1^{n+1} A_i\right) = P\left(\bigcup_1^{n} A_i\right) + P(A_{n+1}) - P\left[\bigcup_1^{n} (A_i \cap A_{n+1})\right].$$

Applying formula (2.7) to $P(\bigcup_1^n A_i)$ and to $P\left[\bigcup_1^n (A_i \cap A_{n+1})\right]$, and then combining the corresponding terms, we obtain (2.7) with n replaced by $n + 1$. □

■ **EXAMPLE 2.6**

Suppose 101 events $A_1, \ldots A_{101}$ are such that $P(A_1) = \cdots = P(A_{101}) = 0.01, P(A_1 \cap A_2) = P(A_1 \cap A_3) = \cdots = P(A_{100} \cap A_{101}) = r$, while every triple intersection is empty. What is the smallest possible value of the probability of intersection r?

SOLUTION. Observe that $1 \geq P(A_1 \cup \cdots \cup A_{101}) = P(A_1) + \cdots + P(A_{101}) - P(A_1 \cap A_2) - \cdots - P(A_{100} \cap A_{101}) = 101/100 - r(100 \times 101)/2 =$. The number of intersections $A_i \cap A_j$ with $i < j$ is $(100 \times 101)/2$ (to see this, we can arrange all pairs into a square table 101×101. The pairs with $i < j$ lie on one side of the diagonal, so their number is $(101^2 - 101)/2) = (100 \times 101)/2)$. There are only two kinds of terms in the sum. Because triple intersections are empty, so are any higher intersections. Thus we have $1 \geq 101/100 - r(101 \times 100)/2$; hence $r \geq 1/505,000$.

PROBLEMS

2.4.1 Let events A, B, and C be such that A contains B and is disjoint from C. Moreover, A is twice as likely as B, three times as likely as C, and half as likely as its complement A^c. Find $P(B \cup C)$.

2.4.2 Events A, B, and C are such that $A \cap B = A \cap C = \emptyset$ while $B \cap C$ is not empty. Determine $P(A \cup B \cup C)$ as a function of x if $P(A) = P(B) = P(C) = 3P(B \cap C) = x$.

2.4.3 Find $P(A \cup B^c)$ for events A and B such that $P(A) = P(B) = 1/2$, and $P(A^c \cap B^c) = 1/3$.

2.4.4 To make the formula $P[(A^c \cap B^c)^c] = 2 - P(A^c) - P(B^c) - x$ valid, x must equal: (1) $P(A \cup B)$. (2) $P(A \cap B)$. (3) $P(A^c \cup B^c)$. (4) None of the above.

2.4.5 Three events A, B, and C are such that C is contained in the intersection $A^c \cap B$, $P(A) = 0.6$, $P(A \cap B) = 0.3$, and $P(C) = P(A^c \cap B^c) = 0.1$. Find: **(i)** $P(B)$. **(ii)** The probability that exactly two of the events A, B, C will occur. **(iii)** The probability that exactly one of the events A, B, C will occur.

2.4.6 Let A, B, C, D, and E be five events such that $P(A) = P(B) = \cdots = P(E) = k$ and $P(A \cap B) = P(A \cap C) = \cdots = P(D \cap E) = p$. Moreover at least one of the events $A, \ldots, E$ must occur, and the intersection of any three events among $A, B, \ldots, E$ is empty. **(i)** Find p if $k = 0.3$. **(ii)** Omit the assumption that at least one of the events must occur and determine all possible values of k if it is known that $p = 0.01$.

2.4.7 Four events A, B, C, and D are such that A and B are disjoint and D is contained in $(A \cup B \cup C)^c$. Moreover, $B \cap C$ is twice as likely as $A \cap C$, while $P(A \cap C^c) = P(B \cap C^c) = 0.2, P(D) = P[(A \cup B \cup C \cup D)^c] = P(C \cap A^c \cap B^c) = 0.1$. Find the probability that out of the events A, B, C, D: **(i)** None will occur. **(ii)** Exactly two will not occur. **(iii)** At most two will occur. **(iv)** Exactly one will occur.

2.4.8 John and Mary have to take a certain statistics course. John attends 40% of the classes, while Mary misses 20% of the classes. They are both present in class only 32% of the time. Find the probability that: **(i)** Only one of them is present in class. **(ii)** They are both absent.

2.4.9 A faulty public phone is such that it returns the coin with probability 60%, it gives you the number you dial with probability 20%, and it takes your coin and does not give you the required connection with probability 30%. Find the probability that you will talk with the number you dial for free.

2.4.10 A certain public phone is such that it returns the coin with probability a, connects you with the number you dial with probability b, and it gives you the connection for free with probability c. Let us agree to say that the phone is *individually honest* if it takes your money if and only if it provides you with the required connection, and that it is *socially honest* if it takes, on average, as many coins as it gives

correct connections (but perhaps from different customers). **(i)** Find conditions for a, b and c under which the phone is individually honest. **(ii)** Find conditions under which the phone is socially honest.

2.4.11 A survey is conducted in a certain city to determine the number of households having electric appliances. It is found that 67% have washing machines (W), 52% have microwave ovens (M), and 48% have dishwashers (D). Furthermore, 32% have (WM), 30% have (MD), 30% have (WD), and 20% have all three. What is the probability that a family has at least one of these appliances?

2.4.12 A regular die and a die with 2, 3, 5, 6, 7, and 8 dots are tossed together, and the total number of dots is noted. What is the probability that the sum is greater than or equal to 10?

2.4.13 A die is loaded in such a way that the probability of j dots on the top face is proportional to j, for j = 1, 2, ..., 6. What is the probability that in one roll of the die an odd number of dots will turn up?

2.5 CLASSICAL PROBABILITY

For the so-called classical or logical interpretation of probability we will assume that *the sample space S contains a finite number N of outcomes and all of these outcomes are equally probable.*

Obviously in this case each of the outcomes has the same probability $1/N$, and for every event A,

$$P(A) = \frac{\text{number of outcomes in } A}{N}. \tag{2.8}$$

In many real situations the outcomes in the sample space reveal a certain symmetry, derived from physical laws, from logical considerations, or simply from the sampling scheme used. In such cases one can often *assume* that the outcomes are equiprobable and use (2.8) as a rule for computing probabilities. Obviously the function P in (2.8) satisfies the axioms of probability.

To use some very simple examples, in tossing a regular die each face has the same probability $1/6$. Then the probability of the event A = "outcome odd" is $P(A) = 3/6 = 1/2$, since there are three odd outcomes among the possible six.

The case above is rather trivial, but considerations of symmetry can sometimes lead to unexpectedly simple solutions of various problems.

■ **EXAMPLE 2.7**

Peter tosses a fair coin n times, and Paul tosses it $n + 1$ times. What is the probability that Paul tosses more heads than Peter?

SOLUTION. Either Paul tosses more heads than Peter (event A) or he tosses more tails than Peter (event B). These two events exclude one another and exhaust all possibilities (since one cannot have ties in number of heads *and* number of tails). Switching the role of heads and tails transforms one of

these events into the other. Thus sample space becomes partitioned into two equiprobable events, and we must have $P(A) = 1/2$.

The use of (2.8) requires techniques for counting the numbers of elements in some sets. These topics, known under the name *combinatorics*, will be discussed in Chapter 3.

PROBLEMS

2.5.1 A coin is tossed 7 times. Assume that each of the $2^7 = 128$ possible outcomes (sequences like HTTHHTH of length 7) is equally likely. Relate each outcome to a binary number by replacing H by 1 and T by 0, for example, THHHTTH is 0111001 = 57. Find the probability that a number generated in this way lies between 64 and 95 (inclusive on both sides).

2.5.2 A die is tossed three times, with outcomes X_1, X_2, and X_3. Assuming that all 216 possible outcomes (x_1, x_2, x_3) are equally likely, find following probabilities: **(i)** $P(X_1 > X_2 = X_3)$. **(ii)** $P(X_1 < X_2 < X_3)$. **(iii)** $P[\max(X_1, X_2, X_3) = 3]$. **(iv)** $P[\min(X_1, X_2, X_3) = 2]$.

2.5.3 Use formula (2.7) to find the number of primes not exceeding 100. [*Hint:* Assume that you sample one of the numbers $1, 2, \ldots, 100$. Let A_i be the event "the number sampled is divisible by i." Determine $p = P(A_2 \cup A_3 \cup A_5 \cup A_7)$. Then the answer to the problem is $100(1 - p) + 3$ (why?).]

2.5.4 A number X is chosen at random from the series 4, 9, 14, 19, ..., and another number Y is chosen from the series 1, 5, 9, 13, ... Each series has 100 terms. Find $P(X = Y)$.

2.6 NECESSITY OF THE AXIOMS*

Looking at Axiom 3, one may wonder why do we need it for the case of countable (and not just finite) sequences of events. Indeed, the necessity of all three axioms, with only finite additivity in Axiom 3, can be easily justified simply by using probability to represent the limiting relative frequency of occurrences of events. Recall the symbol $N(A)$ from Section 2.1 for the number of occurrences of the event A in the first N experiments. The nonnegativity axiom is simply a reflection of the fact that the count $N(A)$ cannot be negative. The norming axiom reflects the fact that event $\mathcal{S}$ is certain and must occur in every experiment so that $N(\mathcal{S}) = N$, and hence $N(\mathcal{S})/N = 1$. Finally (taking the case of two disjoint events A and B), we have $N(A \cup B) = N(A) + N(B)$, since whenever A occurs, B does not, and conversely. Thus, if probability is to reflect the limiting relative frequency, then $P(A \cup B)$ should be equal to $P(A) + P(B)$, since the frequencies satisfy the analogous condition $N(A \cup B)/N = N(A)/N + N(B)/N$.

The need for countable additivity, however, cannot be explained so simply. This need is related to the fact that to build a sufficiently powerful theory, one needs to take limits. If $A_1, A_2, \ldots$ is a monotone sequence of events (increasing or decreasing, i.e., $A_1 \subset A_2 \subset \cdots$ or $A_1 \supset A_2 \cdots$) then $\lim P(A_n) = P(\lim A_n)$, where

the event $\lim A_n$ has been defined in Section 1.4. Upon a little reflection, one can see that such continuity is a very natural requirement. In fact the same requirement has been taken for granted for over 2000 years in a somewhat different context: in computing the area of a circle, one uses a sequence of polygons with an increasing number of sides, all inscribed in the circle. This leads to an increasing sequence of sets "converging" to the circle, and therefore the area of the circle is taken to be the limit of the areas of approximating polygons. The validity of this idea (i.e., the assumption of the continuity of the function $f(A)$ = area of A) was not really questioned until the beginning of the twentieth century. Research on the subject culminated with the results of Lebesgue.

To quote the relevant theorem, let us say that a function P, defined on a class of sets (events), is *continuous from below at the set* A if the conditions $A_1 \subset A_2 \subset \cdots$ and $A = \cup A_n$ imply that $\lim P(A_n) = P(A)$. Similarly, P is continuous *from above at the set* A if the conditions $A_1 \supset A_2 \supset \cdots$ and $A = \cap A_n$ imply $\lim P(A_n) = P(A)$. A function that is continuous at every set from above or from below is simply called *continuous* (above or below). Continuity from below *and* from above is simply referred to as *continuity*.

We may characterize countable additivity as follows:

Theorem 2.6.1 *If the probability P satisfies Axiom 3 of countable additivity, then P is continuous from above and from below. Conversely, if a function P satisfies Axioms 1 and 2, is finitely additive, and is either continuous from below or continuous from above at the empty set $\emptyset$, then P is countably additive.*

Proof. Assume that P satisfies Axiom 3, and let $A_1 \subset A_2 \subset \cdots$ be a monotone increasing sequence. We have

$$\begin{aligned} \bigcup_{n=1}^{\infty} A_n &= A_1 \cup (A_2 \cap A_1^c) \cup (A_3 \cap A_2^c) \cup \cdots \\ &= A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots \end{aligned} \tag{2.9}$$

the events on the right-hand side being disjoint. Since $\cup A_n = \lim A_n$ (see Section 1.5), using (2.9), and the assumption of countable additivity, we obtain

$$\begin{aligned} P(\lim A_n) &= P(\cup A_n) = P(A_1) + \sum_{i=2}^{\infty} [P(A_i) - P(A_{i-1})] \\ &= P(A_1) + \lim_{n\to\infty} \sum_{i=2}^{n} [P(A_i) - P(A_{i-1})] \\ &= P(A_1) + \lim_{n\to\infty} [P(A_n) - P(A_1)] = \lim_{n\to\infty} P(A_n) \end{aligned}$$

(passing from the first to the second line we used the fact that the infinite series is defined as the limit of its partial sums). This proves continuity of P from below. To prove continuity from above, we pass to the complements, and proceed as above.

Let us now assume that P is finitely additive and continuous from below, and let $A_1, A_2, \ldots$ be a sequence of mutually disjoint events. Put $B_n = A_1 \cup \cdots \cup A_n$ so that $B_1 \subset B_2 \subset \cdots$ is a monotone increasing sequence with $\cup A_n = \cup B_n$. We

have then, using continuity from below and finite additivity,

$$\begin{aligned} P(\cup A_n) &= P(\cup B_n) = P(\lim_{n\to\infty} B_n) = \lim_{n\to\infty} P(B_n) \\ &= \lim_{n\to\infty} [P(A_1) + \cdots + P(A_n)] = \sum_{n=1}^{\infty} P(A_n). \end{aligned}$$

again by definition of a numerical series being the limit of its partial sums. This shows that P is countably additive.

Finally, let us assume that P is finitely additive and continuous from above at the empty set $\emptyset$ (impossible event). Taking again a sequence of disjoint events $A_1, A_2, \ldots$, let $C_n = A_{n+1} \cup A_{n+2} \cup \cdots$ We have $C_1 \supset C_2 \supset \cdots$ and $\lim C_n = \bigcap_n C_n = \emptyset$. By finite additivity we obtain

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{n} A_i \cup C_n\right) = P(A_1) + \cdots + P(A_n) + P(C_n). \quad (2.10)$$

Since (2.10) holds for every n, we can write

$$\begin{aligned} P(\cup A_i) &= \lim[P(A_1) + \cdots + P(A_n) + P(C_n)] \\ &= \lim[P(A_1) + \cdots + P(A_n)] + \lim P(C_n) = \sum_{n=1}^{\infty} P(A_i). \end{aligned}$$

Again, by the definition of series and the assumption that $\lim P(C_n) = 0$, P is countably additive, and the proof is complete. □

As an illustration we now prove the following theorem:

Theorem 2.6.2 (First Borel-Cantelli Lemma) *If $A_1, A_2, \ldots$ is a sequence of events such that*

$$\sum_{n=1}^{\infty} P(A_n) < \infty, \quad (2.11)$$

then

$$P(\limsup A_n) = 0.$$

Proof. Recall (1.7) from Chapter 1) where $\limsup A_n$ = "infinitely many events A_j occur" $= \bigcap_{k=1}^{\infty} \bigcup_{i=k}^{\infty} A_i = \lim_{k\to\infty} \bigcup_{i=k}^{\infty} A_i$ (because the unions $\bigcup_{i=k}^{\infty} A_i$, $k = 1, 2, \ldots$ form a decreasing sequence). Consequently, using the continuity of P, subadditivity property (2.3), and assumption (2.11), we have

$$P(\limsup A_n) = \lim_{k\to\infty} P\left(\bigcup_{i=k}^{\infty} A_i\right) \le \lim_{k\to\infty} \sum_{i=k}^{\infty} P(A_i) = 0. \quad \square$$

Paraphrasing the assertion of the lemma, if probabilities of events A_j decrease to zero fast enough to make the series converge, then with probability 1 only finitely many among events A_j will occur. We will prove the converse (under an additional assumption), known as the second Borel-Cantelli lemma, in Chapter 4.

In the remainder of this section we will discuss some theoretical issues related to defining probability in practical situations. Let us start with the observation that the analysis above should leave some more perceptive readers disturbed. Clearly, one should not write a formula without being certain that it is well defined. In particular, when writing $P(\cdots)$ two things ought to be certain: (1) that what appears in the parentheses is a legitimate object of probability, that is, an event, and (2) that the function P is defined unambiguously at this event.

With regard to the first point, the situation is rather simple. All reasonable questions concern events such as $\lim A_n$ and $\limsup A_n$, and hence events obtained by taking countable unions, countable intersections, and complementations of the events $A_1, A_2, \ldots$ Thus the events whose probabilities are discussed belong to the smallest σ-field containing all the events $A_1, A_2, \ldots$ (see Definition 1.4.2 and Theorem 1.4.3). Consequently, to make the formulas at least apparently legitimate, it is sufficient to assume that the class of all the events under considerations is a σ-field, and that probability is a function satisfying the probability axioms defined on this σ-field.

This assumption alone, however, is not enough to safeguard us from possible trouble. To explain the nature of the "danger," let us consider the following hypothetical situation: Suppose that we do not know how to calculate the area of a circle. We could start from the beginning and define the areas of simple figures: first rectangles, then pass to right triangle, and then to arbitrary triangles, which could be reduced to sums and differences of right triangles. From there, the concept of area could be extended to figures that could be triangulated. It is a simple matter to show that the area of such a figure does not depend on how it is triangulated.

From here, we may pass to areas of more complicated figures, the first of these being the circle. The area of the latter could be calculated by inscribing a square in it, and then taking areas of regular polygons with $8, 16, 32, \ldots$ sides and passing to the limit. The result is πr^2. The same result is obtained if we start by inscribing an equilateral triangle, and then take limits of the areas of regular polygons with $6, 12, 24, \ldots$ sides. The same procedure could be tried with an approximation from above, that is, starting with a square or equilateral triangle circumscribed on the circle. Still the limit is πr^2. We could then be tempted to conclude that the area of the circle is πr^2. The result is, of course, true, but how do we know that we will obtain the limit *always* equal to πr^2, regardless of the way of approximating the circle? What if we start, say, from an irregular seven-sided polygon, and then triple the number of sides in each step?

A similar situation occurs very often in probability: Typically we can define probabilities on "simple" events, corresponding to rectangles in geometry, and we can extend this definition without ambiguity to finite unions of the simple events ("rectangles"). The existence and uniqueness of a probability of all the events from the minimal σ-field containing the "rectangles" is ensured by the following theorem, which we state here without proof.

Theorem 2.6.3 *If P is a function defined on a field of events $\mathcal{A}$ satisfying the probability axioms (including countable additivity), then P can be extended in a unique way to a function satisfying the probability axioms, defined on the minimal σ-field containing $\mathcal{A}$.*

This means that if the function P is defined on a field $\mathcal{A}$ of events and satisfies all the axioms of probability, and if $\sigma(\mathcal{A})$ is the smallest σ-field containing all sets in $\mathcal{A}$, then there exists exactly one function P^* defined on $\sigma(\mathcal{A})$ that satisfies the probability axioms, and $P^*(A) = P(A)$ if $A \in \mathcal{A}$.

A comment that is necessary here concerns the question: What does it mean that a function P defined on a field $\mathcal{A}$ satisfies the axioms of probability? Specifically, the problem concerns the axiom of countable additivity, which asserts that if events $A_1, A_2, \ldots$ are disjoint, then

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n). \tag{2.12}$$

However, if P is defined on a field, then there is no guarantee that the left-hand side of formula (2.12) makes sense, since $\bigcup_{n=1}^{\infty} A_n$ need not belong to the field of events on which P is defined. The meaning of the assumption of Theorem 2.6.3 is that formula (2.12) is true whenever the union $\bigcup_{n=1}^{\infty} A_n$ belongs to the field on which P is defined.

The way of finding the probability of some complicated event A is to represent A as a limit of some sequence of events whose probabilities can be computed, and then pass to the limit. Theorem 2.6.3 asserts that this procedure will give the same result, regardless of the choice of sequence of events approximating the event A.

■ **EXAMPLE 2.8 Densities**

A very common situation in probability theory occurs when $\mathcal{S} = (-\infty, +\infty)$. A probability measure P on $\mathcal{S}$ can be defined as follows: let f be a function such that $f(x) \geq 0$ for all x and $\int_{-\infty}^{+\infty} f(x)dx = 1$. We will assume in addition that f is continuous and bounded, although those conditions can be greatly relaxed in general theory.

We now *define* probability on $\mathcal{S}$ by putting

$$P(A) = \int_A f(x)dx \tag{2.13}$$

(in this case f is referred to as a *density* of P). The full justification of this construction lies beyond the scope of this book, but we will give the main points. First, the definition (2.13) is applicable for all intervals A of the form $(a, b), [a, b], (-\infty, b), (a, \infty), [a, \infty)$, and so on. Then we can extend P to finite unions of disjoint intervals by additivity (the class of all such finite unions forms a field). We can easily check that such an extension is unique; that is,

$$P[(a, b)] = \int_a^b f(x)dx = \sum_j \int_{I_j} f(x)dx$$

does not depend on the way interval (a, b) is partitioned into the finite union of nonoverlapping intervals I_j. This provides an extension of P to the smallest field of sets containing all intervals. If we show that P defined this way is continuous on the empty set, then we can claim that there exists an extension of P to the smallest σ-field of sets containing all intervals.

Now, the decreasing sequences of intervals converging to the empty set are built of two kinds of sequences: "shrinking open sets" and "escaping sets," exemplified as

$$I_1 \supset I_2 \supset \cdots \quad \text{with} \quad I_n = (a, a + \epsilon_n), \epsilon_1 > \epsilon_2 > \cdots \to 0$$

and

$$J_1 \supset J_2 \supset \cdots \quad \text{with} \quad J_n = (a_n, \infty), a_1 < a_2 \cdots \to \infty.$$

We have here $\lim I_n = \bigcap I_n = \emptyset$ and $\lim J_n = \bigcap J_n = \emptyset$. In the first case $P(I_n) = \int_a^{a+\epsilon_n} f(x)dx \leq \epsilon_n M \to 0$, where M is a bound for function f. In the second case, $P(J_n) = \int_{a_n}^{\infty} f(x)dx = 1 - \int_{-\infty}^{a_n} f(x)dx \to 1 - \int_{-\infty}^{+\infty} f(x)dx = 0$.

2.7 SUBJECTIVE PROBABILITY*

Let us finally consider briefly the third interpretation of probability, namely as a degree of certainty, or belief, about the occurrence of an event. Most often this probability is associated not so much with an event as with the truth of a proposition asserting the occurrence of this event.

The material of this section assumes some degree of familiarity with the concept of expectation, formally defined only in later chapters. For the sake of completeness, in the simple form needed here, this concept is defined below. In the presentation, we follow more or less the historical development, refining gradually the conceptual structures introduced. The basic concept here is that of a *lottery*, defined by an event, say A, and two objects, say a and b. Such a lottery, written simply aAb, will mean that the participant (X) in the lottery receives object a if the event A occurs, and receives object b if the event A^c occurs.

The second concept is that of *expectation* associated with the lottery aAb, defined as

$$u(a)P(A) + u(b)P(A^c), \tag{2.14}$$

where $u(a)$ and $u(b)$ are measures of how much the objects a and b are "worth" to the participant. When a and b are sums of money (or prices of objects a and b), and we put $u(x) = x$, the quantity (2.14) is sometimes called *expected value*. In cases where $u(a)$ and $u(b)$ are values that person X attaches to a and b (at a given moment), these values do not necessarily coincide with prices. We then refer to $u(a)$ and $u(b)$ as *utilities* of a and b, and the quantity (2.14) is called *expected utility* (EU). Finally, when in the latter case the probability $P(A)$ is the subjective assessment of likelihood of the event A by X, the quantity (2.14) is called *subjective expected utility* (SEU).

First, it has been shown by Ramsay (1926) that the degree of certainty about the occurrence of an event (of a given person) can be measured. Consider an event A, and the following choice suggested to X (whose subjective probability we want to determine). X is namely given a choice between the following two options:

1. *Sure* option: receive some fixed amount $\$u$, which is the same as lottery $(\$u)B(\$u)$, for any event B.

2. A *lottery* option. Receive some fixed amount, say \$100, if A occurs, and receive nothing if A does not occur, which is lottery $(\$100)A(\$0)$. One should expect that if u is very small, X will probably prefer the lottery. On the other hand, if u is close to \$100, X may prefer the sure option.

Therefore there should exist an amount u^* such that X will be indifferent between the sure option with u^* and the lottery option. With the amount of money as a representation of its value (or utility), the expected return from the lottery equals

$$0(1 - P(A)) + 100P(A) = 100P(A),$$

which, in turn, equals u^*. Consequently, we have $P(A) = u^*/100$. Obviously, under the stated assumption that utility of money is proportional to the dollar amount, the choice of \$100 is not relevant here, and the same value for $P(A)$ would be obtained if we chose another "base value" in the lottery option (this can be tested empirically).

This scheme of measurement may provide an assessment of the values of the (subjective) probabilities of a given person, for a class of events. It is of considerable interest that the same scheme was suggested in 1944 by von Neumann and Morgenstern (1944) as a tool for measuring utilities. They assumed that probabilities are known (i.e., the person whose utility is being assessed knows the objective probabilities of events, and his subjective and objective probabilities coincide). If a person is now indifferent between the lottery as above, and the sure option of receiving an object, say q, then the utility $u(q)$ of object q must equal the expected value of the lottery, which is $100P(A)$. This allows one to measure utilities on the scale that has a zero set on nothing (status quo) and "unit" as the utility of \$100. The scheme of von Neumann and Morgenstern was later improved by some authors, culminating with the theorem of Blackwell and Girshick (1954).

Still the disadvantages of both approaches were due to the fact that to determine utilities, one needed to assume knowledge of probabilities by the subject, while conversely, to determine subjective probabilities, one needed to assume knowledge of utilities. The discovery that one can determine *both* utilities and subjective probabilities of the same person is due to Savage (1954). We present here the basic idea of the experiment rather than formal axioms (to avoid obscuring the issue by technicalities).

Let $A, B, C, \ldots$ denote events, and let $a, b, c, \ldots$ denote some objects, whose probabilities $P(A), P(B), \ldots$ and utilities $u(a), u(b), \ldots$ are to be determined (keep in mind that both P and u refer to a particular person X, the subject of the experiment). We now accept the main postulate of the theory, that of the two lotteries, X will prefer the one that has higher SEU.

Suppose that we find an event A with subjective probability $1/2$, so that $P(A) = P(A^c) = 1/2$. If X prefers lottery aAb to lottery cAd, then

$$u(a)P(A) + u(b)P(A^c) > u(c)P(A) + u(d)P(A^c),$$

which means that

$$u(a) - u(c) > u(d) - u(b).$$

A number of experiments on selected objects will allow us to estimate the utilities, potentially with an arbitrary accuracy (taking two particular objects as zero and a

unit of the utility scale). In turn, if we know the utilities, we can determine the subjective probability of any event B. That is, if X is indifferent between lotteries aBb and cBd, we have

$$u(a)P(B) + u(b)(1 - P(B)) = u(c)P(B) + u(d)(1 - P(B)),$$

which gives

$$P(B) = \frac{u(d) - u(b)}{u(a) - u(b) + u(d) - u(c)}.$$

The only problem lies in finding an event A with subjective probability $1/2$. Empirically, an event A has subjective probability $1/2$ if, for any objects a and b, the person is indifferent between lotteries aAb and bAa. Such an event was found experimentally (Davidson et al., 1957). It is related to a toss of a die with three of the faces marked with the nonsense combination ZOJ, and the other three with the nonsense combination ZEJ (these combinations evoked the least number of associations).

Let us remark at this point that the system of Savage involves determining first an event with probability $1/2$, then the utilities, and then the subjective probabilities. Luce and Krantz (1971) suggested an axiom system (leading to an appropriate scheme) that allows *simultaneous* determination of utilities and probabilities. The reader interested in these topics is referred to the monograph by Krantz *et al.* (1971).

A natural question arises: Are the three axioms of probability theory satisfied here (at least in their finite versions, without countable additivity)? On the one hand, this is the empirical question: The probabilities of various events can be determined numerically (for a given person), and then used to check whether the axioms hold. On the other hand, a superficial glance could lead one to conclude that there is no reason why person X's probabilities should obey any axioms: After all, subjective probabilities that do not satisfy probability axioms are not logically inconsistent.

However, there is a reason why a person's subjective probabilities should satisfy the axioms. For any axiom violated by the subjective probability of X (and X accepts the principle of SEU), one could design a bet that appears favorable to X (hence a bet that he will accept), but yet the bet is such that X is sure to lose.

Indeed, suppose first that the probability of some event A is negative. Consider the bet (lottery) $(-c)A(-b)$, (i.e., a lottery in which X pays the sum c if A occurs, and pays the sum b if A does not occur). We have here (identifying, for simplicity, the amounts of money with their utilities)

$$SEU = -cP(A) - bP(A^c),$$

so that SEU is positive for a large enough c if $P(A) < 0$. Thus, following the principle of maximizing SEU, X should accept this lottery over the status quo (no bet) but he will lose in any case—the amount c or the amount b.

Suppose now that $P(\mathcal{S}) < 1$. Consider the bet $(-c)\mathcal{S}b$ whose SEU is $-cP(\mathcal{S}) + bP(\mathcal{S}^c)$. Since $P(\mathcal{S}^c) > 0$, making b large enough, the bet appears favorable to X, yet he is bound to lose the amount c on every trial.

If $P(\mathcal{S}) > 1$ or if the additivity axiom is not satisfied, one can also design bets that will formally be favorable for X (SEU will be positive) but that X will be bound to lose. Determination of these bets is left to the reader.

PROBLEMS

2.7.1 Peter and Tom attend the same college. One day Tom buys a ticket for a rock concert. Tickets are already sold out and are in great demand. Peter, who does not have a ticket, agrees to play the following game with Tom. For a fee of \$25, Peter will toss a coin three times, and receive the ticket if all tosses show up heads. Otherwise, for an additional fee of \$50, Peter will toss a coin two more times and receive the ticket if both tosses show up heads. If not, then for an additional fee of \$100, Peter will toss a coin and receive the ticket if the toss shows up heads. Otherwise, all money will be given to Tom, and he also will keep the ticket. Assuming that the coin is fair, subjective probabilities of various outcomes coincide with objective probabilities, and that Peter's utility is linear in money, show that Peter's utility of the ticket exceeds \$200.

2.7.2 Refer to Problem 2.7.1. Tom would agree on the following conditions: Peter pays him \$50 and tosses a coin, winning the ticket if it comes up heads, and otherwise losing \$50. In such situation, should they both agree that Peter buys the ticket from Tom for \$150?

2.7.3 Suppose that Tom is confronted with the choice between two options: O_1, which is simply to receive \$1,000,000, or O_2, which is to receive \$5,000,000 with probability 0.1, receive \$1,000,000 with probability 0.89, and receive \$0 with the remaining probability 0.01. After some deliberation Tom decides that $O1$ is better, mostly because the outcome \$0, unlikely as it may be, is very unattractive.

Tom is also confronted with a choice between two other options, O_3 and O_4. In O_3 he would receive \$5,000,000 with probability 0.1 and \$0 with probability 0.9. In O_4 he would receive \$1,000,000 with probability 0.11 and \$0 with probability 0.89. Here Tom prefers O_3: the "unattractive" option \$0 has about the same probability in both O_3 and O_4, while the positive outcome, although slightly less probable under O_3, is much more desirable in O_3 that in O_4.

Show that these preferences of Tom are not compatible with the assumption that he has utilities A, B, and C of \$5,000,000, \$1,000,000 and \$0, such that $A > B > C$ (This is known as Allais' paradox; Allais, 1953).

This Page Intentionally Left Blank

CHAPTER 3

COUNTING

3.1 INTRODUCTION

In the classical interpretation of probability all outcomes of the experiment are equally likely, and the probability of an event is obtained as the relative frequency of outcomes that favor this event (imply its occurrence). Simple enumeration of elements in these sets is often not feasible, and therefore practical implementation of this principle requires developing techniques for counting elements of certain sets (e.g., sets of all possible outcomes of an experiment). The branch of mathematics dealing with such methods is called *combinatorics*, or *combinatorial analysis*. In this chapter we introduce some combinatorial principles and illustrate their use in computing probabilities.

While we give here more than the typical material covered by textbooks on probability and statistics, a much more complete presentation of combinatorial methods and their applications to probability can be found in Feller (1968), a textbook thus far unsurpassed in its depth, elegance, and diversity of applications.

3.2 PRODUCT SETS, ORDERINGS, AND PERMUTATIONS

Consider two operations of some sort, that can be performed one after another. Leaving the notion of "operation" vague at the moment, we can make two assumptions:

1. The first operation can be performed in k_1 different ways.
2. For each of the ways of performing the first operation, the second operation can be performed in k_2 ways.

We have the following theorem:

Theorem 3.2.1 *Under assumptions* 1 *and* 2, *a two-step procedure consisting of a first operation followed by the second operation can be performed in* k_1k_2 *distinct ways.*

Proof. Observe that each way of performing the two operations can be represented as a pair (a_i, b_{ij}) with $i = 1, \ldots, k_1$ and $j = 1, \ldots, k_2$, where a_i is the ith way of performing the first operation and b_{ij} is the jth way of performing the second operation if the first operation was performed in ith way. All such pairs can be arranged in a rectangular array with k_1 rows and k_2 columns. □

We will now show some applications of Theorem 3.2.1.

■ EXAMPLE 3.1 Cartesian Products

One of the most common operations on sets is the Cartesian product. If A and B are two sets, their Cartesian product $A \times B$ is defined as the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$. For instance, if A consists of elements x and y while B consists of the digits 1, 2, and 3, then the Cartesian product $A \times B = \{x, y\} \times \{1, 2, 3\}$ contains the six pairs

$$\{(x, 1), (x, 2), (x, 3), (y, 1), (y, 2), (y, 3)\}. \tag{3.1}$$

Observe that the Cartesian product $A \times B$ is an operation quite distinct from the set-theoretical product $A \cap B$. For instance, in the above case, $A \cap B = \emptyset$, since A and B have no elements in common. Also, while $A \cap B = B \cap A$, for Cartesian products $A \times B \neq B \times A$ in general. Indeed, $B \times A$ contains, for example, the pair $(1, x)$, which does not belong to the set (3.1). Observe also that we may have $A = B$. The idempotence law $A \cap A = A$ *does not* hold for Cartesian products. Since the set $A \times A$ consists of *pairs* (x, y) with $x, y \in A$, elements of a different nature than elements of A, then $A \times A \neq A$. In cases when there is no danger of confusion, we will use the term *product* for *Cartesian product.*

Identifying now the first and second operation with "choice of an element from set A" and "choice of an element from set B," we obtain the following consequence of Theorem 3.2.1:

Theorem 3.2.2 (Multiplication Rule) *If* A_1 *and* A_2 *are finite sets consisting, respectively, of* k_1 *and* k_2 *elements, then the Cartesian product* $A_1 \times A_2$ *consists of* k_1k_2 *elements.*

Theorem 3.2.2 allows for an immediate generalization. Namely, we can define Cartesian products of more than two sets. Thus, if $A_1, \ldots, A_n$ are some sets, then

their Cartesian product $A_1 \times A_2 \times \cdots \times A_n$ is the set of all n-tuples $(a_1, a_2, \ldots, a_n)$ with $a_i \in A_i, i = 1, \ldots, n$. By easy induction, Theorem 3.2.2 can now be generalized as follows:

Theorem 3.2.3 *If $A_1, \ldots, A_n$ are finite, with A_i consisting of k_i elements $(i = 1, \ldots, n)$, then $A_1 \times \cdots \times A_n$ contains $k_1 \cdots k_n$ elements.*

■ **EXAMPLE 3.2**

The total number of possible initials consisting of three letters (name, middle name, family name) is 26^3. Each three-letter initial is an element of the set $A \times A \times A$, where A is the alphabet, so $k_1 = k_2 = k_3 = 26$. The total number of possible two- or three- letter initials is the number of the elements in the union $(A \times A) \cup (A \times A \times A)$, equal to $26^2 + 26^3 = 18,252$.

■ **EXAMPLE 3.3 License Plates**

Most states now use the system where a license plate has six symbols. One type (call it A) of such licenses has a prefix of three letters followed by a three-digit number (e.g., CQX 786). Other states use system (call it B) with a two-letter prefix, followed by a four-digit number (e.g., KN 7207). Still other states use system C, a digit and two letters, followed by a three digit number (e.g., 2CP 412). In addition the states try to augment their revenues by allowing (for a special fee) "personalized" plates CATHY3, MYCAR, and the like. Disregarding the personalized plates, which type of the license plate system can register most cars?

SOLUTION. Let A and D stand for the alphabet and for the set of 10 digits: $0, 1, \ldots, 9$. Then a license plate from system A can be regarded as an element of $A \times A \times A \times D \times D \times D$, while a license plate in system B is an element of the set $A \times A \times D \times D \times D \times D$. The numbers of elements in these Cartesian products are $26^3 \times 10^3$ and $26^2 \times 10^4$. The ratio is $26/10 = 2.6$, so in the state using an A system 2.6 times more cars can be registered than in the state with a B system.

Regarding system C, the answer depends whether or not 0 is allowed in the prefix. If the plate such as 0HY 314 is not allowed (e.g., because the digit 0 can be confused with the letter O), then the number of possible license plates is only $9 \times 26 \times 26 \times 10 \times 10 \times 10$, which is 10% less than the number of plates possible in states using system B. If 0 is allowed as the first character, then the numbers of plates of types B and C are the same.

In Examples 3.1 through 3.3 the set of ways of performing the second operations is the same regardless of which option was selected for the first operation. However, Theorem 3.2.1 remains true if the sets of ways of performing the second operation depend on the choice of the first operation. In particular, we can think of the first and second operation as two consecutive choices of an element from *the same set*, without returning the chosen elements. If the set, say A, has n elements, then the first operation (choice of an element) can be performed in n ways. If the chosen

element is not returned, then the second choice can be performed in $n - 1$ ways only, and we have the following:

Corollary 3.2.4 *The number of ordered pairs (x, y) with $x \neq y$ that can be formed out of n distinct elements of a set of size n is $n(n-1)$.*

Instead of thinking in terms of operations, we can still use Cartesian products here. Thus $A \times A$ has n^2 elements by Theorem 3.2.3, of which n are of the form (x, x). The number of pairs with elements distinct is $n^2 - n = n(n-1)$.

We can generalize these considerations as follows.

Definition 3.2.1 An ordered sequence of k elements selected without replacement from a set of n distinct elements $(n \geq k)$ is called a *permutation* of k out of n elements. □

Theorem 3.2.5 *The number of permutations of k out of n, denoted P_n^k, equals*

$$P_n^k = n(n-1)\cdots(n-k+1).$$

Proof: The argument here repeatedly uses the "operation" principle: the first choice can be made in n ways, the second in $n-1$ ways, the kth in $n-(k-1) = n-k+1$ ways. □

If $k = n$, consecutive choices form an ordering of the entire set of size n. We obtain the following:

Corollary 3.2.6 *The set of n elements can be ordered in*

$$P_n^n = n(n-1)\cdots 2 \times 1 \tag{3.2}$$

distinct ways.

The product (3.2) occurs often and has a special symbol:

$$n! = 1 \times 2 \times \cdots \times (n-1)n$$

to be read "n factorial." We have therefore

$$P_n^k = \frac{n!}{(n-k)!}. \tag{3.3}$$

For a reason that will become apparent later, we adopt the convention

$$0! = 1. \tag{3.4}$$

■ **EXAMPLE 3.4**

The letters I, I, I, I, M, P, P, S, S, S, S are arranged at random. What is the probability that the arrangement will spell MISSISSIPPI?

SOLUTION. We can solve this problem treating the choices of consecutive letters as "operations." The first operation must give the letter M; hence there is only one way of choosing it. The next letter (out of the remaining 10) must be an I, and it can be selected in 4 ways. Proceeding in this way, the sequence of consecutive 11 choices leading to the word MISSISSIPPI can be performed in $1 \times 4 \times 4 \times 3 \times 3 \times 2 \times 1 \times 2 \times 2 \times 1 \times 1$ ways, which equals 4!4!2!1!. On the other hand, the total number of ways one can perform the operations of consecutively choosing letters from the set is 11!. Consequently, the required probability equals

$$p = \frac{4!4!2!1!}{11!}. \tag{3.5}$$

In this solution the letters are regarded as distinguishable, as if we had four letters S, labeled S_1, S_2, S_3, and S_4, and similarly for the other letters. In this case, the numerator and denominator are, respectively, the number of ways one can order the set of distinguishable letters so as to form the word MISSISSIPPI and the total number of orderings. Alternatively, one can regard the identical letters as indistinguishable, and in this case we have only one way of ordering them so as to spell the required word, and a total of $11!/(4!4!2!1!)$ distinguishable ways of ordering these letters. Indeed, the denominator here represents the number of ways of permuting letters so as to leave the arrangement invariant. Now

$$p = \left(\frac{11!}{4!4!2!1!}\right)^{-1},$$

which is the same as (3.5).

■ **EXAMPLE 3.5 Birthday Problem**

The following problem has a long tradition and appears in most probability textbooks. If r randomly chosen persons attend a party, what is the probability p_r that none of them will have a birthday on the same day?

SOLUTION. Here we make the following assumption: (1) all years have 365 days (i.e., leap years are disregarded), (2) each day is equally likely to be a birthday of a person (i.e., births occur uniformly throughout the year), and (3) no twins attend the party. To compute p_r, we must find the number of all possible ways in which birthdays can be allocated to r people, and the number of such allocations in which birthdays do not repeat. The first number is 365^r by virtue of Theorem 3.2.3, while the second number is P_{365}^r (assuming $r \leq 365$; if $r > 365$, we must have at least one birthday repeating, so $p_r = 0$). Thus for $r \leq 365$ we have

$$\begin{aligned} p_r &= \frac{P_{365}^r}{365^r} = \frac{365}{365} \times \frac{365-1}{365} \times \cdots \times \frac{365-r+1}{365} \\ &= \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{r-1}{365}\right). \end{aligned}$$

As first approximation, neglecting all products which have denominators of order 365^2 or higher, we can take

$$p_r \approx 1 - \frac{1+2+\cdots+(r-1)}{365} = 1 - \frac{(r-1)r}{730}. \tag{3.6}$$

This approximation works quite well for small r. To get a better approximation, we can use the formula $\log(1-x) \approx -x$ so that

$$\begin{aligned} \log p_r &= \log(1-\frac{1}{365}) + \cdots + \log(1-\frac{r-1}{365}) \\ &\approx -\frac{1}{365} - \cdots - \frac{r-1}{365} = -\frac{r(r-1)}{730}; \end{aligned}$$

hence

$$p_r \approx e^{-\frac{r(r-1)}{730}}. \tag{3.7}$$

It is interesting that for $r = 23$ a repeated birthday is about as likely as no repetition. The smallest r for which p_r is less than 0.01 is 56.

PROBLEMS

3.2.1 A certain set contains n distinct elements. Find n if the number of: **(i)** All possible permutations of length 2 equals 90. **(ii)** Permutations of length 3 is 10 times larger than the number of permutations of length 2.

3.2.2 A skyscraper is 40 stories tall. Five people enter the elevator on the first floor. Assuming each person is equally likely to get off at any of the 39 floors $2, 3, \ldots, 40$, what is the probability that all people will get off at different floors? Find the exact value, and then derive and compute the approximations analogous to (3.6) and (3.7).

3.2.3 A two letter code is to be formed by selecting (without replacement) the letters from a given word. Find the number of possible codes if the word is: **(i)** CHART. **(ii)** ALOHA. **(iii)** STREET.

3.2.4 Determine the number of 0's at the end of 16! and 27!.

3.2.5 Seated at random in a row of n seats are n people, among them John and Mary. Find the probability that: **(i)** John sits next to Mary. **(ii)** John sits next to Mary on her right. **(iii)** John sits somewhere to the right of Mary. **(iv)** John and Mary sit exactly two seats apart.

3.2.6 Seated at random at a round table with n seats are n people, among them John and Mary. **(i)** Answer questions (i)–(iv) of Problem 3.2.5. Anything peculiar about the answer to (iii)? **(ii)** Assume that $n = 2k$. Find the probability that John and Mary sit facing each other (e.g., numbers 1 and 7 on the clock).

3.2.7 Five men and five women are to be seated in a row of ten chairs. Find the number of possible arrangements if: **(i)** The men are required to sit in alternating seats. **(ii)** No two men are to be seated next to each other.

3.2.8 12 girls and 17 boys go to a dance. **(i)** How many possible dancing pairs (boy-girl) may be formed? **(ii)** The dance floor can accommodate at most 11 pairs at a time. If each dance lasts 10 minutes and is followed by a 2 minute break, how much time, at least, will elapse before each boy will have danced with each girl at least once? **(iii)** Answer the same question as in (ii) if the dance floor can accommodate 15 pairs at a time.

3.2.9 Susan has five dresses, three skirts, four blouses, three pairs of shoes, and two hats. She always wears shoes, and either a dress or a blouse and a skirt. She may or may not wear a hat. **(i)** How many different combinations can she wear? **(ii)** Suppose Susan can afford buying either a dress or a hat (but not both). What should she buy to maximize the number of different combinations that she can wear? **(iii)** Suppose that Susan's brown shoes do not match her pink or blue dress, and that the blue hat does not match her yellow blouse. How many matching combinations can she wear?

3.2.10 A restaurant menu has five appetizers, three soups, 15 entrees, and three desserts. **(i)** Assuming you are going to order one item from each group, how many possible dinners can you order? **(ii)** Assume you come to the restaurant with a friend who is as hungry as you are. How many different orders for two full dinners can you place if your friend's choice is not necessarily the same as yours? **(iii)** Answer the question in part (ii) under the constraint that you do not order the same entree and the same dessert as your friend (but the soup and/or appetizer may be the same).

3.2.11 Express the product of odd integers $1 \times 3 \times \cdots \times (2n+1)$ using factorials. This product is sometimes denoted $(2n+1)!!$. [*Hint:* Start with an easier task of using factorials to express the product of even integers $2 \times 4 \times \cdots \times (2n)$.]

3.2.12 Let e_r be the probability that exactly two people in a group of r have the same birthday, and let p_r be the probability that everybody in the group has a different birthday. **(i)** Show that

$$e_r = \frac{r(r-1)}{730} p_{r-1}.$$

(ii) Use the multiplication rule (Theorem 3.2.2) to show that

$$e_r = \frac{r(r-1)}{2} \times \frac{365 \times 364 \times \cdots \times (365-r+2)}{365^r}.$$

3.2.13 Find the number of three-digit integers (i.e., integers between 100 and 999) that have all digits distinct. How many of them are odd?

Problems 3.2.14 through 3.2.20 concern the concept of voting power, as introduced by Shapley and Shubik (1954). Although this concept is not directly related to probability, it provides a good practice in combinatorial analysis. Consider a voting body (a committee, executive board of a company, the US Congress, etc.). Assume, for simplicity, that each member must vote "yes" or "no" (none abstaining). The voting rule specifies whether or not the issue passes, for any configuration of votes of the members (some members will have more than one vote, some will have veto

power, etc.). To define the voting power of various members of the voting body, consider a specific permutation of voting members, and imagine that they all cast a "yes" vote in the order specified by this permutation. There will then be a moment when the issue will pass, regardless of the votes of the remaining members. The last person casting his or her vote before this moment occurs is called a *pivot*. This way in each permutation of voters, exactly one member is a pivot, and the voting power of member x is defined as

$$\pi(x) = \frac{\text{number of permutations in which } x \text{ is a pivot}}{\text{number of all permutations}}. \tag{3.8}$$

Take the example of a committee that consists of four people A, B, C, and D. A has two votes, and the others have one vote each. A simple majority (at least three votes) is needed to pass the issue. In this case the pivot is the person who (in a given permutation) casts the third vote. Thus A will be the pivot in all permutations in which he is in second or third place, such as BACD or DBAC. There are 12 such permutations, that is $\pi(A) = 12/4! = 1/2$. Nevertheless, B will be the pivot if he either appears on the third place in a permutation, provided that A appears in a later place (i.e., in fourth place), or if he appears in second place following A. There are four such permutations: CDBA, DCBA, ABCD, and ABDC, that is π(B) = 4/24 = $1/6$. By symmetry, π(C) =π(D) = $1/6$, and we see that (in this case) having twice as many votes as the others gives three times as much power in making a decision.

3.2.14 Find the power of A in a committee of n persons with A having two votes and every other member having one vote. Assume that simple majority is needed to carry the issue.

3.2.15 Generalize the situation of Problem 3.2.14 assuming that A has k votes (other conditions being the same). What is the smallest k for which A is a dictator (i.e., every other member has zero power)?

3.2.16 Determine the voting powers in a committee consisting of six persons, with A having three votes, B having two votes, all others having one vote, and again, a simple majority being required to pass the issue.

3.2.17 **(i)** Determine the voting power in a committee of five persons, each having one vote, with a simple majority needed to pass the issue, under the additional condition that A has veto power (i.e., positive vote of A is necessary to pass the issue). **(ii)** Assume, in addition, that B has two votes. Is it better to have one vote and veto power or two votes without it?

3.2.18 Find the voting powers in a voting body of $2n$ members, among them the chairperson. Everyone (including the chairperson) has one vote, and the majority carries the issue. In case of a tie, the chair's vote prevails.

3.2.19 The United Nations Security Council consists of five permanent members (China, France, the Russian Federation, the United Kingdom, and the United States) and 10 nonpermanent members. For an issue to pass, it must receive a unanimous vote of all five permanent members (each has veto power) and of at least four nonpermanent members. Disregard the possibility of abstaining from a vote, and determine the voting power of each member of the UN Security Council.

3.2.20 Assume that in a corporation the number of votes of a shareholder equals the number of his shares. If a simple majority is required, determine the voting powers in a corporation in which there are three shareholders, A with 500,000 shares, B with 499,999 shares, and C with 1 share.

3.3 BINOMIAL COEFFICIENTS

The permutations considered in Section 3.2 concerned the ordered choices from a certain set. Often the order in which the elements are selected is not relevant, and we are interested only in the total number of possible choices, regardless of the particular order in which they are obtained. Such choices are referred to as *combinations*. We have the following definition.

Definition 3.3.1 A subset of size k selected from a set of size n (regardless of the order in which this subset was selected) is called a *combination* of k out of n.

Theorem 3.3.1 *The number of combinations of k out of n, C_n^k, is given by*

$$C_n^k = \frac{P_n^k}{k!}. \tag{3.9}$$

Proof. By Theorem 3.2.5 we have P_n^k different permutations of k out of n elements. Each permutation determines the set of k elements selected and their order. Consequently, $k!$ permutations lead to the same combination, which proves (3.9). □

The ratio $P_n^k/k!$ appears in various contexts, and it is convenient to have a special symbol for it.

Definition 3.3.2 The ratio

$$\frac{P_n^k}{k!} = \frac{n(n-1)\cdots(n-k+1)}{k!} \tag{3.10}$$

is called a *binomial coefficient* and is denoted by $\binom{n}{k}$, to be read as "n choose k." □

Using (3.3), we have

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{3.11}$$

Observe, however, that (3.11) requires n to be an integer, whereas in definition (3.10) n can be any real number (k has to be an integer in both cases). We will make use of this distinction later; in this section we will tacitly assume that n is an integer with $n \geq k$.

Observe also that the symbol $\binom{n}{k}$ makes sense for $k = 0$ and $k = n$, in view of the convention that $0! = 1$. Thus we have

$$C_n^k = \binom{n}{k} \tag{3.12}$$

for all integers k, n such that $n \geq k$. For $k = 0$ we have $C_n^0 = 1$, since there is only one empty set, and $C_n^n = 1$, since only one set of size n can be selected out of a set

of size n. Formula (3.12) gives correct values, namely

$$\binom{n}{n} = \binom{n}{0} = 1. \tag{3.13}$$

We will proceed now to study some properties of the binomial coefficients $\binom{n}{k}$. First, note that

$$\binom{n}{k} = \binom{n}{n-k}, \tag{3.14}$$

which follows at once from the symmetry in formula (3.11). One can also prove (3.14) by observing that choosing a set of size k is equivalent to "leaving out" a set of size $n-k$. The number of different sets of size k chosen must be equal to the number of different sets of size $n-k$ "chosen" by leaving them out.

We will now prove the following theorem:

Theorem 3.3.2 (Pascal's Triangle) *The binomial coefficients satisfy the relation*

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}. \tag{3.15}$$

Proof. The formula can be easily proved by "brute force," expressing the left-hand side using (3.9) and reducing it algebraically to get the right-hand side. It is, however, much more instructive to use the following argument, relying on the principal interpretation of the coefficients $\binom{n}{k}$ as C_n^k. The right-hand side of (3.15) counts the number of sets of size k that can be chosen out of a set of size $n+1$. Let us take one element of the latter set and label it somehow. We have then a set of n unlabeled and 1 labeled element. Each subset of size k is of one of the following two categories: (1) subsets that contain only k unlabeled elements, or (2) subsets that contain $k-1$ unlabeled elements and one labeled element. Clearly, the two terms on the left-hand side of (3.15) count the numbers of subsets of the first category and of the second category. □

The name *Pascal's triangle* is connected with a way of computing the coefficients $\binom{n}{k}$, which are useful for small values of n. We build Pascal's triangle starting with the top row (counted as the zeroth row), which consists of the single number 1 (see Figure 3.1). Then we obtain each number in the subsequent rows as a sum of two numbers directly above it (as marked with arrows in the fifth row). The consecutive numbers in the nth row are, reading from the left, the values of

$$\binom{n}{0}, \binom{n}{1}, \binom{n}{2}, \ldots$$

so that, for example, $\binom{6}{3} = 20$, as marked on the triangle in Figure 3.1.

The name *binomial coefficient* is connected to the following formula:

Theorem 3.3.3 (Newton's Binomial Formula) *For any positive integer n and real x, y we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k. \tag{3.16}$$

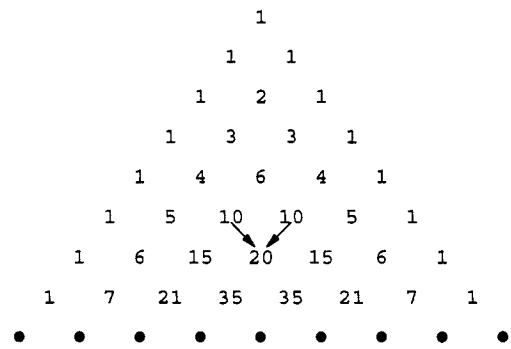


Figure 3.1 Pascal's triangle

Proof. We will prove the theorem by induction. For $n = 1$ the right-hand side equals $\binom{1}{0}x + \binom{1}{1}y = x + y$. Assume now the assertion holds for some n, and multiply both sides of (3.16) by $(x + y)$. Then

$$\begin{aligned}(x+y)^{n+1} &= (x+y)\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k \\ &= \sum_{k=0}^{n}\binom{n}{k}x^{n+1-k}y^k + \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^{k+1} \\ &= \sum_{k=0}^{n}\binom{n}{k}x^{n+1-k}y^k + \sum_{k=1}^{n+1}\binom{n}{k-1}x^{n-(k-1)}y^k.\end{aligned}$$

Separating the term for $k = 0$ in the first sum, and the term for $k = n + 1$ in the last sum, we may write

$$\begin{aligned}(x+y)^{n+1} &= x^{n+1} + \sum_{k=1}^{n}\left[\binom{n}{k} + \binom{n}{k-1}\right]x^{(n+1)-k}y^k + y^{n+1} \\ &= \sum_{k=0}^{n+1}\binom{n+1}{k}x^{(n+1)-k}y^k,\end{aligned}$$

where the last equality is due to Theorem 3.3.2. □

We will now prove the following theorem:

Theorem 3.3.4 *The binomial coefficients satisfy the identities*

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n \tag{3.17}$$

and

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots \pm \binom{n}{n} = 0. \tag{3.18}$$

Proof. It suffices to consider the expansion (3.16) of $(1 + 1)^n$ and $(1 - 1)^n$, leading directly to (3.17) and (3.18). Observe that (3.17) can be shown using C_n^k as a number of distinct subsets of size k that can be chosen out of a set of n elements. The left-hand side of (3.17) equals the total number of all subsets that can be chosen out of a set of size n, including the empty subset and the whole set. The number of all subsets can also be computed differently: we may visualize the process of forming a subset as a process of deciding about each of the elements of the set whether or not to include it in the subset being formed. Each decision here can be made in two ways, and there are n decisions altogether. So the total number of distinct ways of making the string of n decisions is 2^n by Theorem 3.2.3. □

We also have the following theorem:

Theorem 3.3.5 *For every* $n = 1, 2, \ldots$ *and every* $k = 0, 1, \ldots, n$ *the binomial coefficients satisfy the relation*

$$\binom{n}{0}\binom{n}{k} + \binom{n}{1}\binom{n}{k-1} + \cdots + \binom{n}{k}\binom{n}{0} = \binom{2n}{k}. \tag{3.19}$$

Proof. Consider the product $(1 + x)^n(1 + x)^n = (1 + x)^{2n}$. Expanding the right-hand side, we obtain

$$\sum_{k=0}^{2n} \binom{2n}{k} x^k, \tag{3.20}$$

while the left-hand side equals

$$\left[\sum_{i=0}^{n} \binom{n}{i} x^i\right] \times \left[\sum_{j=0}^{n} \binom{n}{j} x^j\right]. \tag{3.21}$$

For $k \leq n$, comparison of the coefficients of x^k in (3.20) and (3.21) gives (3.19). □

As a consequence of (3.19) we obtain a corollary to the theorem.

Corollary 3.3.6

$$\sum_{j=0}^{n} \binom{n}{j}^2 = \binom{2n}{n}.$$

Proof. Take $k = n$ in (3.19) and use the fact that

$$\binom{n}{n-i} = \binom{n}{i}.$$

□

Below we present some examples of the use of binomial coefficient in solving various probability problems, some with a long history.

■ **EXAMPLE 3.6**

In probability theory one often considers a choice without replacement from a finite set containing two categories of objects. If n balls are to be selected

from an urn containing r red and b blue balls, one might want to know the probability that there will be exactly k red balls chosen.

SOLUTION. We apply here the "classical" definition of probability. The choice of n objects without replacement is the same as choosing a subset of n objects from the set of total of $r + b$ objects. This can be done in $\binom{r+b}{n}$ different ways. Since we must have k red balls, this choice can be made in $\binom{r}{k}$ ways. Similarly, $n - k$ blue balls can be selected in $\binom{b}{n-k}$ ways. Clearly, each choice of k red balls can be combined with each of the $\binom{b}{n-k}$ choices of blue balls so that, by Theorem 3.2.2, the total number of choices is the product

$$\binom{r}{k}\binom{b}{n-k}.$$

Consequently the probability in question is

$$P(\text{exactly } k \text{ red balls}) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{r+b}{n}}. \tag{3.22}$$

The next example shows an interesting application of formula (3.22).

■ **EXAMPLE 3.7**

Consider the problem of estimating the number of fish in a lake (the method described below is also used to estimate the sizes of bird or wildlife populations). The lake contains an unknown number N of fish. To estimate N, we first catch c fish, label them, and release them back into the lake. We assume here that labeling does not harm fish in any way, that the labeled fish mix with unlabeled ones in a random manner, and that N remains constant (in practice, these assumptions may be debatable). We now catch k fish, and observe the number, say x, of labeled ones among them. The values c and k are, at least partially, under the control of the experimenter. The unknown parameter is N, while x is the value occuring at random, and providing the key to estimating N. Let us compute the probability $P_N = P_N(x)$ of observing x labeled fish in the second catch if there are N fish in the lake. We may interpret fish as balls in an urn, with labeled and unlabeled fish taking on the roles of red and blue balls. Formula (3.22) gives

$$P_N = \frac{\binom{c}{x}\binom{N-c}{k-x}}{\binom{N}{k}}. \tag{3.23}$$

To estimate N, we can use the principle of *maximum likelihood*, to be explored in detail in Chapter 12. At present, it suffices to say that this principle suggests using as $\hat{N}$, an estimator of N, the value of N that maximizes (3.23). Let us call this value $\hat{N}$. It depends on the observed value x and hence is itself random. Thus $\hat{N}$ is defined by the condition

$$P_{\hat{N}} \geq P_N \quad \text{for all } N$$

and our objective is to find the maximizer of P_N. Since N is a discrete variable, we cannot use methods of finding a maximum based on derivatives. Instead, the method that works in this case is based on the observation that if the function P_N has a maximum (possibly local) at N^*, then $P_{N^*}/P_{N^*-1} > 1$ and $P_{N^*+1}/P_{N^*} < 1$. If at two neighboring arguments the values are equal, the ratio equals 1. Consequently, we should study the ratio P_N/P_{N-1} and find all arguments at which this ratio crosses the threshold 1. After some reduction we have

$$P_N/P_{N-1} = \frac{(N-c)(N-k)}{N(N-c-k+x)}.$$

The above ratio always exceeds 1 if $x = 0$, so in this case the maximum is not attained. Assume now that $x > 0$. The inequality

$$P_N/P_{N-1} \geq 1$$

is equivalent to

$$N \leq \frac{kc}{x}, \tag{3.24}$$

with the equality occuring if and only if $P_N/P_{N-1} = 1$. Thus, the maximum is attained at[5]

$$\hat{N} = \left[\frac{kc}{x}\right],$$

and also at $kc/x - 1$ if the latter value is an integer. Let us observe that the result above is consistent with common intuition: The proportion of labeled fish in the whole lake is c/N, and it should be close to the proportion x/k of labeled fish in the second catch. This gives the approximate equation $c/N \approx x/k$, with the solution $N \approx kc/x$.

■ **EXAMPLE 3.8**

To supplement their revenues, many states are sponsoring number games, or lotteries. The details vary slightly from state to state, but generally, a player who buys a lottery ticket chooses several numbers from a specified set of numbers. We will carry the calculations for the choice of 6 out of 50 numbers $1, 2, \ldots, 50$, which is quite typical. After the sales of tickets close, six winning numbers are chosen at random from the set $1, 2, \ldots, 50$. All those (if any) who chose six winning numbers share the Big Prize; if there are no such winners, the Big Prize is added to the next week's Big Prize. Those who have five winning numbers share a smaller prize, and so on. Let $P(x)$ be the probability that a player has exactly x winning numbers. We will compute $P(x)$ for $x = 6, 5, 4$, and 3. The calculations would be the same if the winning numbers were chosen in advance but remained secret to the players. We can now represent the situation in a familiar scheme of an urn with 6 winning numbers and 44 losing numbers, and the choice of 6 numbers from the urn (without replacement). This is the same problem as that of labeled fish. The total number

[5] The integer part of a, $[a]$, is the largest integer not exceeding a ($[3.21] = 3$, $[-1.71] = -2$, etc.)

of choices that can be made is $\binom{50}{6}$, while $\binom{6}{x}\binom{44}{6-x}$ is the number of choices with exactly x winning numbers. Thus,

$$P(x) = \frac{\binom{6}{x}\binom{44}{6-x}}{\binom{50}{6}}.$$

For $x = 6$ we have

$$P(6) = \frac{1}{\binom{50}{6}} = \frac{6!}{50 \times 49 \times 48 \times 47 \times 46 \times 45} = \frac{1}{15,890,700} = 6.29 \times 10^{-8}.$$

Similarly $P(5) = 1.66 \times 10^{-5}$, $P(4) = 8.93 \times 10^{-4}$ and $P(3) = 0.016669$.

Thus the chances of winning a share in the Big Prize are about one in 16 million. It would therefore appear that there should be, on average, one big winner in every 16 million tickets sold. The weekly numbers of tickets sold are well known, and it turns out that the weekly numbers of winners (of the Big Prize) vary much more than one would expect. For example, in weeks where the number of tickets sold is about 16 million, one could expect no winner, one winner, or two winners; three winners is unlikely. In reality, it is not at all uncommon to have five or more winning tickets in a week with 16 million tickets sold. These observations made some people suspicious about the honesty of the process of drawing the numbers, to the extent that there have been attempts to bring suit against the lottery (e.g., accusing the organizers of biasing the lottery balls with certain numbers so as to decrease their chance of being selected, thus favoring some other numbers).

Actually, the big variability of weekly numbers of winners is to be expected if one realizes that these numbers depend on *two* chance processes: the choice of winning numbers from the urn (which may be, and probably is, quite fair) and the choice of numbers by the players. This choice is definitely not uniform. It favors certain combinations, which seem more "random" to the naive persons than other choices. For instance, the combination $1, 2, 3, 4, 5, 6$ appears less likely than (say) $5, 7, 19, 20, 31$, and 45. As a consequence some combinations are selected more often by the players than others. Each combination has the same chance of being the winning one, but some may have higher numbers of winners associated with them. This point can be illustrated by the following analogy: Imagine that each week a first name (Mary, Susan, George, etc.) is chosen at random from the set of all names used, and all persons in the state with the selected name share the prize. The chances of being chosen are the same for John as for Sebastian, as they depend on the process of sampling names of winners. But if the name Sebastian is chosen, each Sebastian will share the prize with many fewer other winners than if the name John were selected. Here the numbers of winners to share the prize depend on another process, namely that of parents selecting names for their children.

■ **EXAMPLE 3.9**

We have k urns, labeled $1, \ldots, k$, and n identical (indistinguishable) balls. In how many ways can these balls be distributed in k urns?

SOLUTION. There are no restrictions here on the number of balls in an urn, or the number of empty urns. To get the answer, let us identify each possible allocation with a string of $k+1$ bars and n circles, of the form

$$|\bigcirc\bigcirc||\bigcirc\bigcirc\bigcirc\bigcirc|\bigcirc|\cdots\bigcirc|,$$

with the only condition being that the string should start and end with a bar. The spaces between bars represent urns. Thus in the arrangement above the first urn contains 2 balls, the second none, the third 4 balls, and so on. Clearly, the number of distinct arrangements equals $\binom{n+k-1}{n}$—the number of distinct arrangements of $k-1$ bars and n circles. Indeed, we have a string of $n+k-1$ symbols (not counting the two extreme bars), and each arrangement is obtained by specifying n places for the symbol $\bigcirc$.

Example 3.9 shows that the binomial coefficient can be interpreted in two ways. On the one hand, $\binom{a+b}{a}$ is the number C_{a+b}^{a} of distinct sets of size a that can be chosen out of a set of size $a+b$. On the other hand, $\binom{a+b}{a}$ is also the number of distinct strings of a indistinguishable elements of one kind and b indistinguishable elements of another kind. To see this, it suffices to think of the string as being determined by the choice of "a out of total of $a+b$" slots into which we assign elements of the first kind.

■ EXAMPLE 3.10 Matching Problem

A secretary typed n letters and addressed n envelopes. For some reason the letters were put into envelopes at random. What is the probability of at least one match, that is, of at least one letter being put into the correct envelope?

SOLUTION. This problem appears in almost every textbook on probability under various formulations (e.g., of guests receiving their hats at random). One could expect the probability of at least one match to vary greatly with n. However, the contrary is true: this probability is almost independent of n. Let A_i be the event that ith letter is placed in the correct envelope. Using formula (2.7), we have

$$\begin{aligned} P(\text{at least one } A_i) &= P(A_1 \cup \cdots \cup A_n) \\ &= \sum P(A_i) - \sum_{i<j} P(A_i \cap A_j) \\ &\quad + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots \pm P(A_1 \cap \cdots \cap A_n). \end{aligned}$$

By symmetry, the probability of each intersection depends only on the number of events in the intersection,[6] so we let p_r denote the probability of the intersection of r events, $p_r = P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_r})$. Clearly, the numbers of

[6] This property, called exchangeability of events, will be discussed in more detail in Section 4.6.

terms in the consecutive sums are

$$\binom{n}{1}, \binom{n}{2}, \binom{n}{3}, \cdots$$

and

$$P(\text{at least one } A_i) = \binom{n}{1}p_1 - \binom{n}{2}p_2 + \binom{n}{3}p_3 - \cdots \pm \binom{n}{n}p_n. \tag{3.25}$$

To evaluate p_r, we can argue as follows: Assume that the envelopes are ordered in some way. The total number of ways one can order n letters is $n!$. If specific r events, say $A_{i_1}, \ldots, A_{i_r}$, are to occur (perhaps in conjunction with other events), then the letters number $i_1, \ldots, i_r$ must be at their appropriate places in the ordering (to match their envelopes). The remaining $n - r$ letters can appear in any of the $(n - r)!$ orders. Thus

$$p_r = \frac{(n-r)!}{n!}.$$

Consequently, the rth term in the sum (3.25) equals (up to the sign)

$$\binom{n}{r}\frac{(n-r)!}{r!} = \frac{1}{r!},$$

and we obtain

$$P(\text{at least one match}) = \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \cdots \pm \frac{1}{n!}.$$

Since

$$e^{-1} = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!},$$

we have

$$P(\text{at least one match}) \approx 1 - \frac{1}{e},$$

with the accuracy increasing as $n \to \infty$. The approximation is actually quite good for small n. The limiting value is 0.63212056, while the exact values of the probability π_n of at least one match for selected values of n are:

$$\begin{aligned}
\pi_1 &= 1 \\
\pi_2 &= 1 - \frac{1}{2} = 0.5 \\
\pi_3 &= 1 - \frac{1}{2} + \frac{1}{6} = 0.6666667 \\
\pi_4 &= 1 - \frac{1}{2} + \frac{1}{6} - \frac{1}{24} = 0.625 \\
\pi_5 &= 1 - \frac{1}{2} + \frac{1}{6} - \frac{1}{24} + \frac{1}{120} = 0.6333333 \\
\pi_6 &= 1 - \frac{1}{2} + \frac{1}{6} - \frac{1}{24} + \frac{1}{120} - \frac{1}{720} = 0.6319444 \\
\pi_7 &= 1 - \frac{1}{2} + \frac{1}{6} - \frac{1}{24} + \frac{1}{120} - \frac{1}{720} + \frac{1}{5040} = 0.6321429 \\
\pi_8 &= 1 - \frac{1}{2} + \frac{1}{6} - \frac{1}{24} + \frac{1}{120} - \frac{1}{720} + \frac{1}{5040} - \frac{1}{40320} = 0.6321181.
\end{aligned}$$

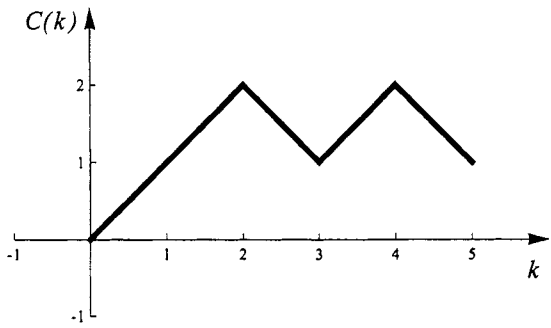


Figure 3.2 Process of counting votes

■ EXAMPLE 3.11 Ballot Problem

Suppose that in an election, candidate A receives a votes while candidate B receives b votes, where $a > b$. Assuming that votes are counted in random order, what is the probability that during the whole process of counting A will be ahead of B?

SOLUTION. Note that other votes, if any, do not matter, and we may assume that $a + b$ is the total number of votes. The process of counting votes is determined by the arrangement of the votes, that is, the arrangement of a symbols A, and b symbols B. Clearly, such an arrangement is uniquely determined by specifying the locations of the A's (or, equivalently, B's). It might be helpful to use a graphical representation: define the function $C(k)$ as the net count for candidate A after inspection of k votes. Thus, if in the first k votes we had r votes for A and $k - r$ votes for B, then $C(k) = r - (k - r) = 2r - k$. We can then represent the process of counting as a polygonal line that starts at the origin and has vertices $(k, C(k)), k = 1, \ldots, a + b$ (see Figure 3.2).

In Figure 3.2 we have the beginning of counting, when the first five votes inspected are AABAB. The problem can now be formulated as finding the probability that the counting function $C(x)$ lies above the x-axis for all $x = 1, 2, \ldots, a + b$. Observe that the first vote counted must be for A (as in Figure 3.2); this occurs with probability $a/(a + b)$.

The remaining votes will give a polygonal line leading from $(1, 1)$ to $(a + b, a - b)$, and we must find the number of such lines that will never touch or cross the x-axis. The number of such lines is equal to the total number of lines from $(1, 1)$ to $(a + b, a - b)$ minus the number of lines from $(1, 1)$ to $(a + b, a - b)$ which touch or cross the x-axis. The total number of lines leading from $(1, 1)$ to $(a + b, a - b)$ is $\binom{a+b-1}{a-1}$, since each such line has $a - 1$ steps "up" and b steps "down," which can be ordered in any manner. Thus it remains to count the number of lines from $(1, 1)$ to $(a + b, a - b)$ that touch or cross the x-axis. Let V be the set of all such lines. Each line in V must touch the x-axis for the first time at some point, say t (see Figure 3.3). If we reflect the part of this line that lies to the left of t with respect to x-axis, we obtain a

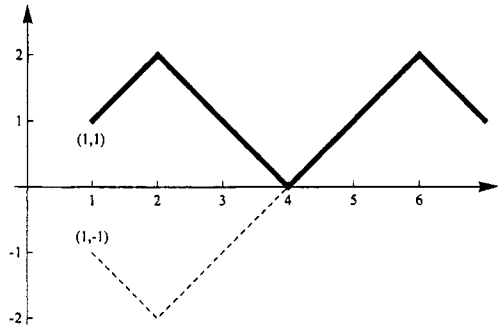


Figure 3.3 Reflection principle

line leading from $(1, -1)$ to $(a+b, a-b)$. Moreover, different lines in V will correspond to different lines leading from $(1, -1)$ to $(a+b, a-b)$, and each line in the latter set will be obtained from some line in V. This means that the set V has the same number of lines as the set of lines leading from $(1, -1)$ to $(a+b, a-b)$. But the latter set contains $\binom{a+b-1}{a}$ lines, since each such line must have a steps "up" and $b-1$ steps "down." Consequently, the required probability equals

$$p = \frac{a}{a+b} \times \frac{\binom{a+b-1}{a-1} - \binom{a+b-1}{a}}{\binom{a+b-1}{a-1}} = \frac{a-b}{a+b}.$$

■ **EXAMPLE 3.12 Poker**

We now consider the probabilities of several poker hands (some students will probably say that finally the book gives some useful information).

In poker, five cards are dealt to a player from a standard deck of 52 cards. The number of possible hands is therefore $\binom{52}{5} = 2,598,960$. The lowest type of hand is that containing one pair (two cards of the same denomination, plus three unmatched cards). To find the number of possible hands containing one pair, one can think in terms of consecutive choices leading to such hand:

(a) The denomination of the cards in a pair can be chosen in $\binom{13}{1}$ ways.

(b) The suits of the pair can be chosen in $\binom{4}{2}$ ways.

(c) The choice of denominations of the remaining three cards can be made in $\binom{12}{3}$ ways.

(d) The suits of those three cards may be chosen in 4^3 ways. Altogether, combining (a)–(d), we have

$$P(\text{one pair}) = \frac{\binom{13}{1}\binom{4}{2}\binom{12}{3}4^3}{\binom{52}{5}}.$$

The next kind of hand is the one containing two pairs. Here the argument is as follows:

(a) The denominations of the two pairs can be selected in $\binom{13}{2}$ ways.

(b) The suits of cards in these two pairs can be selected in $\binom{4}{2} \times \binom{4}{2}$ ways.

(c) The remaining card may be chosen in $\binom{11}{1} \times 4$ ways (two denominations are eliminated).

Combining (a)–(c), we have

$$P(\text{two pairs}) = \frac{\binom{13}{2}\binom{4}{2}^2\binom{11}{1} \times 4}{\binom{52}{5}}.$$

Finally, we calculate the probability of a straight (probabilities of remaining hands are left as exercise). A straight is defined as a hand containing five cards in consecutive denominations but not of the same suit (e.g., 9, 10, jack, queen, and king). An ace can appear at either end, so we could have a straight of the form ace, 2, 3, 4, 5, as well as 10, jack, queen, king, ace.

The number of hands with a straight can be computed as follows: Each such hand is uniquely determined by the lowest denomination (ace, $2, 3, \ldots, 10$) in 10 ways. Then, the suits of five cards are chosen in $4^5 - 4$ ways: 4^5 is the total number of choices of suits, and we subtract 4 selections in which all cards are of the same suit. Thus

$$P(\text{straight}) = \frac{10(4^5 - 4)}{\binom{52}{5}}.$$

PROBLEMS

3.3.1 Which is larger: **(i)** $\binom{1000}{50}$ or $\binom{1000}{51}$. **(ii)** $\binom{1000}{949}$ or $\binom{1000}{950}$. **(iii)** $\binom{1000}{50}$ or $\binom{1001}{50}$.

3.3.2 A committee of size 12 is to be formed out of the US Senate. Find the probability that: **(i)** Both senators from Ohio will be represented. **(ii)** Exactly one senator from Ohio will be represented. **(iii)** No senator from Ohio will be represented.

3.3.3 A committee of size k is to be formed out of the US Senate. How large must k be in order for the probability of at least one senator from Ohio being included to exceed the probability of no senator from Ohio being included?

3.3.4 What is the probability that a randomly chosen committee of 50 senators contains one senator from each of the 50 states?

3.3.5 How many ways can one order the deck of 52 cards so that all four kings are next to each other?

3.3.6 Peter lives at the corner of 2nd Avenue and 72nd Street. His office is in the building at a corner of 7th Avenue and 78th Street. The streets and avenues in the city form a perpendicular grid, with no streets or passages in the middle of the blocks.

Peter walks to work along either street or avenue, always in the direction that gets him closer to his office. He always returns home by subway, so he walks across town only once a day. **(i)** How many different paths can Peter choose to go to work? **(ii)** If Peter makes a list of all possible paths and chooses one of them randomly every morning, how likely it is that he will not walk 4th Avenue between 75th and 76th streets during the next five working days?

3.3.7 (Poker Hands) Find the probability of each of the following hands:
(i) Royal flush (ace, king, queen, jack, and 10 in one suit),
(ii) Straight flush (five cards of one suit in a sequence, but not a royal flush),
(iii) Flush (five cards in one suit, but not a straight flush nor a royal flush),
(iv) Four-of-a-kind (four cards of the same denomination),
(v) Full house (one pair and one triple of the same denomination),
(vi) Three-of-a-kind (three cards of the same denomination plus two cards unmatched).

3.3.8 Find the probability that a poker hand will contain two pairs, one red and the other black.

3.3.9 A poker player has $3\heartsuit, 7\diamondsuit, 8\spadesuit, 9\diamondsuit, Q\clubsuit$. He discards $3\heartsuit$ and $Q\clubsuit$ and obtains 2 cards.[7] **(i)** What is the probability that he will have a straight? **(ii)** Answer the same question if $Q\clubsuit$ is replaced by $J\clubsuit$ (i.e., he discards $3\heartsuit$ and $J\clubsuit$).

3.3.10 A poker player has $3\heartsuit, 7\diamondsuit, 8\diamondsuit, 9\diamondsuit, Q\clubsuit$. She discards $3\heartsuit$ and $Q\clubsuit$ and obtains 2 cards. What is the probability that she will have: **(i)** A straight flush. **(ii)** A flush, but not a straight flush. **(iii)** A straight, but not a straight flush.

3.3.11 A poker player has three-of-a-kind. He discards the two unmatched cards, and obtains two new cards. Find the probability that he will have: **(i)** Three-of-a-kind. **(ii)** Four-of-a-kind. **(iii)** A full house.

3.3.12 **(i)** If n balls are put at random into n boxes, what is the probability of exactly one box remaining empty? **(ii)** Suppose that we distribute n balls in k boxes ($n \geq k$), labeled $1, \ldots, k$. Show that the number of different ways this can be done, so that there is at least one ball in each box, is $\binom{n-1}{k-1}$.

3.3.13 Suppose that n balls are distributed in k boxes ($n \geq k$). Find the probability that there are no empty boxes. [*Hint:* Let $A(n,k)$ be the number of allocations of balls such that no box is empty. Show that $A(n,k) = \sum_{i=1}^{n-k+1} \binom{n}{i} A(n-i, k-1)$. Use boundary condition $A(n,1) = 1$ to find $A(n,k)$.]

3.3.14 Compute probabilities $P(x)$ of winning x numbers in lotteries, where the player chooses: **(i)** 5 out of 44 numbers. **(ii)** 6 out of 55 numbers.

3.3.15 Find the number of polygonal lines with vertices $(x, C(x))$, where $C(x)$ is as in Example 3.11 and with possible edges leading from $(x, C(x))$ to $(x+1, C(x)+1)$ or $(x+1, C(x)-1)$, connecting the points: **(i)** $(0,0)$ and $(10,0)$. **(ii)** $(0,0)$ and $(10,5)$. **(iii)** $(3,-2)$ and $(8,1)$.

3.3.16 Find the number of polygonal lines (as in Problem 3.3.15) that join the points (2,3) and (16, 5) and: **(i)** Never touch the x-axis. **(ii)** Never touch the line $y = 7$.

[7]The discarded cards are not mixed with the deck. Assume that the player receives the replacement of the discarded cards from the unused remainder of the deck.

3.4 EXTENSION OF NEWTON'S FORMULA

As already remarked, in Definition 3.2.1, in the formula

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!}, \tag{3.26}$$

the upper number n need not be an integer.

■ **EXAMPLE 3.13**

As an illustration, let us evaluate $\binom{-1/2}{k}$, which we will use later. We have

$$\begin{aligned}\binom{-1/2}{k} &= \frac{(-\frac{1}{2})(-\frac{1}{2}-1)(-\frac{1}{2}-2)\cdots(-\frac{1}{2}-k+1)}{k!} \\ &= \frac{(-\frac{1}{2})(-\frac{3}{2})(-\frac{5}{2})\cdots(-\frac{2k-1}{2})}{k!} = \frac{(-1)^k 1\times 3\times 5\times(2k-1)}{2^k k!}.\end{aligned}$$

Multiplying the numerator and denominator by $2\times 4\times\cdots\times(2k) = 2^k k!$, we get

$$\binom{-1/2}{k} = \frac{(-1)^k}{2^{2k}}\binom{2k}{k}. \tag{3.27}$$

Newton's binomial formula extends to the following theorem, which can be found in most calculus texts:

Theorem 3.4.1 *For any a we have*

$$(1+x)^a = \sum_{k=0}^{\infty}\binom{a}{k}x^k. \tag{3.28}$$

Proof. It suffices to observe that the right hand side of (3.28) is the Taylor expansion of the left-hand side, $\binom{a}{k}$ equals $\frac{1}{k!}\frac{d^{(k)}}{dx^k}(1+x)^a$ for $x=0$. □

■ **EXAMPLE 3.14**

For $a=-1/2$ and $|x|<1$ we have

$$\frac{1}{\sqrt{1-x}} = \sum_{k=0}^{\infty}\binom{-\frac{1}{2}}{k}(-x)^k.$$

Using (3.27), we obtain, after some algebra,

$$\frac{1}{\sqrt{1-x}} = \sum_{k=0}^{\infty}\binom{2k}{k}\left(\frac{x}{4}\right)^k.$$

■ **EXAMPLE 3.15**

Several well-known formulas for sums of geometric series can be obtained as special cases of (3.28). For instance, upon noting that $\binom{-1}{k} = (-1)^k$, we obtain, for $|x| < 1$,

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots \qquad \text{and} \qquad \frac{1}{1+x} = 1 - x + x^2 - \cdots$$

PROBLEMS

3.4.1 **(i)** Show that

$$\binom{-2}{k} = (-1)^k(k+1).$$

(ii) Show that for every $x > 0$ (integer or not)

$$\binom{-x}{k} = (-1)^k \binom{x+k-1}{k}.$$

3.4.2 Show that

$$\binom{k}{k} + \binom{k+1}{k} + \cdots + \binom{k+r}{k} = \binom{k+r+1}{k+1}.$$

3.4.3 Show that

$$\binom{1/2}{n} = (-1)^{n-1}\frac{1}{n}\binom{2n-2}{n-1}2^{-2n+1}.$$

3.4.4 By integrating the series in Example 3.15, show that

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots. \tag{3.29}$$

3.4.5 Use formula (3.29) to show that

$$\frac{1}{2}\log\frac{1+x}{1-x} = x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots.$$

3.5 MULTINOMIAL COEFFICIENTS

Choosing a subset of size k out of a set of size n is logically equivalent to partitioning the set of size n into two subsets, one of size k and the other of size $n - k$. The number of such partitions is, by definition,

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

The theorem below generalizes this scheme.

Theorem 3.5.1 *Let $k_1, \ldots, k_r$ be positive integers such that $k_1 + \cdots + k_r = n$. The number of ways a set of n elements can be partitioned into r subsets of sizes $k_1, k_2, \ldots, k_r$ equals*

$$\frac{n!}{k_1!k_2!\cdots k_r!}. \tag{3.30}$$

Proof. We can think of a partition above being accomplished in steps: First, we choose k_1 out of n elements, to form the first class of the partition. Next, we choose k_2 elements out of the remaining $n - k_1$ elements, and so on, until we have $n - k_1 - k_2 - \cdots - k_{r-2} = k_{r-1} + k_r$ elements, from which we choose k_{r-1} to form the next-to-last class. The remaining k_r elements form the last class. This can be accomplished, in view of Theorem 3.2.2, in

$$\binom{n}{k_1}\binom{n-k_1}{k_2}\binom{n-k_1-k_2}{k_3}\cdots\binom{n-k_1-\cdots-k_{r-2}}{k_{r-1}} \tag{3.31}$$

ways. Simple algebra shows that formula (3.31) is the same as formula (3.30). □

The ratio (3.30) is called *multinomial coefficient* and is denoted by

$$\binom{n}{k_1, k_2, \ldots, k_r}.$$

As a generalization of Newton's binomial formula, we have

Theorem 3.5.2 *For every integer n,*

$$(x_1 + \cdots + x_r)^n = \sum \binom{n}{k_1, k_2, \ldots, k_r} x_1^{k_1} \cdots x_r^{k_r}, \tag{3.32}$$

where the summation is extended over all r-tuples $(k_1, \ldots, k_r)$ of nonnegative integers with $k_1 + \cdots + k_r = n$.

Proof. In the product $(x_1 + \cdots + x_r)^n$, one term is taken from each factor so that the general term of the sum is of the form $x_1^{k_1} \cdots x_r^{k_r}$ with $k_1 + \cdots + k_r = n$. From Theorem 3.5.1 it follows that the number of times the product $x_1^{k_1} \cdots x_r^{k_r}$ appears equals (3.30). □

In analogy with formula (3.17), the sum of all multinomial coefficients equals r^n, which follows by substituting $x_1 = \cdots = x_r = 1$ in (3.32).

The theorem is illustrated by the following example:

■ **EXAMPLE 3.16**

Suppose that one needs the value of the coefficient of $x^2y^3z^4w$ in the expression $(x + y + z + w)^{10}$. One could argue that in the multiplication $(x + y + z + w) \times \cdots \times (x + y + z + w)$ there are 10 factors, and each term will contain one component from each set of parentheses. Thus choosing x from 2 out of 10 pairs of parentheses, y from 3 out of 10, and so on,

amounts to partitioning 10 pairs of parentheses into four classes, with sizes $k_1 = 2, k_2 = 3, k_3 = 4$, and $k_4 = 1$. The total number of ways such a partition can be accomplished is the coefficient of $x^2y^3z^4w$, and equals $\binom{10}{2,3,4,1} = \frac{10!}{2!3!4!1!} = 12,600$.

An approximation to $n!$ is given by the so-called Stirling's formula.

Theorem 3.5.3 (Stirling's Formula) *We have*

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}, \tag{3.33}$$

where the sign $\sim$ *means that the ratio of the two sides tends to* 1 *as* $n \to \infty$.

We shall not give the proof here, but interested readers can find it in more advanced texts, for example, in Chow and Teicher (1997).

■ **EXAMPLE 3.17**

A group of $2n$ boys and $2n$ girls is divided at random into two equal parts. What is the probability p_n and its approximation for large n, that boys and girls are divided evenly between the two groups?

SOLUTION. Clearly, the number of ways a group can be divided into two equal parts is $\binom{4n}{2n}$. The number of ways $2n$ boys can be divided evenly is $\binom{2n}{n}$ and the same holds for girls. Thus

$$p_n = \frac{\binom{2n}{n}\binom{2n}{n}}{\binom{4n}{2n}} = \frac{[(2n)!]^4}{(n!)^4(4n)!} \tag{3.34}$$

which, based on (3.33), can be approximated by

$$\frac{\left[\sqrt{2\pi}(2n)^{2n+\frac{1}{2}}e^{-2n}\right]^4}{\left[\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}\right]^4\left[\sqrt{2\pi}(4n)^{4n+\frac{1}{2}}e^{-4n}\right]} = \sqrt{\frac{2}{n\pi}}.$$

For example, in the case of 16 boys and 16 girls ($n = 8$), the approximate chances of dividing both sexes evenly are about $\sqrt{1/(4\pi)} = 0.2821$. The exact value is 0.2756.

■ **EXAMPLE 3.18**

Let us find the percentage of seven-digit numbers that have all digits distinct. Clearly, the smallest number that has seven digits is $1,000,000$ and the largest is $9,999,999$. So there are $9,000,000$ distinct seven-digit integers. The number of such integers that have all digits distinct can be computed as follows: The first digit can be selected in nine ways (since 0 is not allowed). The subsequent six digits, if no digits are to repeat, can be selected

in $9 \times 8 \times 7 \times 6 \times 5 \times 4 = 9!/3!$ ways. Altogether the percentage of all seven-digit integers with all digits distinct is

$$p = \frac{9 \times 9!/3!}{9 \times 10^6} \times 100\% = \frac{9!}{6 \times 10^4}\% = 6.048\%.$$

Stirling's formula gives the approximation

$$p \approx \frac{\sqrt{2\pi} \times 9^9 \times \sqrt{9} \times e^{-9}}{6 \times 10^4}\% = \sqrt{\frac{\pi}{2}} \times \frac{9^9}{e^9 10^4}\% = 5.992\%.$$

In this case we could calculate the exact value. However, when the formulas involve factorials of larger numbers, Stirling's approximation may be our only access to numerical values.

PROBLEMS

3.5.1 Show that if $a \leq b \leq c \leq n$, then

$$\binom{n}{c}\binom{c}{b}\binom{b}{a} = \binom{n}{a}\binom{n-a}{b-a}\binom{n-b}{c-b}$$

(i) Use the definition of binomial coefficients as ratios of the factorials. **(ii)** Use directly the interpretation of the binomial coefficients as the number of subsets of a given size. **(iii)** How many ways can one choose an a-element subset from a b-element subset from a c-element subset from a d-element subset from a n element set? (where $a \leq b \leq c \leq d \leq n$).

3.5.2 Find the coefficient of the term $x^4y^5z^3$ in the expansion of $(x - 2y + 3z)^{12}$.

3.5.3 Use the argument analogous to that in Theorem 3.3.2 to show that if $i \geq 1, j \geq 1$, and $k \geq 1$, then

$$\binom{n+1}{i,j,k} = \binom{n}{i-1,j,k} + \binom{n}{i,j-1,k} + \binom{n}{i,j,k-1}.$$

3.5.4 Show that

$$\sum_{k_1+\cdots+k_r=k} \binom{k}{k_1,\ldots,k_r} = \binom{k+r-1}{k}.$$

3.5.5 A committee of 50 is to be chosen from the US Senate. Estimate the numerical value of the probability that every state will be represented.

3.5.6 Use Stirling's formula to approximate the number of ways: **(i)** A set of size $2n$ can be partitioned into two equal parts. **(ii)** A set of size $3n$ can be partitioned into three equal parts.

CHAPTER 4

CONDITIONAL PROBABILITY; INDEPENDENCE

4.1 INTRODUCTION

The formulas in Chapter 2 allow us to handle probabilities of unions of events, expressing them through probabilities of intersections. In this chapter we develop methods useful for computations that are based on a concept of *conditional probability*.

Consider a situation where we want to evaluate the probability $P(A)$ of some event A. Suppose that after finding $P(A)$, we learn that some other event, B, occurred. In many cases such information leads to a change in the assessment of the probability of the event A. The symbol used for this new probability will be $P(A|B)$, to be read "conditional probability of A, given B," or "probability of event A, given that B occurred."

■ **EXAMPLE 4.1**

Conditional probabilities are most easily interpreted as probabilities in subpopulations. Consider an attribute such as color blindness, known to occur much more often among men than among women. If D is the event "a randomly selected person is color blind," then $P(D)$ refers to the chance of color blindness in the whole population. Suppose now that the person selected is

known to be a woman (event W). This information changes the assessment of probability of color blindness to $P(D|W)$, which is now the probability of color blindness in the subpopulation of women.

Questions that might arise here are:

1. How to use data on probabilities of color blindness separately among men and among women to find the overall chance of color blindness, that is, to find $P(D)$ if we know $P(D|W)$ and $P(D|M)$?

2. How to find the probability that a randomly selected color blind person is a woman, that is, $P(W|D)$?

The first of these questions requires using the weighted average, usually referred to as the formula for total probability (Section 4.3). To answer the second question one has to use the Bayes' formula (Section 4.4).

The examples and exercises in this chapter are designed to provide practice in recognizing, from the description of the problem, which probabilities are conditional and which are not. The key phrases here for indicating conditional probabilities would be "the proportion (frequency, percentage) of ... *among* ...;" and "*if* ... *occurs*, then the probability of ... is."

PROBLEMS

4.1.1 A computer file contains data on households in a certain city. Each entry line in this file contains various information about one household: income, socioeconomic status, number of children, their ages, and so on. The computer is programmed so that it selects one entry at random, each with the same probability of being selected. Consequently, probabilities of various events are interpretable as relative frequencies of entries in the data file with the corresponding property.

Let X, Y, and Z be, respectively, the numbers of boys, girls, and cars in the households sampled. Let A be the event that a household has a TV set, and let B be the event that it has a swimming pool.

(i) Interpret the probabilities below as relative frequencies of occurrence of some attributes in certain subpopulations. (a) $P(A)$. (b) $P(A|Z > 0)$. (c) $P(Z > 0|A)$. (d) $P(X = 0|X + Y = 3)$. (e) $P(B|A^c)$. (f) $P[(X + Y = 0)^c|A \cap B]$. (g) $P(XY = 0|X + Y > 1)$.

(ii) Use symbols to express probabilities corresponding to the following relative frequencies: (a) Relative frequency of households with two cars. (b) Relative frequency of households with no children among households with at least one car. (c) Relative frequency of households that have both a swimming pool and a TV set, among those who have either a swimming pool or a TV set and have at least one car.

4.2 CONDITIONAL PROBABILITY

The conditional probability of event A given event B is be defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0. \tag{4.1}$$

The definition of conditional probability in cases when the conditioning event B has probability zero will discussed in later chapters. In this chapter it is always assumed, even if not stated explicitly, that the event appearing in the condition has a positive probability.

A motivation of the definition (4.1), based on the frequential interpretation of probability, is the following. As in Chapter 2, let $N(\cdot)$ denote the number of occurrences of the event in parentheses in initial N repetitions of the experiment. Then $P(A|B)$ is to be approximated by the frequency of occurrence of A among those cases where B occurred. Now B occurred in $N(B)$ cases, and the number of occurrences of A among them is $N(A \cap B)$. Consequently, $P(A|B)$ should be the limit of $N(A \cap B)/N(B) = [N(A \cap B)/N]/[N(B)/N]$, and the latter converges to $P(A \cap B)/P(B)$.

For a fixed event B the conditional probabilities $P(A|B)$ can be regarded as a function of event A. Observe that these conditional probabilities satisfy the axioms of probability from Chapter 2. Indeed, nonnegativity holds because $P(A|B)$ is a ratio of two nonnegative numbers, $P(A \cap B)$ and $P(B)$. Next $P(\mathcal{S}|B) = P(\mathcal{S} \cap B)/P(B) = P(B)/P(B) = 1$. Finally, if the events $A_1, A_2, \ldots$ are mutually exclusive, the same is true for the events $A_1 \cap B$, $A_2 \cap B, \ldots$ so that

$$P\left(\bigcup_k A_k|B\right) = \frac{P(\bigcup_k A_k \cap B)}{P(B)} = \frac{\sum_k P(A_k \cap B)}{P(B)} = \sum_k P(A_k|B).$$

A simple consequence of (4.1) is the formula

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \tag{4.2}$$

Notice that the first equality is equivalent to (4.1). The second equality follows from the observation that the left-hand side of (4.2) remains the same when we interchange the roles of A and B. Such an interchange applied to the middle term gives the right-hand side term.

Formula (4.1) shows how to find conditional probability if we have the corresponding unconditional probabilities. Formula (4.2), on the other hand, shows that one may find the probability of an intersection of two events as the product of the conditional probability of one event given the second, and the unconditional probability of the second event. These two ways of using conditional probability will now be illustrated by simple examples.

■ **EXAMPLE 4.2**

Consider families with two children. Assuming that each of the four combinations of sexes, BB, BG, GB, and GG, is equally likely, find the probability that a family has two boys, given that at least one child is a boy.

SOLUTION. We have

$$P(\text{two boys} \mid \text{at least one boy}) = \frac{P([\text{two boys}] \cap [\text{at least one boy}])}{P(\text{at least one boy})}. \tag{4.3}$$

Since the event "two boys" implies (is contained in) the event "at least one boy," their intersection is the event "two boys." Hence the probability in the

numerator equals $1/4$. In the denominator we have the event $\{BB, BG, GB\}$, and its probability is $3/4$. Thus, the answer is $1/3$. Notice that

$$P(\text{two boys} \mid \text{older child is a boy}) = \frac{P(BB)}{P(BG \text{ or } BB)} = \frac{1}{2}. \tag{4.4}$$

The answers in (4.3) and (4.4) are not the same. To grasp the reason why they are not, observe that the two probabilities refer to families with two boys in different subpopulations. In the first case we consider all families that have at least one boy, so we eliminate families with two girls. In the second case we are interested in all families whose older child is a boy, so we again eliminate families with two girls, but also families with boy being the younger child (combination GB).

■ **EXAMPLE 4.3**

Assume that in some population the ratio of the number of men to the number of women is r. Assume also that color blindness occurs among men with frequency p, while among women with frequency p^2. If you choose a person at random, what is the probability of selecting a woman who is not color blind? (Incidentally, color blindness is a sex-linked attribute, and this is why the frequency of its occurrence among females is the square of the frequency of its occurrence among males; see Example 4.9.)

SOLUTION. Observe first that the answer is not $1 - p^2$, the latter being the conditional probability of a selected person not being color blind if it is known that this person is a woman. Let M, W, and D denote the events "man selected," "woman selected," and "selected person is color blind." Our objective is to find the probability $P(W \cap D^c)$. Using (4.2), we write $P(W \cap D^c) = P(D^c|W)P(W) = P(W|D^c)P(D^c)$. The third term is useless, since the data provide directly neither $P(D^c)$ nor $P(W|D^c)$. Now, using the middle term, we have $P(D^c|W) = 1 - p^2$. To determine $P(W)$, we note that $P(M)/P(W) = [1 - P(W)]/P(W) = r$, which gives the solution $P(W) = 1/(1 + r)$. Consequently $P(W \cap D^c) = (1 - p^2)/(1 + r)$.

As an immediate consequence of the definition of conditional probability, we have the following theorem, often called the *chain rule*:

Theorem 4.2.1 (Chain Rule) *For any events $A_1, A_2, \ldots, A_n$ we have*

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_n) \\ &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}), \end{aligned} \tag{4.5}$$

provided $P(A_1 \cap \cdots \cap A_{n-1}) > 0$ (which implies that all conditional probabilities appearing in (4.5) *are well defined).*

Proof. It suffices to write each of the conditional probabilities on the right-hand side as the corresponding ratio of unconditional probabilities according to (4.1). The product cancels then to the probability on the left-hand side. □

It might be helpful to observe that the chain rule is closely related to the counting formula in sampling without replacement in Section 3.2.

■ **EXAMPLE 4.4**

A man has N keys of which only one opens the door. For some reason he tries them at random (eliminating the keys that have already been tried). What is the probability that he opens the door on kth attempt?

SOLUTION. Let A_i be the event "ith attempt unsuccessful." The problem is then to find $P(A_1 \cap A_2 \cap \cdots \cap A_{k-1} \cap A_k^c)$. Applying the chain rule, we obtain

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_{k-1} \cap A_k^c) \\ &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k^c|A_1 \cap \cdots \cap A_{k-1}). \end{aligned} \tag{4.6}$$

We have here $P(A_1) = (N-1)/N$. Next, $P(A_2|A_1) = (N-2)/(N-1)$, since after A_1 occurs there are $N-1$ keys to try, of which $N-2$ do not open the door. Generally, if $A_1 \cap A_2 \cap \cdots \cap A_{j-1}$ occurs, then $j-1$ keys have been tried and eliminated. This means that there are $N-(j-1) = N-j+1$ keys to be tried, of which one fits the door; hence

$$P(A_j|A_1 \cap A_2 \cap \cdots \cap A_{j-1}) = (N-j)/(N-j+1).$$

Substitution to (4.6) gives the answer $1/N$. It may be surprising that the answer does not depend on k. A success close to the beginning or to the end of the search is as likely as a success closer to the middle. Let us see another solution of the problem. Trying the keys at random without repetition is the same as ordering the keys into a sequence, and trying them in this order. The chances that the correct key will appear at the kth place are the same as for any other place; hence the probability must be the same for success at any trial.

A convenient way of using the chain rule is related to computing probabilities through *event trees*. This technique is applicable when the outcomes in the sample space S are sequences of events, typically with the preceding events affecting the probabilities of the subsequent ones. The set of such outcomes can then be depicted as a tree, and probabilities of the outcomes ("branches" of a tree) are calculated according to the chain rule as products of probabilities assigned to consecutive segments of the branch. This general description will be illustrated by an example of an urn scheme, a convenient device for analyzing many real-life problems.

■ **EXAMPLE 4.5**

An urn initially contains r red balls and g green balls. A ball is selected at random and returned to the urn. If the chosen ball is red, then a red and b green balls are added. If the chosen ball is green, then c red and d green balls are added. Next, a second ball is drawn, returned, and the process of adding

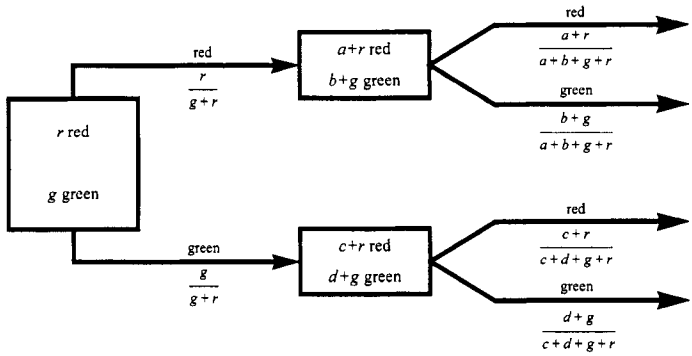


Figure 4.1 Possible results of the two first draws in Example 4.2.4.

balls is repeated. Finally, the third ball is drawn. Find the probabilities for each possible sequence of colors on three draws: (red, green, green), (red, red, red), and so on.

SOLUTION. Let us use the notation R_1, G_1, R_2, G_2, R_3, and G_3 for the events that occur at the first, second, and third drawings. The "natural" choice of the sample space $\mathcal{S}$ is

$$\mathcal{S} = \left\{ \begin{array}{cccc} R_1 \cap R_2 \cap R_3, & R_1 \cap R_2 \cap G_3, & R_1 \cap G_2 \cap R_3, & R_1 \cap G_2 \cap G_3, \\ G_1 \cap R_2 \cap R_3, & G_1 \cap R_2 \cap G_3, & G_1 \cap G_2 \cap R_3, & G_1 \cap G_2 \cap G_3 \end{array} \right\}.$$

We need to find the probabilities such as $P(R_1 \cap G_2 \cap G_3)$. Using the chain rule, we obtain $P(R_1 \cap G_2 \cap G_3) = P(R_1)P(G_2|R_1)P(G_3|R_1 \cap G_2)$. The outcomes are represented as branches of the tree in Figure 4.1, labeled with the corresponding probabilities. By the chain rule, each branch has a probability equal to the product of the probabilities assigned to its segments, for example,

$$P(R_1 \cap G_2 \cap G_3) = \frac{r}{r+g} \times \frac{g+b}{r+g+a+b} \times \frac{g+b+d}{r+g+a+b+c+d}.$$

Observe that we do not make any assumptions about the signs of a, b, c, and d, which makes this scheme quite flexible. Therefore, if we want to consider the scheme where the balls sampled are not returned, and no other balls are added, we can put $a = -1, b = c = 0$, and $d = -1$.

PROBLEMS

4.2.1 Let events A, B, and C be such that $P(A) > 0, P(B) > 0$, and $P(C) > 0$. Label the following statements as true or false: **(i)** The conditional probability $P(A|B)$ can never exceed the unconditional probability $P(A)$. **(ii)** If $P(A|B) = P(A|C)$ then $P(B) = P(C)$. **(iii)** If A and B are disjoint then $P(A|A \cup B) = P(A)/[P(A)+P(B)]$. **(iv)** $P(A|B)+P(A^c|B) = 1$. **(v)** $P(A|B)+P(A|B^c) = 1$.

4.2.2 Assume that A and B are disjoint. Find $P(A^c|B)$ and $P(A \cup B|A^c)$.

4.2.3 Assume that $P(A \cap B) > 0$ and determine $P(A)/[P(A)+P(B)]$ as a function of a, where $a = P(B|A)/P(A|B)$.

4.2.4 Find $P(A \cap B)$ if $P(A) = 3/4, P(B) = 1/2, P(A|B) - P(B|A) = 2/9$.

4.2.5 $P(A) = P(B) = 1/2, P(A|B) = 1/3$. Find $P(A \cap B^c)$.

4.2.6 Find $P(A|B)$, if $P(A^c) = 2P(A)$ and $P(B|A^c) = 2P(B|A) = 0.2$.

4.2.7 Find $P[A \cap B \cap C|(A \cap B) \cup (A \cap C)]$ if $P(A) = 0.8, P(B) = 0.4, P(C) = 0.4, P(A \cup B) = 1, P(A \cup C) = 0.9$, and $P(B \cup C) = 0.6$.

4.2.8 Events A, B, C are such that at least one of them occurs. Moreover, $P(A|B) = P(B|C) = P(C|A) = 1/2, P(B \cap C) = 2P(A \cap B) = 4P(C \cap A)/3 = 2P(A \cap B \cap C)$. Find the probability that: **(i)** Exactly one of events A, B, and C occurs. **(ii)** Only B occurs.

4.2.9 Three distinct integers are chosen at random from the set $\{1, 2, \ldots, 15\}$. Find the probability that: **(i)** Their sum is odd. **(ii)** Their product is odd. **(iii))** The sum is odd if it is known that product is odd. **(iv)** The product is odd if it is known that the sum is odd.
(v) Answer questions (i)–(iv) if three integers are selected from $\{1, \ldots, 20\}$.
(vi) Generalize answers to (i)–(iv) in the case of selection from $\{1, \ldots, n\}$.
(vii) Answer questions (i)–(iv) by drawing from $\{1, \ldots, n\}$, and passing to the limit with $n \to \infty$.

4.2.10 A deck of eight cards contains four jacks and four queens. A pair of cards is drawn at random. Find the probability that both cards are jacks if: **(i)** At least one of the cards is jack? **(ii)** At least one of the cards is a red jack? **(iii)** One of the cards is a jack of hearts?

4.2.11 A fair coin is tossed until a head appears. Given that the first head appeared on an even-numbered toss, what is the probability that it appeared on the $2n$th toss?

4.2.12 A tennis player has the right to two attempts at a serve: If he misses his first serve, he can try again. A serve can be played "fast" or "slow." If a serve is played fast, the probability that it is good (the ball hits opponent's court) is A; the same probability for a slow serve is B. Assume that $0 < A < B$; that is, the fast serve is more difficult (but not impossible) to make.

If a serve is good, the ball is played, and eventually one of the players wins a point. Let a be the probability that the server wins the point if the serve is fast (and good), and let b be the same probability for a slow serve. Assume that $0 < b < a$; that is, a fast serve gives a certain advantage to the server (the ball is harder to return, etc.).

The server has four possible strategies: FF, FS, SF, and SS, where FF is "play first serve fast; if missed, play second serve fast," FS is "play first serve fast; if missed play second serve slow," and so on.

(i) Determine the probabilities P_{FF}, P_{FS}, P_{SF}, and P_{SS} of winning the point by the server under the four strategies. **(ii)** Show that the strategy SF is always inferior to the strategy FS. (*Hint:* Consider the difference $P_{FS} - P_{SF}$.) **(iii)** Show that if $Aa > Bb$, then strategy FF is better than FS. **(iv)** Show, by choosing appropriate numbers A, B, a, and b (with $A < B$ and $a > b$), that each of the strategies FF and SS may be best among the four. Explain why such cases do not occur among top players, for whom the best strategy is FS.

4.2.13 Three cards are drawn without replacement from an ordinary deck of cards. Find the probability that: **(i)** The first heart occurs on the third draw. **(ii)** There will be more red than black cards drawn. **(iii)** No two consecutive cards will be of the same value.

4.2.14 An urn contains three red and two green balls. If a red ball is drawn, it is not returned, and one green ball is added to the urn. If a green ball is drawn, it is returned, and two blue balls are added. If a blue ball is drawn, it is simply returned to the urn. Find the probability that in three consecutive draws from the urn, there will be exactly one blue ball drawn.

4.3 PARTITIONS; TOTAL PROBABILITY FORMULA

In this section we introduce some formulas, important and useful in computing certain probabilities. We begin with the concept of *partition* (used already in Section 1.4).

Definition 4.3.1 We say that a class of events $\mathcal{H} = \{H_1, H_2, \ldots\}$ forms a *partition* (of the sample space $\mathcal{S}$) if these events exclude one another and one of them must occur. Formally, this means that

$$H_i \cap H_j = \emptyset \quad \text{for all } i \neq j \tag{4.7}$$

and

$$H_1 \cup H_2 \cup \cdots = \mathcal{S}. \tag{4.8}$$

□

We say that a partition is *finite* or *countable*, depending on whether it contains a finite or a countably infinite number of events. In later chapters we also consider uncountable partitions.

Definition 4.3.2 Given two partitions

$$\mathcal{H} = \{H_1, H_2, \ldots\} \quad \text{and} \quad \mathcal{K} = \{K_1, K_2, \ldots\},$$

we say that $\mathcal{H}$ is *finer* than $\mathcal{K}$ (or: $\mathcal{H}$ is a *refinement* of $\mathcal{K}$, or sometimes, $\mathcal{H}$ is *contained* in $\mathcal{K}$) if for every event H_i in $\mathcal{H}$ there exists an event K_j in $\mathcal{K}$ such that $H_i \subset K_j$. □

■ **EXAMPLE 4.6**

For any event B, $\mathcal{K} = \{B, B^c\}$ is always a partition of $\mathcal{S}$. A pair of events A and B allows us to define a partition $\mathcal{H}$ into four sets: $H_1 = A \cap B$, $H_2 = A \cap B^c$, $H_3 = A^c \cap B$, and $H_4 = A^c \cap B^c$. We have here $\mathcal{H} \subset \mathcal{K}$, as one can easily verify. Generally, if $\mathcal{H} = \{H_1, H_2, \ldots\}$ and $\mathcal{K} = \{K_1, K_2, \ldots\}$ are two partitions, then the class of all sets $H_i \cap K_j$ for H_i in $\mathcal{H}$ and K_j in $\mathcal{K}$ is a partition, called the *intersection* of $\mathcal{H}$ and $\mathcal{K}$ and denoted $\mathcal{H} \times \mathcal{K}$. It is a refinement of both $\mathcal{H}$ and $\mathcal{K}$.

Finer partitions correspond to more detailed (richer) information, whereas larger partitions are used to represent general information.

■ **EXAMPLE 4.7**

Consider a data file on employees of some company. Let $\mathcal{H} = \{H_0, H_1, \ldots\}$ be the partition of data by level of education, and let $\mathcal{K} = \{K_0, K_1, \ldots\}$ be the partition of data corresponding to the number of full years of employment in the company. Knowledge of the category in $\mathcal{H}$ to which a data entry belongs gives information about the education level of the employee. Knowledge of the category in the intersection $\mathcal{H} \times \mathcal{K}$ gives information about both education level and number of years of employment, hence gives more information than either of $\mathcal{H}$ or $\mathcal{K}$ alone.

Since in the following theorems, sets in partitions will appear as conditions in conditional probabilities, it will be convenient to introduce the following definition:

Definition 4.3.3 A partition $\mathcal{H} = \{H_1, H_2, \ldots\}$ is called *positive*, if $P(H_i) > 0$ for all H_i in $\mathcal{H}$. □

We can now formulate the following simple theorem, called the *formula for total probability*. This formula will provide a valuable tool for computing probabilities.

Theorem 4.3.1 *Let $\mathcal{H} = \{H_j, j = 1, 2, \ldots\}$ be a positive (finite or infinite) partition and let A be an arbitrary event. Then*

$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \cdots + P(A|H_n)P(H_n) \quad (4.9)$$

in the case of a finite partition $\mathcal{H}$, or

$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \cdots \quad (4.10)$$

in the case of a partition $\mathcal{H}$ into a countable number of sets (infinite).

Proof. We can write, using (4.7) and (4.8), as well as (4.1),

$$\begin{aligned} P(A) &= P((A \cap H_1) \cup (A \cap H_2) \cup \cdots) \\ &= P(A \cap H_1) + P(A \cap H_2) + \cdots \\ &= P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \cdots . \end{aligned}$$

The summation may be finite or not, depending on whether the partition is finite. □

One may think of the conditional probability $P(A|H_k)$ as the relative frequency of attribute A in the subpopulation corresponding to event H_k. Then formulas (4.9) and (4.10) give simply the overall frequency of attribute A in the whole population, expressed as the weighted average, with weights $P(H_k), k = 1, 2, \ldots$ being the contributions of the kth subpopulation to the whole population.

In the simplest case of a positive partition into two sets ($\{B, B^c\}$), formula (4.9) reduces to

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|B^c)P(B^c) \\ &= P(A|B)P(B) + P(A|B^c)(1 - P(B)). \end{aligned}$$

■ **EXAMPLE 4.8**

Recall Example 4.3. In a certain group of people the ratio of the number of men to the number of women is r. It is known that the incidence of color blindness among men is p, and the incidence of color blindness among women is p^2. What is the probability that a person randomly chosen from this group is color blind?

SOLUTION. We use the formula for total probability, with events M and W (choice of man and choice of woman) as a partition. The probability of the event D (the person selected is color blind) is then $P(D) = P(D|M)P(M) + P(D|W)P(W)$. Since $P(M) = r/(1 + r)$ and $P(W) = 1/(1 + r)$ (see Example 4.3), the answer is $P(D) = pr/(1+r)+p^2/(1+r) = p(p+r)/(1+r)$.

■ **EXAMPLE 4.9**

The reason why the frequency of color blindness among females is the square of the corresponding frequency among males is that the gene responsible for color blindness (as well as the gene responsible for other sex-linked attributes, e.g., hemophilia) is located on the X chromosome. There are two sex chromosomes, X and Y, and every person has a pair of such chromosomes, with individuals of the type XX being females and XY being males.

The color blindness gene has two forms (alleles), say C and c, with form c being recessive and causing color blindness. Because every man has one X chromosome, he therefore is either of category C or c, the latter being color blind. Women, having a pair of X chromosomes, are of one of three categories: (1) CC, which we will call "normal;" (2) Cc, the so-called carriers (i.e., persons who are not color blind but are capable of transmitting the color blindness gene c to their offspring), and (3) cc, women who are color blind.

Let p be the relative frequency of the form c in the population of genes in question. Then p is also the relative frequency of color blind men (since each man carries one color blindness gene, either C or c). Now let u and v denote the relative frequency of carriers and of color blind women, respectively. To find u and v, we proceed as follows: So far we know that a woman is color

blind if her father is color blind and either (1) her mother is color blind, or (2) her mother is a carrier and transmits the gene c to the daughter. This gives, using the formula for total probability, the relation

$$v = p\left(v + \frac{1}{2}u\right). \tag{4.11}$$

On the other hand, a woman is a carrier if either (1) her father is color blind and her mother is "normal," (2) her father is color blind and her mother is a carrier who transmits to her the gene C, (3) her father is "normal" and her mother is a carrier who transmits gene c, or (4) her father is "normal" and her mother is color blind.

This gives the relation

$$u = p\left[(1 - u - v) + \frac{1}{2}u\right] + (1 - p)\left[\frac{1}{2}u + v\right]. \tag{4.12}$$

The solution of the pair of equations (4.11) and (4.12) gives $v = p^2, u = 2p(1 - p)$. The relative frequency of women who are "normal" (i.e., neither color blind nor a carrier) is $1 - u - v = (1 - p)^2$, while the relative frequency of color blind women is p^2, as asserted.

■ EXAMPLE 4.10 Secretary Problem

The following problem, also known in the literature as the *marriage problem*, or *dowry problem*, provides a nice illustration of the use of conditioning with respect to a partition.

In response to an announcement about a vacant secretarial position, n candidates appear. They are interviewed successively in random order (all permutations assumed equally likely). Each candidate may be ranked with respect to those interviewed before her, but not with respect to the others. After the interview the candidate may be accepted or rejected, and no change of decision is allowed later. In particular, a rejected candidate cannot be offered a job after any subsequent candidates have been interviewed. The problem is to devise a policy that would maximize the probability of employing the best candidate.

To put it formally, let $1, 2, \ldots, n$ be the true ranks of the candidates (no ties), with 1 standing for the best candidate and n for the worst. Let $x_1, x_2, \ldots x_n$ be a permutation of ranks $1, 2, \ldots, n$, and for $j = 1, 2, \ldots, n$ let y_j be the relative rank of jth candidate x_j with respect to candidates who appeared at places $1, 2, \ldots, j$ in a given permutation. For instance, if $n = 6$ and the permutation of true ranks is 3, 6, 2, 4, 1, 5, then the relative ranks y_i are 1, 2, 1, 3, 1, 5. Indeed the first candidate is always "best so far," so $y_1 = 1$. Next, 6 is inferior to 3, so $y_2 = 2$ (second best in the set $\{3, 6\}$). Then $y_3 = 1$, since 2 is best among $\{3, 6, 2\}$, and so forth.

We may observe $y_1, y_2, \ldots$ (but not $x_1, x_2, \ldots$), and the objective is to find a policy (i.e., a rule that tells us whether or not to hire the jth candidate knowing only $y_1, \ldots, y_j$) that would maximize the probability of stopping at j such that $x_j = 1$.

One can show (e.g., see Chow et al., 1971) that the optimal policy belongs to the class of policies of the following form: *Interview $r-1$ candidates without employing any of them, and then employ the first candidate superior to all the preceding ones* (i.e., stop at the first $k \geq r$ with $y_k = 1$). Let us call such a policy π_r, and let p_r be the probability of employing the best candidate under policy π_r. The problem is then reduced to finding an r that maximizes p_r.

SOLUTION 1. For $r = 1$ the policy π_1 leads to employing the first candidate and $p_1 = 1/n$. Assume now that $r > 1$.

In order for policy π_r to be successful, the true rank 1 must appear in a permutation at one of the places $r, r+1, \ldots, n$; otherwise, there will be no $y_k = 1$ with $k \geq r$. This observation, incidentally, shows that the probability of policy π_r leading to nobody getting employed is $(r-1)/n$, since 1 is equally likely to appear at any of n places.

If 1 appears at place $k \geq r$, then the policy π_r leads to winning, provided that there is no $y_j = 1$ for $j = r, r+1, \ldots, k-1$. In other words, the minimum in the sequence $x_1, \ldots, x_{r-1}, x_r, \ldots, x_{k-1}$ must occur at one of the places $1, \ldots, r-1$. The chances of that are $(r-1)/(k-1)$. Summing over k, we obtain

$$p_r = \sum_{k=r}^{n} \frac{r-1}{k-1} \times \frac{1}{n}. \tag{4.13}$$

Observe that we used formula (4.9) for partition into events $B_1, \ldots, B_n$, where B_k is the event $\{x_k = 1\}$. Moreover, $P(W|B_k) = 0$ for $k < r$ and $P(W|B_k) = (r-1)/(k-1)$ otherwise, where W denotes winning. To find the optimal r for a given (large) n, we let $r/n = x$ and use the fact that

$$1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \log n.$$

Then

$$\begin{aligned} p_r &= \frac{r-1}{n} \sum_{k=r}^{n} \frac{1}{k-1} = \frac{r-1}{n} \left(\sum_{k=1}^{n} \frac{1}{k-1} - \sum_{k=1}^{r-1} \frac{1}{k-1} \right) \\ &\approx \left(x - \frac{1}{n} \right) [\log(n-1) - \log(r-2)] \approx -x \log x. \end{aligned}$$

We look for the maximum of this function for $0 < x \leq 1$. Easy differentiation gives $x = 1/e$, which means that the optimal policy, for a large n, is as follows: *Interview $r = [n/e]$, which is about 37% candidates, without employing any of them. Then stop the first candidate that appears better than any of the preceding ones.*

The probability that this optimal policy will not lead to employing anybody is $(r-1)/n \sim 1/e \approx 0.3678$. Thus the probability of a correct decision is approximately equal the probability of no decision. Consequently, the probability of a wrong decision (employing an inferior candidate) is about 0.2644.

SOLUTION 2. We can use another partition (Bartoszyński, 1974) to arrive at an alternative formula for p_r. Let $T = \min(x_1, \ldots, x_{r-1})$ be the lowest true rank among the first $r-1$ candidates. Then

$$P(T = t) = \frac{\binom{n-t}{r-2}}{\binom{n}{r-1}}, \quad t = 1, 2, \ldots, n-r+2.$$

Here the total number of choices of true ranks for the first $r-1$ candidates is $\binom{n}{r-1}$. For the rank t to be the lowest, the remaining $r-2$ ranks should be selected out of ranks $t+1, t+2, \ldots, n$. This can be done in $\binom{n-t}{r-2}$ ways.

Now, if $T = t$, the true ranks $1, 2, \ldots, t-1$ will appear among candidates at places $r, r+1, \ldots, n$. The *first* among them will be employed, and therefore the chances of winning are $1/(t-1)$. Adding for $t \geq 2$ (for $T = 1$ nobody will be employed!), we obtain

$$p_r = \sum_{t=2}^{n-r+2} \frac{\binom{n-t}{r-2}}{\binom{n}{r-1}} \times \frac{1}{t-1}.$$

The maximum in this formula occurs at the same r as in the formula (4.13), but the calculations are now much more complicated.

The formula for total probability can be extended to conditional probabilities as follows:

Theorem 4.3.2 *Let $\mathcal{H} = \{H_1, H_2, \ldots\}$ be a positive partition of the sample space, and let A, B be two events with $P(B) > 0$. Then*

$$P(A|B) = \sum_j P(A|B \cap H_j) P(H_j|B). \tag{4.14}$$

Proof. The right-hand side can be written as[8]

$$\begin{aligned} \sum_j \left(\frac{P(A \cap B \cap H_j)}{P(B \cap H_j)} \right) \left(\frac{P(B \cap H_j)}{P(B)} \right) &= \frac{1}{P(B)} \sum_j P(A \cap B \cap H_j) \\ &= \frac{1}{P(B)} P\left((A \cap B) \cap \bigcup_j H_j \right) \\ &= \frac{P(A \cap B)}{P(B)} = P(A|B). \end{aligned}$$

□

As was mentioned earlier, partitions can be used to represent information, with event $H_i \in \mathcal{H}$ signifying the situation where we know that the outcome is in H_i, but do not have any more specific information.

[8]Formally, the summation in (4.14) extends only over those j for which $P(B \cap H_j) > 0$, since otherwise the conditional probability is not defined. This does not matter, however, since $P(B \cap H_j) = 0$ implies $P(H_j|B) = 0$.

The conditional probability of an event can be treated as a function defined on a partition. In such a function, call it $P_{\mathcal{H}}(A|\cdot)$, the arguments are sets of the partition. Then $P(A|H_k)$ is a *random quantity* whose value depends on H_k.

It is natural to consider next the situation of "coarsening" a partition $\mathcal{H}'$ by grouping its sets, thus forming a new partition $\mathcal{H}$. This corresponds to less precise information, represented by the partition $\mathcal{H}'$.

Theorem 4.3.3 *Let $\mathcal{H}, \mathcal{H}'$ be two partitions, with $\mathcal{H}$ being a coarsening of $\mathcal{H}'$. Then for any event A, the conditional probabilities with respect to partitions $\mathcal{H}$ and $\mathcal{H}'$ are related as follows: if $H_i = H'_{i_1} \cup \cdots \cup H'_{i_k}$, with $H_i \in \mathcal{H}$, $H'_{i_j} \in \mathcal{H}'$, then*

$$P_{\mathcal{H}}(A|H_i) = \sum_{j=1}^{k} P_{\mathcal{H}'}(A|H'_{i_j})P(H'_{i_j}|H_i).$$

Thus the conditional probability with respect to a coarser partition is the weighted average of probabilities with respect to the finer partition.

Proof. The proof is obtained by simple algebra. If $\{H'_{i_j}, \ldots, H_{i_k}\}$ is a partition of H_i, then

$$\begin{aligned}
P(A|H_i) &= P(A \cap H_i|H_i) = P(A \cap \bigcup_{j=1}^{k} H'_{i_j}|H_i) \\
&= \sum_{j=1}^{k} P(A \cap H'_{i_j}|H_i) = \sum_{j=1}^{k} \frac{P(A \cap H'_{i_j} \cap H_i)}{P(H_i)} \\
&= \sum_{j=1}^{k} \frac{P(A \cap H'_{i_j})}{P(H_i)} = \sum_{j=1}^{k} \frac{P(A|H'_{i_j})P(H'_{i_j})}{P(H_i)} \\
&= \sum_{j=1}^{k} P(A|H'_{i_j}) \frac{P(H'_{i_j} \cap H_i)}{P(H_i)} = \sum_{j=1}^{k} P(A|H'_{i_j})P(H'_{i_j}|H_i). \qquad \square
\end{aligned}$$

PROBLEMS

4.3.1 An event W occurs with probability 0.4. If A occurs, then the probability of W is 0.6; if A does not occur but B occurs, the probability of W is 0.1. However, if neither A nor B occurs, the probability of W is 0.5. Finally, if A does not occur, the probability of B is 0.3. Find $P(A)$.

4.3.2 Suppose that initially the urn contains one red and two green balls. We draw a ball and return it to the urn, adding three red, one green, and two blue balls if a red ball was drawn, and three green and one blue ball if a green ball was drawn. Then a ball is drawn from the urn. Find probability that both selected balls are: **(i)** Blue. **(ii)** Red. **(iii)** Green. **(iv)** Of the same color.

4.3.3 Suppose that in Problem 4.3.2 we return the second ball to the urn, and add new balls as described, with the condition that if the second ball is blue, we add one

ball of each color. Then we draw the third ball. What is the probability that the third ball is: **(i)** Blue. **(ii)** Blue if the first ball was red. **(iii)** Blue if the second ball was red. **(iv)** Blue if no blue ball was drawn on any of the preceding draws?

4.3.4 Let A and B be two events with $P(B) > 0$, and let $C_1, C_2, \ldots$ be a possible partition of a sample space. Prove or disprove the following formulas:

$$P(A|B) = \sum_i P(A|B \cap C_i)P(C_i),$$

$$P(A|B) = \sum_i P(A|B \cap C_i)P(B|C_i)P(C_i).$$

4.3.5 (Tom Sawyer Problem) You are given a task, say painting a fence. The probability that the task will be completed if k friends are helping you is p_k $(k = 0, 1, \ldots)$. If j friends already helped you, the probability that the $(j+1)$st will also help is π_j $(j = 0, 1, \ldots)$. On the other hand, if the jth friend did not help, then the $(j+1)$st will not help either. **(i)** Find the probability that the task will be completed. **(ii)** Solve part (i) if $p_k = 1 - \lambda^k$, $\pi_j = \mu$ for all j.

4.3.6 Recall Example 4.9. Find the probability that the mother is a carrier if: **(i)** Both father and son are color blind, and the mother is not. **(ii)** It is known only that the son is color blind. **(iii)** The son is color blind, but the parents are not.

4.4 BAYES' FORMULA

Let us consider now a following question: If it is known that a certain event A occurred, and its conditional probabilities $P(A|H_1), P(A|H_2), \ldots$ for partition H_1, $H_2, \ldots$ are known, can we determine the probability of the event H_k of the partition? The answer is contained in the following time-honored theorem, dating back to the eighteenth century.

Theorem 4.4.1 (Bayes' Formula) *Let $\mathcal{H} = \{H_1, H_2, \ldots\}$ be a positive partition of $\mathcal{S}$, and A be an event with $P(A) > 0$. Then for any event H_k of the partition $\mathcal{H}$,*

$$P(H_k|A) = \frac{P(A|H_k)P(H_k)}{P(A|H_1)P(H_1) + \cdots + P(A|H_n)P(H_n)}$$

in the case of a finite partition $\mathcal{H}$, and

$$P(H_k|A) = \frac{P(A|H_k)P(H_k)}{P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \cdots}$$

when partition $\mathcal{H}$ is countably infinite.

Proof. Using formula (4.1) twice, we write

$$P(H_k|A) = \frac{P(H_k \cap A)}{P(A)} = \frac{P(A|H_k)P(H_k)}{P(A)}.$$

The theorem follows if one replaces the denominator on the right-hand side by the formulas (4.9) and (4.10). □

In the case of a partition (positive) into two events, $\mathcal{H} = \{B, B^c\}$, and any event A with $P(A) > 0$, we have

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

One of the interpretations of Bayes' theorem is when the partition $\mathcal{H}$ represents all possible mutually exclusive conditions (states of nature, hypotheses, etc.) that are logically possible. The probabilities $P(H_1), P(H_2), \ldots$ represent the prior knowledge, experience, or belief about the likelihood of $H_1, H_2, \ldots$. An event A is then observed, and this fact usually modifies the probabilities of H_i's. Accordingly $P(H_i)$ and $P(H_i|A)$ are often called *prior* and *posterior* probabilities of H_i.

■ **EXAMPLE 4.11**

Returning to Example 4.8, suppose that the person that we randomly selected is color blind. Intuitively, this should increase the chance that this person is a man. Indeed, we have here

$$P(M|D) = \frac{P(D|M)P(M)}{P(D|M)P(M) + P(D|W)P(W)} = \frac{\frac{pr}{1+r}}{\frac{pr}{1+r} + \frac{p^2}{1+r}} = \frac{r}{r+p},$$

a quantity close to 1 if p is small.

As mentioned above, Bayes' theorem can be used to reassess the probabilities of "states of nature" or "hypotheses" based on the additional evidence. How should one reassess the probabilities when the evidence comes sequentially? Specifically, if the evidence comes in two portions, say A' followed by A'', should one modify the prior probabilities given $A' \cap A''$, or should one first obtain posterior probabilities given A' and then use these posteriors as new priors, to be modified given A''?

The answer, as might be expected, is that it does not matter. We have the following theorem:

Theorem 4.4.2 (Updating the Evidence) *Let $\mathcal{H} = \{H_1, H_2, \ldots\}$ be a partition, and let A' and A'' be two events. If $P(A' \cap A'') > 0$, then for every event H_k in partition $\mathcal{H}$, we have*

$$\begin{aligned} P(H_k|A' \cap A'') &= \frac{P(A' \cap A''|H_k)P(H_k)}{\sum P(A' \cap A''|H_j)P(H_j)} \\ &= \frac{P(A''|H_k \cap A')P(H_k|A')}{\sum P(A''|H_j \cap A')P(H_j|A')}. \end{aligned} \tag{4.15}$$

Proof. The middle term is Bayes' formula applied to the left-hand side. We write

$$\begin{aligned} P(A' \cap A''|H_i)P(H_i) &= P(A' \cap A'' \cap H_i) \\ &= P(A''|A' \cap H_i)P(A' \cap H_i) \\ &= P(A''|H_i \cap A')P(H_i|A')P(A') \end{aligned}$$

to show the equality of the middle and right-hand-side terms. □

■ **EXAMPLE 4.12**

An urn contains two coins: One is a regular coin, with heads and tails, while the other has heads on both sides. One coin is chosen at random from the urn and tossed n times. The results are all heads. What is the probability that the coin tossed is a two-headed one?

SOLUTION. Intuitively, for a large n we expect the probability that the coin selected has two heads to be close to 1, since it is increasingly unlikely to get n heads in a row with a regular coin. Let H_1 and H_2 be the events "regular coin was chosen" and "coin with two heads was chosen." Clearly, H_1 and H_2 form a partition. Let the prior probabilities be $P(H_1) = P(H_2) = 1/2$, and let A_n be the event "n heads in a row"; our objective is to find $P(H_2|A_n)$. Since $P(A_n|H_2) = 1$ for all n (this coin will only give heads), and $P(A_n|H_1) = 1/2^n$, by Bayes' theorem we have

$$P(H_2|A_n) = \frac{P(A_n|H_2)P(H_2)}{P(A_n|H_1)P(H_1) + P(A_n|H_2)P(H_2)} = \frac{2^n}{2^n + 1}. \tag{4.16}$$

As expected, the probability (4.16) does approach 1 as n increases.

Suppose now that after A_n was observed, an additional m tosses again produced only heads (event B_m). Because $A_n \cap B_m$ is the same as A_{n+m}, the posterior probability of the two-headed coin (H_2) given A_{n+m} is $2^{n+m}/(2^{n+m}+1)$, after we replace $n+m$ for n in (4.16). Using the second part of formula (4.15), and the fact that $P(B_m|H_i \cap A_n) = P(B_m|H_i), i = 1, 2$, we obtain

$$\begin{aligned} P(H_2|A_n \cap B_m) &= \frac{P(B_m|H_2)P(H_2|A_n)}{P(B_m|H_1)P(H_1|A_n) + P(B_m|H_2)P(H_2|A_n)} \\ &= \frac{1 \times \frac{2^n}{2^n+1}}{\frac{1}{2^m} \times \frac{1}{2^n+1} + 1 \times \frac{2^n}{2^n+1}} = \frac{2^{n+m}}{2^{n+m}+1}, \end{aligned}$$

which agrees with the result of updating "all at once."

It might seem at first that this problem is purely artificial, invented for the sole purpose of providing practice for students. This is not so; the problem is of importance in breeding and selection. Consider a gene with two forms (alleles) A and a. Assume that A is dominant and a is recessive. so that we have two phenotypes: Individuals aa can be distinguished from the others, but individuals AA and Aa cannot be told apart. Moreover, assume that allele a is undesirable and we want to eliminate it, ultimately producing a pure strain of individuals of type AA only. The problem then lies in eliminating individuals Aa. One of the ways to do it is to allow crossbreeding between a tested individual (of type AA or Aa) with an individual of type aa (you may think here of plants that can easily be cross-pollinated). If the tested individual is of the type AA ("coin with two heads"), all offsprings will be of type Aa. However, if the tested individual is of the type Aa ("regular coin"), about half of the offspring will be of type Aa and half of type aa (corresponding to results of tosses being heads or tails with probability $1/2$). Now, if n offspring are of type Aa, the posterior probability that the tested individual is AA can be

computed as above, except that the prior probability need not be $1/2$ (it may be assessed from genetic theory and other information about the tested individual). Usually the breeder will accept an individual as a pure strain AA if the posterior probability that it is AA exceeds a threshold such as 0.99 or so, hence, if sufficiently many offspring of type Aa have been observed.

■ **EXAMPLE 4.13**

In a Monty Hall TV game there are three curtains, say A, B, and C, of which two hide nothing while behind the third there is a Big Prize. The Big Prize is won if it is guessed correctly which curtain hides it. You choose one of the curtains, say A. Before curtain A is pulled to reveal what is behind it, the game host pulls one of the two other curtains, say B, and shows that there is nothing behind it. He then offers you the option to change your decision (from curtain A to curtain C). Should you stick to your original choice (A) or change to C (or does it matter)? The answer to this question is counterintuitive and has stirred a lot of controversy. The common conviction is that it does not matter: There are two closed curtains, and the chances of the Big Prize being behind either of them must be fifty–fifty, so it is irrelevant whether or not you change your choice.

Actually, the answer is that *you should* switch to C; this way you double your chance of winning.

We assume that (1) the game host pulls one of the two curtains (B or C) that does not hide the Big Prize (this implies that the host knows where the Big Prize is), and (2) if both curtains B and C have nothing behind them, the host selects one at random.

Let A, B, and C be the events "Big Prize is behind curtain A" (respectively, B and C). We assume the original choice was curtain A, and let B^* be the event "host shows that there is nothing behind curtain B." We want $P(A|B^*)$. By Bayes' formula, taking A, B, C as a partition and assuming $P(A) = P(B) = P(C) = 1/3$,

$$\begin{aligned} P(A|B^*) &= \frac{P(B^*|A)P(A)}{P(B^*|A)P(A) + P(B^*|B)P(B) + P(B^*|C)P(C)} \\ &= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3}} = \frac{1}{3}. \end{aligned}$$

Since $P(B|B^*) = 0$, we must have $P(C|B^*) = 2/3$, so the chances of finding the Big Prize behind curtain C (i.e., winning if one changes the original choice) is $2/3$ and *not* $1/2$.

Here is an argument that might help convince people who claimed originally that the chances were $1/2$. Imagine that John and Peter are going to play the game a large number of times, say 300 times each. Each originally chooses A, and after being given a choice to switch, John never switches while Peter always switches.

John wins only if his original choice was correct, which happens in about 33% times (we assume that the Big Prize is each time randomly located behind

one of the curtains A, B, and C). He may expect to win about 100 times out of 300.

On the other hand, Peter wins whenever his original choice was *wrong* (because then switching is a correct choice). Consequently he wins in about 200 cases out of 300.

PROBLEMS

4.4.1 Suppose that medical science has developed a test for a certain disease that is 95% accurate, on both those who do and those who do not have the disease. If the incidence rate of this disease in the population is 5%, find the probability that a person: **(i)** Has the disease when the test is positive. **(ii)** Does not have the disease when the test is negative.

4.4.2 Two different suppliers, A and B, provide the manufacturer with the same part. All supplies of this part are kept in a large bin. In the past 2% of all parts supplied by A and 4% of parts supplied by B have been defective. Moreover, A supplies three times as many parts as B. Suppose that you reach into the bin and select a part. **(i)** Find the probability that this part is defective. **(ii)** If the part is nondefective, find the probability that it was supplied by B?

4.4.3 An urn originally contains three blue and two green chips. A chip is chosen at random from the urn, returned, and four chips of the opposite color are added to the urn. Then a second chip is drawn. Find the probability that: **(i)** The second chip is blue. **(ii)** Both chips are of the same color. **(iii)** The first chip was green if the second chip is blue.

4.4.4 One box contains six red and three green balls. The second box has six red and four green balls. A box is chosen at random. From this box two balls are selected and found to be green. Find the probability that the pair was drawn from the first box if the draws are: **(i)** Without replacement. **(ii)** With replacement.

4.4.5 Suppose that box A contains four red and five green chips and box B contains six red and three green chips. A chip is chosen at random from box A and placed in box B. Finally, a chip is chosen at random from those now in box B. What is the probability that a green chip was transfered given that a red chip was drawn from box B?

4.4.6 We have three dice, each with numbers $x = 1, \ldots, 6$, and with probabilities as follows: die 1: $p(x) = 1/6$, die 2: $p(x) = (7 - x)/21$, die 3: $p(x) = x^2/91$. A die is selected, tossed, and the number 4 appears. What is the probability that it is die 2 that was tossed?

4.4.7 Players A and B draw balls in turn, without replacement, from an urn containing three red and four green balls. A draws first. The winner is the person who draws the first red ball. Given that A won, what is the probability that A drew a red ball on the first draw?

4.4.8 A prisoner is sentenced to life in prison. One day the warden comes to him and offers to toss a fair coin for either getting free or being put to death. After some

deliberation the prisoner refuses, on the ground that it is too much risk: He argues that he may escape, or be pardoned, and so on. The warden asks him if he would agree to play such a game if the odds for death were 1:9 or less. The prisoner agrees.

Here is the game: An urn contains coins, labeled with digits on both sides. There is one coin labeled 1 and 2. There are nine coins labeled 2 and 3, and there are 81 coins labeled 3 and 4. A coin is to be selected at random by the warden and tossed. The prisoner can see the upper face and has to guess correctly the other face of the coin.

The prisoner decides to use the following guessing strategy: if 1 shows up, he would guess that the other side is 2 (and win). Similarly, if 4 shows up, he would guess that the other side is 3, and win. If 2 shows up, he would guess that the other side is 3, since there is one coin with 1 against nine coins with 3 on the other side. Similarly, if 3 shows up, he would guess 4, since there are nine coins with 2, against 81 coins with 4.

A coin was chosen, tossed and 2 appeared. The prisoner was about to say "3," when the following doubts occurred to him: "Suppose that this coin has indeed 3 on the other side. It is therefore a 2–3 coin. But this coin, before it was tossed, had equal chances of showing 2 and showing 3. But if it had shown 3, I would have guessed 4, and be put to death. So I played a game with fifty–fifty chances for death, which I decided not to play at the start."

As the story goes, the prisoner decided not to play. So he spent the rest of his life in prison contemplating whether or not he did the right thing. Find the probability that the prisoner would go free computed: **(i)** Before a coin is selected. **(ii)** After a coin of a given kind is selected. **(iii)** After a given face shows up.

4.4.9 One of three prisoners, A, B, and C, is to be executed the next morning. They all know about it, but they do not know who is going to die. The warden knows, but he is not allowed to tell them until just before the execution.

In the evening, one of the prisoners, say A, goes to the warden and asks him: "Please, tell me the name of one of the two prisoners, B and C, who is not going to die. If both are not to die, tell me one of their names at random. Since I know anyway that one of them is not going to die, you will not be giving me any information."

The warden thought about it for a while, and replied: "I cannot tell you who is not going to die. The reason is that now you think you have only $1/3$ chance of dying. Suppose I told you that B is not to be executed. You would then think that you have a $1/2$ chance of dying, so, in effect, I would have given you some information."

Was the warden right or was the prisoner right?

4.5 INDEPENDENCE

The notion of the conditional probability $P(A|B)$ introduced in Section 4.4 concerned the modification of the probability of an event A in light of the information that some other event B has occurred. Obviously an important special case here when is such information is irrelevant: Whether or not B has occurred, the chances of A remain the same. In such a case we say that the event A is *independent* of the event B. As we will see, the relation of independence defined in this way is symmetric: when A is independent of B, then B is also independent of A.

The essence of the idea above is that A is independent of B whenever $P(A|B) = P(A)$. Using (4.1), this implies that $P(A \cap B)/P(B) = P(A)$, and multiplying by $P(B)$—which we may to do, since $P(B) > 0$ by assumption—we obtain $P(A \cap B) = P(A)P(B)$. This relation is symmetric in A and in B, as asserted; moreover it holds also if one or both events have probability zero.

Consequently, we introduce the following definition:

Definition 4.5.1 We say that two events, A and B, are *independent* if their probabilities satisfy the *multiplication rule:*

$$P(A \cap B) = P(A)P(B). \tag{4.17}$$

□

In practice, (4.17) is used in two ways. First, we can compute both sides separately and compare them to check whether or not the two events are independent. More often, however, we *assume* that A and B are independent and use (4.17) for determining the probability of their intersection (joint occurrence). Typically the assumption of independence is justified on intuitive grounds (e.g., when the events A and B do not influence each other).

■ **EXAMPLE 4.14**

A box contains r red and g green balls. We draw a ball at random from the box, and then we draw another ball. Let R_1 and R_2 be the events "red ball on the first (second) draw." The question now is whether the events R_1 and R_2 are independent.

SOLUTION. The answer depends crucially on what happens after the first draw: Is the first ball returned to the box before the second draw or not? More generally, is the content of the box modified in any way by the first draw?

Suppose first that the ball drawn is returned to the box (the scheme known under the name *sampling with replacement*). Then $P(R_1) = P(R_2) = r/(r+g)$; hence $P(R_1)P(R_2) = [r/(r+g)]^2$. To check the independence, we have to compute $P(R_1 \cap R_2)$. Recall that the two draws with replacement may produce $(r+g)^2$ distinct results, by the product rule of the preceding chapter. The number of ways one can draw two red balls is—again, by the product rule—equal to r^2. Consequently $P(R_1 \cap R_2) = r^2/(r+g)^2$, which is the same as $P(R_1)P(R_2)$. This shows that events R_1 and R_2 are independent.

If we do not return the ball to the box (*sampling without replacement*), we can proceed as follows: We have $P(R_1 \cap R_2) = P(R_1)P(R_2|R_1)$ by (4.2), and $P(R_2|R_1) = (r-1)/(r+g-1)$, since if R_1 occurs, there remain $r+g-1$ balls in the box, of which $r-1$ are red. Consequently, $P(R_1 \cap R_2) = r(r-1)/(r+g)(r+g-1)$. To verify whether or not this last quantity equals $P(R_1)P(R_2)$, we need to evaluate $P(R_2)$. Observe that the probability of a red ball on the second draw is a random quantity depending on the outcome of the first draw. Taking events R_1 and G_1 as a partition and using the total probability formula, we can write $P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1)$, which equals $r/(r+g)$, so that $P(R_1 \cap R_2)$ is not the same as $P(R_1)P(R_2)$. Therefore the events in question are dependent.

We will prove next an important property of independent events that will facilitate many calculations.

Theorem 4.5.1 *If the events A and B are independent, so are the events A and B^c, A^c, and B, as well as A^c and B^c.*

Proof. It is sufficient to prove only the independence of A and B^c. Indeed, the independence of the second pair will follow by symmetry, and the independence of the third pair will follow by successive application of the two already proved statements.

Thus assume that $P(A \cap B) = P(A)P(B)$, and consider the event $A \cap B^c$. Since $A = (A \cap B) \cup (A \cap B^c)$, we have $P(A) = P(A \cap B) + P(A \cap B^c)$ so that

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) = P(A) - P(A)P(B) \\ &= P(A)(1 - P(B)) = P(A)P(B^c), \end{aligned}$$

which had to be shown. □

As shown in Example 4.14, the difference between sampling with and without replacement lies in the fact that the first leads to independent outcomes, while the second does not. However, when the population is large, this difference is negligible.

Specifically, we may regard two events A and B as "almost independent" if the difference between $P(A \cap B)$ and $P(A)P(B)$ is small. Let us compute the difference

$$Q = |P(R_1 \cap R_2) - P(R_1)P(R_2)|$$

in the case of sampling without replacement (obviously, for sampling with replacement this difference is zero). We have here

$$Q = \left| \binom{r}{r+g} \left[\binom{r-1}{r+g-1} - \binom{r}{r+g} \right] \right| = \frac{rg}{(r+g)^2(r+g-1)},$$

a quantity that is small when either r or g is large (or both are).

It appears that we might take $|P(A \cap B) - P(A)P(B)|$ as a measure of degree of dependence between events A and B. This is indeed the case, except that we have to take into account the fact that this difference may be small simply because one of the events A or B has small probability. Consequently we need to "standardize" the difference. This leads to the following definition:

Definition 4.5.2 Assume that $0 < P(A) < 1$ and $0 < P(B) < 1$. The quantity

$$r(A, B) = \frac{P(A \cap B) - P(A)P(B)}{\sqrt{P(A)(1 - P(A))P(B)(1 - P(B))}} \tag{4.18}$$

will be called the *coefficient of correlation* between events A and B. □

The coefficient of correlation is a measure of dependence in the sense specified by the following theorem:

Theorem 4.5.2 *The coefficient of correlation $r(A, B)$ is zero if and only if A and B are independent. Moreover, we have $|r(A, B)| \leq 1$, with $r(A, B) = 1$ if and only if $A = B$, and $r(A, B) = -1$ if and only if $A = B^c$.*

Proof. The proof is by elementary checking, except for the "only if" parts of the last two statements. These parts follow from the general theorem on the properties of correlation coefficients for random variables, to be discussed in Chapter 7. □

We can extend the concept of independence to the case of more than two events.

Definition 4.5.3 We say that the events $A_1, A_2, \ldots, A_n$ are *independent* if for any set of indices $i_1, i_2, \ldots i_k$ with $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ we have the following multiplication rule:

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}). \tag{4.19}$$

The events in an infinite sequence $A_1, A_2, \ldots$ are called independent if for every n, the first n events of the sequence are independent. □

Thus in the case of n events we have one condition for every subset of the size $k \geq 2$, that is, for $k = 1$, condition (4.19) is satisfied. Since the number of all subsets of a set with n elements is 2^n, the number of conditions in (4.19) is $2^n - n - 1$.

In the case of three events, A, B, C, the definition (4.19) represents four conditions: $P(A \cap B \cap C) = P(A)P(B)P(C)$, and three conditions for pairs, $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$ and $P(B \cap C) = P(B)P(C)$. In case of four events A, B, C, and D, the multiplication rule (4.19) must hold for the quadruplet (A, B, C, D), for four triplets (A, B, C), (A, B, D), (A, C, D), and (B, C, D), and for six pairs (A, B), (A, C), (A, D), (B, C), (B, D), and (C, D).

The question arises: Do we really need so many conditions? Taking the simplest case of $n = 3$ events, it might seem that independence of all three possible pairs, namely (A, B), (A, C) and (B, C), should imply the condition $P(A \cap B \cap C) = P(A)P(B)P(C)$, a direct analogue of the defining condition for pairs of events.

Conversely, one might expect that the multiplication rule $P(A \cap B \cap C) = P(A)P(B)P(C)$ implies independence of pairs (A, B), (A, C), and (B, C), as well as conditions of the kind $P(A \cap B^c \cap C^c) = P(A)P(B^c)P(C^c)$, and so on. In fact, none of these implications is true, as will be shown by examples.

■ **EXAMPLE 4.15**

Independence in pairs does not imply the multiplication rule for more events. We will use the case of three events A, B, and C, and show that the condition $P(A \cap B \cap C) = P(A)P(B)P(C)$ need not hold, even if all three pairs of events are independent.

Suppose that we toss a die twice and let A = "odd outcome on the first toss," B = "odd outcome on the second toss," and C = "sum odd." We have $P(A) = P(B) = P(C) = 1/2$. By simple counting, we obtain $P(A \cap B) = P(A \cap C) = P(B \cap C) = 1/4$, so each of the possible pairs is independent. However, $P(A)P(B)P(C) = 1/8$, while $P(A \cap B \cap C) = 0$, since the events A, B, and C are mutually exclusive. If A and B hold, then the sum of outcomes must be even so that C does not occur.

■ **EXAMPLE 4.16**

Multiplication rule does not imply pairwise independence. We will now show that independence in pairs is not implied by the condition

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

We take two events A and B that are dependent, and then take C with $P(C) = 0$. Let the sample space be $S = \{a, b, c, d, \ldots\}$, and let $A = \{a, d\}$, $B = \{b, d\}$, and $C = \{c, d\}$. Moreover, let $P(a) = P(b) = P(c) = p - p^3$, and $P(d) = p^3$, where $p > 0$ satisfies the inequality $3(p - p^3) + p^3 \leq 1$. The remaining points (if any) have arbitrary probabilities, subject to the usual constrain that the sum of probabilities is 1. We have here $P(A) = P(B) = P(C) = (p - p^3) + p^3 = p$; hence $P(A \cap B \cap C) = P(A)P(B)P(C) = p^3$. However, since $A \cap B = A \cap C = B \cap C = A \cap B \cap C$, we have $P(A \cap B) = p^3$ while $P(A)P(B) = p^2$, and similarly for other pairs.

The possibility of events being pairwise independent, but not totally (mutually) independent, is mainly of theoretical interest. Unless explicitly stated otherwise, whenever we speak of independent events, we will always mean mutual independence in the sense of Definition 4.5.3.

Let us state here the following analogue of Theorem 4.5.1 for the case of n events.

Theorem 4.5.3 *If the events $A_1, A_2, \ldots, A_n$ are independent, the same is true for events $A_1', A_2', \ldots, A_n'$, where for each k the event A_k' stands for either A_k or its complement A_k^c.*

The following theorem provides one of the most commonly used "tricks of the trade" in probability theory:

Theorem 4.5.4 *If the events $A_1, A_2, \ldots, A_n$ are independent, then*

$$P(A_1 \cup \cdots \cup A_n) = 1 - [1 - P(A_1)][1 - P(A_2)] \cdots [1 - P(A_n)].$$

Proof. The proof uses the fact that can be summarized as $P(\text{at least one}) = 1 - P(\text{none})$. By De Morgan's law, $(A_1 \cup A_2 \cup \cdots \cup A_n)^c = A_1^c \cap A_2^c \cap \cdots A_n^c$. Using Theorem 4.5.3, we write

$$\begin{aligned} 1 - P(A_1 \cup \cdots \cup A_n) &= P(A_1^c \cap A_2^c \cap \cdots \cap A_n^c) \\ &= P(A_1^c)P(A_2^c) \cdots P(A_n^c) \\ &= [1 - P(A_1)][1 - P(A_2)] \cdots [1 - P(A_n)], \end{aligned}$$

as was to be shown. □

■ **EXAMPLE 4.17**

Consider a sequence of independent experiments such that in each of them an event A (usually labeled "success") may or not occur. Let $P(A) = p$ for every experiment, and let A_k be the event "success at the kth trial." Then $A_1 \cup \cdots \cup A_n$ is the event "at least one success in n trials," and consequently

$$P(\text{at least one success in } n \text{ trials}) = 1 - (1 - p)^n.$$

At the end of this section we will prove the theorem complementing Theorem 2.6.2 of Chapter 2.

Theorem 4.5.5 (Second Borel-Cantelli Lemma) *If $A_1, A_2, \ldots$ is a sequence of independent events such that $\sum_{n=1}^{\infty} P(A_n) = \infty$, then*

$$P(\limsup A_n) = 1.$$

Proof. We have $\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$ = {infinitely many A_i's will occur}. By De Morgan's Law, $[\limsup A_n]^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c$. Since the intersections $\bigcap_{k=n}^{\infty} A_k^c$ increase with n, we have

$$P([\limsup A_n]^c) = \lim_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k^c\right) = \lim_{n\to\infty} \lim_{N\to\infty} P\left(\bigcap_{k=n}^{N} A_k^c\right).$$

If A_k are independent, so are A_k^c. Then, using the inequality $1 - x \leq e^{-x}$, we obtain

$$\begin{aligned} \lim_{N\to\infty} P\left(\bigcap_{k=n}^{N} A_k^c\right) &= \lim_{N\to\infty} [1 - P(A_n)] \cdots [1 - P(A_N)] \\ &\leq \lim_{N\to\infty} e^{-[P(A_n)+\cdots+P(A_N)]} = 0 \end{aligned}$$

in view of the assumption that the series $\sum P(A_n)$ diverges. Consequently,

$$P([\limsup A_n]^c) = 0,$$

as was to be shown. □

PROBLEMS

4.5.1 Label the statements true or false.
(i) The target is to be hit at least once. In three independent shots at the target (instead of one shot) you triple the chances of attaining the goal (assume each shot has the same positive chance of hitting the target).
(ii) If A and B are independent, then $P(A^c|B^c) = 1 - P(A)$.
(iii) If A and B are independent, then they must be disjoint. **(iv)** If A and B are independent, $P(A) = P(B)$, and $P(A \cup B) = 1/2$, then $P(A) > 1/4$.

4.5.2 Suppose that A and B are independent events such that $P(A \cap B^c) = 1/3$ and $P(A^c \cap B) = 1/6$. Find $P(A \cap B)$.

4.5.3 Events A and B are independent, $P(A) = kP(B)$, and at least one of them must occur. Find $P(A^c \cap B)$.

4.5.4 Events A and B are independent, A and C are mutually exclusive, and B and C are independent. Find $P(A \cup B \cup C)$ if $P(A) = 0.5, P(B) = 0.25$, and $P(C) = 0.125$.

4.5.5 If disjoint events A and B have positive probabilities, check independence of events in the following pairs : $\emptyset$ and A, A and B, A and S, A and $A \cap B$, $\emptyset$ and A^c.

4.5.6 Let X be the number on the ball randomly selected from a box containing 12 balls, labeled 1 through 12. Check pairwise independence of events A, B, and C, defined as: X is even, $X \geq 7$, and $X < 4$, respectively.

4.5.7 The probability that a certain event A occurs at least once in three independent trials exceeds the probability that A occurs twice in two independent trials. Find possible values of $P(A)$.

4.5.8 Suppose that a point is picked at random from the unit square $0 \leq x \leq 1, 0 \leq y \leq 1$. Let A be the event that it falls in the triangle bounded by the lines $y = 0, x = 1$, and $x = y$, and let B be the event that it falls into the rectangle with vertices $(0,0), (1,0), (1,1/2)$ and $(0,1/2)$. Find all statements that are true: **(i)** $P(A \cup B) = P(A \cap B)$. **(ii)** $P(A|B) = 1/8$. **(iii)** A and B are independent. **(iv)** $P(A) = P(B)$. **(v)** $P(A \cap B) = 3/8$.

4.5.9 Events A and B are such that $3P(A) = P(B) = p$, where $0 < p < 1$. Find the correct answers in parts (i) and (ii).
(i) The relation $P(B|A) = 3P(A|B)$ is: (a) True. (b) True only if A and B are disjoint. (c) True only if A and B are independent. (d) False.
(ii) The relation $P(A \cap B^c) \leq \min(p/3, 1-p)$ is: (a) True. (b) False.

4.5.10 A coin is tossed six times. Find the probability that the number of heads in the first three trials is the same as the number of heads in the last three trials.

4.5.11 Two people take turns rolling a die. Peter rolls first, then Paul, then Peter again, and so on. The winner is the first to roll a six. What is the probability that Peter wins?

4.5.12 A coin with probability p of turning up heads is tossed until it comes up tails. Let X be the number of tosses required. You bet that X will be odd, and your opponent bets that X will be even. For what p is the bet advantageous to you? Is there a p such that the bet is fair?

4.5.13 Find the probability that in repeated tossing of a pair of dice, a sum of 7 will occur before a sum of 8.

4.5.14 Three people, A, B, and C, take turns rolling a die. The first one to roll 5 or 6 wins, and the game is ended. Find the probability that A will win.

4.5.15 Consider a die in which the probability of a face is proportional to the number of dots on this face. What is the probability that in six independent throws of this die each face appears exactly once?

4.5.16 A machine has three independent components, two that fail with probability p and one that fails with probability 0.5. The machine operates as long as at least two parts work. Find the probability that the machine will fail.

4.5.17 The French mathematician Jean D'Alembert claimed that in tossing a coin twice, we have only three possible outcomes: "two heads," "one head," and "no heads." This is a legitimate sample space, of course. However, D'Alembert also claimed that each outcome in this space has the same probability $1/3$. **(i)** Is it possible to have a coin biased in such a way so as to make D'Alembert's claim true? **(ii)** Is it possible to make two coins with different probabilities of heads, confirm D'Alembert's claim?

4.5.18 Is it possible to bias a die in such a way that in tossing the die twice, each sum $2, 3, \ldots, 12$ has the same probability?

4.5.19 An athlete in a high jump competition has the right of three attempts at each height. Suppose that his chance of clearing the bar at height h is equal to $p(h)$, independently of the results of previous attempts. The heights to be attempted are set by the judges to be $h_1 < h_2 < \cdots$. An athlete has the right to skip trying a given height. Let Y be his final result, that is, the highest h that he actually cleared. Let A stand for the strategy "try all heights," and let B stand for the strategy "skip the first height and then try every second height." Find probability $P(Y = h_{2k})$ under strategies A and B.

4.6 EXCHANGEABILITY; CONDITIONAL INDEPENDENCE

At the end of this chapter we introduce an important type of dependence between events.

Definition 4.6.1 Let A, B, and H be three events, and let $P(H) > 0$. We say that events A and B are *conditionally independent given* H, if

$$P(A \cap B|H) = P(A|H)P(B|H). \qquad \square$$

Definition 4.6.2 Let $\mathcal{H} = \{H_1, H_2, \ldots\}$ be a positive partition (finite or countably infinite). We say that events A and B are *conditionally independent given* $\mathcal{H}$, if A and B are conditionally independent given any set H_i in partition $\mathcal{H}$. $\square$

■ EXAMPLE 4.18

In order to understand why events that are conditionally independent with respect to every set of the partition may be dependent, consider the following simple situation: Let the partition consist of two events, H and H^c, with $P(H) = 0.6$. Suppose that if H occurs, then both A and B are likely, say $P(A|H) = 0.8, P(B|H) = 0.9$, and conditional independence requires that $P(A \cap B|H) = 0.8 \times 0.9 = 0.72$. On the other hand, if H^c occurs, then both A and B are rather unlikely, say $P(A|H^c) = P(B|H^c) = 0.1$. Again, conditional independence requires that $P(A \cap B|H^c) = (0.1)^2 = 0.01$

It is easy to check that the events A and B are dependent. Indeed, $P(A) = P(A|H)P(H) + P(A|H^c)P(H^c) = 0.8 \times 0.6 + 0.1 \times 0.4 = 0.52, P(B) = P(B|H)P(H) + P(B|H^c)P(H^c) = 0.9 \times 0.6 + 0.1 \times 0.4 = 0.58$. On

the other hand, $P(A \cap B) = P(A \cap B|H)P(H) + P(A \cap B|H^c)P(H^c) = 0.72 \times 0.6 + 0.01 \times 0.4 = 0.436$, which is not equal to $P(A)P(B) = 0.3016$.

What is happening here is that the occurrence or nonoccurrence of one of the events allows us to make rather reliable predictions about the other event. In the case under considerations, the events A and B are positively correlated: If one occurs, the other is likely to occur also, and vice versa. By Definition 4.5.2 we have

$$r(A, B) = \frac{0.436 - 0.3016}{\sqrt{0.52 \times 0.48 \times 0.58 \times 0.42}} = 0.545.$$

The nature of dependence that we have here might better be visualized by a real-life situation. Imagine some animals, such as birds, that raise their young every year. If the conditions in a given year are hard (a draught, severe winter, etc.), often all the young die, and typically very few survive. If the conditions are favorable, the number of young that make it to next year is typically higher. The fates of different families *in any given year* are independent, in the sense that survival of offspring in one family does not depend on the survival of offspring in the other. If A and B are events occurring in two different families, and H_i's are events describing possible types of conditions in any given year, then A and B are conditionally independent with respect to any given events H_i.

Unconditionally, however, events A and B are dependent, the dependence arising from the fact that all families are subject to the same conditions. Thus, if one family fares badly, then we can expect that others will also, simply because the conditions are likely to be hard for them all.

The definition of conditional independence given an event, and therefore also definition of conditional independence given a partition, extends naturally to the case of n events. The extension consists of first defining the conditional independence of events $A_1, A_2, \ldots, A_n$ given an event H_i. We will not give the definition here. It is a repetition of Definition 4.5.3, the only difference being that the unconditional probabilities in formula (4.19) are replaced by conditional probabilities given H_i. Definition 4.6.2 remains unchanged.

Now let $A_1, A_2, \ldots, A_n$ be a set of events, and let $N = \{1, 2, \ldots n\}$ be the set of their indices. We introduce the following definition:

Definition 4.6.3 The events $A_1, A_2, \ldots, A_n$ are called *exchangeable* if for any subset K of $\{1, 2, \ldots, n\}$ the probability

$$P\Big(\bigcap_{i \in K} A_i \cap \bigcap_{j \notin K} A_j^c\Big)$$

of joint occurrence of all events with indices in K and nonoccurrence of all events with indices not in K depends only on the size of the set K. □

This means that for exchangeable events, the probability of occurrence of exactly m of the events does not depend on which events are to occur and which are not. For instance, in the case of three events A, B, and C, their exchangeability means that

$P(A \cap B^c \cap C^c) = P(A^c \cap B \cap C^c) = P(A^c \cap B^c \cap C)$ and also $P(A \cap B \cap C^c) = P(A \cap B^c \cap C) = P(A^c \cap B \cap C)$.

We have the following theorem:

Theorem 4.6.1 *If the events $A_1, A_2, \ldots, A_n$ are conditionally independent given a partition $\mathcal{H} = \{H_1, H_2, \ldots\}$, and $P(A_i|H_k) = P(A_j|H_k)$ for all i, j, k, then they are exchangeable.*

Proof. Let $P(A_i|H_k) = w_k$ (by assumption, this probability does not depend on i). The probability that, say, the first r events will occur and the remaining ones will not (by the law of total probability and conditional independence) is

$$\begin{aligned} & P(A_1 \cap A_2 \cap \cdots A_r \cap A_{r+1}^c \cap \cdots \cap A_n^c) \\ &= \sum_k P(A_1 \cap A_2 \cap \cdots \cap A_r \cap A_{r+1}^c \cap \cdots \cap A_n^c | H_k) P(H_k) \\ &= \sum_k w_k^r (1 - w_k)^{n-r} P(H_k). \end{aligned} \tag{4.20}$$

Clearly, (4.20) is also the probability of occurrence of *any* r among the events $A_1, \ldots, A_n$ and nonoccurrence of the remaining ones. □

As can be expected, in case of conditional independence of A' and A'' with respect to the partition $\mathcal{H}$, Theorem 4.4.2 on updating the evidence will take a simpler form. We have

Theorem 4.6.2 *Let $\mathcal{H} = \{H_1, H_2, \ldots\}$ be a partition and let A', A'' be two events conditionally independent with respect to $\mathcal{H}$. If $P(A' \cap A'') > 0$, then for every event H_k in partition $\mathcal{H}$ we have*

$$\begin{aligned} P(H_k|A' \cap A'') &= \frac{P(A'|H_k)P(A''|H_k)P(H_k)}{\sum_j P(A'|H_j)P(A''|H_j)P(H_j)} \\ &= \frac{P(A''|H_k)P(H_k|A')}{\sum_j P(A''|H_j)P(H_j|A')}. \end{aligned} \tag{4.21}$$

Proof. The middle part shows how the conditional probability of H_k is computed if the "evidence" $A' \cap A''$ is taken jointly. The last part shows how the updating is done successively, if first one updates probabilities using A' and then one uses A'' for further modification of posterior probabilities.

For the first equation, observe that the middle part of (4.21) is just the middle part of (4.15) in the case of conditional independence.

To prove the second equation, observe that the factor $P(A''|H_k \cap A')$ in the right-hand side of (4.15) can be written as

$$\begin{aligned} P(A''|H_k \cap A') &= \frac{P(A' \cap A'' \cap H_k)}{P(A' \cap H_k)} = \frac{P(A' \cap A''|H_k)P(H_k)}{P(A'|H_k)P(H_k)} \\ &= \frac{P(A'|H_k)P(A''|H_k)}{P(A'|H_k)} = P(A''|H_k). \end{aligned}$$ □

PROBLEMS

4.6.1 A subset S of size t, $1 \leq t \leq N$, is selected at random from the set $\{1, \ldots, N\}$ and event A_i, $i = 1, \ldots, n$, is defined as: "Element i was among the elements selected." If $S = \{i_1, \ldots, i_t\}$ is chosen, we say that events $A_{i_1}, \ldots, A_{i_t}$ occur, while the remaining events do not. **(i)** Show that events $A_1, \ldots, A_N$ are exchangeable. **(ii)** Find the correlation coefficient $r = r(A_i, A_j), i \neq j$ between two events. Give an intuitive explanation why $r < 0$.

4.6.2 Generalizing the scheme of Problem 4.6.1, let p_t be the probability of choosing size t for subset $S_t \subset \{1, \ldots, N\}$, $1 \leq t \leq N$. After choosing t, subset S_t is selected at random, and all events with indices in S_t occur while other events do not. **(i)** Argue that events $A_1, \ldots, A_N$ are exchangeable. **(ii)** Find probabilities $P(A_1^c)$ and $P(A_1 \cap A_2)$.

CHAPTER 5

MARKOV CHAINS*

5.1 INTRODUCTION AND BASIC DEFINITIONS

Thus far we have learned only a mere beginning of the conceptual structure of probability theory, namely: sample spaces in Chapter 1, axioms and the simplest laws that can be deduced from them in Chapter 2, and the concept of conditional probability with accompanying law of total probability and Bayes' rule in Chapter 4. Additionally, in Chapter 3, we learned how to compute or assign probabilities in discrete setup. It may therefore come as a surprise that even with such limited tools, it is already possible to develop an extensive and powerful theory with numerous practical applications and cognitive consequences, namely the theory of Markov chains.

Generally, the term *chain* will denote a sequence of events, typically dependent one on another in some way. It will be convenient to use the terminology referring to time: we can think of events as occuring one after another, so that the event whose occurrence is actually observed is the "present" event, while the events following it belong to the "future" and the remaining ones to the "past." This way we obtain a description of a process of dynamic changes (of something that we generally call a *system*). Quite naturally the main problem will be to develop tools for making some inference (prediction, etc.) of the future on the basis of the knowledge of the past.

Specifically, consider a system that evolves in a random manner. The observations are taken at some fixed times $t_0 < t_1 < t_2 < \cdots$, so in effect we record

a sequence of states at which we find the system at the times of the observations. Without loss of generality, we let $t_n = n$ so that the observations are taken every unit of time starting at $t = 0$.

The notions of "system" and "state" in the description above are left unspecified in order to have flexibility in applying the theory to various situations. Before proceeding any further, it is worthwhile to consider some examples that will later guide our intuition.

■ EXAMPLE 5.1

Imagine a gambler who plays a sequence of games. In each game she may win or lose some amount of money. Let w_n denote his net winnings in nth game (loss, if w_n is negative). If we are interested only in the financial aspects of the situation, we can regard the gambler as a "system," with the state of the system identified with the gambler's fortune. Thus the state in this case is a number. Letting $s(n)$ be the state at the time immediately after nth game, and letting $s(0)$ be the initial fortune, we have $s(n) = s(0) + w_1 + w_2 + \cdots + w_n$.

This scheme has to be made specific by adding appropriate assumptions. First, we need assumptions about the nature of the winnings w_k. In the simplest case, the gambler may play for the same fixed amount, say \$1, so that $s(n)$ either increases or decreases by 1 with each game, and the outcomes of consecutive games are independent. Another kind of assumption is that of "boundary" conditions, which specify what happens when the gambler becomes ruined, that is, when $s(n)$ reaches the value 0. For example, the game may end, with $s(n)$ remaining 0 forever, or it may continue from the state $s(n) = 1$ if the gambler borrows a dollar to continue playing, and so forth.

Such an evolving scheme is often called a "random walk" because one can interpret the state $s(n)$ as a position of a pawn that moves randomly along the x-axis, being shifted at the nth move by the amount w_n.

The interesting question is: What is the probability that after the nth move at time t, the distance $s(n) - s(0)$ from the initial location will satisfy the inequality $a < s(n) - s(0) < b$? We will return to this problem in later chapters (especially Chapter 10) as we develop tools for analyzing the properties of random sums of the form $w_1 + \cdots + w_n$, representing total displacement after n hits.

■ EXAMPLE 5.2

Consider the following scheme aimed at describing the evolution of an epidemic. Imagine a group of subjects, of whom some may be infected with the disease and therefore capable of spreading it to others (even though they may not yet be aware of being infectious), and some other persons may be susceptible. The spread of the disease is generated by the process of contact whereby proximity between an infective and a susceptible subject can lead to an infection. The state of the system can be described by a pair of numbers $s(n) = [x(n), y(n)]$, where $x(n)$ is the number of infectives at time $t = n$, and $y(n)$ is the number of susceptibles at time $t = n$.

By taking a sufficiently small time between observations, one can assume that the only possible changes from $[x(n), y(n)]$ to $[x(n+1), y(n+1)]$ are as follows:

1. $x(n+1) = x(n)+k, y(n+1) = y(n)-k$ (infection of a group of k susceptibles, $k = 1, 2, \ldots$).

2. $x(n+1) = x(n) - 1, y(n+1) = y(n)$ ("removal" of an infective, which may correspond to death, isolation in the hospital, recovery with immunity, etc.).

3. "Status quo," that is, $x(n+1) = x(n), y(n+1) = y(n)$.

The probabilities of these three types of transitions depend on the nature of the contact process and the mode of transmission of the given disease (in particular, the process simplifies somewhat if we assume that k is always 1). By specifying these probabilities, one may obtain models of spread of various infectious diseases.

Examples 5.1 and 5.2 show that the notions of the *system* and *state* are rather flexible and can be interpreted in a number of ways. Basically, *system* here means any fragment of reality that we want to model and analyze, while *state* is the information about the system that is relevant for the purpose of the study.

In this chapter we will regard time as discrete. In general theory one also considers systems evolving in continuous time, when changes can occur and be observed at any time. The theory designed to describe and analyze randomly evolving systems is called the *theory of stochastic processes*. While such theory lies beyond the scope of the present book, we will occasionally introduce some elements of it.

PROBLEMS

5.1.1 Customers arrive at a service station (a taxi stand, a cable car lift at a skiing resort, etc.) and form a queue. At times $t = 1, 2, \ldots$ the first m customers ($m \geq 1$) in the queue are served (if there are that many). Let $Y_1, Y_2, \ldots$ denote the numbers of customers arriving during the time intervals between services, and let X_n denote the number of customers in the system at the time immediately preceding the nth service. Argue that $X_{n+1} = (X_n - m)^+ + Y_n$, where $U^+ = \max(U, 0)$. Assuming that the events $\{Y_{n_1} = k_1\}, \ldots, \{Y_{n_j} = k_j\}$ are independent for any $j, k_1, \ldots, k_j$ and any distinct $n_1, \ldots, n_j$, and that $P\{Y_n = k\} = p_k$, are independent of n, find $P(X_{n+1} = j | X_n = i)$ for all possible $i, j = 0, 1, 2, \ldots$.

5.1.2 After some time spent in a bar, Peter starts to walk home. Suppose that the streets form a rectangular grid. Peter always walks to the nearest corner and then decides on the direction of the next segment of his walk (so that he never changes direction in the middle of the block). Define the "state of the system" in such a way as to accommodate conveniently the assumption that Peter never goes directly back to the corner he just left.

5.2 DEFINITION OF A MARKOV CHAIN

It ought to be clear from the examples in Section 5.1 that a sample space that may be useful for describing the evolution of a system consists of all possible sequences of states at the observation times. Each such sequence represents a possible history of the process. Formally, let $\mathcal{E}$ be the set of all possible states of the system, with elements of $\mathcal{E}$ denoted by letter e, possibly with identifying subscripts or superscripts, such as e_1, e'. The sample space $\mathcal{S}_N$, which will represent the history of the system at times $t = 0, 1, \ldots, N$, will consist of all possible sequences $s_N = [s(0), \ldots, s(N)]$, where $s(i)$ is the element of $\mathcal{E}$ describing the state of the system at time $t = i$.

In this chapter we will assume that the set $\mathcal{E}$ is either finite or countably infinite. We may then label the states by integers so that $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$ or $\mathcal{E} = \{e_1, e_2, \ldots\}$. Under this assumption each of the sample spaces $\mathcal{S}_N$ is also either finite or countably infinite, and we will define the probabilities on individual sample points of $\mathcal{S}_N$ by explicit formulas. At the same time, however, we will consider certain limiting passages, with the lengths of the sequences of states increasing to infinity. The space of all infinite sequences of states in $\mathcal{E}$ is uncountable whether or not $\mathcal{E}$ is finite. The details of the construction of probability of such spaces (although sketched in Chapter 2) lie beyond the scope of this book. However, we will use simple limiting passages with the lengths of sequences increasing to infinity, basically treating each space $\mathcal{S}_N$ as a space of all beginnings of length N of infinite sequences.

The main types of events that we will consider are of the form "state e at time $t = n$ ". In Chapter 1, events were identified with subsets of the sample space. Accordingly, the event above can be defined as

$$\begin{aligned} \text{"state } e \text{ at time } t = n\text{"} &= \{s(n) = e\} \\ &= \text{set of all sequences} [s(0), \ldots, s(N)] \\ &\quad \text{with } s(n) \text{ equal } e. \end{aligned}$$

Typically the probability of the next state depends in some way on the preceding states. So, by the chain formula (4.5), for all $n \leq N$ we have

$$\begin{aligned} &P[s(0) = e_{i_0}, s(1) = e_{i_1}, \ldots, s(n) = e_{i_n}] = P[s(0) = e_{i_0}] \qquad (5.1) \\ &\times P[s(1) = e_{i_1} | s(0) = e_{i_0}] \times \cdots \times P[s(n) = e_{i_n} | s(n-1) = e_{i_{n-1}}, \ldots, s(0) = e_{i_0}]. \end{aligned}$$

The commas in formula (5.1) signify the operation of an intersection of events, thus replacing a rather clumsy notation $P[s(0) = e_{i_0} \cap s(1) = e_{i_1} \cap \cdots \cap s(n) = e_{i_n}]$. For convenience, the events in the conditions are written in the order that starts from the most recent one. Formula (5.1) expresses the probability of a sequence of consecutive states through the probability of the "initial" state $s(0)$, and conditional probabilities involving all preceding states, and thus encompasses longer and longer fragments of the past. By the definition below, such conditional probabilities for Markov chains can be simplified by reaching only to the most recent state.

Definition 5.2.1 The evolution of a system is said to form a *Markov chain* if for every n and every vector $(e_{i_0}, e_{i_1}, \ldots, e_{i_{n-1}}, e_{i_n})$ of states, the conditional proba-

bilities satisfy the following relation, called the *Markov property:*

$$\begin{aligned} &P[s(n) = e_{i_n} | s(n-1) = e_{i_{n-1}}, s(n-2) = e_{i_{n-2}}, \ldots, s(0) = e_{i_0}] \\ &\quad = P[s(n) = e_{i_n} | s(n-1) = e_{i_{n-1}}]. \end{aligned} \tag{5.2}$$

We have therefore the next theorem.

Theorem 5.2.1 *If the sequence of transitions between states constitutes a Markov chain, then*

$$\begin{aligned} &P[s(0) = e_{i_0}, \ldots, s(n) = e_{i_n}] \\ &\quad = P[s(0) = e_{i_0}] \prod_{j=1}^{n} P[s(j) = e_{i_j} | s(j-1) = e_{i_{j-1}}]. \end{aligned} \tag{5.3}$$

Generally, the transition probability $P[s(j) = e_{i_j} | s(j-1) = e_{i_{j-1}}]$ from state $e_{i_{j-1}}$ to state e_{i_j} at time $t = j$ depends on j. If this is *not* the case, the situation can be greatly simplified, and this special case is the starting point of a theory. This theory is practically and cognitively useful, leads to rather deep mathematical results and fruitful generalizations, and is pleasantly elegant as well. Accordingly we introduce the following definition:

Definition 5.2.2 A Markov chain will be said to have *stationary transition probabilities*, or be *time homogeneous*, if for all states e_i, e_j the probability

$$P[s(t) = e_j | s(t-1) = e_i]$$

does not depend on t.

From now on, *all Markov chains under consideration in this chapter will have stationary transition probabilities*. Writing $p_{ij} = P[s(t) = e_j | s(t-1) = e_i]$ and $r_i = P[s(0) = e_i]$, we may recast (5.3) as

$$P[s(0) = e_{i_0}, s(1) = e_{i_1}, \ldots, s(n) = e_{i_n}] = r(e_{i_0}) p_{i_0 i_1} \cdots p_{i_{n-1} i_n}.$$

Clearly, $\{r_i, i = 1, 2, \}$, called the *initial probability distribution*, and the set of probabilities p_{ij}, called the *transition probability* matrix, must satisfy the following conditions:

$$\sum_i r_i = 1 \tag{5.4}$$

and

$$\sum_j r_{ij} = 1 \quad \text{for every } i. \tag{5.5}$$

The last condition means simply that if the system is in some state e_i, it must pass to one of the states e_j (possibly remaining in the same state e_i) in the next step.

Since square matrices (finite or not) with nonnegative entries and row sums equal to 1 are called *stochastic matrices*, we see—in view of (5.5)—that every transition probability matrix of a Markov chain is a stochastic matrix.

■ EXAMPLE 5.3 Gambler's Ruin

Continuing Example 5.1 about the gambler, assume that each game is played for a unit stake, with the gambler's probability of winning a game being p. Moreover, assume that the games are independent. In such a case the sequence $s(0), s(1), \ldots$ of the gambler's fortunes after consecutive games is a Markov chain. Indeed, letting $P[s(n+1) = j | s(n) = i] = p_{ij}$, we have for all $k > 0$,

$$p_{k,k+1} = p, \; p_{k,k-1} = 1 - p,$$

and $p_{kj} = 0$ for all j other than $k + 1$ or $k - 1$. These probabilities do not depend on the history $s(n-1), s(n-2), \ldots$ preceding the nth game. The transition probabilities for $k = 0$ depend on the assumption about what happens when the gambler becomes ruined. For instance, if upon reaching $k = 0$ the game stops, we can put p_{0j} equal to 1 for $j = 0$ and equal to 0 for all other j. If the process starts with the gambler having initial capital M, then r_j equals 1 for $j = M$ and equals 0 for all other j. It seems plausible (and in fact it is true) that if $p < 1/2$ (the games are unfavorable for the gambler), then regardless of the initial state, the state 0 (gambler's ruin) will sooner or later be reached. Less obvious, but still true, is that the same holds for the fair games ($p = 1/2$) no matter what the initial fortune is. An interesting question is to determine the probability of the gambler ever reaching 0 (i.e., of becoming eventually ruined) if she plays favorable games, that is, if $p > 1/2$. This probability depends on the initial state M, and may be shown to be $[(1-p)/p]^M$.

■ EXAMPLE 5.4 Division of Stake in Gambler's Ruin

Two players, A and B, play a sequence of games, each game for a unit stake, until one of them loses all his or her money. Games are independent; in each the probability of A winning is p (and the probability of A losing is $1 - p$, which means that there are no ties). For some reason the contest is interrupted (and is not going to be resumed) when A has x dollars and B has $M - x$ dollars (so that the total stake is M). The question is: How can the stake M be divided between the players in a fair way? One of the ways (arguably just) is to divide the stake in proportion to the chances of eventually winning the contest if it was to continue. Consequently, the problem lies in determining the probabilities of eventually winning the contest for each of the two players. We can interpret the situation in terms of the gambler's problem as follows: Let $i = 0, 1, \ldots, M$ represent the state of the system, defined as the capital of A. Then the changes of states constitute a Markov chain, as in Example 5.3, the only difference being that now $p_{00} = p_{MM} = 1$ (i.e., one cannot leave states 0 and M). The objective is to determine $u(x)$, equal to the probability of A ultimately winning the contest if the starting state is x. We will study this kind of problem in more generality in the next sections. To see the technique, observe that for $0 < x < M$ the probabilities $u(x)$ must satisfy the equations

$$u(x) = pu(x+1) + (1-p)u(x-1), \tag{5.6}$$

with boundary conditions $u(0) = 0, u(M) = 1$. Actually, (5.6) is a special case of the total probability formula (4.9) for the event "ultimate winning by A" and the partition into the two events "next game won by A" and "next game lost by A." This technique of using the partition obtained by considering the possible results of the next transition will be used again and again in the analysis of Markov chains. In the present case a linear combination of solutions of (5.6) is also a solution of (5.6). Consequently, we need two linearly independent solutions to determine the constants in their linear combinations using the two boundary conditions.

Obviously $u(x) = 1$ is always a solution of (5.6). To find another solution, let $u(x) = s^x$. This gives $s = ps^2 + q$, where we let $q = 1 - p$. The solutions of the quadratic equation are $[1 \pm (1 - 4pq)^{1/2}]/2p = [1 \pm |p - q|]/2p$, which equal 1 and q/p. Thus, if $p \neq q$ (hence $p \neq 1/2$), a general solution of (5.6) is $u(x) = A + B(q/p)^x$. Using the boundary conditions, we obtain

$$u(x) = \frac{(q/p)^x - 1}{(q/p)^M - 1}.$$

If $p = q = 1/2$, the solutions $u(x) = (q/p)^x$ and $u(x) = 1$ coincide. In this case as another (linearly independent) solution we can take $u(x) = x$, and the general solution of (5.6) will be $u(x) = A + Bx$. Using again the boundary conditions, we obtain $u(x) = x/M$ as the probability of winning in the case of fair games.

■ **EXAMPLE 5.5**

Let us now return to Example 5.2 of an epidemic. The state of the system is represented by a pair of integers (x, y), with x being the number of infectives and y being the number of susceptibles. As was argued, by taking sufficiently small time intervals between transitions, we can assume that apart from remaining in the same state, the only possible transitions from (x, y) are to $(x - 1, y)$, representing removal (death, isolation, recovery with immunity, etc.) of one infective, and to $(x + 1, y - 1)$, representing infection of one susceptible. These transitions are depicted in Figure 5.1 as a random walk over the integer lattice on the plane, with the two types of transitions being steps in the western and southeastern directions. We now have to define the transition probabilities for the two kinds of transitions above (removal and infection) so as to obtain a meaningful approximation to the real conditions of an epidemic. Here one could argue that it is reasonable to assume that the probability of the first transition (removal) is proportional to x, while the probability of the second transition (infection) is proportional to both x and y. Thus the transition probabilities between states will be of the form $p[(x - 1, y)|(x, y)] = ax, p[(x + 1, y - 1)|(x, y)] = bxy$ for suitably chosen a and b.

The nonlinear term (proportional to the product xy) is largely responsible for the fact that the analysis of an epidemic process is extremely difficult. From the point of view of applications, two characteristics of epidemics are of major interest: first, the time untill the termination of the epidemic, and

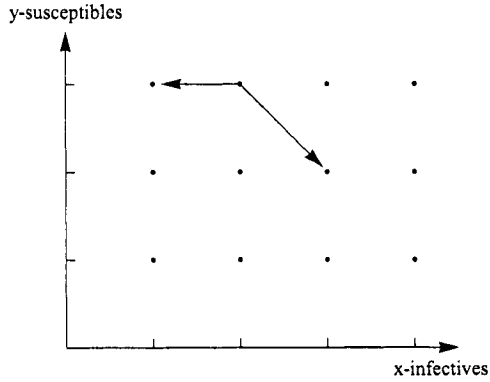


Figure 5.1 Transitions in model of epidemic

second, its total size, that is, the total number of persons who will become infected (and subsequently removed).

The first characteristic may be defined formally as

$$N = \text{first } n \text{ such that } x(n) = 0.$$

Here the event $N = k$ corresponds to $x(0) > 0, x(1) > 0, \ldots, x(k-1) > 0$, and $x(k) = 0$. Graphically, N is the first time when the walk depicted in Figure 5.1 reaches the vertical axis.

The second characteristic cannot be expressed in terms of the variables introduced thus far. The handling of the case is an illustration of the point made repeatedly in Chapter 1 about the nonuniqueness of the choice of the sample space. One could simply enrich the state space by including another count that keeps track of the removals. Thus one could consider the states as being described by three variables, $x(n)$, $y(n)$, and $z(n)$, where x and y are the numbers of infectives and susceptibles as before, while z is the number of removals. The two kinds of transitions considered so far are replaced by transitions from (x, y, z) to $(x-1, y, z+1)$, representing the removal of an infective, and from (x, y, z) to $(x+1, y-1, z)$, representing a new infection. Assuming that the epidemic starts with $z = 0$, the total size Z of epidemic is defined as $Z = z(N)$, where $N =$ the first n with $x(n) = 0$.

As mentioned above, finding the probabilities $P\{N = n\}$ and $P\{Z = k\}$ for given initial conditions of the Markov chain is extremely difficult, and we will not deal with this problem here.

The examples above illustrate the problem of *absorption* that will be analyzed in Section 5.5.

PROBLEMS

5.2.1 **(i)** Modify Example 5.3 by assuming that in each game the player may win with probability p, lose with probability q, or draw (so that his fortune does not

change) with probability r, where $p + q + r = 1$. Find the transition probability matrix in this case. **(ii)** Modify Example 5.4 same way as in (i). Write the analogue to the equation (5.6) for the probability of ruin and solve it.

5.2.2 Suppose the results of an election in a certain city are found to depend only on the results of the last two elections. Specifically, letting R and D denote Republican and Democratic victories, the state before any election may be RR, RD, DR, DD, the letters signifying respectively the outcomes of the next-to-last and last elections. Generally, assume that a, b, c, and d are the probabilities of a Republican victory in each of the four states listed above. Find the transition probability matrix of the resulting Markov chain.

5.2.3 A college professor teaches a certain course year after year. He has three favorite questions, and he always uses one of them in the final exam. He never uses the same question twice in a row. If he uses question A in one year, then the next year he tosses a coin to choose between question B and C. If he uses question B, he tosses a pair of coins, and chooses question A if both coins show tails. Finally, if he uses question C, then he tosses three coins and uses question A if all of them show heads. Find the transition probability matrix for the resulting Markov chain.

5.2.4 (Dog Fleas, or the Ehrenfest Model of Diffusion) Consider two urns (or dogs), and N balls (or fleas), labeled $1, \ldots, N$, allocated between the urns. At times $t = 1, 2, \ldots$, a number 1 through N is chosen at random, and the ball with the selected number is moved to the other urn. Let $s(n)$ be the number of balls in the first urn at time n. Find the transition probability matrix for the Markov chain $s(n)$.

5.2.5 Consider a specific kind of part needed for the operation of a certain machine (e.g., the water pump of a car). When the part breaks down, it is replaced by a new one. The probability that a new part will last for exactly n days is $r_n, n = 1, 2, \ldots$. Let the state of the system be defined as the age of the part currently in the machine. Find the transition probability matrix of the resulting chain.

5.3 N-STEP TRANSITION PROBABILITIES

We will now find the probability

$$p_{ij}^{(n)} = P\{s(t+n) = e_j | s(t) = e_i\} \tag{5.7}$$

of passing from e_i at time t to e_j at time $t+n$, called the n-step transition probability. Obviously, for Markov chains with stationary transition probabilities, the quantity (5.7) does not depend on t. We have the following theorem:

Theorem 5.3.1 *The n-step transition probabilities satisfy the relations*

$$p_{ij}^{(0)} = 1 \text{ or } 0, \quad \textit{depending whether or not } j = i, \tag{5.8}$$

and for all $m, n \geq 0$

$$p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)}. \tag{5.9}$$

Proof. We proceed by induction: Formula (5.8) covers the case $m = n = 0$. In zero steps the system cannot change so it must remain, with probability 1, in state e_i. We prove formula (5.9) for $n = 1$. If the system passes from e_i to e_j in $m + 1$ steps, then in step m it must be in some state e_k. The events "state e_k after m steps" form a partition, and the total probability formula (4.9) gives in this case $p_{ij}^{(m+1)} = \sum_k p_{ik}^{(m)} p_{kj}$, as was to be shown, since $p_{ij} = p_{ij}^{(1)}$ by definition. The extension to the case of arbitrary n is immediate. □

In matrix notation Theorem 5.3.1 states simply that $\mathbf{P}^{(m+n)} = \mathbf{P}^{m+n}$, with $\mathbf{P}^0 = \mathbf{I}$, the identity matrix.

One can expect that as the number of transitions increases, the system "forgets" the initial state; that is, the effect of the initial state on the probability of a state at time n gradually wears out. Formally, we can expect that the probabilities $p_{ij}^{(n)}$ converge, as n increases, to some limits independent of the starting state e_i. In Section 5.4 we explore conditions under which this is true, and we also find the limits of the probabilities $p_{ij}^{(n)}$. At present let us consider the following example.

■ **EXAMPLE 5.6**

The weather on Markov Island is governed by the following laws. There are only three types of days: sunny, rainy, and cloudy, with the weather on a given day depending only on the weather on the preceding day. There are never two rainy days in row, and after a rainy day, a cloudy day is twice as likely as a sunny day. Fifty percent of days following a cloudy day are also cloudy, while 25% are rainy. Finally, after a sunny day each type of weather is equally likely. How often, on average, is it cloudy on Markov Island?

SOLUTION. The fact that the tomorrow's weather depends only on the weather today makes the process of the weather change a Markov process. Using the obvious notation R, S, and C for the three types of weather, we have $p_{RR} = 0$, $p_{RC} = 2/3$, $p_{RS} = 1/3$, $p_{CC} = 1/2$, and $p_{CR} = p_{CS} = 1/4$. Finally, $p_{SS} = p_{SC} = p_{SR} = 1/3$, so the transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3}. \end{bmatrix} \tag{5.10}$$

Let us assume that the limits of the n-step transition probabilities exist and do not depend on the starting state. Thus we have three limits, u_R, u_S, and u_C, where

$$u_R = \lim p_{RR}^{(n)} = \lim p_{CR}^{(n)} = \lim p_{SR}^{(n)},$$

and similarly for u_S and u_C. We can then write, using (5.9) for $n = 1$,

$$\begin{aligned} p_{RR}^{(m+1)} &= p_{RR}^{(m)} p_{RR} + p_{RC}^{(m)} p_{CR} + p_{RS}^{(m)} p_{SR}, \\ p_{CC}^{(m+1)} &= p_{CR}^{(m)} p_{RC} + p_{CC}^{(m)} p_{CC} + p_{CS}^{(m)} p_{SC}, \\ p_{SS}^{(m+1)} &= p_{SR}^{(m)} p_{RS} + p_{SC}^{(m)} p_{CS} + p_{SS}^{(m)} p_{SS}. \end{aligned} \tag{5.11}$$

Passing to the limit in (5.11), we obtain therefore a system of three linear equations:

$$\begin{aligned} u_R &= u_R p_{RR} + u_C p_{CR} + u_S p_{SR} \\ u_C &= u_R p_{RC} + u_C p_{CC} + u_S p_{SC} \\ u_S &= u_R p_{RS} + u_C p_{CS} + u_S p_{SS}. \end{aligned}$$

Using the probabilities from the matrix (5.10), we obtain

$$\begin{aligned} -u_R + \frac{u_C}{4} + \frac{u_S}{3} &= 0 \\ \frac{2u_R}{3} - \frac{u_C}{2} + \frac{u_S}{3} &= 0 \\ \frac{u_R}{3} + \frac{u_C}{4} - \frac{2u_S}{3} &= 0. \end{aligned} \tag{5.12}$$

This is a homogeneous system, so an additional equation is needed to determine the solution. We have here the identity

$$p_{RR}^{(n)} + p_{RC}^{(n)} + p_{RS}^{(n)} = 1, \tag{5.13}$$

and passing to the limit with n, we get

$$u_R + u_C + u_S = 1. \tag{5.14}$$

The solution of (5.12) and (5.14) is $u_R = 9/41, u_C = 20/41$, and $u_S = 12/41$, which means that almost 50% of days on Markov Island are cloudy.

The equations of this example are derived under the assumption that (1) the limits $p_{ij}^{(n)}$ exist and are independent of the starting state i, and (2) the number of states is finite. The latter assumption allowed us to pass to the limit in the sums in (5.11) and (5.13), obtaining (5.12) and (5.14). The first assumption here is true under some conditions. Basically we have to exclude two obvious cases when the probabilities $p_{ij}^{(n)}$ either cannot converge, or if they do, the limits depend on the starting state i.

■ EXAMPLE 5.7

Consider a Markov chain with two states, e_1 and e_2, such that $p_{12} = p_{21} = 1$ (so that $p_{11} = p_{22} = 0$). Clearly, the system must alternate between states 1 and 2 in a deterministic way. Consequently, $p_{11}^{(n)} = 0$ or 1, depending whether n is even or odd, and the sequence $p_{11}^{(n)}$ does not converge.

■ EXAMPLE 5.8

Consider again a Markov chain with two states e_1 and e_2. Assume now that the transition probabilities are $p_{11} = p_{22} = 1$ (so that $p_{12} = p_{21} = 0$). In this case the system must forever remain in the initial state, and we have $p_{11}^{(n)} = 1$ and $p_{21}^{(n)} = 0$ for all n. The limits $p_{ij}^{(n)}$ exist but depend on the initial state i.

These two examples, in essence, exhaust all possibilities that may prevent the existence of the limits $p_{ij}^{(n)}$ and their independence of the initial state i. One may summarize them as *periodicity* (in Example 5.7), and the *impossibility of reaching a certain state from some other state or states* (in Example 5.8). In the next section we introduce definitions pertaining to the classification of states in a Markov chain. These definitions will allow us to formulate the conditions under which the n-step transition probabilities converge to certain limits, regardless of the starting state.

PROBLEMS

5.3.1 Find all two-step transition probabilities for: **(i)** The Markov chain described in Problem 5.2.2. **(ii)** The dog flea model of Problem 5.2.4.

5.3.2 Assume that a man's occupation can be classified as professional, skilled laborer, or unskilled laborer. Assume that of the sons of professional men, a percent are professional, the rest being equally likely to be skilled laborers as unskilled laborers. In the case of sons of skilled laborers, b percent are skilled laborers, the rest being equally likely to be professional men or unskilled laborers. Finally, in the case of unskilled laborers, c percent of the sons are unskilled laborers, the rest again divided evenly between the other two categories. Assume that every man has one son. Form a Markov chain by following a given family through several generations. Set up the matrix of transition probabilities, and find: **(i)** The probability that the grandson of an unskilled laborer is a professional man. **(ii)** The probability that a grandson of a professional man is an unskilled laborer.

5.3.3 In Problem 5.3.2 it was assumed that every man has one son. Assume now that the probability that a man has a son is r. Define a Markov chain with four states, where the first three states are as in Problem 5.3.2, and the fourth state is entered when a man has no son. This state cannot be left (it corresponds to a male line of the family dying out). Find the probability that an unskilled laborer has a grandson who is a professional man.

5.4 THE ERGODIC THEOREM

The main objective of this section is to formulate and sketch the proof of the *ergodic theorem*, which asserts that the nth-step transition probabilities converge in the manner described in the Section 5.3. We begin with some definitions.

Definition 5.4.1 We say that a set C of states is *closed* if for every state e_i in C we have

$$\sum_{e_j \in C} p_{ij} = 1. \qquad \square$$

The condition means that a one-step transition from a state in C leads always to a state in C. By induction, the same must be true for any number of steps. So, if C is closed, then for every state e_i in C and every $n = 1, 2, \ldots$ we have $\sum_{e_j \in C} p_{ij}^{(n)} = 1$ (it is not possible to leave a closed set of states).

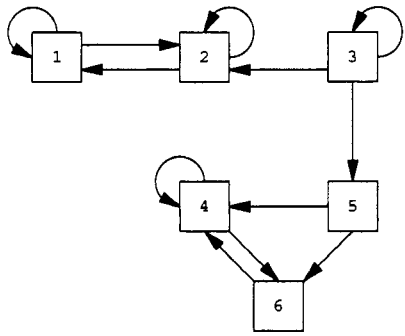


Figure 5.2 Scheme of transitions

■ **EXAMPLE 5.9**

Consider a Markov chain with the transition probability matrix given below (in this matrix "+" stands for a strictly positive probability):

$$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{array} \begin{array}{c} \begin{array}{cccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{array} \\ \begin{bmatrix} + & 0 & + & 0 & 0 & 0 \\ 0 & + & 0 & 0 & + & 0 \\ + & 0 & + & 0 & 0 & 0 \\ 0 & + & 0 & 0 & + & 0 \\ 0 & + & 0 & 0 & 0 & 0 \\ 0 & 0 & + & + & 0 & + \end{bmatrix} \end{array}$$

The scheme of possible one-step transitions is shown in Figure 5.2. Notice that in addition to the whole set of states, which is always closed in view of (5.5), the closed sets are $\{e_1, e_3\}$, $\{e_2, e_5\}$, and $\{e_2, e_4, e_5\}$.

Definition 5.4.2 A Markov chain in which the only closed set of states is the set of all states is called *irreducible*. We then say that state e_j is *accessible* from state e_i if $p_{ij}^{(n)} > 0$ for some n (i.e., one can reach e_j in some number of steps starting from e_i). Moreover, we say that states e_i and e_j *communicate* if each of them is accessible from the other (not necessarily in the same number of steps). Last we say that state e_i is *transient* if there exists a state e_j such that e_j is accessible from e_i but e_i is not accessible from e_j. □

Let us remark here that a more general definition of transient states, suitable for chains with infinitely many states, is: A state is transient, if the probability of never returning to it is strictly positive.

We may now characterize irreducibility as follows:

Theorem 5.4.1 *A chain is irreducible if and only if all states communicate.*

Proof. If two states communicate, and C is a closed class of states, then either both states belong to C or neither of them does. It follows that if all states communicate, then they form the unique closed set, and hence the chain is irreducible.

Conversely, suppose that the states e_i and e_j do not communicate. This means that at least one of them is not accessible from the other. Assume, for instance, that e_j is not accessible from e_i. For every k, if state k is accessible from e_i, then e_j cannot be accessible from e_k. Indeed, if it were not so, we would have $p_{ik}^{(n)} > 0$ for some n, and also $p_{kj}^{(m)} > 0$ for some m. But then $p_{ij}^{(n+m)} \geq p_{ik}^{(n)} p_{kj}^{(m)} > 0$; hence e_j would be accessible from e_i, contrary to the assumption. It follows that the class C of sets accessible from e_i (closed by definition) does not contain state e_j, and the chain is not irreducible. □

We will say that state e_i is *periodic* if there exists $d > 1$ such that whenever $p_{ii}^{(n)} > 0$, then n is divisible by d. This means that a return to state e_i is possible only in a number of steps that is a multiple of d. The smallest d with this property is called the *period* of state e_i. If no such d exists, then state e_i is called *aperiodic*. One can show that in an irreducible chain all states must have the same period, or all be aperiodic. In this case we simply say that the chain is aperiodic.

We can now formulate the following theorem:

Theorem 5.4.2 *Let* $\mathbf{P} = [p_{ij}], i, j = 1, \ldots, M$ *be the transition probability matrix of an aperiodic irreducible Markov chain with a finite number* M *of states. Then the limits*

$$\lim_{n\to\infty} p_{ij}^{(n)} = u_j \tag{5.15}$$

exist for every $j = 1, \ldots, M$, *and are independent of the initial state* e_i. *Moreover,*

$$u_k = \sum_j u_j p_{jk}, \quad k = 1, \ldots, M, \tag{5.16}$$

$$\sum_k u_k = 1. \tag{5.17}$$

Proof. Assume first that the limits (5.15) exist. As in Example 5.6 we have

$$p_{ij}^{(n+1)} = \sum_{k=1}^{M} p_{ik}^{(n)} p_{kj}.$$

Passing to the limit on both sides, we obtain

$$u_j = \lim_{n\to\infty} \sum_{k=1}^{M} p_{ik}^{(n)} p_{kj} = \sum_{k=1}^{M} \lim_{n\to\infty} p_{ik}^{(n)} p_{kj} = \sum_{k=1}^{M} u_k p_{kj},$$

which proves (5.16). Next, since $\sum_j p_{ij}^{(n)} = 1$ for every n, we again pass to the limit on both sides, obtaining equation (5.17). In both cases the interchange of order of summation and limiting passage is allowed in view of the finiteness of the number of terms in the sum. It remains to prove the existence of the limits. While the original proof was purely algebraic (based on an analysis of the powers of a stochastic matrix), we will outline here a proof based on the concept of *coupling*, a surprisingly powerful probabilistic technique for analyzing Markov processes of various kinds. Let $u_k, k = 1, \ldots, M$, be a solution of the system (5.16)–(5.17). If a

Markov chain starts with initial probabilities $r_i = u_i, i = 1, \ldots, M$, then the chain is *stationary* in the sense that $P[s(n) = e_j] = u_j$ for every $j = 1, \ldots, M$ and for every n. Indeed, the statement is true for $n = 0$ by definition. If it holds for some n, then (taking the possible states at the nth step as a partition) we obtain, using (5.16),

$$P\{s(n+1) = e_j\} = \sum_{i=1}^{M} P\{s(n) = e_i\} p_{ij} = \sum_{i=1}^{M} u_i p_{ij} = u_j.$$

Consider now two chains running independently in parallel, both with transition probability matrix $\mathbf{P}$. One chain starts from state i and the other starts from the initial state chosen at random according to the distribution $\{u_j\}$. Let $s(n)$ and $s'(n)$ be the states at the nth transition of the two chains. We let T denote the first time when the two chains are in the same state, that is, $s(T) = s'(T)$. We say that T is the time of *coupling* of the two chains.

The following argument is a good example of making use of conditioning with respect to events from a partition: As the partition we take the class of events $\{T = m\}$ for $0 \leq m \leq n$, together with the event $\{T > n\}$. These events are clearly disjoint and cover all possibilities. We have

$$\begin{aligned} p_{ij}^{(n)} &= P\{s(n) = e_j | s(0) = e_i\} \qquad (5.18) \\ &= \sum_{m=0}^{n} P\{s(n) = e_j | s(0) = e_i, T = m\} \times P\{T = m | s(0) = e_i\} \\ &\quad + P\{s(n) = e_j | s(0) = e_i, T > n\} \times P\{T > n | s(0) = e_i\}. \end{aligned}$$

However,

$$\begin{aligned} &P\{s(n) = e_j | s(0) = e_i, T = m\} \qquad (5.19) \\ &= P\{s(n) = e_j | s(0) = e_i, s'(k) \neq s(k), k \leq m-1, s'(m) = s(m)\} \\ &= P\{s'(n) = e_j | s(0) = e_i, s'(k) \neq s(k), k \leq m-1, s'(m) = s(m)\} \\ &= P\{s'(n) = e_j | s'(m) = s(m)\} = P\{s'(n) = e_j\} = u_j \end{aligned}$$

(by stationarity of the chain s'). Substituting (5.19) into (5.18), we obtain

$$\begin{aligned} p_{ij}^{(n)} &= \sum_{m=0}^{n} u_j P\{T = m | s(0) = e_i\} \\ &\quad + P\{s(n) = e_j | T > n, s(0) = e_i\} \times P\{T > n | s(0) = e_i\} \\ &= u_j P\{T \leq n | s(0) = e_i\} \\ &\quad + P\{s(n) = e_j | T > n, s(0) = e_i\} \times P\{T > n | s(0) = e_i\}. \end{aligned}$$

To show that the sum of the last two terms converges to u_j as asserted, it suffices to show that for every i, the term $P\{T > n | s(0) = e_i\}$ converges to zero as n increases. We will not give a formal proof here, but intuitively it ought to be clear that the probability that T exceeds n, that is, the probability of no coupling of the two processes during the first n transitions, tends to zero. The only two ways in which the coupling could be avoided indefinitely is (1) if the two chains are periodic and "out of phase" or (2) if the chain is reducible, with two disjoint closed sets of states. Both possibilities are excluded by the assumption. □

■ **EXAMPLE 5.10**

Consider a Markov chain with two states, e_1 and e_2. Let the transition matrix be

$$\begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix},$$

where $0 < a < 1, 0 < b < 1$. Then the chain is irreducible (all states communicate) and aperiodic. The probabilities u_1, u_2 satisfy the system of equations

$$u_1 = au_1 + (1-b)u_2, \qquad u_2 = (1-a)u_1 + bu_2, \qquad u_1 + u_2 = 1.$$

The first two equations are the same. The solution is easily found to be

$$u_1 = \frac{1-b}{2-a-b}, \qquad u_2 = \frac{1-a}{2-a-b}. \tag{5.20}$$

If both a and b are close to zero, then the probability of remaining in either of the states is small; hence a typical sequence of states will look like 12121212212121211212121211..., with the states alternating most of the time. The probability of finding the system in state 1 after a large number of steps is close to $1/2$, so that we can expect u_j to be close to $1/2$ for $i = 1, 2$, as follows from (5.20).

Next, if a is close to 0 and b is close to 1, then the system tends to leave state 1 fast, and tends to remain in state 2. Thus a typical sequence of states would be like 12222212222222211222222221.... Consequently, u_1 should be close to 0 and u_2 should be close to 1, which agrees with (5.20).

Finally, if a and b are both close to 1, the system tends to remain in its present state, and consequently a typical sequence of states would be like 1112222111222.... The average durations of this series of identical terms depend on $(1-a)/(1-b)$, which is the ratio of the probabilities of change.

■ **EXAMPLE 5.11 Quality Inspection Scheme**

Consider a machine (production line, etc.) that produces some items. Each of these items may be good or defective, the latter event occurring with probability p and independently of the quality of other items.

The items are inspected according to the scheme (see Taylor and Karlin, 1984) described below, and every item found defective is replaced by a good one. The inspection scheme is described by two positive integers, say A and B. For brevity, we will be using the term (A, B) *scheme*. Inspection in the (A, B) scheme is of two kinds: *full inspection* (where every item is inspected), and *random inspection*, where one item is chosen for inspection at random from a group of B items. The rules are as follows:

1. The process starts with a full inspection.

2. Each full inspection period continues until one encounters a sequence of A good items. Then a period of random inspection starts.

3. Each period of random inspection lasts until one finds a defective item. In this case one starts the next period of full inspection.

Thus the periods of full inspection alternate with the periods of random inspection. The first type of periods are costly but result in full elimination of defective items. The second type of periods are less expensive but leave some defective items.

To apply the theory of Markov chains to the analysis of the inspection scheme described above, let us define state i ($i = 0, 1, \ldots, A-1$) as the state when system is in a period of full inspection and the last item inspected was the ith in a string of consecutive good items. Thus we are in state 0 if the last item was defective, in state 1, if the last item was good, but the preceeding was defective, and so on. Moreover, let A be the state when the system is in a period of random inspection.

While in practical implementation the (A, B) inspection scheme may involve collecting B items and sampling one of them at random for inspection, such a procedure is not very convenient to analyze analytically. We will therefore approximate the random sampling scheme used in real situations by another scheme, which leads to a Markov chain. We will carry out the analysis assuming that while in state A every item is inspected, but the probability of it being defective is p/B, not p. The probability of finding no defective item among B items inspected is $(1 - p/B)^B \approx 1 - p$, which is the same as the probability of finding a good item if the choice is random from a set of B items.

Thus we obtain the transition matrix: for $i = 0, 1, \ldots, A-1$,

$$p_{i,j} = \begin{cases} 1-p & \text{for } j = i+1 \\ p & \text{for } j = 0 \\ 0 & \text{for all other } j, \end{cases}$$

while for $i = A$,

$$p_{A,j} = \begin{cases} 1-p/B & \text{for } j = A \\ p/B & \text{for } j = 0 \\ 0 & \text{for all other } j. \end{cases}$$

Clearly, since all states communicate, the chain is irreducible. Also we have $p_{A,A} > 0$, which means that (see the remarks preceeding Theorem 5.4.2) the chain is aperiodic. Thus the limits u_i exist and satisfy the system of equations

$$\begin{aligned} u_0 &= pu_0 + pu_1 + \cdots + pu_{A-1} + (p/B)u_A, \\ u_i &= (1-p)u_{i-1}, \quad i = 1, 2, \ldots, A-1, \\ u_A &= (1-p)u_{A-1} + (1-p/B)u_A, \\ 1 &= u_0 + u_1 + \cdots + u_A. \end{aligned} \tag{5.21}$$

The second of equations (5.21) gives, by induction, the formula

$$u_i = (1-p)^i u_0, \quad i = 1, \ldots, A-1. \tag{5.22}$$

The third equation of (5.21) gives $u_A = (B/p)(1-p)u_{A-1}$; hence

$$u_A = \frac{B}{p}(1-p)^A u_0. \tag{5.23}$$

Using equation in (5.21), we obtain

$$u_0\left\{1 + (1-p) + (1-p)^2 + \cdots + (1-p)^{A-1} + \frac{B}{p}(1-p)^A\right\} = 1;$$

hence $u_0 = p\{1 + (B-1)(1-p)^A\}^{-1}$, and generally,

$$\begin{aligned} u_i &= \frac{p(1-p)^i}{1+(B-1)(1-p)^A}, \qquad i = 0, 1, \ldots, A-1, \\ u_A &= \frac{B(1-p)^A}{1+(B-1)(1-p)^A}. \end{aligned}$$

Since we inspect all items as long as the system is in states $0, \ldots, A-1$ and only one item out of B is inspected in state A, the average fraction of inspected items is

$$f = u_0 + \cdots + u_{A-1} + \frac{1}{B}u_A = \frac{1}{1+(B-1)(1-p)^A}.$$

This fraction is the major component in the cost (per item produced) of inspection. If $p = 1$, the cost is 1 (all items inspected). If $p = 0$, the cost is $1/B$ because only one in each B items is then inspected.

On the other hand, a fraction of noninspected items is $1 - f$, and out of those, a fraction p is defective. Since all other defective items are detected and replaced by good ones, the average quality of product (measured by fraction of defectives) resulting from (A, B) scheme is $(1-f)p$, which is equal to

$$Q = \frac{(B-1)(1-p)^A p}{1+(B-1)(1-p)^A}.$$

In practice, p may be not known (or may be subject to change). Then finding (A, B) optimal against some specific p (with optimality suitably defined through costs and the final quality) is not practicable. One can, however, determine the maximum of the quality Q (i.e., $Q^* = \max\{Q : 0 \le p \le 1\}$). This value depends on the constants A and B of the inspection plan. Similarly one can find the maximum inspection cost $f^* = \max\{f : 0 \le p \le 1\}$. The knowledge of Q^* and f^*, as functions of (A, B), allows finding the optimal sampling scheme, as an acceptable compromise between cost of inspection and loss due to the resulting quality of the product. The values Q^* and f^* are conservative in the sense that they provide guaranteed quality and guaranteed cost, not to be exceeded, on average, in real situations characterized by an unknown p.

Since u_j is the limiting frequency of visits in state e_j, one can expect that the reciprocal $1/u_j$ should be equal to the average number of steps between

two consecutive visits in state e_j. We will return to this problem in subsequent chapters.

Theorem 5.4.2 covers the case of an irreducible chain. We also have the following theorem which we present without proof:

Theorem 5.4.3 *If the chain has a finite number of states, and there exist m and j such that $p_{ij}^{(m)} > 0$ for all i, then the assertions of Theorem* 5.4.2 *hold: the limits* (5.15) *exist and satisfy the system of equations* (5.16)–(5.17).

Observe that in this case the chain need not be irreducible, in the sense that some states may be transient (if e_j is transient, then $\lim p_{ij}^{(n)} = 0$ for all i).

■ **EXAMPLE 5.12**

Suppose that the transition matrix is

$$\begin{bmatrix} 1 & 0 \\ p & 1-p \end{bmatrix},$$

where $0 < p < 1$. Clearly, state 1 is absorbing and state 2 is transient. The chain has two closed sets of states $\{1\}$ and $\{1, 2\}$. The assumptions of Theorem 5.4.3 (there must be a positive column in some power of the transition matrix) hold for $j = 1$ and $m = 1$. The system of equations is $u_1 = u_1 + pu_2, u_2 = (1 - p)u_2, u_1 + u_2 = 1$, and the solution is $u_1 = 1, u_2 = 0$, as expected. State 1 will eventually be reached.

PROBLEMS

5.4.1 Determine the period of the dog flea model of Problem 5.2.4.

5.4.2 Argue that if the number of states is M, and state e_j is accessible from state e_i, then it is accessible in no more than $M - 1$ steps.

5.4.3 A stochastic matrix $\mathbf{P} = [p_{ij}]$ is called *doubly* stochastic if the sums of its columns are 1. Show that if an irreducible and aperiodic Markov chain with M states has a doubly stochastic transition matrix, then $u_j = 1/M$ for all j.

5.4.4 Show that the probabilities u_j for the dog flea model of diffusion are given by the formula

$$u_j = \binom{N}{j} 2^{-N}, \qquad j = 0, 1, \ldots, N.$$

5.4.5 Another model of diffusion, intended to represent the diffusion of noncompressible substances (e.g., liquids) is as follows. There are N red and N green balls, distributed evenly between urns A and B, so that each urn contains exactly N balls. At each step one ball is selected from each urn at random, and the balls are interchanged. The state of the system is defined as the number of red balls in urn A. Find the transition probability matrix, show that the limiting probability distribution u_j exists, and find it.

5.4.6 Let $s(n)$ be a stationary Markov chain with transition probability matrix $\mathbf{P} = [p_{ij}]$ and $P[s(n) = j] = u_j$. Moreover, let $q_{ij} = P[s(n-1) = j | s(n) = i]$ so that $\mathbf{Q} = [q_{ij}]$ is also a transition probability matrix (why?). The chain with matrix $\mathbf{Q}$ is obtained from the chain with matrix $\mathbf{P}$ by *reversing time*. If $\mathbf{P} = \mathbf{Q}$, we say that the chain is *time reversible*. **(i)** Determine the matrix $\mathbf{Q}$ in terms of $\mathbf{P}$. **(ii)** Check if any of the two diffusion models in Problems 5.2.4 and 5.4.5 is time reversible.

5.5 ABSORPTION PROBABILITIES

Let T be the set of all transient states, and assume that the remaining states may be partitioned into disjoint closed classes $C_1, C_2, \ldots, C_r$ with $r > 1$.

For $e_j \in T$, let $q_j(C_k)$ be the probability of eventual absorption by the class C_k, that is,

$$q_j(C_k) = P\{s(n) \in C_k \text{ for some } n | s(0) = e_j\}.$$

Once the system reaches one of the states in C_k, it will remain there forever, and consequently

$$q_j(C_1) + q_j(C_2) + \cdots + q_j(C_r) \leq 1.$$

Here $1 - [q_j(C_1) + \cdots + q_j(C_r)]$ is the probability of the system remaining forever in transient states (this probability is zero if the number of states is finite; however, in case of an infinite number of states, it may be possible that the system will never leave the class T of transient states). We are interested in determining the probabilities $q_j(C_k)$ for any e_j in T and $k = 1, \ldots, r$. As it turns out, one can find these probabilities for each class separately. To simplify the notation, we can therefore omit the index k and look for generic values $q_j(C)$ for some closed class C. We have the following theorem:

Theorem 5.5.1 *The probabilities $q_j(C)$ satisfy the system of linear equations:*

$$q_i(C) = \sum_{e_j \in T} p_{ij} q_j(C) + \sum_{e_j \in C} p_{ij}, \; e_i \in T. \tag{5.24}$$

Proof. The proof is immediate. This equation states that from a state e_i in T one can pass in one step to another state e_j in T and then the absorption probability becomes $q_j(C)$, or one can pass to C, the probability of this being the second sum in (5.24). All other transitions lead to closed sets different than C, and then transition to C becomes impossible. □

■ **EXAMPLE 5.13**

Let us consider a basic unit of a tennis match, namely a game. During the game the same player always serves, and the winner of the game is the first player to win four points and to be at least two points ahead of the other player. Suppose that the players are A and B, and that A (e.g., the server) has a probability p of winning each point (i.e., the probability of B winning a point is $1 - p$). Moreover, assume that the points are played independently. By tradition, the

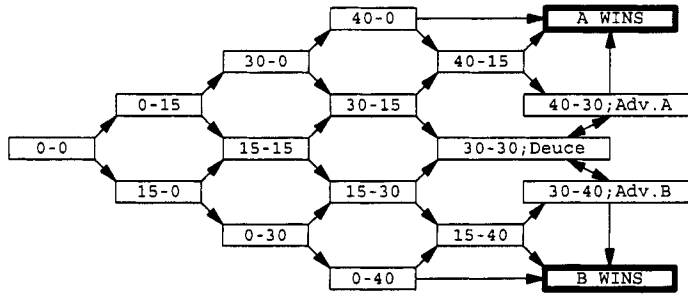


Figure 5.3 Scheme of a game in tennis

first two points are worth 15 each, and the third one is worth 10. So the tennis game may be regarded as a process of transitions over the partial scores, as depicted in Figure 5.3. The score corresponding to each of the players having won two points is $(30 - -30)$, and it is equivalent to the score corresponding to three points won by each $(40 - -40)$, or to any number $k > 3$ of balls won by each (called *deuce*). Equivalence here means that from this score it is necessary to win two points more than the opponent in order to win the game. One can regard the game of tennis as a random walk on the set of states listed in Figure 5.3. The possible transitions and their probabilities are as marked, with the probability of going "up" being p and the probability of going "down" being $1 - p$. There are two closed sets, one consisting of the state labeled "A wins" and the other consisting of the state labeled "B wins." All the remaining states are transient. We want to find the probabilities $q_j(A)$ of player A winning the game if the score is j. In particular, we are interested in the value $q_{0-0}(A)$. For C consisting of the single state A, the system of equations (5.24) takes the following form: For states (scores) from which an immediate victory or loss is not yet possible, we have

$$\begin{aligned} q_{0-0}(A) &= pq_{15-0}(A) + (1-p)q_{0-15}(A), \\ q_{15-0}(A) &= pq_{30-0}(A) + (1-p)q_{15-15}(A), \\ &\vdots \\ q_{30-30}(A) &= pq_{40-30}(A) + (1-p)q_{30-40}(A). \end{aligned} \tag{5.25}$$

Next, for states from which an immediate win by A is possible,

$$\begin{aligned} q_{40-0}(A) &= p + (1-p)q_{40-15}(A), \\ q_{40-15}(A) &= p + (1-p)q_{40-30}(A), \\ q_{40-30}(A) &= p + (1-p)q_{30-30}(A). \end{aligned} \tag{5.26}$$

Finally, for states from which an immediate loss by A is possible, we have

$$\begin{aligned} q_{0-40}(A) &= pq_{15-40}(A), \\ q_{15-40}(A) &= pq_{30-40}(A), \\ q_{30-40}(A) &= pq_{30-30}(A). \end{aligned} \tag{5.27}$$

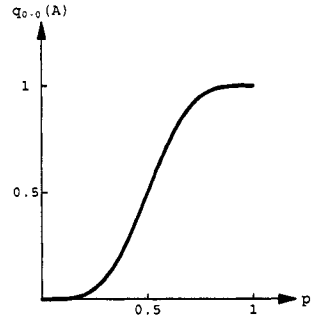


Figure 5.4 Probability $q_{0-0}(A)$ of winning a game by A as a function of p

To solve this system, we have to first solve the set of equations (5.25), (5.26), and (5.27) and then proceed recursively backward, finally to obtain the formula

$$p_{0-0}(A) = (15p^4 - 34p^5 + 28p^6 - 8p^7)/(1 - 2p + 2p^2)$$

(see Bartoszyński and Puri, 1981). The graph of this probability, regarded as a function of probability p of player A winning a single point, is given in Figure 5.4. Several observations here may be of some interest. First, the graph is fairly close to being linear for p between 0.35 and 0.65, with values $q_{0-0}(A)$ close to 0 or 1 outside this range. This means that tennis is really interesting if it is played by similar players. If one of them is much stronger than the other (e.g., the probability p of winning a point is above 70%), then the opponent has very little chance of winning a game, let alone a set or a match. Second, the derivative of the function $q_{0-0}(A)$ at $p = 1/2$ is $5/2$, so the slope of $q_{0-0}(A)$ in the nearly linear (middle) part of the graph is about 2.5. This shows the role of practice training: for evenly matched players, every 1% increase in the probability of winning a point pays off in about a 2.5% increase in probability of winning a game. This is a substantial increase if one considers that a match (between men) consists of at least 18 games.

PROBLEMS

5.5.1 Find the probability of winning a game in tennis directly, using the fact that when the deuce (or 30–30) is attained, the probability of winning is $\sum_{n=0}^{\infty}(2pq)^n p^2$.

5.5.2 Consider the following simple model of evolution: On a small island there is room for 1000 members of a certain species. One year a favorable mutant appears. We assume that in each subsequent generation either the mutants take one place from the regular members of the species with probability 0.6, or the opposite happens. Thus, for example, the mutation disappears in the very first generation with a probability of 0.4. What is the probability that the mutants will eventually take over?

CHAPTER 6

RANDOM VARIABLES: UNIVARIATE CASE

6.1 INTRODUCTION

In Chapter 1 probability theory was presented as techniques to describe, analyze, and predict random phenomena. We then introduced the concept of sample space, identified events with subsets of this space, and developed some techniques of evaluating probabilities of events.

Random variables, each defined as numerical function on some sample space $\mathcal{S}$, will initially be regarded merely as useful tools for describing events. Then, if X stands for a random variable, inequality such as $X \leq t$ determines a set of all outcomes s in $\mathcal{S}$ satisfying the condition $X(s) \leq t$. We will postulate[9] that $\{X \leq t\}$ is an event for each t. This way we gain a powerful tool for describing events in addition to the techniques used thus far, such as specifying subsets of sample space

[9]The reader may wonder why we use the term *postulate* here: the set $\{s \in \mathcal{S} : X(s) \leq t\}$ is a subset of $\mathcal{S}$ and consequently, is an event. However, as explained in Section 2.4, in the case of nondenumerable sample spaces $\mathcal{S}$ we may be unable to define probability P on the class of *all* subsets of $\mathcal{S}$. We must therefore restrict considerations to some class of subsets of $\mathcal{S}$, forming a σ-field, say $\mathcal{A}$, of events. In defining random variables, we postulate that $\{s \in \mathcal{S} : X(s) \leq t\} \in \mathcal{A}$ for every t. In accepted terminology, we say that random variables X are functions on $\mathcal{S}$ that are *measurable* with respect to the σ-field $\mathcal{A}$. From now on we will always tacitly assume that each set of the form $\{X \leq t\}$ is an event, and therefore it is permissible to speak of its probability.

S by listing their elements or by providing a verbal description.

(*) However, in Chapter 1, it was repeatedly stressed that the concept of sample space is, to a large extent, a mathematical construction, and that one can have several sample spaces, all of them equally acceptable for the same phenomenon. The problem then is how to reconcile the inherent lack of uniqueness of the sample space used to describe a given phenomenon with the idea of a random variable being a function defined on the sample space. At first glance it would appear that the concept of a random variable, being based on another concept (sample space $\mathcal{S}$) that allows subjective freedom of choice, must itself be tainted by subjectivity and therefore be of limited use.

Specifically, suppose that there are several sample spaces, say $\mathcal{S}, \mathcal{S}', \ldots$, with the associated probability measures $P, P', \ldots$ that are equally acceptable for describing a given phenomenon of interest, and let $X, X', \ldots$ be random variables defined, respectively, on $\mathcal{S}, \mathcal{S}', \ldots$.

As we will explain, the random variables that are useful for analyzing a given phenomenon must satisfy some "invariance" principle that make them largely independent of the choice of the sample space. The general idea is that a random variable is exactly what the name suggests: *a number depending on chance*. Phrases of this sort were commonly used as "definitions" of the concept of random variable before the probability theory was built on the notion of sample space. What we want to achieve is to define random variables in a way that makes them associated with the phenomenon studied rather than with a specific sample space, hence invariant under the choice of sample space. This is accomplished by introducing the following principle: (*)

Invariance Principle for Random Variables
Suppose that X' and X'' are two random variables defined on two sample spaces $\mathcal{S}'$ and $\mathcal{S}''$ used to describe the same phenomenon. We say that these two random variables are *equivalent* if for every t the event $\{X' \leq t\}$ occurs if and only if the event $\{X'' \leq t\}$ occurs, and moreover if these events have the same probability.

Clearly, the equivalence of random variables defined above satisfies the logical requirements for relation of equivalence: reflexivity, symmetry, and transitivity. We can therefore consider equivalence classes of random variables and speak of representatives of these classes. Random variables are particularly useful in describing and analyzing random phenomena: they do not depend on a particular choice of a sample space; hence they are free of the subjectivity and arbitrariness involved in the selection of $\mathcal{S}$.

6.2 DISTRIBUTIONS OF RANDOM VARIABLES

Although formally each random variable is a numerical function on some sample space $\mathcal{S}$, we usually suppress the dependence on elements of $\mathcal{S}$ in notation, referring to it only when needed to avoid confusion. As symbols for random variables, we will use capital letters from the end of the alphabet, X, Y, Z, possibly with subscripts.

With every random variable we will associate its *probability distribution*, or simply *distribution*.

Definition 6.2.1 By a *distribution* of the random variable X we mean the assignment of probabilities to all events defined in terms of this random variable, that is, events of the form $\{X \in A\}$, where A is a set of real numbers. □

Formally, the event above is a subset of the sample space $\mathcal{S}$. We have

$$\{X \in A\} = \{s \in \mathcal{S} : X(s) \in A\} \subset \mathcal{S}.$$

The basic type of events that we will consider is an interval, that is,

$$\{a < X < b\}, \{a \leq X < b\}, \{a \leq X \leq b\}, \{a < X \leq b\}, \tag{6.1}$$

where $-\infty \leq a \leq b \leq \infty$.

If we can compute probability of each of the events (6.1) for all $a \leq b$, then using the rules from Chapter 2, we can compute probabilities of more complex events. For example,

$$\begin{aligned} P[\{a < X \leq b\}^c] &= P\{X \leq a \text{ or } X > b\} = P\{X \leq a\} + P\{X > b\} \\ &= P\{X \leq a\} + 1 - P\{X \leq b\}. \end{aligned}$$

Similarly, for the probability of the intersection $\{a < X \leq b\} \cap \{c < X \leq d\}$, the answer depends on the mutual relationship between a, b, c, and d (in addition to assumptions $a \leq b$ and $c \leq d$). If, for instance, $a < c < b < d$, then the event in question reduces to $\{c < X \leq b\}$, and so on.

Actually, it turns out that it is sufficient to know the probabilities of only one type of event (6.1) in order to determine the probabilities of remaining types. We will prove one such relation; the proofs of others are similar.

Theorem 6.2.1 *The probabilities of the events of the form* $\{a < X \leq b\}$ *for all* $-\infty \leq a \leq b \leq \infty$ *uniquely determine the probabilities of events of the form* $\{a < X < b\}, \{a \leq X \leq b\}$, *and* $\{a \leq X < b\}$.

Proof. We have

$$\begin{aligned} \{a < X < b\} &= \bigcup_n \left\{a < X \leq b - \frac{1}{n}\right\}, \\ \{a \leq X \leq b\} &= \bigcap_n \left\{a - \frac{1}{n} < X \leq b\right\}, \\ \{a \leq X < b\} &= \bigcap_n \bigcup_m \left\{a - \frac{1}{n} < X \leq b - \frac{1}{m}\right\}. \end{aligned} \tag{6.2}$$

We will prove the last of these identities. Let s be a sample point belonging to the left-hand side so that $a \leq X(s) < b$. Then for every n we have $a - 1/n < X(s)$. Also, if $X(s) < b$, then $X(s) \leq b - 1/m$ for some m (actually for all m sufficiently

large). Thus s belongs to the set $\bigcap_n \bigcup_m \{a - 1/n < X \leq b - 1/m\}$. On the other hand, if s belongs to the right-hand side of the last equality in (6.2), then we have

$$a - \frac{1}{n} < X(s) \leq b - \frac{1}{m}$$

for all n and for some m. Passing to the limit with $n \to \infty$, we obtain $a \leq X(s) \leq b - 1/m < b$, which means that s belongs to the set $\{a \leq X < b\}$. □

Thus each of the sets $\{a < X < b\}, \{a \leq X \leq b\}$ and $\{a \leq X < b\}$ can be represented as union or intersection (or both) of a sequence of events of the form $\{a_n < X \leq b_n\}$.[10]

Let us observe that $\{a < X \leq b\} = \{X \leq b\} \setminus \{X \leq a\}$. Consequently,

$$P\{a < X \leq b\} = P\{X \leq b\} - P\{X \leq a\}$$

and the probabilities of all events of the form $\{a < X \leq b\}$ are determined by probabilities of events of the form $\{X \leq t\}$ for $-\infty < t < \infty$. This justifies the following important definition:

Definition 6.2.2 For any random variable X, the function of real variable t defined as

$$F_X(t) = P\{X \leq t\} \tag{6.3}$$

is called the *cumulative probability distribution function,*[11] or simply cumulative distribution function (cdf) of X. □

The following two examples will illustrate the concept of a cdf:

■ **EXAMPLE 6.1**

The experiment consists of shooting once at a circular target T of radius R. Assume that it is certain that the target will be hit, and that the probability of hitting a particular section A contained in T is given by $|A|/|T|$, where $|\cdot|$ is the area of the section. We will determine the cdf of a random variable X defined as the distance between the point of hitting and the center of the target.

SOLUTION. A natural sample space in this case is just T. Without loss of generality, we can put the center of coordinate system in the center of T so that sample points $s = (x, y)$ satisfy $x^2 + y^2 \leq R^2$. If the target is hit at point $s = (x, y)$, then $X = X(s) = \sqrt{x^2 + y^2}$ is the distance of the point of hit from the origin. Clearly, we have $0 \leq X \leq R$.

[10] In terminology introduced in Chapter 1, we could say that each of the three types of events above belongs to the smallest σ-field generated by the class of all events of the form $\{a < X \leq b\}$. Equivalently, we could say that the intervals on the real line of the kinds (a, b), $[a, b)$, and $[a, b]$ belong to the smallest σ-field of sets of real numbers that contains all intervals of the form $(a, b]$.

[11] According to the footnote at the beginning of this chapter, the right-hand side of (6.3) is well defined for each t (i.e., $\{X \leq t\}$ is an event).

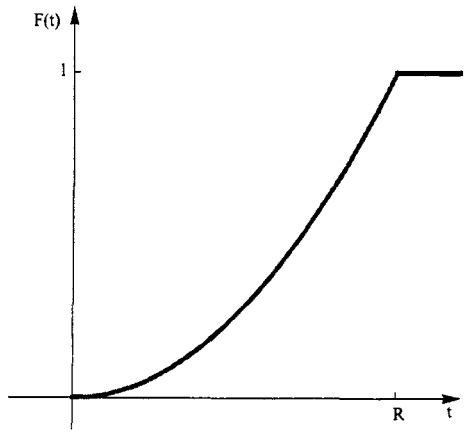


Figure 6.1 Cdf of the distance from the center of the target

Now let $F_X(t) = P\{X \leq t\}$. Obviously, if $t > R$, then $F_X(t) = 1$, and if $t < 0$, then $F_X(t) = 0$. For t with $0 \leq t \leq R$, we have

$$\begin{aligned} F_X(t) &= P\{X \leq t\} \\ &= P\{\text{point } s \text{ falls in the circle of radius } t \text{ centered at the origin}\} \\ &= \frac{\pi t^2}{\pi R^2} = \left(\frac{t}{R}\right)^2. \end{aligned}$$

Thus

$$F_X(t) = \begin{cases} 0 & t < 0 \\ \left(\frac{t}{R}\right)^2 & 0 \leq t \leq R \\ 1 & t > R. \end{cases} \tag{6.4}$$

The graph of $F_X(t)$ is given in Figure 6.1.

Remark A simple fact, useful in determining cdf's of random variables, is that if the random variable X is bounded from above, then $F_X(t) = 1$ for all $t > M$, where M is the upper bound for X. Indeed, if $t > M$, then

$$\begin{aligned} F_X(t) &= P\{X \leq t\} \\ &= P\{X \leq M\} + P\{M < X \leq t\} = P\{X \leq M\} = 1. \end{aligned}$$

Similarly, if X is bounded from below, then $F_X(t) = 0$ for $t < m$, where m is the lower bound for X.

■ **EXAMPLE 6.2**

Let the experiment consist of three tosses of a coin, and let X be total number of heads obtained. The sample space consists of the eight sample points listed below, together with their associated values of X:

S	X	S	X
HHH	3	TTH	1
HHT	2	THT	1
HTH	2	HTT	1
THH	2	TTT	0

The random variable X satisfies the condition $0 \leq X \leq 3$ so that $F_X(t) = 0$ if $t < 0$ and $F_X(t) = 1$ if $t > 3$. Moreover, since X can take on only values $0, 1, 2$, and 3, we have $P\{X \in A\} = 0$ for every set A that does not contain any of the possible values of X. Finally, simple counting gives $P\{X = 0\} = P\{X = 3\} = 1/8$ and $P\{X = 1\} = P\{X = 2\} = 3/8$. Thus for $0 \leq t < 1$ we write

$$\begin{aligned} F_X(t) &= P\{X \leq t\} = P\{X < 0\} + P\{X = 0\} + P\{0 < X \leq t\} \\ &= P\{X = 0\} = \frac{1}{8}. \end{aligned}$$

Similarly, if $1 \leq t < 2$, then

$$\begin{aligned} F_X(t) &= P\{X \leq t\} \\ &= P\{X < 0\} + P\{X = 0\} + P\{0 < X < 1\} \\ &\quad + P\{X = 1\} + P\{1 < X \leq t\} \\ &= P\{X = 0\} + P\{X = 1\} = \frac{1}{2}. \end{aligned}$$

Proceeding in a similar way with $2 \leq t < 3$, we get

$$F_X(t) = \begin{cases} 0 & \text{for } t < 0 \\ \frac{1}{8} & \text{for } 0 \leq t < 1 \\ \frac{1}{2} & \text{for } 1 \leq t < 2 \\ \frac{7}{8} & \text{for } 2 \leq t < 3 \\ 1 & \text{for } 3 \leq t. \end{cases}$$

The graph of $F_X(t)$ is given in Figure 6.2.

We will next investigate general properties of cdf's in some detail.

Theorem 6.2.2 *Every cdf F has the following properties:*
(a) *F is nondecreasing.*
(b) $\lim_{t \to -\infty} F(t) = 0, \quad \lim_{t \to \infty} F(t) = 1.$
(c) *F is continuous on the right.*

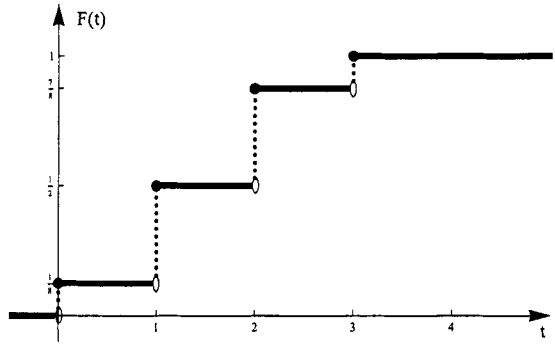


Figure 6.2 Cdf of the number of heads in 3 tosses of a coin

Proof. Let F be the cdf of random variable X. To prove (a), let $t_1 < t_2$. Then

$$F(t_2) - F(t_1) = P\{X \leq t_2\} - P\{X \leq t_1\} = P\{t_1 < X \leq t_2\} \geq 0.$$

Consequently, $F(t_1) \leq F(t_2)$, as was to be shown.

To prove the second relation in (b), we have to show that $\lim_{n\to\infty} F(t_n) = 1$ for any sequence $\{t_n\}$ such that $t_n \to \infty$. Without a loss of generality, we can assume that $t_1 < t_2 < \cdots \to \infty$. The events $\{X \leq t_1\}, \{t_1 < X \leq t_2\}, \{t_2 < X \leq t_3\}, \ldots$ are disjoint, and their union is the whole sample space $\mathcal{S}$. By the second and third axioms of probability, we have

$$\begin{aligned} 1 &= P(\mathcal{S}) = P\{X \leq t_1\} + \sum_{k=2}^{\infty} P\{t_{k-1} < X \leq t_k\} \\ &= F(t_1) + \sum_{k=2}^{\infty}[F(t_k) - F(t_{k-1})] = F(t_1) + \lim_{n\to\infty} \sum_{k=2}^{n}[F(t_k) - F(t_{k-1})] \\ &= \lim_{n\to\infty} \{F(t_1) + [F(t_2) - F(t_1)] + \cdots + [F(t_n) - F(t_{n-1})]\} \\ &= \lim_{n\to\infty} F(t_n), \end{aligned}$$

as was to be shown. The proof that $\lim_{t\to-\infty} F(t) = 0$ is analogous.

To prove (c), let $\{t_n\}$ be a sequence such that $t_1 > t_2 > \cdots \to t^*$. We have to show that $\lim_{n\to\infty} F(t_n) = F(t^*)$.

The events $\{X \leq t_n\}$ form a monotonically decreasing sequence, with $\bigcap_{n=1}^{\infty}\{X \leq t_n\} = \{X \leq t^*\}$. So, by Theorem 2.6.1, we have

$$\begin{aligned} F(t^*) &= P\{X \leq t^*\} = P\left[\bigcap_{n=1}^{\infty}\{X \leq t_n\}\right] \\ &= P[\lim_{n\to\infty}\{X \leq t_n\}] = \lim_{n\to\infty} P\{X \leq t_n\} = \lim_{n\to\infty} F(t_n). \end{aligned}$$ □

One of the consequences of the foregoing properties of cdf's is the following fact: for every x

$$P\{X = x\} = F(x) - F(x - 0), \tag{6.5}$$

where $F(x-0) = \lim_{h \downarrow 0} F(x-h)$ is the left-hand-side limit of F at x. Indeed, we have

$$\{X = x\} = \bigcap_{n=1}^{\infty} \{x - h_n < X \leq x\} \tag{6.6}$$

for every decreasing sequence of positive numbers h_n with $h_n \to 0$. Since the sets in the product (6.6) are decreasing, by the continuity property of probability,

$$\begin{aligned} P\{X = x\} &= \lim_{n \to \infty} P\{x - h_n < X \leq x\} = \lim_{n \to \infty} [F(x) - F(x - h_n)] \\ &= F(x) - \lim_{n \to \infty} F(x - h_n) = F(x) - F(x - 0). \end{aligned}$$

Sometimes it is necessary to identify values of a random variable corresponding to a given value of cdf. These values are called quantiles.

Definition 6.2.3 Let X be a random variable with cdf F, and let $0 < p < 1$. The *pth quantile* ξ_p of X is defined as any solution of the simultaneous inequalities

$$P\{X \leq x\} \geq p, \qquad P\{X \geq x\} \geq 1 - p. \tag{6.7}$$

The inequalities (6.7) are equivalent to

$$F(x) \geq p, \qquad F(x-0) \leq p. \qquad \square$$

To illustrate this definition, we assume now that X is a random variable with continuous cdf $F(x)$. For any p with $0 < p < 1$ there exists[12] a point ξ_p satisfying the relation $F(\xi_p) = p$. This means that $P\{X \leq \xi\} = p$. If F is continuous, then $P\{X \geq \xi_p\} = P\{X > \xi_p\} = 1 - F(\xi_p) = 1 - p$. So ξ_p satisfies (6.7). The point ξ_p need not be unique though; F can satisfy the relation $F(x) = p$ for an interval of values x, and each of them can serve as ξ_p. For the case of random variables with continuous cdf, the condition (6.7) would then reduce to

$$P\{X \leq \xi_p\} = p, \qquad P\{X \geq \xi_p\} = 1 - p, \tag{6.8}$$

with each of the relations (6.8) implying the other.

■ **EXAMPLE 6.3**

For the random variable X in Example 6.1, the quantiles can be determined from the relation $(\xi_p/R)^2 = p$; hence $\xi_p = R\sqrt{p}$. For instance, if $p = 0.25$, then $\xi_{0.25} = R/2$: chances of hitting the target at a distance less than $1/2$ of the radius are 0.25.

The reason for using inequalities in condition (6.7) instead of the simpler condition (6.8) lies in the fact that the equation $F(x) = p$ may have no solution, as illustrated by the following example:

[12]The existence of such a point follows from the continuity (hence also the Darboux property) of function F; it must assume every value between its lower bound 0 and upper bound 1.

■ **EXAMPLE 6.4**

Let X assume values 1, 2, and 3 with probabilities 0.2, 0.6, and 0.2, respectively. Then $F(x) = 0.8$ for $2 \leq x < 3$, and $F(x) = 1$ for $x \geq 3$. Thus the equation $F(x) = p$ is solvable only for $p = 0.2$ and $p = 0.8$. To find, for example, a $\xi_{0.5}$, we need to use (6.7), looking for points such that $P\{X \leq x\} \geq 0.5$ and $P\{X \geq x\} \geq 0.5$. The first inequality gives $x \geq 2$, while the second gives $1 - F(x-0) \geq 0.5$. Hence $x \leq 2$, and both inequalities are satisfied only for $x = 2$. So we have $\xi_{0.5} = 2$.

Certain quantiles have special names. For example, $\xi_{0.5}$ is called the *median*, $\xi_{0.25}$ and $\xi_{0.75}$ are called *lower* and *upper* quartiles, respectively. Quantiles may also be called percentiles, pth quantile being the same as $100p$th percentile.

We have the following theorem:

Theorem 6.2.3 *If X is a random variable with continuous cdf and $0 < \alpha < \beta < 1$, then*

$$P\{\xi_\alpha \leq X \leq \xi_\beta\} = \beta - \alpha, \tag{6.9}$$

where ξ_α and ξ_β are any quantiles of order α and β, respectively.

Proof. The left-hand side of (6.9) equals $F(\xi_\beta) - F(\xi_\alpha - 0)$, which equals $F(\xi_\beta) - F(\xi_\alpha)$ by continuity of F, and hence equals $\beta - \alpha$ by the definition of quantiles. □

■ **EXAMPLE 6.5**

For example, if F is continuous, there is always 50% probability that a random variable with cdf F will assume a value between its upper and lower quartile.

Some comments about the concept of cdf are now in order. First, we defined the cdf of a random variable X using nonstrict inequality, that is, $F(t) = P\{X \leq t\}$. It is equally permissible to define cdf by the formula $F^*(t) = P\{X < t\}$. The only difference between F and F^* is that the latter is left continuous. Such a definition of the cdf actually appears throughout in Russian and Eastern European statistical literature.

The second comment concerns the sufficiency of conditions (a)–(c) in Theorem 6.2.2. To put it differently, if a function F satisfies conditions (a)–(c), does there exist a random variable X such that F is the cdf of X, that is, $F(t) = P\{X \leq t\}$?

The following theorem provides the answer. We give it here without proof. Interested readers can see advanced texts on probability, for example, Chow and Teicher (1997).

Theorem 6.2.4 *Any function F that satisfies conditions* (a)–(c) *of Theorem* 6.2.2 *is a cdf of some random variable.*

The importance of this theorem (as in the case of Theorem 2.6.3) is that it guarantees that some phrases, such as "let F be a cdf," make sense, with no additional assumptions about F besides (a)–(c) of Theorem 6.2.2.

The next question, quite naturally, is: Do cdf's determine uniquely the random variables associated with them? The answer here is negative: there are many random variables (associated with different phenomena, or even associated with the same phenomenon) that have the same cdf. Thus it is a random variable that determines its cdf, and not conversely. To see why this is so, let us consider the following example:

■ **EXAMPLE 6.6**

Continuing Example 6.2, let us consider the experiment consisting of three tosses of a coin, and two random variables: X = total number of heads and Y = total number of tails. We have here, as before,

S	X	Y	S	X	Y
HHH	3	0	TTH	1	2
HHT	2	1	THT	1	2
HTH	2	1	HTT	1	2
THH	2	1	TTT	0	3

A simple count shows that $P\{X = k\} = P\{Y = k\}$ for all k, which implies that the cdf's $F_X(t)$ and $F_Y(t)$ are identical.

This example illustrates the fact that there exist two distinct variables X and Y defined on *the same* sample space $\mathcal{S}$, and that these two variables have the same distribution function. Obviously, if X and Y are defined as above, but refer to two distinct triplets of tosses, their distributions will still be the same.

On the cumulative distribution function of a random variable X, recall from Definition 6.2.1 that the distribution of random variable X is defined as an assignment of probabilities $P\{X \in A\}$, where A is a set of real numbers. The cumulative distribution function F_X then provides probabilities of intervals open on the left and closed on the right:

$$P\{X \in (a, b]\} = P\{a < X \leq b\} = F_X(b) - F_X(a). \tag{6.10}$$

The question now is: Does F_X also determine the probabilities $P\{X \in A\}$ for sets A other then intervals $(a, b]$, and if so, what is the class of these sets? Before answering, let us observe that the leftmost member of (6.10) can be written as $P\{s \in \mathcal{S} : X(s) \in (a, b]\}$, so that the symbol P refers to the subsets of sample space $\mathcal{S}$, as it should according to the definition of probability. On the other hand, the rightmost member depends only on a and b, and assigns a number to an interval $(a, b]$. We may therefore say that we are dealing with probability on sets of real numbers (at least on intervals). According to the footnote at the beginning of this chapter, the left-hand side is well defined: We know that events (sets on which P is defined) form a σ-field, and we also know that sets $\{X \leq t\}$ are events so that

$$\{a < X \leq b\} = \{X \leq b\} \setminus \{X \leq a\}$$

is also an event.

The proof of the following extension theorem can be skipped in the first reading.

Theorem 6.2.5 *If F is a cumulative distribution function, then the function m_F defined on the class $\mathcal{B}_0$ of all intervals $(a, b]$, $-\infty \leq a \leq b \leq +\infty$, by the formula*

$$m_F(a, b] = F(b) - F(a) \tag{6.11}$$

can be extended uniquely to a probability measure m_F on the smallest σ-field $\mathcal{B}$ containing $\mathcal{B}_0$.

(*) *Proof.* According to Theorem 2.6.3 any σ-additive function on a field $\mathcal{G}$ has a unique extention to a σ-additive function on the smallest σ-field containing $\mathcal{G}$. Therefore, to prove the theorem, it suffices to show that (1) m_F can be extended uniquely to the smallest *field* containing all intervals $(a, b]$, (i.e., to the smallest field, say $\mathcal{B}_1$, containing $\mathcal{B}_0$), and that (2) the function m_F extended to $\mathcal{B}_1$ is σ-additive on $\mathcal{B}_1$.

To prove (1), observe that the complement of $(a, b]$ is the union of $(-\infty, a]$ and $(b, \infty]$, while the intersection $(a, b] \cap (c, d]$ is empty if $b \leq c$ and otherwise equals $(\max(a, c), \min(b, d)]$. Consequently, each set in $\mathcal{B}_1$ can be represented as a finite union of disjoint intervals $(a, b]$ from $\mathcal{B}_0$. Using generally the symbol I for sets in $\mathcal{B}_0$, we can write each set $A \in \mathcal{B}_1$ as

$$A = \bigcup_{j=1}^{m} I_j, \tag{6.12}$$

where $I_j \cap I_k = \emptyset$, if $j \neq k$. We then define

$$m_F(A) = \sum_{j=1}^{m} m_F(I_j), \tag{6.13}$$

with $m_F(I_j)$ as defined in (6.11). We only need to show that definition (6.13) is unambiguous; that is, it does not depend on the choice of representation (6.12). Indeed, suppose that

$$A = \bigcup_{j=1}^{m} I_j = \bigcup_{k=1}^{n} I_k^*, \tag{6.14}$$

where $I_j \cap I_r = I_k^* \cap I_s^* = \emptyset$ if $j \neq r, k \neq s$. We have to show that

$$\sum_{j=1}^{m} m_F(I_j) = \sum_{k=1}^{n} m_F(I_k^*). \tag{6.15}$$

We may write, using (6.14)

$$I_j = I_j \cap A = I_j \cap \bigcup_{k=1}^{n} I_k^* = \bigcup_{k=1}^{n} (I_j \cap I_k^*)$$

and similarly

$$I_k^* = A \cap I_k^* = \left(\bigcup_{j=1}^{m} I_j \right) \cap I_k^* = \bigcup_{j=1}^{m} (I_j \cap I_k^*).$$

Consequently,

$$\begin{aligned}\sum_{j=1}^{m} m_F(I_j) &= \sum_{j=1}^{m} m_F\left(\bigcup_{k=1}^{n}(I_j \cap I_k^*)\right) \\ &= \sum_{j=1}^{m}\sum_{k=1}^{n} m_F(I_j \cap I_k^*) = \sum_{k=1}^{n}\sum_{j=1}^{m} m_F(I_j \cap I_k^*) \\ &= \sum_{k=1}^{n} m_F\left(\bigcup_{j=1}^{m}(I_j \cap I_k^*)\right) = \sum_{k=1}^{n} m_F(I_k^*), \qquad (6.16)\end{aligned}$$

which shows (6.15).

It remains now to show that the function m_F is σ-additive on $\mathcal{B}_1$. In fact, argument in (6.15) remains valid for infinite unions, provided that we show m_F to be σ-additive on $\mathcal{B}_0$, that is, that for all $a \leq b$ the condition

$$(a, b] = \bigcup_{k=1}^{\infty} I_k,\ I_k \cap I_j = \emptyset \text{ for } k \neq j$$

implies

$$F(b) - F(a) = \sum_{k=1}^{\infty} m_F(I_k).$$

For every fixed n we have

$$\bigcup_{k=1}^{n} I_k \subset (a, b].$$

Rearranging, if necessary, the intervals $I_k = (a_k, b_k]$, we can assume that for every n,

$$a \leq a_1 \leq b_1 \leq a_2 \leq b_2 \leq \cdots \leq b_{n-1} \leq a_n \leq b_n \leq b. \qquad (6.17)$$

Consequently,

$$\begin{aligned} m_F\left(\bigcup_{k=1}^{n} I_k\right) &= \sum_{k=1}^{n} m_F(I_k) = \sum_{k=1}^{n}[F(b_k) - F(a_k)] \\ &\leq \sum_{k=1}^{n}[F(b_k) - F(a_k)] + \sum_{k=1}^{n-1}[F(a_{k+1}) - F(b_k)] \\ &= F(b_n) - F(a_1) \leq F(b) - F(a) = m_F(a, b]. \end{aligned}$$

Letting $n \to \infty$, we write

$$\sum_{k=1}^{\infty} m_F(I_k) \leq m_F(a, b].$$

It remains to prove the reverse inequality. At this point it may be appropriate to make the following comment, without which the proof below might appear incomprehensible. In fact, one could say that the proof is "obvious." Since $(a, b]$ is partitioned into a countable number of disjoint intervals, their measures m_F can simply

be added by canceling negative and positive terms in expressions for contiguous intervals, according to the formula, valid for $a < b < c$,

$$[F(b) - F(a)] + [F(c) - F(b)] = F(c) - F(a).$$

The trouble is that *sometimes intervals I_k cannot be arranged into a sequence of contiguous intervals.*

To visualize such a possibility, consider the partition of interval $(-1, 1]$ by points

$$-1, -\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4}, \ldots, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1$$

so that

$$(-1, 1] = \bigcup_{n=1}^{\infty} \left(\frac{-1}{n}, \frac{-1}{n+1} \right] \cup \bigcup_{n=1}^{\infty} \left(\frac{1}{n+1}, \frac{1}{n} \right].$$

There is no rightmost term in the first union, and no leftmost term in the second union. So no cancelation occurs in passing from one sum to the other. The situation can be much more complicated because there could be infinitely many such accumulation points as 0 in the example above.

To continue with this proof, let us exclude the trivial case $a = b$ and choose ϵ such that $0 < \epsilon < b - a$. Let the sets I_k in (6.16) be $I_k = (a_k, b_k]$ (observe that no monotonicity of sequences $\{a_k\}$ and $\{b_k\}$ is assumed).

Since function F is continuous on the right, for every n there exists $\beta_n > 0$ such that

$$F(b_n + \beta_n) - F(b_n) < \frac{\epsilon}{2^n}. \tag{6.18}$$

Let $I_n^{\epsilon} = (a_n, b_n + \beta_n)$. We have then

$$[a + \epsilon, b] \subset \bigcup_{n=1}^{\infty} I_n^{\epsilon}.$$

By the Heine-Borel lemma,[13] there exists a finite N such that

$$[a + \epsilon, b] \subset \bigcup_{n=1}^{N} I_n^{\epsilon}. \tag{6.19}$$

Let $n_1 \leq N$ be such that $b \in I_{n_1}^{\epsilon}$. If $a + \epsilon < a_{n_1}$, choose $n_2 \leq N$ such that $a_{n_1} \in I_{n_2}^{\epsilon}$. We continue in this way until k such that $a_{n_k} \leq a + \epsilon$. Such a k must exist in view of (6.19). Renumbering the chosen intervals, if necessary, we have $[a + \epsilon, b] \subset \bigcup_{n=1}^{k} I_n^{\epsilon}$ and

$$\begin{aligned} &a_1 < b < b_1 + \beta_1, \\ &a_{i+1} < a_i < b_{i+1} + \beta_{i+1}, \qquad i = 1, 2, \ldots, k-1, \\ &a_k \leq a + \epsilon < b_k + \beta_k. \end{aligned}$$

[13] Heine-Borel lemma asserts that from any covering of a compact set by open sets one can choose a finite covering. In the present case the closed interval $[a + \epsilon, b]$ is compact, and covering is by open sets $(a_n, b_n + \beta_n)$.

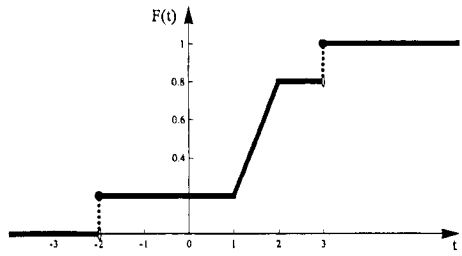


Figure 6.3 Cdf of random variable X

Consequently, using (6.18), we have

$$m_F[a+\epsilon, b] \leq m_F(a_k, b_1+\beta_1] = m_F(a_1, b_1+\beta_1] + \sum_{j=2}^{k} m_F(a_j, a_{j-1}]$$

$$\leq \sum_{j=1}^{k} m_F(a_j, b_j+\beta_j] \leq \sum_{j=1}^{\infty} m_F(a_j, b_j] + \epsilon.$$

Letting $\epsilon \to 0$, we obtain

$$m_F(a, b] \leq \sum_{j=1}^{\infty} m_F(a_j, b_j] = \sum_{j=1}^{\infty} m_F(I_j),$$

which completes the proof. **(*)**

PROBLEMS

6.2.1 In the statements below, F and G stand for cdf's of random variables X and Y, respectively. Classify each of the statements below as true or false:
(i) If X is always strictly positive, then $F(t)$ is strictly positive for all t.
(ii) If $F(37) = F(45)$, then $P(40 < X < 42) = P(43 < X < 44)$.
(iii) If $Y = X + 3$, then $G(t) \leq F(t)$ for all t.
(iv) If $G(17) - F(17) = 1$, then both X and Y are always less than 17.
(v) If $G(17) - F(17) = 1$, then $P(Y > X) = 0$.
(vi) If $G(17) \times F(17) = 1$, then $P(\max(X, Y) \leq 17) = 1$.
(vii) If $G(17) \times F(17) = 0$, then $P(\min(X, Y) \leq 17) = 1$.
(viii) If $|F(t) - G(t)| < \epsilon$ for all t, then $|X - Y| < \epsilon$.
(ix) If $P(X \leq Y) = 1$, then $F(t) \geq G(t)$ for all t.

6.2.2 Figure 6.3 shows the cdf of a random variable X. Find: **(i)** $P(X = -2)$, $P(X = 0)$. **(ii)** $P(X \leq 3)$, $P(X < 3)$, $P(X < 0.13)$. **(iii)** $P(X > 2)$, $P(X > 2.79)$. **(iv)** $P(-1 < X \leq 0.7)$, $P(-2 \leq X < 1)$. **(v)** $P(1 \leq |X| \leq 2)$.

6.2.3 Let X be a random variable with cdf given by $F_X(x) = 0$ for $x < 0$ and $F_X(x) = 1 - 0.3e^{-\lambda x}$ for $x \geq 0$. Determine: **(i)** $P(X = 0)$. **(ii)** λ if $P(X \leq 3) = 3/4$. **(iii)** $P(|X| \leq 5)$ using results of (ii).

6.2.4 Determine the medians and lower and upper quartiles for random variables with the following cdf's:

(i) $F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ kx^3 & \text{for } 0 \le x \le 1/\sqrt[3]{k}, \quad k > 0 \\ 1 & \text{for } x > 1/\sqrt[3]{k}, \end{cases}$

(ii) $F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \alpha e^{-x} & \text{for } x \ge 0. \end{cases}$ (consider all possible α)

6.2.5 A point is chosen at random from a square with side a. Let X be the distance from the selected point to the nearest corner of the square. Find and graph $F_X(x)$.

6.2.6 A coin of diameter d is dropped on a floor covered with square tiles with side length $D > d$. Let X be the number of tiles which intersect with the coin. **(i)** Find the distribution of X. **(ii)** Determine the median of X as a function of D and d.

6.2.7 Prove the first part of assertion (b) of Theorem 6.2.2.

6.3 DISCRETE AND CONTINUOUS RANDOM VARIABLES

Although the cdf of a random variable X provides all the information necessary to determine probabilities $P\{X \in A\}$ for a large class of sets A, there exist wide and practically important classes of random variables whose distributions may be described in simpler ways. Two such classes will be discussed in this section. Accordingly we introduce the following definition:

Definition 6.3.1 A random variable X will be called *discrete*, if there exists a finite or countably infinite set of real numbers $U = \{x_1, x_2, \ldots\}$ such that

$$P\{X \in U\} = \sum_n P\{X = x_n\} = 1. \tag{6.20}$$

■ **EXAMPLE 6.7**

Let $U = \{1, 2, \ldots, n\}$ for some n, and let $P\{X = j\} = 1/n$ for $j \in U$. Condition (6.20) clearly holds, so the values of X are restricted to U. This example describes the selection (assumed fair) of the number of the winning lottery ticket, where n is the total number of tickets.

A discrete random variable with a finite set U of values, all of them having the same probability of occurrence, is called (discrete) *uniform*. We will later analyze this distribution in more detail.

■ **EXAMPLE 6.8 Binomial Distribution**

The binomial distribution plays a central role in probability theory and statistical modeling, and it will often be used throughout this book. Here we just introduce a definition and basic formulas.

Definition 6.3.2 The binomial random variable is defined as a total number of successes in n independent experiments, each experiment leading to success with probability p. □

We have encountered special cases of binomial random variable in the preceding chapters (e.g., in analyzing the distribution of the "total number of heads in 3 tosses of a coin"). The set of possible values of X is $\{0, 1, \ldots, n\}$, since the number of successes is an integer, and at best equals n (the number of trials) and at worst equals 0 (if all repetitions lead to failure). The probability of k successes and $n-k$ failures in any specific order $SFFS \ldots S$ equals $p(1-p)(1-p)p \ldots p = p^k(1-p)^{n-k}$. Since the probability of this string does not depend on its order, we obtain $P\{S_n = k\}$ by taking $p^k(1-p)^{n-k}$ as many times as there are different orders of k letters S and $n-k$ letters F. This number is $\binom{n}{k}$, since each such order is completely specified by selecting the set of locations for letter S among n slots. Thus

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n, \tag{6.21}$$

and the Newton binomial formula (3.16) shows that, as expected,

$$\sum_{k=0}^{n} P\{X = k\} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1.$$

In the sequel, we shall use the symbol BIN(n, p) to denote binomial distribution with parameters n and p. This way we say that X has distribution BIN(n, p) or simply $X \sim \text{BIN}(n, p)$. We also let the individual probabilities in distribution BIN(n, p) be denoted by $b(k; n, p)$ so that

$$b(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n. \tag{6.22}$$

In the examples above, the set U was finite, with $U = \{1, \ldots, n\}$ in Example 6.7 and $U = \{0, 1, \ldots, n\}$ in Example 6.8. In the example below, the set U is infinite.

■ **EXAMPLE 6.9**

Consider a sequence of independent tosses of a die. Let X be the number of tosses until the first ace. Clearly, X can assume values $1, 2, 3, \ldots$, and the event $\{X = k\}$ occurs if the kth toss gives an ace (chances are $1/6$) and the first $k-1$ tosses are all different from an ace [chances are $(5/6)^{k-1}$]. Thus

$$P\{X = k\} = \frac{1}{6}\left(\frac{5}{6}\right)^{k-1}, \qquad k = 1, 2, \ldots \tag{6.23}$$

We can easily check that

$$\sum_{k=1}^{\infty} P\{X = k\} = \frac{1}{6}\sum_{k=1}^{\infty}\left(\frac{5}{6}\right)^{k-1} = \frac{1}{6} \times \frac{1}{1 - \frac{5}{6}} = 1,$$

so condition (6.20) is satisfied.

This random variable is generally described as "waiting time for first success" (in this case success being an ace). Distribution (6.23) is an example of *geometric* distribution, denoted GEO(p), which will be discussed in detail in Chapter 9.

■ **EXAMPLE 6.10 Poisson Distribution**

We say that the random variable X has *Poisson* distribution with parameter $\lambda > 0$, denoted POI(λ), if the possible values of X are nonnegative integers $0, 1, 2, \ldots$ and

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots. \tag{6.24}$$

The Poisson distribution is often applied in practice, so we will study its properties in some detail in following sections. At present, observe that (6.24) is correctly defined, since

$$\sum_{k=0}^{\infty} P\{X = k\} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \times e^{\lambda} = 1.$$

In the examples above the set U of possible values of X consisted of integers. This is the most frequent case, since typically discrete random variables represent counts involved in observing random phenomena. In other words, discrete random variables are typically obtained as "the number of..." One should remember, however, that the definition allows any values in the set U, and not necessarily integers.

The distribution of a discrete random variable X is determined by the set $U = \{x_1, x_2, \ldots\}$ of its possible values, and the assignment of probabilities $P\{X = x_i\}$ to all $x_i \in U$, the only condition being (6.20) (i.e., the sum of all probabilities must be 1). The function p_X, defined by the formula

$$p_X(x) = \begin{cases} P\{X = x\} & \text{if } x \in U \\ 0 & \text{if } x \notin U, \end{cases}$$

is often called the *probability mass function*, or the *probability function* of random variable X, or (in engineering-oriented texts) a probability *density* function. We will often use the first two of these terms. The term *density* can be confusing in this context, so we will reserve it for continuous random variables, discussed later in this section.

The cdf of a discrete random variable can now be easily determined. We have

$$F(t) = P\{X \le t\} = \sum_{x_i \le t} P\{X = x_i\}. \tag{6.25}$$

■ **EXAMPLE 6.11**

In the simplest cases it is convenient to represent the distribution of a discrete random variable X as a double array of elements of U with corresponding probabilities, such as

Values	-3	$\frac{1}{2}$	5
Probability	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

We have here $U = \{-3, 1/2, 5\}$ and $P\{X = -3\} = 1/2, P\{X = 1/2\} = 1/3, P\{X = 5\} = 1/6$.

According to (6.25), $F(t) = 0$ for all $t < -3$, and for $-3 \leq t < 1/2$,

$$F(t) = P\{X \leq t\} = P\{X = -3\} = \frac{1}{2}.$$

For $1/2 \leq t < 5$, we have

$$P\{X \leq t\} = P\{X = -3\} + P\left\{X = \frac{1}{2}\right\} = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}.$$

Finally, for $t \geq 5$, we have $P\{X \leq t\} = P\{X \leq 5\} = 1$.

It is worthwhile to point out that the cdf is defined for *all* real arguments, and not only for the possible values of the random variable. For example, one can find $F_X(2.27)$ if X is the result of a single toss of a regular die.

■ **EXAMPLE 6.12**

The random variable X in Example 6.9 was defined as the number of tosses of a die, up to and including the first ace. Here the cdf is as follows: First, since $X \geq 1$, we have $F(t) = 0$ for all $t < 1$. Second, if $k \leq t < k + 1$, where $k = 1, 2, \ldots$, then we have

$$F(t) = P\{X \leq t\} = \sum_{j=1}^{k} P\{X = j\} = \sum_{j=1}^{k} \frac{1}{6}\left(\frac{5}{6}\right)^{j-1} = 1 - \left(\frac{5}{6}\right)^{k}.$$

In both examples the cumulative distribution function is a step function, with cdf increasing at every point of the set U and constant between the steps. Although this is the property characterizing most discrete random variables, there also exist discrete random variables such that their cdf's are not constant on *any* interval, as shown in the example below.

■ **EXAMPLE 6.13**

Let U be the set of all rational numbers. It is well known that U is countable, that is, all elements of U can be arranged into a sequence $x_1, x_2, \ldots$ (but U *cannot* be arranged in a sequence $x_1, x_2, \ldots$ with $x_n < x_{n+1}$ for all n). Let X be a random variable such that $P\{X = x_n\} = 1/2^n, n = 1, 2, \ldots$. Then the cdf of X is not constant on any interval. Indeed, for $t_1 < t_2$ we have

$$F(t_2) - F(t_1) = \sum_{t_1 < x_n \leq t_2} P\{X = x_n\}, \tag{6.26}$$

and the right-hand side is positive, since there exists a rational number x_n between any two distinct real numbers t_1 and t_2. Consequently, F increases between any two points.

Formula (6.26) shows how to calculate the probability of any interval $(t_1, t_2]$. Generally, for any set A we have

$$P\{X \in A\} = \sum_{x_n \in A} P\{X = x_n\}.$$

■ **EXAMPLE 6.14**

The random variable X has the following distribution:

Value	-3	-2	0	1	2
Probability	$\frac{2}{9}$	$\frac{1}{9}$	p	$\frac{2}{9}$	p^2

and we want to find $P(|X+1| > 1)$.

SOLUTION. First, we use (6.20) to find the value of p. We have here

$$\frac{2}{9} + \frac{1}{9} + p + p^2 + \frac{2}{9} = 1,$$

which gives the equation $p^2 + p - 4/9 = 0$. The solutions are $p = 1/3$ and $p = -4/3$, of which only the first is admissible as a probability. Consequently $P\{X = 0\} = 1/3, P\{X = 1\} = 1/9$ and

$$\begin{aligned} P\{|X+1| > 1\} &= P\{X = -3\} + P\{X = 1\} + P\{X = 2\} \\ &= \frac{2}{9} + \frac{2}{9} + \frac{1}{9} = \frac{5}{9}. \end{aligned}$$

Let us now introduce another large class of random variables.

Definition 6.3.3 The random variable X will be called *continuous* if there exists a nonnegative function f, called the *density* of X, such that

$$F_X(t) = \int_{-\infty}^{t} f(x)dx \tag{6.27}$$

for $-\infty < t < \infty$. □

■ **EXAMPLE 6.15 Uniform Distribution**

Let $A < B$, and let the density $f(x)$ have the form:

$$f(x) = \begin{cases} 0 & \text{if } x < A \\ c & \text{if } A \le x \le B \\ 0 & \text{if } x > B. \end{cases} \tag{6.28}$$

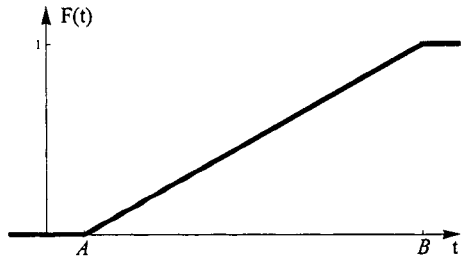


Figure 6.4 Cdf of distribution uniform on $[A, B]$

If $t < A$, then the integrand in (6.27) is 0, so $F(t) = 0$. For $A \leq t \leq B$ we have

$$F(t) = \int_{-\infty}^{A} f(x)dx + \int_{A}^{t} f(x)dx = \int_{-\infty}^{A} 0dx + \int_{A}^{t} cdx = c(t - A).$$

Finally, if $t > B$, then

$$F(t) = \int_{-\infty}^{A} f(x)dx + \int_{A}^{B} f(x)dx + \int_{B}^{t} f(x)dx = \int_{A}^{B} cdx = c(B - A).$$

This means that $F(t)$ is of the form presented in Figure 6.4. Clearly, since $\lim_{t\to\infty} F(t) = 1$ by Theorem 6.2.2, we must have $c(B - A) = 1$; hence $c = 1/(B - A)$. In other words, if function (6.28) is a density of a random variable, then c is uniquely determined by the length of the interval $[A, B]$. Thus the density of the distribution (continuous) *uniform* on $[A, B]$, denoted $U[A, B]$, is

$$f(x) = \begin{cases} 0 & \text{if } x < A \\ \frac{1}{B-A} & \text{if } A \leq x \leq B \\ 0 & \text{if } x > B. \end{cases} \tag{6.29}$$

In general, by letting $t \to \infty$ in (6.27), we see that every density function f must satisfy the relation

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \tag{6.30}$$

Also, it follows from (6.27) that $F(t)$ is a continuous function, and consequently for every x,

$$P(X = x) = F(x) - F(x - 0) = 0. \tag{6.31}$$

Formula (6.27) combined with (6.31) leads to the following theorem:

Theorem 6.3.1 *If random variable X has density f, then for all $a < b$ we have*

$$\begin{aligned} P\{a \leq X \leq b\} &= P\{a < X \leq b\} = P\{a < X < b\} \\ &= P\{a \leq X < b\} = F(b) - F(a) = \int_{a}^{b} f(x)dx. \end{aligned}$$

This formula implies that in dealing with continuous random variables one can afford the luxury of being sloppy in handling inequalities. In particular, one can treat an event such as $\{X \leq a\}$ as (equivalent to) the complement of the event $\{X \geq a\}$, and so on. This is in sharp contrast with discrete random variables, where the events $\{X \leq a\}$ and $\{X < a\}$ may have different probabilities.

A question arises as to the extent to which the cdf of a continuous random variable determines its density. Formula (6.27) suggests that we have

$$F'(t) = f(t), \tag{6.32}$$

and consequently, since F is a nondecreasing function (by Theorem 6.2.2), we must have

$$f(t) \geq 0. \tag{6.33}$$

In fact formulas (6.32) and (6.33) need not be valid for all points t. The reason is that the density f determines the probabilities of intervals through integration. This means that f is not defined uniquely. Indeed, the two densities f_1 and f_2—which differ only at a single point or on a finite set of points—will satisfy the condition

$$\int_a^b f_1(x)\,dx = \int_a^b f_2(x)\,dx$$

for all a and b. The same will be true if f_1 and f_2 differ on some set of measure zero.

Consequently, we may only claim that formulas (6.32) and (6.33) are valid almost everywhere, that is, except on a set of measure zero.

■ **EXAMPLE 6.16**

The random variable X, with density given in Example 6.15, is called *uniform* on interval $[A, B]$. It is clear, however, that if we modify the definition of f at boundaries A and B (e.g., by putting $f(A) = f(B) = 0$), the cdf will remain unchanged. A particular consequence of this observation is as follows: in real situations we often deal with discontinuous densities defined by "broken formulas," that is, functions given by different formulas in different intervals. In such cases it does not matter how the density is defined at the endpoints.

The last remark is true in regard to calculating probabilities. One should nevertheless be careful with the interpretation of density. For example, since for $\Delta x > 0$, we have

$$P\{x \leq X \leq x + \Delta x\} = \int_x^{x+\Delta x} f(t)dt,$$

we can approximate the last integral, for *continuous densities f*, by $f(x)\Delta x$. This leads to an "engineer's" interpretation, according to which $f(x)dx$ is the "probability that random variable X will assume the value in the infinitesimal interval $(x, x + dx)$." Here one has to be careful not to apply such an interpretation at points where f is not continuous. For instance, in the Example 6.15 we had $f(B) = 1/(B - A)$; hence the probability of X assuming a value in the interval $[B, B + h]$ for some small $h > 0$ may be taken as $h/(B - A)$. But this value is positive, whereas in fact the probability of X exceeding B by *any* amount is zero.

Let us now introduce an important type of continuous distributions called an exponential distribution.

Definition 6.3.4 (Exponential Distribution) The distribution with the density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases} \tag{6.34}$$

where $\lambda > 0$, will be called *exponential* with parameter λ, denoted EXP(λ).

To check that (6.34) is indeed a density, we write

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{+\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x}\,|_0^{\infty} = 1.$$

According to the remark made in Example 6.16, the value of the density at $x = 0$ plays no role. Thus we could have defined $f(x)$ in Definition 6.3.4 as $\lambda e^{-\lambda x}$ for $x \geq 0$ and 0 for $x < 0$.

■ **EXAMPLE 6.17**

Let us compute $P\{|X - 1/2| > 1/4\}$ for a random variable X that has EXP(2) distribution.

SOLUTION. Since the density determines probabilities through integration,

$$\begin{aligned} P\left\{|X - \frac{1}{2}| > \frac{1}{4}\right\} &= \int_{|x-1/2|>1/4} f(x)\, dx \\ &= \int_{x<1/4} f(x)\, dx + \int_{x>3/4} f(x)dx \\ &= \int_{-\infty}^{0} 0\, dx + \int_{0}^{1/4} 2e^{-2x}\, dx + \int_{3/4}^{\infty} 2e^{-2x}\, dx \\ &= 1 - e^{-1/2} + e^{-3/2} = 0.6166. \end{aligned}$$

Example 6.17 should serve as a warning. It happens often that density is defined by a formula consisting of several parts, like (6.34). In calculating probabilities, one has to integrate f over some set, and special care should be taken to use the proper part of the formula for f on appropriate parts of the domain of integration.

We will always choose a version of density that is "most regular." In particular, this means choosing a continuous or piecewise continuous version if possible. When density is discontinuous, the choice of values at breakpoints is irrelevant.

■ **EXAMPLE 6.18 Normal Distribution**

One of the most common distributions encountered in both probability theory and statistics, useful in modeling real-life phenomena, is the *normal* distribution, with density defined by

$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \tag{6.35}$$

where μ and $\sigma > 0$ are two parameters whose interpretation will be given later. The distribution (6.35) is usually denoted $\mathrm{N}(\mu, \sigma^2)$. The distribution $\mathrm{N}(0, 1)$, with density

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{6.36}$$

is called *standard normal*.

We will show that the function (6.35) is a density; that is, it integrates to 1. Indeed, letting $z = (x - \mu)/\sigma$, so that $\sigma dz = dx$, we have

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-z^2/2} dz.$$

Next we need to prove that a standard normal density (6.36) integrates to 1.

The function $e^{-x^2/2}$ does not have an antiderivative that can be written in a closed form. Thus, to compute the integral

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} dx,$$

we apply a trick and compute I^2:

$$\begin{aligned} I^2 &= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} dx \right) \times \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-y^2/2} dy \right) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)/2} dx\, dy. \end{aligned} \tag{6.37}$$

Using polar coordinates (r, θ), with $x = r\cos\theta, y = r\sin\theta$, we write the Jacobian of the transformation as

$$J = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r.$$

Then we change the variables in (6.37) to obtain

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_0^{\infty} r\, e^{-r^2/2}\, dr = 1;$$

hence $I = 1$, as was to be shown.

As already mentioned, the density function $\varphi(x)$ of standard normal distribution (as well as any other normal density function $f(x; \mu, \sigma)$) does not have an antiderivative expressible in closed form. Therefore to find probabilities for normal random variables, one needs to use either statistical tables or the available statistical software.

Suppose that we have to find

$$P\{a \le X \le b\} = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \tag{6.38}$$

for random variable $X \sim N(\mu, \sigma^2)$. As we see, in many families of distributions that depend on parameters, one needs an extensive set of tables. Fortunately, in the case of any normal distribution, one table—that of the standard normal distribution—is sufficient. To see why, consider the probability (6.38) and use the change of variable $z = (x - \mu)/\sigma$ in the integral. We have $dx = \sigma dz$; hence after substitution to (6.38) we obtain

$$P\{a \leq X \leq b\} = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2} dz = \Phi(z_2) - \Phi(z_1),$$

where $z_1 = (a - \mu)/\sigma$, $z_2 = (b - \mu)/\sigma$, and Φ is the cdf of the standard normal distribution (Table A.2). This table is prepared for $z > 0$, but for $z < 0$ one may use the fact that $\varphi(z)$ is symmetric about 0 so that

$$\Phi(-z) = 1 - \Phi(z). \tag{6.39}$$

The procedure is illustrated by the following example:

■ **EXAMPLE 6.19**

Suppose that $X \sim N(-0.7, 4)$ and that U is defined as an integer nearest to X. Find $P(U = -1)$.

SOLUTION. In this case $\mu = -0.7$ and $\sigma^2 = 4$; hence $\sigma = 2$. Consequently,

$$\begin{aligned} P(U = -1) &= P(-1.5 < X < 0.5) \\ &= \Phi\left(\frac{-0.5 - (-0.7)}{2}\right) - \Phi\left(\frac{-1.5 - (-0.7)}{2}\right) \\ &= \Phi(0.1) - \Phi(-0.4). \end{aligned}$$

Using (6.39) to reduce to positive arguments of Φ, we have

$$\begin{aligned} P(U = -1) &= \Phi(0.1) - (1 - \Phi(0.4)) = \Phi(0.1) + \Phi(0.4) - 1 \\ &= 0.5398 + 0.6554 - 1 = 0.1952, \end{aligned}$$

where the numerical values are found in Table A2.

At the end of this section it is necessary to point out that discrete and continuous random variables do not exhaust all possibilities. First, we may have practical situations of random variables of *mixed* type, partially discrete and partially continuous. Second, we may also have random variables that are neither continuous nor discrete. This second possibility may appear at first as a mathematical pathology of some sort; nevertheless, there are random variables occuring in practice that have such pathological distributions.

■ **EXAMPLE 6.20**

An example of a random variable of mixed type may be as follows. We purchase a piece of equipment (e.g., a light bulb). We denote its lifetime by T.

The lifetime is typically a continuous random variable in the sense that T can assume any value from some interval; thus any particular value $T = t$ from this interval has probability zero. In addition the value $T = 0$ can be assumed with positive probability. In other words, the bulb may either be broken at the time of purchase (in which case $T = 0$), or it may break at some future time $t > 0$, at which time the event $\{T = t\}$ has probability zero. Consequently, the cdf of T is a function that equals 0 for all negative t, and is continuously increasing to 1 for positive t. At $t = 0$ the cdf is discontinuous, with $F(0) > 0$ being the probability of purchasing a broken light bulb.

Such mixtures of continuous and discrete distributions still do not exhaust all possibilities. This is illustrated by the following example, which is of theoretical interest:

■ **EXAMPLE 6.21**

We will now construct a cdf $F(t)$ that is continuous, increases from 0 to 1, and is such that $F'(t) = 0$, except on a set of measure 0. The latter condition excludes the existence of density; that is, if the density exists, then it equals $F'(t)$ almost everywhere. Thus we have $f(t) = 0$ almost everywhere; hence $\int_{-\infty}^{+\infty} f(t)\, dt = 0$. which means that F' is not a density.

The construction is based on the *Cantor set*. We let $F(t) \equiv 0$ for $t \leq 0$ and $F(t) \equiv 1$ for $t \geq 1$. Next we let $F(t) \equiv 1/2$ for $1/3 \leq t < 2/3$. On middle parts of intervals $[0, 1/3]$ and $[2/3, 1]$, that is, for $1/9 \leq t < 2/9$ and for $7/9 \leq t < 8/9$, we let $F(t) \equiv 1/4$ and $F(t) \equiv 3/4$, respectively. This process is continued, and at each step, $F(t)$ is the average of values on neighboring intervals in the middle one-third of the "gap." The total length of intervals in $[0, 1]$ where F is constant (hence where $F' = 0$) is

$$L = \frac{1}{3} + 2 \times \frac{1}{9} + 4 \times \frac{1}{27} + \cdots = \sum_{n=0}^{\infty} \frac{2^n}{3^{n+1}} = \frac{1}{3} \times \frac{1}{1 - \frac{2}{3}} = 1.$$

Moreover, one can easily show that F is continuous at each point. Thus we have constructed a cdf of a random variable which is neither discrete nor continuous (it is called *singular*).

PROBLEMS

6.3.1 A die is biased in such a way that the probability of obtaining k dots ($k = 1, \ldots, 6$) is proportional to k^2. Which number of dots is more likely: odd or even?

6.3.2 You have 5 coins in your pocket: 2 pennies, 2 nickels, and a dime. Three coins are drawn at random. Let X be the total amount drawn (in cents). Find: **(i)** The distribution of X. **(ii)** $P(X \leq 10 | X \leq 15)$. **(iii)** The probabilities that two pennies are drawn, if it is known that $X \leq 11$.

6.3.3 Let X have the density $f(x) = Ce^{-0.4|x|}$, $-\infty < x < +\infty$ (such distribution is called *Laplace* or *double exponential*). Find C and then obtain: **(i)** $P(X > -2)$. **(ii)** $P(|X + 0.5| < 1)$.

6.3.4 Let X have EXP(1) distribution. Moreover, let $Y = [X]$ be the integer part of X, and let Z be the integer nearest to X. Find: **(i)** The distributions of Y and Z. **(ii)** $P(Y = Z)$. **(iii)** $P(Y = 3|Z = 4)$. **(iv)** $P(Z = 4|Y = 3)$. **(v)** $P(Y = 4|Z = 3)$. **(vi)** $P(Z = 3|Y = 4)$.

6.3.5 Let X have EXP(λ) distribution, and let Y and Z be defined as in Problem 6.3.4. **(i)** Find $P(Y = Z)$. **(ii)** Show that $P(Y = k|Z = k + 1) = P(Y = Z)$ for all $k = 0, 1, \ldots$. **(iii)** Find $P(Z = k + 1|Y = k)$ for $k = 0, 1, \ldots$.

6.3.6 Let X have EXP(λ) distribution. Show that for $s, t > 0$ the following *memoryless property* holds: $P\{X > s + t|X > s\} = P\{X > t\}$.

6.3.7 Let X_n be the difference (possibly negative) between the number of heads and the number of tails in n tosses of a coin. Find: **(i)** The distribution of X_4. **(ii)** The cdf of X_4 at point $x = -0.6$. **(iii)** The probability that X_n is positive given that it is nonnegative for (a) $n = 4$ and (b) $n = 5$.

6.3.8 Let X have the density $f(x) = Cx$ for $0 \leq x \leq 1, f(x) = C(2 - x)/2$ for $1 < x \leq 2$, and $f(x) = 0$ otherwise. Find C and $F(x)$. Compute the following probabilities and show them on the graphs of $f(x)$ and $F(x)$: **(i)** $P(X \geq 3/2)$. **(ii)** $P(|X - 1| \leq 1/2)$. (*Hint.* The problem can be solved without integration, just using simple geometry.)

6.3.9 Let random variable X with the cdf F be uniformly distributed over the union of intervals $(0, a)$ and $(a + 2, b)$. Assuming that $F(4) = 0.2$ and $F(a + 1) = 0.25$, find: **(i)** a and b. **(ii)** $F(8.39)$. **(iii)** $P(3.01 \leq X \leq 9.14)$.

6.3.10 An oscillator sends the wave $X(t) = A\cos(2\pi t)$, where $A = 1$ or 2 with equal probabilities. We observe the value of $X(t)$ at the point chosen at random from the U[n, n+1] distribution for some n. Find: **(i)** $P(X(t) \leq 1)$. **(ii)** $P(|X(t)| > 3/2)$. **(iii)** $P(X(t) > 0)$.

6.4 FUNCTIONS OF RANDOM VARIABLES

In practical situations we often deal with functions (or *transformations*) of random variables. Given the original distributions, we face the problem of determining the distributions of transformed random variables. The examples abound, starting from the simplest cases, such as a change of unit of measurement or a representation on the logarithmic scale. More complicated cases, discussed in Chapter 7, involve pairs of random variables such as conversion of a pair of Cartesian coordinates of a random point on the plane to polar coordinates of this point, or ratios of coordinates. The latter case occurs, for instance, in determining the distribution of velocity, calculated as the ratio of distance to time, both subject to variability (either in the actual measured velocity or due to errors of measurements).

Another compelling reason to study transformations of random variables could be the generation of random numbers that follow a specific distribution. A typical method here involves transformations. A computer (and often even pocket calculators) can generate random numbers from distribution uniform on $[0, 1]$. Then the

desired random variables are obtained as suitable transformations of uniform random variables. Obviously to apply such a method, it is necessary to develop techniques of determining the distributions of functions of random variables, at least of those uniformly distributed.

Finally, the third reason is related to statistics. Observations such as experimental results and values recorded in the sample are regarded as values of random variables. These are often summarized into global indices, such as the average. Each such index (generally referred to as a *statistic*) is a function of the sample values, hence a function of random variables, and for statistical inference it is vital to know the distributions of such indices.

We begin with the conceptually and technically simplest case of transformations of one discrete random variable.

Let X assume values in the set $U = \{x_1, x_2, \ldots\}$, with corresponding probabilities

$$p_i = P\{X = x_i\},$$

such that $\sum p_i = 1$. Then $Y = \varphi(X)$, where φ is a real-valued function, also has a discrete distribution. The following example illustrates how the distribution of $Y = \varphi(X)$ can be obtained:

■ **EXAMPLE 6.22**

Let us consider a simple numerical example. Suppose that X has the distribution

Value	-2	-1	0	1	2	3	4
Probability	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$

If $\varphi(x) = x^2$, then $Y = X^2$, and the distribution of Y is

Value	0	1	4	9	16
Probability	$\frac{1}{10}$	$\frac{1}{10} + \frac{2}{10}$	$\frac{1}{10} + \frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$

Function φ is not one to one, and therefore

$$P\{Y = 4\} = P\{X^2 = 4\} = P\{X = 2\} + P\{X = -2\} = \frac{1}{10} + \frac{1}{10}.$$

$P\{Y = 1\}$ can be obtained similarly.

In general, we have

$$P\{Y = y\} = P\{\varphi(X) = y\} = P\{x : \varphi(x) = y\} = \sum_{x_i : \varphi(x_i) = y} P\{X = x_i\}.$$

An analogous formula can be used for functions of two variables, that is, random variables of the form $Z = \varphi(X, Y)$, where the distribution of Z is expressed in terms of the distribution of X and Y. The principle here is extremely simple, so it is usually best to start "from scratch," as in the next example.

■ **EXAMPLE 6.23**

Let the experiment consist of two tosses of a regular die, and let X and Y be the result on the first and second toss, respectively. We want to find the distribution of $Z = \max(X, Y)$. Thus the value of the function $\varphi(x, y)$ equals x or y depending on whether $x \geq y$ or $y \geq x$, and the random variable Z may be described as "the best out of two tosses."

Clearly, Z may be $1, 2, \ldots, 6$, and it is simplest to find the distribution of Z by listing all of the outcomes associated with each specific value of Z. Thus $Z = 1$ only if $(X, Y) = (1, 1)$; hence $P\{Z = 1\} = 1/36$. Next $Z = 2$ if the outcome is $(2, 1), (1, 2)$, or $(2, 2)$ so that $P\{Z = 2\} = 3/36$. Proceeding in this way, we get

Z	1	2	3	4	5	6
Probability	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

It can be seen how "giving the second chance" improves the score. Larger values are much more likely than smaller values.

■ **EXAMPLE 6.24**

Let us continue Example 6.23, generalizing it to the case of n random variables (to be considered formally in Chapter 7). Let us imagine that an adult plays some game with a child, and in order to give the child some advantage the adult allows the child to use the "best of n tosses" of a die. How large should n be to give the child a 99% or higher chance to score at least 4?

SOLUTION. Let $X_1, X_2, \ldots, X_n$ denote the result of consecutive tosses, and let $Z = \max(X_1, \ldots, X_n)$. We want the smallest n with $P\{Z \geq 4\} \geq 0.99$. Let us therefore determine the distribution of Z, exhibiting its dependence of n. Direct enumeration was feasible for $n = 2$, as in the preceding example, but it is cumbersome for a larger n. However, it is easy to find $P\{Z \leq k\}$ for $k = 1, \ldots, 6$. Indeed, $P\{Z \leq 1\} = P\{Z = 1\} = 1/6^n$, since only one outcome among the total of 6^n outcomes gives the maximum score of 1. Next

$$\{Z \leq 2\} = \{\max(X_1, \ldots, X_n) \leq 2\} = \{X_1 \leq 2, X_2 \leq 2, \ldots, X_n \leq 2\},$$

and we have

$$P\{Z \leq 2\} = \frac{2^n}{6^n}.$$

In general,

$$P\{Z \leq k\} = \frac{k^n}{6^n} \qquad \text{for } k = 1, \ldots, 6;$$

hence

$$P\{Z = k\} = P\{Z \le k\} - P\{Z \le k-1\} = \frac{k^n}{6^n} - \frac{(k-1)^n}{6^n}.$$

Since $P\{Z \ge 4\} \ge 0.99$ means that

$$P\{Z \le 3\} = \left(\frac{3}{6}\right)^n \le 0.01,$$

we have to solve the inequality $(1/2)^n \le 0.01$. Taking $n \log(1/2) \le \log 0.01$, we obtain

$$n \ge \frac{\log 0.01}{\ln(1/2)} = \frac{\log 100}{\log 2} = 6.64.$$

Thus $n = 7$ tosses "practically guarantees" that at least one toss will lead to 4, 5, or 6.

We now consider the case of transformation of a single continuous random variable. Let F and f denote the cdf and density of X and let $Y = \varphi(X)$, where φ is assumed to be at least piecewise differentiable.

Regardless of whether we want the cdf or the density, the best strategy is to start by finding the cdf of Y. This method, occasionally referred to as the *cdf technique*, will be illustrated by some examples.

We begin with the simplest case, when φ is a strictly monotone function. If φ is increasing, we write for the cdf of Y:

$$F_Y(y) = P\{Y \le y\} = P\{\varphi(X) \le y\} = P\{X \le \psi(y)\} = F_X(\psi(y)),$$

where ψ is the function inverse to φ.

The density is obtained by differentiating cdf, and consequently

$$f_Y(y) = \frac{d}{dy} F_X(\psi(y)) = f_X(\psi(y))\psi'(y). \tag{6.40}$$

If φ is monotonically decreasing, ψ must be a decreasing function too, and

$$F_Y(y) = P\{\varphi(X) \le y\} = P\{X \ge \psi(y)\} = 1 - F_X(\psi(y)).$$

Therefore

$$f_Y(y) = -f_X(\psi(y))\psi'(y), \tag{6.41}$$

a quantity that is positive, since now $\psi'(y) < 0$. Together, (6.40) and (6.41) lead to the following theorem:

Theorem 6.4.1 *If φ is a continuous differentiable function with inverse ψ and X is a continuous random variable with density f_X, then the density of $Y = \varphi(X)$ is*

$$f_Y(y) = f_X(\psi(y))|\psi'(y)|. \tag{6.42}$$

■ **EXAMPLE 6.25 Linear Transformations**

Let $Y = aX + b$. If $a > 0$, we can write

$$F_Y(y) = P\{aX + b \le y\} = P\left\{X \le \frac{y-b}{a}\right\} = F_X\left(\frac{y-b}{a}\right)$$

and

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

If $a < 0$, we have

$$F_Y(y) = P\{aX + b \le y\} = P\left\{X \ge \frac{y-b}{a}\right\} = 1 - F_X\left(\frac{y-b}{a}\right)$$

and

$$f_Y(y) = -\frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

So for any $a \neq 0$,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

The following transformation is important, both theoretically and practically. Let X have cdf F and density f, and let F be strictly increasing, so that F^{-1} exists. We consider the transformation $Y = F(X)$. Obviously, $0 < Y < 1$, so the distribution of Y is concentrated on $[0, 1]$. We therefore have $F_Y(y) = 0$ for $y \le 0$ and $F_Y(y) = 1$ for $y \ge 1$, while for $0 < y < 1$ we write

$$F_Y(y) = P\{F(X) \le y\} = P\{X \le F^{-1}(y)\} = F(F^{-1}(y)) = y.$$

Consequently, $f_Y(y) = \frac{d}{dy} F_Y(y) = 1$ on $[0, 1]$, and we proved the next theorem.

Theorem 6.4.2 (Probability Integral Transform) *If X has continuous strictly increasing cdf F, then the distribution of $Y = F(X)$ is uniform on $[0, 1]$.*

This theorem may be formulated as follows: *If Y has $U[0, 1]$ distribution, then $X = F^{-1}(Y)$ has a distribution with the cdf F.* Thus we have obtained a method of generating random variables with given continuous invertible cdf F, using random variables uniform on $[0, 1]$.

■ **EXAMPLE 6.26**

The assumption of monotonicity of φ (and hence also ψ) is crucial for the validity of (6.42), except for the obvious remark that what is really needed is monotonicity "in the domain where it really matters." Specifically, if C is the set of all points x at which $f(x) > 0$ (called the *support* of X), then it suffices that φ is monotone only on C.

Consider the distribution with density

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0, \end{cases}$$

and let $\varphi(x) = x^2$. Then φ is not monotone for all x, but it is monotone on the support of X, $C = [0, \infty)$. Thus $\psi(x) = \sqrt{x}, \psi'(x) = 1/(2\sqrt{x})$, and the density of $Y = X^2$ is

$$f_Y(y) = \begin{cases} \frac{\alpha}{2\sqrt{y}} e^{-\alpha\sqrt{y}} & \text{for } y > 0 \\ 0. & \text{for } y \leq 0, \end{cases}$$

In the case where φ is not monotone, we still have

$$F_Y(y) = P\{Y \leq y\} = P\{\varphi(X) \leq y\},$$

but this time φ has no inverse, and the inequality $\varphi(X) \leq y$ is usually not equivalent to a single inequality for X. Still the right-hand side can, in most cases, be represented through the cdf F_X evaluated at some points dependent on y. Differentiating, we can recover the density f_Y of Y.

This principle will now be illustrated by few examples.

■ **EXAMPLE 6.27**

Let X be a random variable with the density

$$f_X(x) = \begin{cases} \frac{5}{2}x^4 & \text{for } |x| \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{6.43}$$

Let $\varphi(x) = x^2$, so that now φ is not monotone, and consider $Y = \varphi(X)$. Since $Y \geq 0$, we have $F_Y(y) = 0$ for $y \leq 0$, while for $y > 0$ we write

$$\begin{aligned} F_Y(y) &= P\{X^2 \leq y\} = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Differentiating, we get for $y > 0$,

$$f_Y(y) = F_Y'(y) = [f_X(\sqrt{y}) + f_X(-\sqrt{y})] \times \frac{1}{2\sqrt{y}},$$

so that (remembering that $f_X = 0$ if $x < -1$ or $x > 1$)

$$f_Y(y) = \begin{cases} \frac{5}{2}y^{3/2} & \text{for } 0 < y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

■ **EXAMPLE 6.28 Square of a Normal Random Variable**

Let X have a normal distribution with the density

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

and let $\varphi(x) = x^2$ so that now φ is not monotone. We have $F_Y(y) = 0$ for $y \leq 0$, while for $y > 0$ we write

$$F_Y(y) = P\{X^2 \leq y\} = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Differentiating, we obtain

$$f_Y(y) = F_Y'(y) = [f_X(\sqrt{y}) + f_X(-\sqrt{y})] \times \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} \quad (6.44)$$

for $y > 0$, and $f_Y(y) = 0$ for $y \leq 0$.

This is a special case of a *gamma* density, given by the formula $Cy^{a-1}e^{-by}(y > 0)$, where $a > 0, b > 0$, and C is the normalizing constant. Density (6.44) has $a = 1/2, b = 1/2$. Gamma distributions with $b = 1/2$ and a such that $2a$ is a positive integer form a family of *chi-square* distributions—very important in statistical inference. Both gamma and chi-square distributions will be discussed in more detail in following chapters.

■ **EXAMPLE 6.29 Folded Normal Distribution**

Let X be as in Example 6.28, and let now $\varphi(x) = |x|$. Then for $y > 0$,

$$F_Y(y) = P\{|X| \leq y\} = P\{-y \leq X \leq y\} = F_X(y) - F_X(-y),$$

and

$$f_Y(y) = f_X(y) + f_X(-y) = \sqrt{\frac{2}{\pi}} e^{-y^2/2}, \quad (6.45)$$

and $f_Y(y) = 0$ for $y \leq 0$. This distribution is sometimes also called *half-normal*.

PROBLEMS

6.4.1 If X is the result of tossing a balanced die, find the distribution of: **(i)** $Y = (X-1)^2$. **(ii)** $Z = |X - 2.5|$.

6.4.2 Let X have the Poisson distribution with parameter λ, and let $Y = 2X$. Find the distribution of Y.

6.4.3 Let X have a continuous distribution with cdf F_X and density f_X, such that $F_X(0) = 0$. Find the cdf and density of random variables: **(i)** $\sqrt{X}$. **(ii)** $\log X$. **(iii)** $1/X$. **(iv)** e^X.

6.4.4 Let X be U[0, 1]. Find φ such that $Y = \varphi(X)$ has EXP(λ) distribution.

6.4.5 Let X have EXP(λ) distribution, and let $Y = \sqrt{X}$. Find:
(i) The cdf and density of Y. **(ii)** The lower quartile of Y.

6.4.6 Assume that X has the standard normal distribution. Find the density of: **(i)** $Y = X^3$. **(ii)** $Y = (X-1)^3$. **(iii)** $Y = e^X$.

6.4.7 Find the density of $Y = X(1 - X)$ if X has U[0, 1] distribution.

6.4.8 Random variable X has density $f_X(x) = Cx^4$ for $-2 \leq x \leq 1$ and 0 otherwise. Find the density of $Y = X^2$ (follow Example 6.27).

6.4.9 Let X have U[-1, 1] distribution. Find the distribution of: **(i)** $Y = |X|$. **(ii)** $Z = 2|X| - 1$.

6.4.10 The duration (in days) of the hospital stay following a certain treatment is a random variable $Y = 4 + X$, where X has a density $f(x) = 32/(x+4)^3$ for $x > 0$. Find: **(i)** The density of Y. **(ii)** The probability that a randomly selected patient will stay in the hospital for more than 10 days following the treatment.

6.4.11 Let X have density $f(x) = 2(1 - x)$ for $0 \leq x \leq 1$. Find the density of: **(i)** $Y = X(1 - X)$. **(ii)** $W = \max(X, 1 - X)$.

6.4.12 Suppose that the measured radius R is a random variable with density

$$f_R(x) = \begin{cases} 12x^2(1-x) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

In a circle with radius R find the cdf of: **(i)** The diameter. **(iii)** The area.

6.4.13 The speed of a molecule of gas at equilibrium is a random variable X with density

$$f(x) = \begin{cases} kx^2e^{-bx^2} & \text{for } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where k is a normalizing constant and b depends on the temperature of the gas and the mass of the molecule. Find the probability density of the kinetic energy $E = mX^2/2$ of the molecule.

6.5 SURVIVAL AND HAZARD FUNCTIONS

In this section we consider a special class of random variables that can serve as possible models for *lifetimes*, of living organisms or of some equipment. They may denote the actual lifetime (i.e., time until death or equipment failure) or time until some event (e.g., recovery from a disease). Such random variables, denoted usually by T, are continuous and nonnegative.

Let F and f denote the cdf and density of T so that for $t \geq 0$,

$$F_T(t) = P\{T \leq t\} = \int_0^t f(u)du \tag{6.46}$$

[the integration starts at 0, since nonnegativity of T implies that $P\{T \leq 0\} = 0$, and hence $f(t) = 0$ for $t < 0$]. Moreover, let

$$S(t) = 1 - F(t) = P\{T > t\} = \int_t^\infty f(u)du \tag{6.47}$$

be the *survival distribution function* (or *survival function*) of the random variable T.

We introduce the following definition:

Definition 6.5.1 The function

$$h(t) = -\frac{d \ln S(t)}{dt}, \tag{6.48}$$

defined for $t > 0$ and $S(t) > 0$, is called the *hazard rate* (or *intensity rate*) *function* of random variable T. □

By differentiating $-\ln(1 - F(t))$, it follows from (6.48) that at almost all points t we have

$$h(t) = \frac{f(t)}{S(t)}. \tag{6.49}$$

Consequently, we have the following theorem expressing the cdf through the hazard function:

Theorem 6.5.1 *If $h(t)$ is a hazard function of a random variable T, then its cdf equals*

$$F(t) = 1 - e^{-\int_0^t h(u)du}.$$

Proof. The proof is immediate, by integrating (6.49) between 0 and t. □

The interpretation of hazard functions is as follows: $h(t)\delta t$ can be approximated by

$$\frac{f(t)\delta t}{S(t)} \approx \frac{P\{t < T \leq t + \delta t\}}{P\{T > t\}};$$

hence, by the definition of conditional probability,

$$\begin{aligned} h(t)\delta t &\sim P\{t < T \leq t + \delta t | T > t\} \\ &= P\{\text{ lifetime ends before } t + \delta t | \text{ lifetime longer than } t\}. \end{aligned}$$

In other words, $h(t)$ is the death rate at t of those who survived until t (are "at risk" at t). Thus the hazard function describes the process of aging in terms of changes of risk of death with current age.

■ **EXAMPLE 6.30**

Let the random variable X have EXP(α) distribution. Then $F(t) = 1 - e^{-\alpha t}$, $S(t) = e^{-\alpha t}$, and

$$h(t) = \frac{\alpha e^{-\alpha t}}{e^{-\alpha t}} = \alpha.$$

Thus the exponential distribution describes the lifetime distribution in the case of lack of aging, when the risk of death does not change with age.

■ **EXAMPLE 6.31**

Suppose that you buy a new car, and let T be the time of the first breakdown. Typically $h(t)$ is initially high, and then declines to a constant. It remains at this level for several years, eventually beginning to increase.

The reason is that early failures are typically caused by hidden faults of material, undetected in factory control. If they do not show up immediately, then there are probably no faults, and risk of failure remains constant for some time. The later increase is due to the wearing out of various parts.

In an engineering context, especially in problems of reliability of equipment, the hazard function is often called *failure rate* function.

Assuming that $F(t) < 1$ for all t (which is the most important case of interest), we can write

$$F(t) = 1 - e^{-H(t)}, \tag{6.50}$$

where $H(t) = -\log[1 - F(t)]$. On the other hand, in view of Theorem 6.5.1, we also have $H(t) = \int_0^t h(u)du$, which gives (6.32). Since $F(t) \to 1$ as $t \to \infty$, we must have

$$\int_0^\infty h(u)du = \infty. \tag{6.51}$$

Example 6.30 shows that the exponential distribution has a constant hazard, interpreted as lack of aging. If hazard is increasing, we have the phenomenon of aging ("new better than old"), while the opposite is true in case of decreasing hazard ("old better than new"). A flexible model of both situations is given by the following definition:

Definition 6.5.2 A nonnegative random variable T with the hazard rate

$$h(t) = Kt^\gamma, \qquad t > 0, K > 0, \gamma > -1 \tag{6.52}$$

is said to have a *Weibull distribution.* □

Observe first that the condition $\gamma > -1$ ensures that the relation (6.51) holds so that T is a genuine random variable in the sense that

$$\lim_{t\to\infty} P\{T \le t\} = 1$$

(i.e., $T = \infty$ is excluded). For reasons of tradition and convenience, one usually puts $\alpha = K/(\gamma+1), \beta = \gamma+1$, so that the hazard rate of Weibull distribution takes the form

$$h(t) = \alpha\beta t^{\beta-1}, \qquad \alpha > 0, \beta > 0, t > 0. \tag{6.53}$$

PROBLEMS

6.5.1 Find the hazard function of the U[0, 1] distribution. Explain why $h(t)$ is unbounded.

6.5.2 Find the density and survival function of the distribution with hazard rate $h(t) = a + bt$ for $t > 0, a > 0, b > 0$.

6.5.3 Let X be a random variable with density

$$f(x) = \begin{cases} 0 & x < 0 \\ 0.5 & 0 \le x \le 1 \\ qe^{-\alpha x} & x > 1. \end{cases}$$

(i) Find q if α is known. **(ii)** Find hazard $h(t)$ and survival function $S(t)$.

6.5.4 Find the cdf and density of Weibull distribution with the hazard function (6.53).

6.5.5 Assume that the fuel pumps in a certain make of cars have lifetimes with a Weibull hazard rate $2/\sqrt[3]{t}$ (t measured in years). Find the probability that a fuel pump is still working after 5 months.

6.5.6 The series system is built in such a way that it operates only when all its components operate (so it fails when at least one component fails). Assuming that the lifetime of each component has EXP(1) distribution and that the components operate independently, find the distribution and survival function of the system's lifetime T.

6.5.7 A cancer specialist claims that the hazard function of random variable T_c = "age at death due to cancer" is a bounded function which for large t has the form $h_c(t) = k/t^2$ (where t is the age in years and k is some constant).

Assume that $h(t)$ is the hazard of the time of death due to other reasons than cancer (other diseases, accidents, old age, etc.). If T_c and T are times of death with hazards $h_c(t)$ and $h(t)$, assume that the observed time of death is $T^* = \min(T_c, T)$, with $\max(T_c, T)$ being unobservable. To get an insight into the feasibility of the assumption of the cancer specialist in question, imagine that *all* reasons of death other than cancer have been eliminated (i.e., $H(t) \equiv 0$) and find the probability that a person will live forever.

CHAPTER 7

RANDOM VARIABLES: MULTIVARIATE CASE

7.1 BIVARIATE DISTRIBUTIONS

The considerations of Section 6.2 can be extended to the case of several random variables analyzed at once, or equivalently, to the analysis of vector-valued random variables.

In the simplest case we have a pair of random variables (X, Y), that is, a pair of functions on the sample space $\mathcal{S}$.

■ **EXAMPLE 7.1**

Let the experiment consist of three tosses of a coin, and let X = number of heads in all three tosses and Y = number of tails in the last two tosses. The sample space $\mathcal{S}$ and corresponding values of X and Y are then

$\mathcal{S}$	X	Y	$\mathcal{S}$	X	Y
HHH	3	0	TTH	1	2
HHT	2	1	THT	1	1
HTH	2	1	HTT	1	1
THH	2	0	TTT	0	2

We can summarize all possible values of (X, Y) and their probabilities in the following table:

X/Y	0	1	2
0	0	0	$\frac{1}{8}$
1	0	$\frac{2}{8}$	$\frac{1}{8}$
2	$\frac{1}{8}$	$\frac{2}{8}$	0
3	$\frac{1}{8}$	0	0

The entries in the table represent the corresponding probabilities. For instance, $P\{X = 2, Y = 1\} = 1/4$ was obtained by counting the number of points s of the sample space such that $X(s) = 2$ and $Y(s) = 1$ (there are two such points: HHT and HTH). To simplify the notation, we will be using commas to denote the intersection of events. Thus we write $P\{X = x, Y = y\}$ instead of rather clumsy $P[\{X = x\} \cap \{Y = y\}]$.

In a natural way this example leads to a definition of a *discrete* bivariate random variable.

Definition 7.1.1 We say that the pair (X, Y) of random variables has a *discrete* distribution if there exist finite or countable sets A and B such that

$$P\{(X, Y) \in A \times B\} = \sum_{x \in A, y \in B} P\{X = x, Y = y\} = 1. \qquad \square$$

In Example 7.1 we have $A = \{0, 1, 2, 3\}$ and $B = \{0, 1, 2\}$. Obviously $P\{X = x, Y = y\} = 0$ if any of the values x or y lie outside the set A (respectively, B), but it is also possible that $P\{X = x, Y = y\} = 0$ for some $x \in A$ and $y \in B$. For instance, in Example 7.1 we have $P\{X = 0, Y = 1\} = 0$. In other words, the values (x, y) that have positive probability can form a proper subset of $A \times B$.

The continuous bivariate distributions, as may be expected, are defined through their densities.

Definition 7.1.2 Random variables (X, Y) are *jointly continuous* (or have *joint continuous distribution*) if there exists a function $f(x, y)$ such that for every rectangle

$$C = \{(x, y) : a \le x \le b, c \le y \le d\}$$

with $-\infty \le a < b \le \infty, -\infty \le c < d \le \infty$ we have

$$P\{(X, Y) \in C\} = \int_C \int f(x, y)\, dx\, dy. \tag{7.1}$$

The function f is called *joint* or *bivariate* density of (X, Y). $\square$

Some comments are in order here. First, as in the univariate case the density is defined only up to sets of measure zero (e.g., single points or arcs on the plane).

Second, the obvious consequences of Definition 7.1.2 are

$$\int_{R^2}\int f(x,y)\,dx\,dy = 1, \tag{7.2}$$

where R^2 is the plane, and

$$f(x,y) \geq 0 \quad \text{almost everywhere.} \tag{7.3}$$

These relations are obvious analogues of (6.30) and (6.33) for univariate densities.

The third comment is as follows: Formula (7.1) only covers rectangles. In the analogy with Theorem 6.2.5, it can be shown that probability P defined on the class of rectangles by (7.1) determines its unique extension to all sets in the plane that can be approximated through countable operations on rectangles, in particular to all figures that can be triangulated, as well as to circles, ellipses, and so on. In other words, formula (7.1) holds for a much wider class of sets C on the plane.

Conditions (7.2) and (7.3) allow us to use the Fubini theorem and replace the double integral (7.1) by the iterated integral, thus making actual integration possible. In particular, if C is the rectangle specified in Definition 7.1.2, then

$$\begin{aligned} P\{(X,Y)\in C\} &= \int\int_C f(x,y)\,dx\,dy \\ &= \int_a^b \left[\int_c^d f(x,y)\,dy\right] dx = \int_c^d \left[\int_a^b f(x,y)\,dx\right] dy. \end{aligned}$$

In general, we have

$$\int\int_C f(x,y)\,dx\,dy = \int_{C_2} \varphi(y)dy = \int_{C_1} \psi(x)dx.$$

Here C_1 and C_2 are "shadows" of C on the x-axis and y-axis, that is, $C_1 = \{x : (x,y)\in C\}$ for some y and $C_2 = \{y : (x,y)\in C\}$ for some x. Also

$$\psi(x) = \int_{C_x} f(x,y)\,dy, \qquad \varphi(y) = \int_{C_y} f(x,y)\,dx$$

with $C_x = \{y : (x,y)\in C\}$ and $C_y = \{x : (x,y)\in C\}$ being sections of C at points x and y, respectively (see Figure 7.1). To understand well the difference between the double integral

$$\int_{R^2}\int f(x,y)\,dx\,dy$$

and iterated integrals

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy,$$

interpreted as

$$\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} f(x,y)\,dx\right] dy \text{ and } \int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} f(x,y)\,dy\right] dx,$$

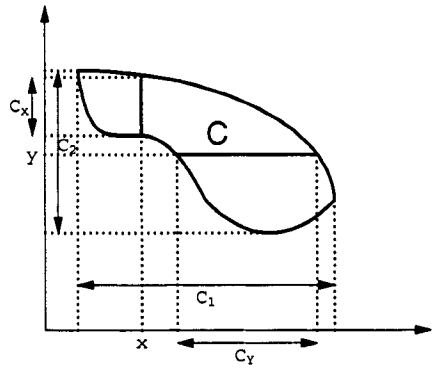


Figure 7.1 Shadows and sections of domain C of integration

the readers are advised to consult a good calculus text for the respective definitions. The Fubini theorem gives conditions under which all three integrals are equal (we will discuss these topics in Chapter 8). It is both easy and worthwhile to explain the issues involved here, using a simplified situation of a series (instead of integrals). Thus $\sum_{n=1}^{\infty} a_n$ is *defined* as $\lim_{n\to\infty} \sum_{k=1}^{n} a_k$, provided that this limit (of a well-defined numerical sequence) exists. By the same argument, the double sum $\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} a_{mn}$ is defined unambiguously (this sum is an analogue of the iterated integral). However, the symbol $\sum_{m,n=1}^{\infty} a_{mn}$ (the analogue of a double integral) is *not* well defined, since it does not specify the order in which the two-dimensional array $\{a_{mn}\}$ is to be added. The point here is that the sum of an infinite sequence of numbers may depend on the order of summation. Thus *some* assumptions about $\{a_{mn}\}$ are needed to make the last sum independent of the order of summation. Under these assumptions (e.g., nonnegativity) the double sum and both iterated sums are either all infinite or all equal to the same finite number.

A similar kind of difficulty appears in the case of double and iterated integrals and is resolved by the Fubini theorem.

We will now illustrate the calculation of probabilities by some examples.

■ EXAMPLE 7.2

The density $f(x, y)$ is given by the formula

$$f(x,y) = \begin{cases} cx(x+y) & \text{if } x \geq 0,\ y \geq 0,\ x+y \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{7.4}$$

Find $P\{X \leq 1/2\}$.

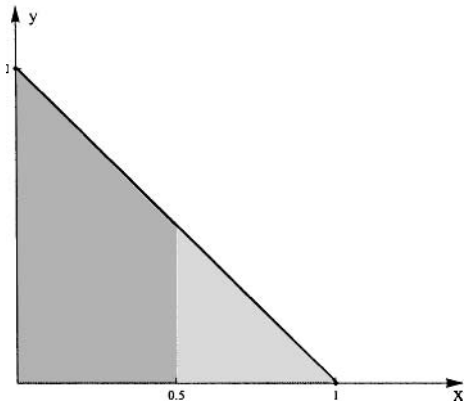


Figure 7.2 Support of density f and the set $\{X \leq 1/2\}$

SOLUTION. The first objective is to determine the constant c in formula (7.4). As usual, the normalizing condition (7.2) provides the key here:

$$\begin{aligned} 1 &= \iint cx(x+y)\,dx\,dy = c\int_0^1 \left[\int_0^{1-y} c(x^2+xy)\,dx\right] dy \\ &= c\int_0^1 \left(\tfrac{x^3}{3} + \tfrac{x^2}{2}y\Big|_0^{1-y}\right) dy = c\int_0^1 \left[\tfrac{(1-y)^3}{3} + \tfrac{(1-y)^2 y}{2}\right] dy = \tfrac{c}{8}, \end{aligned}$$

so $c = 8$.

Now, $P\{X \leq 1/2\}$ is the integral of the density $f(x,y)$ over the dark area in Figure 7.2. This time it is simpler to integrate over y first,

$$\begin{aligned} P\{X \leq 1/2\} &= \int_D\int 8x(x+y)\,dx\,dy = 8\int_0^{1/2}\left[\int_0^{1-x}(x^2+xy)\,dy\right] dx \\ &= 8\int_0^{1/2}\left(x^2y + x\frac{y^2}{2}\Big|_0^{1-x}\right) dx \\ &= 8\int_0^{1/2}\left[x^2(1-x) + \frac{1}{2}x(1-x)^2\right] dx = \frac{7}{16}. \end{aligned}$$

Although used not as often as in the univariate case, the cumulative distribution function (cdf) is still the most general way of specifying probabilities in the bivariate case, regardless of whether the distribution is discrete, continuous, or neither. We introduce the following definition:

Definition 7.1.3 The function of two real variables, defined as

$$F(x,y) = P\{X \leq x, Y \leq y\}$$

is called the *cumulative distribution function* (cdf) of the pair (X,Y) of random variables. □

The relation between cdf and the probability mass function in the discrete case, or the density function in the continuous case, is similar to those for univariate random variables:

$$F(x,y) = \sum_{x_i \le x, y_j \le y} P\{X = x_i, Y = y_j\}$$

in the discrete case, and

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\, dv\, du \tag{7.5}$$

in the continuous case. In particular, from (7.5) it follows that for almost all (x,y) we have

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}.$$

This brings us to the following analogue of Theorem 6.2.2:

Theorem 7.1.1 *Every bivariate cumulative distribution function $F(x,y)$ has the following properties:*
(a) $\lim_{x,y \to +\infty} F(x,y) = 1$.
(b) *For every y, the function $F(x,y)$ is nondecreasing and continuous from the right in x.*
(c) *For every x, the function $F(x,y)$ is nondecreasing and continuous from the right in y.*
(d) $\lim_{x \to -\infty} F(x,y) = 0$ *for every y, and* $\lim_{y \to -\infty} F(x,y) = 0$ *for every x.*
(e) *For all $x_1 < x_2$ and $y_1 < y_2$,*

$$F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1) \ge 0. \tag{7.6}$$

We omit the proof here, leaving proofs of some of the properties as exercises. The following comments, however, are important. While conditions (a)–(d) are direct analogue of the properties of univariate cdf's, condition (e) has no counterpart. The question therefore arises whether conditions (a)–(d) alone characterize bivariate cdf's. In other words, is every function satisfying (a)–(d) a cdf of some pair (X,Y) of random variables? The answer is negative, as shown by the following example:

■ **EXAMPLE 7.3**

Let $F(x,y)$ be defined by the formula

$$F(x,y) = \begin{cases} 1 & \text{if } x+y \ge 0 \\ 0 & \text{if } x+y < 0. \end{cases} \tag{7.7}$$

It is easy to check that function (7.7) satisfies conditions (a)–(d) given above. Suppose that $F(x,y)$ is the cdf of a pair of random variables, and let us compute the probability (see Figure 7.3), $P\{-1 < X \le 2, -1 < Y \le 2\}$. Since $F(x,y)$ is the probability of the pair (X,Y) taking a value "southwest" of (x,y) [i.e., to the left and below (x,y)], we have

$$F(2,2) - F(-1,2) - F(2,-1) + F(-1,-1) = 1 - 1 - 1 + 0 = -1,$$

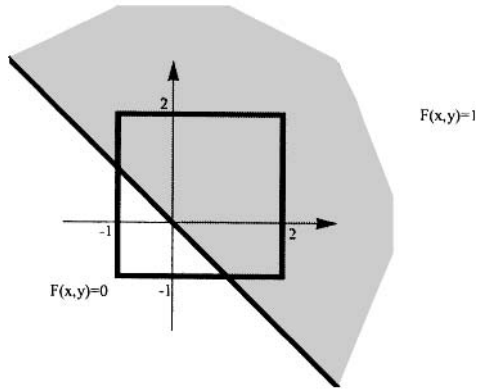


Figure 7.3 A function that is not a cdf but satisfies (a)–(d)

so that $F(x, y)$ cannot be a cdf.

Condition (e) is necessary, since the left-hand side of (7.6) equals $P\{x_1 < X \leq x_2, y_1 < Y \leq y_2\}$ and hence must be nonnegative.

It turns out (we omit the proof here) that (a)–(e) characterize a bivariate cdf. This means that any function satisfying (a)–(e) is a cdf of some pair of random variables.

PROBLEMS

7.1.1 A regular die is tossed twice. Find: **(i)** The joint distribution of variables $X =$ the total of outcomes and $Y =$ the best of the two outcomes. **(ii)** $P(X \leq 8, Y \leq 5)$; $P(X = 9, Y \leq 2)$, $P(4 \leq X \leq 7, 1 \leq Y \leq 3)$. **(iii)** $P(Y = 3|X = 4)$, $P(Y < 6|X = 7)$, $P(4 < Y \leq 6|X \leq 8)$.

7.1.2 Let X, Y have the joint distribution given by the following table:

X/Y	2	3	4
0	$\frac{1}{48}$	0	b
1	0	$\frac{5}{48}$	$\frac{8}{48}$
2	a	0	$\frac{11}{48}$
3	$\frac{5}{48}$	$\frac{12}{48}$	0

(i) Find a and b if it is known that $P(X = Y) = 1/3$. **(ii)** Find $P(XY = 0)$. **(iii)** If F is the cdf of (X, Y), find $F(-1.5, 3)$, $F(0.7, 2.11)$, and $F(1.5, 18)$.

7.1.3 A regular die is tossed twice. Let X be the number of times that 1 came up, and Y be the number of times 2 came up. Find: **(i)** The joint distribution of X and Y. **(ii)** The correlation coefficient between events $\{X = 1\}$ and $\{Y = 2\}$. [*Hint:* See formula (4.18) in Definition 4.5.2.]

7.1.4 Let the joint cdf of random variables X, Y be $F(x,y) = (1/48)xy(x+2y)$ for $0 \le x \le 2, 0 \le y \le 3$. Find the density $f(x,y)$.

7.1.5 The joint density of X and Y is $f(x,y) = y^2(xy^3+1)$ on the rectangle $0 \le x \le k, 0 \le y \le 1$. Find: **(i)** k. **(ii)** $P(X \le Y)$.

7.1.6 Assume that X, Y have density $f(x,y) = x+y$ for $0 \le x \le 1$ and $0 \le y \le 1$, and $f(x,y) = 0$ otherwise. Find $P\{Y \le \sqrt[3]{X}\}$.

7.1.7 Assume that (X,Y) have the joint density $f(x,y) = cxy^2$ for $0 \le x \le 1, 0 \le y \le 1$, and $f(x,y) = 0$ otherwise. Find: **(i)** c. **(ii)** $P\{X^2 \le Y \le X\}$. **(iii)** The cdf of (X,Y).

7.1.8 Let the joint density of random variables X, Y be $f(x,y) = cx^3y^2$ for $0 \le x \le 1, x^2 \le y \le 1$ and $f(x,y) = 0$ otherwise. Find $P(X < Y)$.

7.1.9 Variables X and Y have the joint density $f(x,y) = 1/y$ for $0 < x < y < 1$ and $f(x,y) = 0$ otherwise. Find: **(i)** $P(X+Y > 0.5)$. **(ii)** $P(Y > 2X)$.

7.2 MARGINAL DISTRIBUTIONS; INDEPENDENCE

One can naturally expect that the bivariate distribution (in the form of cdf, joint density, or probability mass function, as the case might be) contains more information than the univariate distributions of X and Y separately. Given a bivariate distribution, we are able to recover both univariate distributions of X and of Y, but not conversely.

We begin with the case of discrete bivariate distributions. Let $A = \{x_1, x_2, \ldots\}$ and $B = \{y_1, y_2, \ldots\}$ be the sets of possible values of X and Y, respectively, and let

$$p_{ij} = P\{X = x_i, Y = y_j\}. \tag{7.8}$$

Our objective is to express the distributions of X and of Y through p_{ij}. Since the events $\{Y = y_1\}, \{Y = y_2\}, \ldots$ form a partition (i.e., in the sense defined in Chapter 4, these events are mutually exclusive and one of them has to occur), we may write for every i,

$$\{X = x_i\} = \bigcup_j [\{X = x_i\} \cap \{Y = y_j\}] = \bigcup_j \{X = x_i, Y = y_j\}.$$

Since the events on the right-hand side are disjoint, we have

$$P\{X = x_i\} = \sum_j P\{X = x_i, Y = y_j\} = \sum_j p_{ij}.$$

In a similar way $P\{Y = y_j\} = \sum_i p_{ij}$. We now introduce:

Definition 7.2.1 Given the joint distribution (7.8), the distributions of X alone and Y alone, calculated from the formulas

$$P\{X = x_i\} = \sum_j p_{ij} = p_{i+}, \quad P\{Y = y_j\} = \sum_i p_{ij} = p_{+j},$$

will be referred to as *marginal* distributions. □

If we think of numbers p_{ij} as arranged into a matrix, then $P\{X = x_i\}$ and $P\{Y = y_j\}$ are sums of its corresponding rows (or columns). Since the sum of all p_{ij} equals 1, both marginal distributions satisfy the condition that the sum of their probabilities equals 1.

■ **EXAMPLE 7.4**

In Example 7.1 we considered three tosses of a coin, with X being the number of heads in all three tosses, and Y being the number of tails in the last two tosses. The joint distribution of X and Y is summarized by the following table:

X/Y	0	1	2	X
0	0	0	$\frac{1}{8}$	$\frac{1}{8}$
1	0	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{3}{8}$
2	$\frac{1}{8}$	$\frac{2}{8}$	0	$\frac{3}{8}$
3	$\frac{1}{8}$	0	0	$\frac{1}{8}$
Y	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

In the margins we have the row sums and column sums. The distribution of X is

$$P\{X = 0\} = P\{X = 3\} = \frac{1}{8}, \quad P\{X = 1\} = P\{X = 2\} = \frac{3}{8},$$

while the distribution of Y is

$$P\{Y = 0\} = P\{Y = 2\} = \frac{1}{4}, \quad P\{Y = 1\} = \frac{1}{2}.$$

The adjective *marginal* refers to the way in which the distribution was obtained; it implies nothing about the properties of the distribution. The definition of marginal distribution for continuous random variables is analogous, with summation replaced by integration:

Definition 7.2.2 If (X, Y) is a pair of continuous random variables with bivariate density $f(x, y)$, then the functions

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y)\, dy \quad \text{and} \quad f_2(y) = \int_{-\infty}^{+\infty} f(x, y)\, dx$$

are called the *marginal densities* of variables X and Y, respectively. □

The justification of Definition 7.2.2 consists of two parts. First, we need to show that f_1 and f_2 are densities, meaning that they are nonnegative for almost all arguments (i.e., for all arguments, except a set of arguments of measure zero) and that f_1 and f_2 integrate to 1. These properties are immediate consequences of the fact

that $f(x, y) \geq 0$ except possibly on a set of measure zero. Consequently, using the Fubini theorem, we can write

$$\begin{aligned} 1 &= \int\int f(x, y)\, dx\, dy = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} f(x, y)\, dx \right] dy \\ &= \int_{-\infty}^{+\infty} f_2(y)\, dy, \end{aligned}$$

and similarly for f_1.

The second part of justification of Definition 7.2.2 consists in showing that, for instance,

$$P\{a \leq X \leq b\} = \int_a^b f_1(x)\, dx.$$

Here we can write

$$\begin{aligned} \int_a^b f_1(x)\, dx &= \int_a^b \left[\int_{-\infty}^{+\infty} f(x, y)\, dy \right] dx \\ &= \iint_{a \leq x \leq b, -\infty < y < +\infty} f(x, y)\, dx\, dy \\ &= P\{a \leq X \leq b, -\infty < Y < +\infty\} = P\{a \leq X \leq b\}. \end{aligned}$$

■ **EXAMPLE 7.5**

A man shoots at a circular target. Assume that his skills are such that he is certain to hit the target. However, he is unable to aim with any more precision, so that the probability of hitting a particular part of the target is proportional to the area of this part. Let X and Y be the horizontal and vertical distances from the center of the target to the point of impact. We want to find the distribution of X. SOLUTION. Without a loss of generality, we can introduce the coordinate system with the origin in the center of the target. Let us assume that its radius is R. Then the target is described as $x^2 + y^2 \leq R^2$. From the conditions of the problem it is seen that the density of the point of impact is constant on the target so that

$$f(x, y) = \begin{cases} c & \text{if } x^2 + y^2 \leq R^2 \\ 0 & \text{otherwise.} \end{cases}$$

Next the condition $\iint_{R^2} f(x, y)\, dx\, dy = 1$ implies that c must be the reciprocal of the area of the target; hence $c = 1/\pi R^2$. If $|x| > R$, then $f(x, y) = 0$ for all y (see Figure 7.4). For $|x| \leq R$, we have $f(x, y) = 0$ when $x^2 + y^2 > R^2$ (hence if $|y| > \sqrt{R^2 - x^2}$), and $f(x, y) = 1/\pi R^2$ when $x^2 + y^2 \leq R^2$ (so $|y| \leq \sqrt{R^2 - x^2}$). Consequently, for $|x| \leq R$,

$$f_1(x) = \int_{-\sqrt{R^2 - x^2}}^{\sqrt{R^2 - x^2}} \frac{1}{\pi R^2}\, dy = \frac{2\sqrt{R^2 - x^2}}{\pi R^2}.$$

The density $f_1(x)$ has the shape given in Figure 7.4.

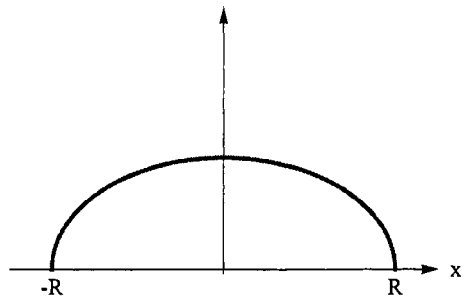


Figure 7.4 Marginal density

At this point we need to comment on notation. The symbols $f_1(x)$ and $f_2(y)$ for marginal densities of X and Y are used in the literature together with more readable symbols such as $f_X(x)$ and $f_Y(y)$, with the name of the variable appearing as a subscript. This latter system becomes cumbersome when the variables are labeled X_1 and X_2 instead of X and Y, since logically one should then use $f_{X_1}(x)$ and $f_{X_2}(y)$. We will deliberately avoid keeping rigidly to any fixed system of notation, and use whichever notation appears most natural in a given instance. Another problem concerns the use of an argument in the density or cdf. It appears natural to label the argument x in density or cdf of X, label it y in density of Y, and so on. However, it is *not possible* to use such notation consistently: Indeed, if $F(x)$ and $f(x)$ are to be used as the cdf and density of X, then we have $F(x) = \int_{-\infty}^{x} f(u)\,du$, and in the integrand we can use almost any symbol *except* x, since the symbol $\int_{-\infty}^{x} f(x)\,dx$ is unacceptably ambiguous.

We will now define the concept of independence of two random variables. We recall, from Chapter 4 that the independence of two events A and B is defined by the product rule $P(A \cap B) = P(A)\,P(B)$. A random variable X allows us to define events of the form $\{X \in A\}$, where A is some set of real numbers, and the same is true for the random variable Y. It seems natural to require that these events be independent for independent random variables X and Y. Thus we have

Definition 7.2.3 (Intention of the Concept) We say that random variables X and Y are *independent* if events $\{X \in A\}$ and $\{Y \in B\}$ are independent for all sets A, B, that is, if

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}. \tag{7.9}$$

□

This definition, as spelled out in its label, concerns the "final effect" of the concept: independence of *every pair* of events in a very large class of such pairs. Clearly, using this definition to check independence would be very difficult if we were to verify the product rule (7.9) for every pair A, B. Consequently, it becomes necessary to find a condition that is verifiable and strong enough to imply independence in the sense of Definition 7.2.3. It turns out that it is sufficient to require only a seemingly weaker condition.

Definition 7.2.4 (Verifiable Definition of Independence) The random variables X and Y are said to be *independent* if their joint and marginal cdf's satisfy the following condition: For every x, y we have

$$F(x, y) = F_1(x)F_2(y), \tag{7.10}$$

where $F_1(x)$ and $F_2(y)$ are marginal cdf's of X and Y, respectively. □

In the case of discrete random variables with p_{ij} being the probability $P\{X = x_i, Y = y_j\}$, condition (7.10) is implied by

$$p_{ij} = p_{i+}\, p_{+j} \qquad \text{for every } i, j. \tag{7.11}$$

For continuous random variables the condition implying (7.10) calls for marginal densities f_1 and f_2 to be such that

$$f(x, y) = f_1(x)\, f_2(y) \tag{7.12}$$

for all x, y, except possibly for a set of points (x, y) of measure zero.

The intuition behind the concept of independence of random variables (analogous to the intuition behind the concept of independence of events) is that the information about the value of one of them provides no information about the value of the other. Random variables for which independence conditions are not satisfied will be called dependent.

■ **EXAMPLE 7.6**

Let us go back to Example 7.1. In the table of joint and marginal distribution we had $P\{X = 3, Y = 2\} = 0$ while $P\{X = 3\} \times P\{Y = 2\} = \frac{1}{8} \times \frac{1}{4} \neq 0$. This means that X and Y are dependent random variables.

■ **EXAMPLE 7.7**

Let us consider again the case of three tosses of a coin. Let U be the number of heads in the first two tosses, and let V = be the number of tails in the last toss. The sample points are

S	U	V	S	U	V
HHH	2	0	HTT	1	1
HHT	2	1	THT	1	1
HTH	1	0	TTH	0	0
THH	1	0	TTT	0	1

The joint and marginal probability distributions are

U/V	0	1	U
0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{2}$
2	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
V	$\frac{1}{2}$	$\frac{1}{2}$	1

A direct check shows that we have

$$P\{U = x, V = y\} = P\{U = x\} \times P\{V = y\}$$

for every cell (x, y) in the table above. This shows that U and V are independent.

Verification that two random variables are independent requires checking the multiplicative property (7.11) or (7.12) for *all* x and y. In the last example this required comparing joint probabilities with the products of probabilities in marginal distributions for all six cells in the joint distribution table. Such a direct verification is not feasible except for discrete variables with small sets of possible values. To handle more complicated cases, we typically must have some algebraic formula for the joint distribution from which we can calculate the marginal distributions and verify the product rule algebraically. On the other hand, to show that two variables are *dependent* (i.e., that they are *not independent*), it is enough to find *one* pair (x, y) for which the product rule does not hold.

A practical consequence here is that it is generally worthwhile to try to determine based on the meaning of the variables in question, whether or not we can expect them to be independent. This determines the strategy: Do we aim at checking that variables are independent, or do we aim at showing that they are dependent? These two goals may require somewhat different types of technique.

To illustrate the point, in Example 7.6 we could have expected the variables to be dependent: the more tails in the last two tosses (Y), the lower one can expect the total number of heads (X) to be. In particular, we may be able to find a "pure exclusion," such as the fact that one cannot have $X = 3$ and $Y = 2$ simultaneously, while separately each of these events has positive probability.

On the other hand, in Example 7.7 we could have expected U and V to be independent, since the values of U and V were determined by nonoverlapping sets of tosses. In particular, the following criterion is useful in showing that two variables are *not independent*.

Theorem 7.2.1 *If the table of joint probabilities for* (X, Y) *contains a zero entry, then* X *and* Y *are dependent.*

Proof. Suppose that $P\{X = x_0, Y = y_0\} = 0$. Since x_0 and y_0 are the possible values of X and Y, respectively, the row sum and column sum at (x_0, y_0) must be positive. Hence $P(X = x_0) \times P(Y = y_0)$ is positive and (7.11) is not satisfied. □

This criterion provides a quick "visual" test for lack of independence. Of course, it works only in one direction: if there are no zeros in the table of joint distribution, the variables may or may not be dependent, and further checking is necessary.

We will now consider some examples of applications of Definition 7.2.4 for the case of continuous random variables.

■ **EXAMPLE 7.8**

The joint density of random variables X and Y is

$$f(x,y) = \begin{cases} cx^n e^{-\alpha x - \beta y} & \text{for } x > 0, y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The question is whether these random variables are independent.

SOLUTION. It is interesting that the answer here does not require any calculations. In particular, we do not need to calculate c. We can simply represent the joint density on the positive quadrant as

$$f(x,y) = c_1 x^n e^{-\alpha x} \times c_2 e^{-\beta y} \tag{7.13}$$

with c_1 and c_2 such that

$$c_1 \int_0^\infty x^n e^{-\alpha x} dx = 1 \quad \text{and} \quad c_2 \int_0^\infty e^{-\beta y} dy = 1$$

(then automatically $c = c_1 c_2$). The two factors on the right of (7.13) must be marginal densities (why?) and we showed independence of X and Y. However, for the joint density $f(x,y)$ to factor into the product of two functions, each depending on one variable only, is not enough to warrant the independence of variables X and Y, since the marginal densities depend on the shape of the support of the joint density (i.e., the set of points where $f(x,y)$ is positive). To see it, consider the following example.

■ **EXAMPLE 7.9**

Let the joint density be

$$f(x,y) = \begin{cases} cxy^2 & \text{if } x \geq 0, y \geq 0, x + y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

At first glance the situation here is similar to that in Example 7.8, but it is not. The marginal distributions are *not* $c_1 x$ and $c_2 y^2$ for any c_1 and c_2. Indeed, (see Figure 7.2 for limits of integration):

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{+\infty} f(x,y)\, dy = \int_{-\infty}^{0} 0\, dy + \int_0^{1-x} cxy^2\, dy + \int_{1-x}^{\infty} 0\, dy \\ &= \frac{cx(1-x)^3}{3} \end{aligned}$$

for $0 < x < 1$, and $f_1(x) = 0$ for $x > 1$ and $x < 0$. Similarly $f_2(y) = 0$ if $y > 1$ or $y < 0$, and for $0 < y < 1$,

$$f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\, dx = \frac{Cy^2(1-y)^2}{2}.$$

Since $f(x,y) \neq f_1(x)f_2(y)$, the random variables are dependent.

We can now formulate

Theorem 7.2.2 *Assume that the joint density $f(x,y)$ is continuous, and let $f(x,y) > 0$ for $(x,y) \in A$. If A is not a Cartesian product $A = A_X \times A_Y$, then variables X and Y are dependent.*

Proof. The marginal densities f_1 and f_2 are strictly positive in the "shadows" A_X and A_Y of the set A on the x-axis and y-axis (see Figure 7.5). If A is not equal to the Cartesian product $A_X \times A_Y$, then there exists a point (x_0, y_0) with $x_0 \in A_X, y_0 \in A_Y$, and $(x_0, y_0) \notin A$. If $x_0 \in A_X$, then there exists y^* such that $(x_0, y^*) \in A$, and therefore $f(x_0, y^*) > 0$. Since f is continuous, it must be positive in some neighborhood of (x_0, y^*), and therefore $f_1(x_0) = \int f(x_0, y)\,dy > 0$. Similarly $f_2(y_0) > 0$. Consequently, $0 = f(x_0, y_0) \neq f_1(x_0)f_2(y_0) > 0$. The continuity of $f(x,y)$, and hence of $f_1(x)f_2(y)$, implies now that $f(x,y) \neq f_1(x)f_2(y)$ on a set of a positive probability. □

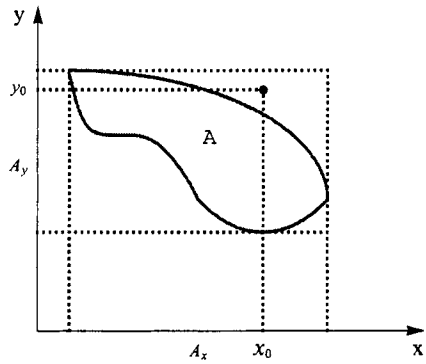


Figure 7.5 Condition for dependence

While the continuity of $f(x,y)$ is sufficient for the criterion given by Theorem 7.2.2, it is not necessary. However, it ought to be mentioned here that this criterion should be applied with some caution if f is not continuous, since $f(x,y)$ is defined only up sets of measure zero.

■ **EXAMPLE 7.10**

In Example 7.9 the set A is a triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$, so it can be inferred without any calculations that X and Y are dependent.

At the end of this section let us make the following important remark. The formulas (7.9) through (7.12) from Definitions 7.2.3 and 7.2.4 can be used in two ways.

The first is as illustrated so far: to determine whether or not two variables are independent. However, a more frequent use of these formulas is to find the joint distribution (joint density, joint cdf, etc.) of independent variables X and Y.

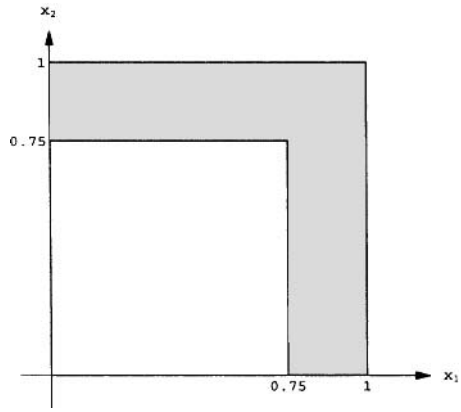


Figure 7.6 Probability of better of two attempts exceeding 0.75

■ **EXAMPLE 7.11**

A man makes two attempts at some goal. His performance X at the first attempt, measured on the scale from 0 to 1, is a random variable with density $f_1(x) = 12x^2(1-x)$. His performance Y at the second attempt is independent of X, and generally tends to be lower; its density is $f_2(y) = 6y(1-y), 0 \leq y \leq 1$. What is the probability that the man exceeds level 0.75 in the better of the two attempts?

SOLUTION. The joint density of (X, Y) is

$$f(x,y) = \begin{cases} 72x^2(1-x)y(1-y) & \text{for } 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The required probability is obtained as the integral of $f(x, y)$ over the shaded area in Figure 7.6 hence equals

$$1 - \int_0^{0.75} \int_0^{0.75} 72x_1^2(1-x_1)x_2(1-x_2)\,dx_1\,dx_2 = 0.3771.$$

PROBLEMS

7.2.1 The joint probability function of variables X and Y is $f(x, y) = c|x - y|$ for $x = 0, 1, 2, 3$, and $y = 0, 1, 2$. Find: **(i)** c. **(ii)** $P(X = Y)$. **(iii)** $P(X > Y)$. **(iv)** $P(|X - 1| \leq 1)$. **(v)** $P(X + Y \leq 3)$.

7.2.2 Two cards are drawn at random from the ordinary deck of cards. Let X be the number of aces and let Y be the number of hearts obtained. **(i)** Find the joint probability function of (X, Y). **(ii)** Find the marginal distribution of X and Y. **(iii)** Are X and Y independent?

7.2.3 An urn contains five balls, two of them red and three green. Three balls are drawn without replacement. Let X and Y denote the number of red (X) and green (Y) balls drawn. **(i)** Find the joint distribution of (X, Y). **(ii)** Find the marginal distributions of X and Y. **(iii)** Are X and Y independent? **(iv)** Find the joint distribution of X and $Z = Y - X$.

7.2.4 A box contains three coconut candies, five hazelnut chocolates, and two peanut butter chocolates. A sample of four sweets is chosen from the box. Let X, Y, and Z be the number of coconut candies, hazelnut chocolates, and peanut butter chocolates in the sample, respectively. **(i)** Find the joint distribution of (X, Y). **(ii)** Find the marginal distributions of X, Y, and Z. **(iii)** Are X and Y independent? Are X and Z independent?

7.2.5 Let X and Y have the joint distribution $P\{X = x, Y = y\} = c\lambda^{x+y}/(x!y!)$ for $x = 0, 1, \ldots, y = 0, 1, \ldots$, and $\lambda > 0$. **(i)** Find c. **(ii)** Find the marginal distribution of X. **(iii)** Are X and Y independent?

7.2.6 Random variables X and Y have joint distribution given by the following table:

X/Y	1	2	3
1	$\frac{1}{3}$	a	$\frac{1}{6}$
2	b	$\frac{1}{4}$	c

Show that X and Y are dependent, regardless of values a, b, and c.

7.2.7 Random variables have joint distribution given by the table

X/Y	1	2	3
1	a	$2a$	$3a$
2	b	c	d

Find a, b, c, d if X, Y are independent, and $P(X = 2) = 2P(X = 1)$.

7.2.8 Consider a system consisting of three components connected as in Figure 7.7. Let Y_1, Y_2, Y_3 be independent lifetimes of components 1, 2, and 3, respectively, each with EXP(α) distribution. If T is the lifetime of the whole system, find: **(i)** The cdf of T. **(ii)** The hazard function of T.

7.2.9 Let T_1, T_2 be independent random variables with hazard functions $h_1(t)$ and $h_2(t)$, respectively. **(i)** Show that the variable with the hazard function $h(t) = h_1(t) + h_2(t)$ has the same distribution as $\min(T_1, T_2)$. **(ii)** Express $P(T_1 < T_2)$ through h_1 and h_2.

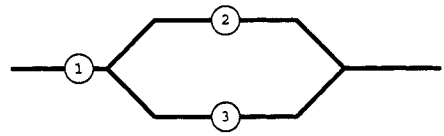


Figure 7.7 Three-component system

7.2.10 Let (X, Y) have the distribution given by the table

X/Y	3	4	5	6
1	$\frac{1}{24}$	$\frac{1}{2}$	$\frac{1}{24}$	$\frac{1}{24}$
2	$\frac{1}{24}$	0	$\frac{1}{12}$	$\frac{1}{12}$
3	0	$\frac{1}{12}$	$\frac{1}{12}$	0

Find the probability distribution of independent random variables (X', Y') such that X' and Y' have the same marginals as X and Y.

7.2.11 Let X and Y be the lifetimes of two components of a machine. Their joint distribution is given by the density $f(x, y) = xe^{-x(1+y)}$ for $x \geq 0$, $y \geq 0$ and zero otherwise. **(i)** Find $P(X \geq 5)$. **(ii)** Find the probability that $\max(X, Y) > 2$. **(iii)** Check the independence of X and Y using their marginal densities.

7.2.12 Random variables X and Y have joint density

$$f(x, y) = \begin{cases} k(ax + by) & 0 < x < 1, 0 < y < 2 \\ 0 & \text{otherwise,} \end{cases}$$

where $a > 0, b > 0$. Find: **(i)** k (as a function of a and b). **(ii)** The marginal distributions of X and Y. **(iii)** The cdf of (X, Y).

7.2.13 Assume that in shooting in a target, the coordinates (X, Y) of the point of impact are independent random variables, each with a N(0, σ^2) distribution. Find the density of D, the distance of the point of impact from the center of the target.

7.2.14 An ecologist has to randomly select a point inside a circular region with radius R. She first samples the direction from the center of the region according to a uniform distribution on $[0^o, 360^o]$, and then samples the distance from the center according to U[0, R]. Find: **(i)** The density $f(x, y)$ of the chosen points in (x, y) coordinates. **(ii)** The marginal distribution of X.

7.2.15 Assume that X and Y are independent random variables with EXP(a) and EXP(b) distributions, respectively. Assume that it is not possible to observe both X and Y but that one can observe

$$U = \min(X, Y) \text{ and } Z = \begin{cases} 1 & \text{if } X < Y \\ 0 & \text{otherwise.} \end{cases}$$

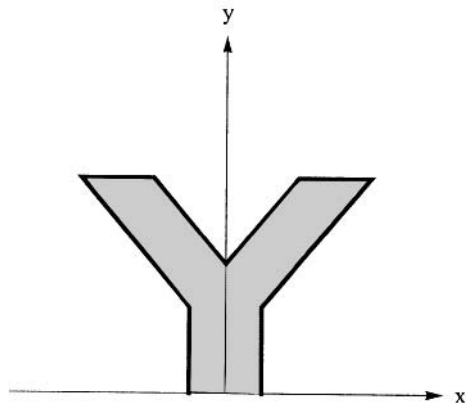


Figure 7.8 Joint distribution of X and Y

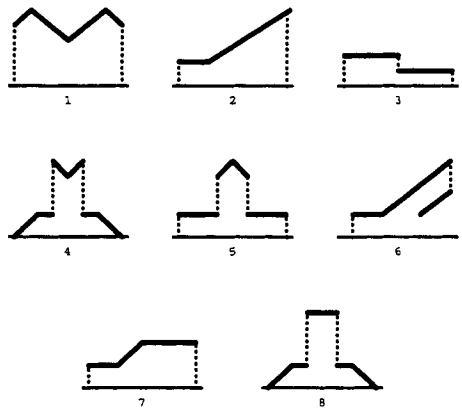


Figure 7.9 Options for marginal densities of X and Y

This kind of situation, called *censoring*, occurs often in medicine and engineering. For instance, X and Y may be the times of death due to heart failure (X) and death due to other causes (Y). Then U is the observed lifetime, and Z is the indicator of the cause of death. **(i)** Find the joint distribution of (U, Z). **(ii)** Prove that U and Z are independent. (*Hint:* Note that U is continuous, Z is discrete. It suffices to show that $P\{U \leq u | Z = i\}$ is independent of i, $i = 0, 1$.)

7.2.16 Let X, Y be independent, continuous random variables with a symmetric (but possibly different) distribution around 0. Show that Y/X and $Y/|X|$ have the same distribution. (*Hint:* Compare the cdf's of $W = X/Y$ and $V = X/|Y|$.)

7.2.17 Let X and Y have distribution uniform in the shape of the letter Y (see Figure 7.8). Identify the shapes of the marginal densities of X and Y in Figure 7.9.

7.3 CONDITIONAL DISTRIBUTIONS

A natural question concerning two random variables is how to handle situations where the information about the value of one of the variables affects the distribution of the other.

In symbols, the objective now is to determine the probabilities of the form

$$P(X \in Q_1 | Y = y) \text{ or } P(Y \in Q_2 | X = x) \tag{7.14}$$

for various Q_1, Q_2, x, and y.

We begin with the case of discrete random variables.

Let $V = \{y_1, y_2, \ldots\}$ be the set of all possible values of Y. Then the event $\{Y = y\}$ appearing as the condition in (7.14) has positive probability only if $y \in V$; otherwise, $P\{Y = y\} = 0$. Here we can use the theory developed in Chapter 4, where we defined the conditional probability $P(A|B)$ of event A given that the event B occurred as $P(A|B) = P(A \cap B)/P(B)$, provided that $P(B) > 0$. If $P(B) = 0$, the probability $P(A|B)$ was left undefined. It is clear that to evaluate (7.14) in the case of discrete random variables, we do not need any new concepts. If the condition has probability zero (i.e., if $P\{Y = y\} = 0$), we leave $P\{X \in Q | Y = y\}$ undefined. For $y_j \in V$, we have $P\{Y = y_j\} > 0$, and

$$P(X \in Q | Y = y_j) = \frac{P(X \in Q, Y = y_j)}{P(Y = y_j)}. \tag{7.15}$$

Let $U = \{x_1, x_2, \ldots\}$ be the set of possible values of X. If we write $p(x_i, y_j)$ or p_{ij} for $P\{X = x_i, Y = y_j\}$, the denominator in (7.15) is simply the marginal probability $P\{Y = y_j\} = \sum_i p_{ij} = p_{+j}$, and therefore

$$P\{X \in Q_1 | Y = y_j\} = \frac{\sum_{i: x_i \in Q_1} p_{ij}}{p_{+j}}.$$

An analogous formula with the role of X and Y interchanged is

$$P\{Y \in Q_2 | X = x_i\} = \frac{\sum_{j: y_j \in Q_2} p_{ij}}{p_{i+}}.$$

■ EXAMPLE 7.12

Let the experiment consist of two tosses of a die. We want to find conditional probabilities of the result of the first toss given the absolute difference of two tosses, and conversely, the conditional probabilities for the absolute difference given the results of the first toss. Accordingly, we let X_1 and X_2 denote the result of the first and the second toss, respectively, and put $Z = |X_1 - X_2|$.

The sample space $\mathcal{S}$ is naturally represented by the cells of the following table (where the values of Z are written in the cells):

X_1/X_2	1	2	3	4	5	6
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

The joint distribution of X_1 and Z is obtained by simple counting:

X_1/Z	0	1	2	3	4	5	X_1
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
3	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
4	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	0	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
Z	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	1

We can now answer various questions concerning conditional probabilities. For instance,

$$P(X_1 \text{ odd} \,|\, Z = 5) = \frac{P\{X_1 = 1, Z = 5\}}{P\{Z = 5\}} = \frac{\frac{1}{36}}{\frac{2}{36}} = \frac{1}{2},$$

while

$$P(Z > 2 | X_1 = 5) = \frac{P\{Z = 3, X_1 = 5\}}{P\{X_1 = 5\}} + \frac{P\{Z = 4, X_1 = 5\}}{P\{X_1 = 5\}} = \frac{1}{3}.$$

In the case of continuous random variables, determining a quantity such as $P\{X \in A | Y = y\}$ cannot rely upon the concepts of Chapter 4, since the conditioning event $\{Y = y\}$ has probability zero. We therefore *define* the conditional density and then verify that it has all the necessary properties.

Definition 7.3.1 The conditional densities g_{12} and g_{21} are defined by

$$g_{12}(x|y) = \frac{f(x, y)}{f_2(y)} \tag{7.16}$$

provided that $f_2(y) > 0$, and

$$g_{21}(y|x) = \frac{f(x, y)}{f_1(y)} \tag{7.17}$$

provided that $f_1(x) > 0$, where $f(x, y)$ is the joint density of (X, Y) and f_1 and f_2 are the marginal densities of X and Y, respectively. □

If the denominators in (7.16) or (7.17) are zero, the left-hand sides remain undefined. To check that formulas (7.16) and (7.17) define densities, observe first that both functions, which are regarded as functions of the *first* variable (i.e., x in $g_{12}(x|y)$ and y in $g_{21}(y|x)$), are nonnegative. Moreover, we have

$$\int_{-\infty}^{+\infty} g_{12}(x|y)\,dx = \int_{-\infty}^{+\infty} \frac{f(x,y)}{f_2(y)}dx = \frac{1}{f_2(y)}\int_{-\infty}^{+\infty} f(x,y)\,dx = \frac{f_2(y)}{f_2(y)} = 1$$

and similarly for g_{21}. To justify the definition on "semantic" grounds, assume for simplicity that $f(x,y)$, and hence also $f_2(y)$, are continuous, and let us consider, for some $h > 0$, the probability $P\{X \in A | y \le Y \le y+h\}$. For small h we have

$$\begin{aligned} P\{X \in A | y \le Y \le y+h\} &= \frac{P\{X \in A, y \le Y \le y+h\}}{P\{y \le Y \le y+h\}} \\ &= \frac{\int_A \int_y^{y+h} f(x,u)\,du\,dx}{\int_y^{y+h} f_2(u)\,du} \approx \frac{\int_A f(x,y)\,h\,dx}{f_2(y)\,h} \\ &= \int_A g_{12}(x|y)\,dx. \end{aligned}$$

This shows that $g_{12}(x|y)$ is a well-defined density of X given that $Y = y$.

■ **EXAMPLE 7.13**

Let the joint density of (X, Y) be given by

$$f(x,y) = \begin{cases} cxy^2 & \text{if } 0 \le x \le 1,\ 0 \le y \le 1,\ x+y \ge 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus the density is positive on the triangle with the vertices $(1,1)$, $(1,0)$, and $(0,1)$; hence X and Y are dependent (by the criterion for dependence of continuous random variables from Section 7.2). The marginal densities are as follows (we do not need to determine the numerical value of c; it will cancel in densities g_{12} and g_{21}): If $0 < x < 1$, then

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{+\infty} f(x,y)\,dy = \int_{1-x}^{1} cxy^2\,dy = cx\frac{y^3}{3}\Big|_{1-x}^{1} \\ &= \frac{c}{3}(3x^2 - 3x^3 + x^4). \end{aligned}$$

Similarly, for $0 < y < 1$,

$$f_2(y) = \int_{1-y}^{1} cxy^2\,dx = cy^2\frac{x^2}{2}\Big|_{1-y}^{1} = \frac{c}{2}y^3(2-y).$$

Thus

$$g_{12}(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{cxy^2}{(c/2)y^3(2-y)} = \frac{2x}{y(2-y)};$$

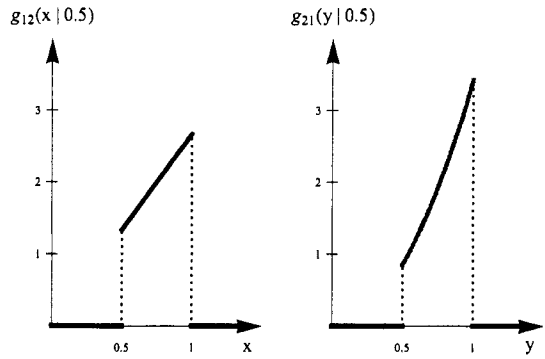


Figure 7.10 Conditional densities.

hence

$$g_{12}(x|y) = \begin{cases} \frac{2x}{2y-y^2} & \text{for } 1-y < x < 1,\ 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

For instance, if $Y = 0.5$, then for $0.5 < x < 1$,

$$g_{12}(x\,|0.5) = g_{12}(x\,|Y = 0.5) = \frac{8}{3}x.$$

Similarly, for $1 - x < y < 1$,

$$g_{12}(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{xy^2}{(c/3)(3x^2 - 3x^3 + x^4)} = \frac{3y^2}{x(3 - 3x + x^2)}$$

and equals 0 otherwise. Thus, if $X = 0.5$, for $0.5 < y < 1$ the density $g_{21}(y|0.5) = 24y^2/7$ (see Figure 7.10).

In Chapter 4 we repeatedly stressed the fact that the formula $P(A|B) = P(A \cap B)/P(B)$ can be used to determine the probability $P(A \cap B)$ of two events occurring jointly, given the probability of one of those events and the appropriate conditional probability, such as $P(B)$ and $P(A|B)$. The same situation is true in the case of conditional distributions in a continuous, discrete, or mixed case; given one marginal distribution and one conditional distribution, one can determine the joint distribution and hence also the marginal distribution of the other variable. We will illustrate this technique by three examples.

■ **EXAMPLE 7.14**

An animal lays a certain number X of eggs, where X is random and has the Poisson distribution

$$P\{X = n\} = \frac{\lambda^n}{n!}e^{-\lambda}, \quad n = 0, 1, 2, \ldots$$

(see Example 6.10). Each egg hatches with a probability p, independent of the hatching of other eggs. Determine the distribution of Y = the number of eggs that hatch.

SOLUTION. Here the randomness of Y has two sources: first, the number X of eggs laid is random (varies from animal to animal), and second, even for animals that laid the same number of eggs, the randomness in the process of hatching may make the numbers of offspring different.

Our solution strategy is as follows: We are given the distribution of X. The two assumptions of the problem will allow us to determine the conditional distribution of Y given X. These assumptions allow us to determine the joint distribution of (X, Y), and the distribution of Y will be obtained as marginal from the joint distribution of (X, Y).

We now need to determine the conditional distribution of Y given X, that is, $P\{Y = j|X = n\}$. First, it is clear that $0 \leq Y \leq X$ (the number of eggs that hatch is nonnegative and cannot exceed the number of eggs laid). We assume that eggs hatch with the same probability and independently. Thus Y must have binomial distribution if we regard the process of incubation as an "experiment over an egg" with "success" identified with hatching. Therefore,

$$P\{Y = j|X = n\} = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, 1, \ldots, n.$$

By formula (7.15) the joint distribution of (X, Y) is

$$\begin{aligned} P\{X = n, Y = j\} &= P\{Y = j|X = n\} \times P\{X = n\} \\ &= \binom{n}{j} p^j (1-p)^{n-j} \times \frac{\lambda^n}{n!} e^{-\lambda}, \end{aligned}$$

where $n = 0, 1, 2, \ldots$ and $j = 0, 1, \ldots, n$ (for all other values of n and j, the joint probability is zero).

We will next find the marginal probability $P\{Y = j\}$ for $j = 0, 1, 2, \ldots$. Clearly, for $X \geq Y$, we have $n \geq j$. Hence

$$\begin{aligned} P\{Y = j\} &= \sum_{n=j}^{\infty} P\{X = n, Y = j\} = \sum_{n=j}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} \binom{n}{j} p^j (1-p)^{n-j} \\ &= \sum_{n=j}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j} \\ &= \frac{(\lambda p)^j}{j!} e^{-\lambda} \sum_{n=j}^{\infty} \frac{[\lambda(1-p)]^{n-j}}{(n-j)!} \\ &= \frac{(\lambda p)^j}{j!} e^{-\lambda} \sum_{\nu=0}^{\infty} \frac{[\lambda(1-p)]^{\nu}}{\nu!} = \frac{(\lambda p)^j}{j!} e^{-\lambda} e^{\lambda(1-p)} \\ &= \frac{(\lambda p)^j}{j!} e^{-\lambda p}. \end{aligned}$$

The marginal distribution of Y is again Poisson, except that the parameter has changed from λ to λp.

The process that leads from X to Y in this example is sometimes called *binomial thinning*. One can visualize it, in general, as a random process that gives the value of X (by assumption, an integer), with some distribution. We can think of X as the number of objects of some kind that are produced, a number of elements sampled, a number of events of some kind that occur, and so on. The process of thinning causes some of the X objects (events, etc.) to disappear (not be counted), due to a certain random process of selection. For instance, some of the X objects produced are defective, some of the X elements sampled are not acceptable or have certain other characteristic, some of the events are unobservable, and so on. Such process of elimination of some X's and the selection of others is called *thinning*. We say that the process of thinning is binomial if the inclusions of X-elements as Y-elements are independent and occur with the same probability. The present example shows that binomial thinning of Poisson random variables leads again to Poisson random variables, with appropriately modified parameters.

We will now consider similar situation for continuous distributions.

■ **EXAMPLE 7.15**

A point X is chosen at random from the interval $[A, B]$ according to the uniform distribution (see Example 6.15), and then a point Y is chosen at random, again with uniform distribution, from the interval $[X, B]$. Find the marginal distribution of Y.

SOLUTION. Since

$$f_1(x) = \begin{cases} \frac{1}{B-A} & \text{for } A \le x \le B \\ 0 & \text{otherwise} \end{cases}$$

and

$$g_{21}(y|x) = \begin{cases} \frac{1}{B-x} & \text{for } x \le y \le B \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$f(x,y) = f_1(x)\, g_{21}(y|x) = \begin{cases} \frac{1}{B-A} \times \frac{1}{B-x} & \text{for } A \le x \le y \le B \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, for fixed y $(A \le y \le B)$,

$$f_2(y) = \int_A^y \frac{1}{B-A} \times \frac{1}{B-x} dx = \frac{1}{B-A} \log \frac{B-A}{B-y}.$$

The values close to B are more likely than those close to A; the density of Y is in fact unbounded as we approach the upper boundary B of the range of Y.

■ **EXAMPLE 7.16**

Finally, we will consider the case of a mixed distribution, with X being continuous and Y being discrete.

Assume that Y takes one of $n \geq 2$ integer values with $P\{Y = i\} = p_i$ and $p_1 + p_2 + \cdots + p_n = 1$. In the simplest case, $n = 2$, we have

$$p_1 = P\{Y = 1\} = p, \; p_2 = P\{Y = 2\} = 1 - p.$$

Next assume that for a given $Y = i$, the random variable X has a continuous distribution with density $\varphi_i(x)$. Thus we have

$$P\{a \leq X \leq b, Y = i\} = p_i \int_a^b \varphi_i(x)\, dx.$$

To find the marginal distribution of X, we can use the formula for total probability (4.9) from Chapter 4, with events $\{Y = i\}$ as the partition. Then

$$\begin{aligned} P\{a \leq X \leq b\} &= \sum_{i=1}^{n} P\{a \leq X \leq b | Y = i\} \times P\{Y = i\} \\ &= \sum_{i=1}^{n} p_i \int_a^b \varphi_i(x)\, dx = \int_a^b \sum_{i=1}^{n} p_i \varphi_i(x)\, dx. \end{aligned} \tag{7.18}$$

It follows that X is a continuous random variable with the density

$$f_X(x) = \sum_{i=1}^{n} p_i \varphi_i(x),$$

called the *mixture* of densities φ_i with *mixing coefficients* p_i.

If we have only two values of Y, and therefore only two densities φ_1 and φ_2 of X, then

$$f_X(x) = p\varphi_1(x) + (1 - p)\varphi_2(x).$$

The formulas above remain valid if Y assumes one of infinitely many values $1, 2, \ldots$, with probabilities $p_1, p_2, \ldots$ such that we have $\sum p_i = 1$. The only potential source of trouble is the interchange of integration and summation in the last step in (7.18). But this interchange is permissible because the terms of the sum are all nonnegative.

PROBLEMS

7.3.1 Suppose that three cards are drawn without replacement from an ordinary deck. Let X be the number of aces among the cards drawn and Y be the number of red cards among them. Find: **(i)** The joint distribution of (X, Y). **(ii)** The conditional distribution of the number of aces if it is known that all three cards selected are red.

7.3.2 Let X and Y have the joint density $f(x, y) = \lambda^2 e^{-\lambda y}$ for $0 \leq x \leq y$ and $f(x, y) = 0$ otherwise. Find: **(i)** The joint cdf of (X, Y). **(ii)** The marginal densities of X and Y. **(iii)** The conditional density of Y given X.

7.3.3 Two parts of a document are typed by two typists. Let X and Y be the numbers of typing errors in the two parts of the paper. Assuming that X and Y are

independent and have Poisson distributions with parameters λ_1 and λ_2, respectively, find the probability that: **(i)** The paper (i.e., two combined parts) has at least two typing errors. **(ii)** The total number of typing errors is m. **(iii)** The first part of the paper has k typing errors given that there are n typing errors altogether.

7.3.4 Let variables X and Y be independent, each with U[0, 1] distribution. Find: **(i)** $P(X + Y \leq 0.5 \mid X = 0.25)$. **(ii)** $P(X + Y \leq 0.5 \mid X \geq 0.25)$. **(iii)** $P(X \geq Y \mid Y \geq 0.5)$.

7.3.5 Let X and Y have joint density of the form

$$f(x, y) = \begin{cases} A(y - x)^{\alpha} & \text{for } 0 \leq x < y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find: **(i)** The values of α such that f can be a density function. **(ii)** The value of A for α specified in part (i). **(iii)** The marginal densities of X and Y. **(iv)** The conditional densities $g_{12}(x|y)$ and $g_{21}(y|x)$.

7.3.6 The phrase "A stick is broken at random into three pieces" can be interpreted in several ways. Let us identify the stick with interval $[0, 1]$ and let $0 < X < Y < 1$ be the breaking points. Some of the possible ways of generating X, Y are as follows: **(i)** A point (U, V) is chosen from the unit square with uniform distribution, and we let $X = \min(U, V)$ and $Y = \max(U, V)$. **(ii)** The point U is chosen from $[0, 1]$ with uniform distribution. If $U < 1/2$, then V is chosen with uniform distribution from $[U, 1]$, whereas if $U \geq 1/2$, then V is chosen with uniform distribution on $[0, U]$. Then (X, Y) are defined as in (i). **(iii)** X is chosen from $[0, 1]$ according to the uniform distribution, and then Y is chosen with uniform distribution on $[X, 1]$. **(iv)** U is chosen with uniform distribution on $[0, 1]$. Next, one of the intervals $[0, U]$ or $[U, 1]$ is chosen at random, with probability U and $1 - U$, respectively. Then V is chosen with uniform distribution from the chosen interval, and again, $X = \min(U, V)$ and $Y = \max(U, V)$.

In each of the cases (i)–(iv) find the joint density of (X, Y) and the marginal densities of X and Y. Which of the ways (i)–(iv) are equivalent?

7.3.7 A fast-food restaurant has a dining room and a drive-thru window. Let X and Y be the fractions of time (during a working day) when the dining room (X) and the drive-thru window (Y) are busy. The joint density of (X, Y) is $f(x, y) = k(2x^2 + y^2)$ for $0 \leq x \leq 1, 0 \leq y \leq 1$ and $f(x, y) = 0$ otherwise. Find: **(i)** k. **(ii)** The marginal densities of X and Y. **(iii)** The probability that the drive-thru window will be busy more than 75% of the time on a day when the dining room is empty less than 10% of the time. **(iv)** Do you find anything disturbing in this problem? If so, explain.

7.4 BIVARIATE TRANSFORMATIONS

In Chapter 6 we considered transformations of single random variables. Here we will consider functions of two continuous random variables X and Y with the joint density $f(x, y)$ with support C (so that $f(x, y) > 0$ on C and $f(x, y) = 0$ outside

C). The objective is to find the density of $Z = \varphi(X, Y)$, where φ is a differentiable function of two real arguments. By far the simplest here is the cdf technique, introduced in Section 6.4 for the case of a single variable. It may also be applied in the multivariate case if we can obtain $P\{Z \leq z\} = P\{\varphi(X, Y) \leq z\}$ in a closed form as a function of z. Density can then be obtained by differentiation. This methods works especially well for $\varphi(X, Y) = \max(X, Y)$ and $\varphi(X, Y) = \min(X, Y)$, when X and Y are independent. For example, $P\{\max(X, Y) \leq z)\} = P\{X \leq z, Y \leq z\} = F_X(z)F_Y(z)$; hence the density of $Z = \max(X, Y)$ is $f_X(z)F_Y(z) + F_X(z)f_Y(z)$.

Now we will present a technique that may be applied to a wider class of cases. It will be given in the form of an algorithm, and its use will be illustrated by several of examples. A formal proof will not be given, since the algorithm is in fact based on a change of variables in two-dimensional integrals, which can be found in advanced-level calculus texts.

Determination of densities of bivariate (and multivariate) transformations is typically regarded by students as challenging. We hope that by presenting it as a purely mechanical procedure—which it largely is—we will alleviate, or perhaps eliminate, the terror. It is true that the procedure requires attention and some level of algebraic skills, but the difficulties are closer to those of proofreading a telephone directory then to those of playing a game of chess.

The algorithm is as follows:

1. Choose a "companion" function, say $w = \eta(x, y)$, such that the pair of equations

$$z = \varphi(x, y) \quad \text{and} \quad w = \eta(x, y) \tag{7.19}$$

can be solved, leading to

$$x = \alpha(z, w) \quad \text{and} \quad y = \beta(z, w). \tag{7.20}$$

2. Determine the image of the support C of density $f(x, y)$ in the (z, w) plane under transformation (7.19). Let this image be D.

3. Find the Jacobian of transformation (7.19), that is, the determinant

$$J = \begin{vmatrix} \frac{\partial \alpha}{\partial z} & \frac{\partial \alpha}{\partial w} \\ \frac{\partial \beta}{\partial z} & \frac{\partial \beta}{\partial w} \end{vmatrix}. \tag{7.21}$$

4. Determine the joint density $g(z, w)$, of random variables $Z = \varphi(X, Y)$, $W = \eta(X, Y)$, given by the formula

$$g(z, w) = \begin{cases} f(\alpha(z, w), \beta(z, w))|J| & \text{for } (z, w) \in D \\ 0 & \text{for } (z, w) \notin D. \end{cases}$$

5. Compute the density of $Z = \varphi(X, Y)$ as the marginal density of the joint density $g(z, w)$:

$$g_Z(z) = \int_{-\infty}^{+\infty} g(z, w)\, dw = \int_{D_z} f(\alpha(z, w), \beta(z, w))|J|\, dw,$$

where $D_z = \{w : (z, w) \in D\}$.

Out of these five steps, only step 1 requires some moderate amount of thinking (or, at least some experience). The reason is that the choice of the companion transformation η affects all subsequent steps, making the calculations easy, difficult, or perhaps even impossible.

■ EXAMPLE 7.17 Sum of Random Variables

The operation of addition of random variables appears so often that it is worthwhile to derive general formulas here. For $z = \varphi(x, y) = x + y$ we have a possible choice of a companion transformation, $w = \eta(x, y) = x$. Hence the inverse transformation (7.20) is

$$x = w \quad \text{and} \quad y = z - w,$$

so that $\alpha(z, w) = w$, and $\beta(z, w) = z - w$. Thus

$$J = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix}$$

and $|J| = 1$. If (X, Y) have joint density $f(x, y)$, then the joint density of (Z, W) is $f(w, z - w)$. The density of Z is

$$g_Z(z) = \int_{-\infty}^{+\infty} f(w, z - w)\, dw. \tag{7.22}$$

Now we have only to determine the effective limits of integration, that is, the set of values w (for given z) at which the integrand is positive.

■ EXAMPLE 7.18 Sum of Exponential Random Variables

When X and Y are independent, and both have exponential distributions with the same parameter λ, their joint density is

$$f(x, y) = \begin{cases} \alpha^2 e^{-\alpha(x+y)} & \text{if } x \geq 0,\ y \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7.23}$$

Consequently, the density $g_Z(z)$, as given by (7.22), is

$$g_Z(z) = \int_{-\infty}^{+\infty} f(w, z - w)\, dw = \int_{D_z} \alpha^2 e^{-\alpha z}\, dw$$

where D_z is the set $\{w : w \geq 0, z - w \geq 0\}$, since both arguments must be positive for f, given by (7.23), to be positive. Since we must then have $z > 0$ and $0 \leq w \leq z$,

$$g_Z(z) = \alpha^2 e^{-\alpha z} \int_0^z dw = \alpha^2 z e^{-\alpha z}, z \geq 0 \tag{7.24}$$

and $g_Z(z) = 0$ for $z < 0$. We recognize the density (7.24) as a gamma density with parameters 2 and α (see Example 6.28).

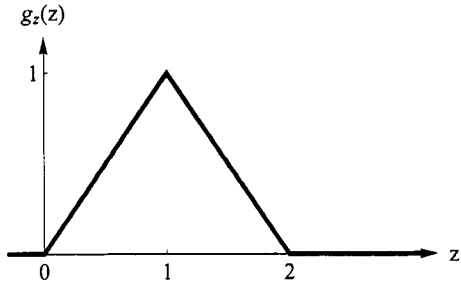


Figure 7.11 Triangular density

■ EXAMPLE 7.19 Sum of Two Uniform Random Variables

Suppose
now that two independent variables each have U(0, 1) distribution. Their joint density is then

$$f(x,y) = \begin{cases} 1 & \text{if } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Formula (7.22) gives the density of variable $Z = X + Y$ as

$$g_Z(z) = \int_{D_z} 1 dw, \tag{7.25}$$

where $D_z = \{w : 0 < w < 1, 0 \leq z - w \leq 1\}$. The pair of inequalities defining D_z, namely $0 < w < 1$ and $z - 1 \leq w \leq z$, can be written as

$$\max(0, z-1) \leq w \leq \min(z, 1),$$

and $0 \leq Z \leq 2$, since $0 \leq X \leq 1$ and $0 \leq Y \leq 1$. For $0 \leq z \leq 1$, the set D_z is the interval $0 \leq w \leq z$, while for $1 \leq z \leq 2$ we have the interval $z - 1 \leq w \leq 1$. Consequently (7.25) gives (see Figure 7.11)

$$g_Z(z) = \begin{cases} z & \text{for } 0 \leq z \leq 1 \\ 2 - z & \text{for } 1 \leq z \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

This distribution is called *triangular*.

■ EXAMPLE 7.20

Formula (7.22) was obtained by taking a specific companion transformation $w = \eta(x, y) = x$. There is no compelling reason for this choice. We could, of course, have chosen $w = \eta(x, y) = y$, getting a very similar formula. However, we could also have chosen something more fancy, for example, $w = \eta(x, y) = x/(x + 5y)$.

The system of equations

$$z = x + y, \qquad w = \frac{x}{x + 5y},$$

has the solution

$$x = \frac{5wz}{1 + 4w}, \qquad y = z\frac{1 - w}{1 + 4w}.$$

Consequently, the Jacobian equals

$$J = \begin{vmatrix} \frac{5w}{1+4w} & \frac{5z}{(1+4w)^2} \\ \frac{1-w}{1+4w} & \frac{-5z}{(1+4w)^2} \end{vmatrix} = -\frac{5z}{(1 + 4w)^2}.$$

Thus the density of the sum $Z = X + Y$ can be obtained as the integral

$$g_Z(z) = \int_{-\infty}^{+\infty} f\left(\frac{5wz}{1 + 4w}, z\frac{1 - w}{1 + 4w}\right) \left|\frac{5z}{(1 + 4w)^2}\right| dw, \tag{7.26}$$

where again the effective range of integration depends on the support C of the joint density f.

This example is given here to show that there is no such thing as "the" formula for density of sum of random variables: Once the integration is carried out, (7.22) and (7.26) will both lead to the same final form of $g_Z(z)$. We cannot even say that (7.22) is simpler than (7.26), since the simplicity of integration depends here on the form of f and of its support C.

■ EXAMPLE 7.21 Product of Two Random Variables

We have $\varphi(x, y) = xy$, and we want to find the density of the random variable $Z = XY$.

Let us again choose the companion function $\eta(x, y) = x$ so that the system of equations (7.19) is $z = xy, w = x$, and its solution (7.20) is

$$x = \alpha(z, w) = w, \qquad y = \beta(z, w) = \frac{z}{w}.$$

The Jacobian of this transformation is

$$J = \begin{vmatrix} 0 & 1 \\ \frac{1}{w} & -\frac{z}{w^2} \end{vmatrix} = -\frac{1}{w};$$

hence $|J| = 1/|w|$. The joint density of (Z, W) is now

$$f\left(w, \frac{z}{w}\right) \frac{1}{|w|};$$

and the density of Z is given by

$$g_Z(z) = \int_{-\infty}^{+\infty} f\left(w, \frac{z}{w}\right) \frac{1}{|w|} dw.$$

Again, the effective limits of integration depend on the support C of density $f(x, y)$ and, consequently, the sets D_z.

We will now give an example that provides an algorithm of generating random variables with normal distribution.

Theorem 7.4.1 *If X and Y are independent, uniformly distributed on $(0, 1)$, then the random variables Z and W*

$$Z = \sqrt{-2\log X}\sin(2\pi Y), \qquad W = \sqrt{-2\log X}\cos(2\pi Y),$$

are independent, and each has the standard normal distribution.

Proof. Variables X and Y are independent, each with U(0, 1) distribution. Their joint density is

$$f(x, y) = \begin{cases} 1 & \text{if } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

For $z = \sqrt{-2\log x}\sin(2\pi y)$ and $w = \sqrt{-2\log x}\cos(2\pi y)$ we have $z^2 + w^2 = -2\log x$, and $x = e^{-(z^2+w^2)/2}$. On the other hand, $z/w = \tan(2\pi y)$, which gives

$$y = \frac{1}{2\pi}\arctan\left(\frac{z}{w}\right).$$

The Jacobian equals

$$J = \begin{vmatrix} -ze^{-(z^2+w^2)/2} & -we^{-(z^2+w^2)/2} \\ \frac{1/w}{2\pi(1+z^2/w^2)} & \frac{-z/w^2}{2\pi(1+z^2/w^2)} \end{vmatrix}, \tag{7.27}$$

which, after some algebra, reduces to

$$\frac{1}{\sqrt{2\pi}}e^{-z^2/2} \times \frac{1}{\sqrt{2\pi}}e^{-w^2/2}.$$

The unit square, equal to the support of $f(x, y)$, is mapped into the whole plane (z, w). It follows that the joint density of (Z, W), equal in this case to the Jacobian (7.27), is a product of two normal densities. □

As already mentioned, the conditional densities given a specific value of a random variable cannot be calculated according to the principles of Chapter 4, since the conditioning event has probability zero. As a warning we present an example, seemingly paradoxical, where by following the rules as explained in this chapter, one obtains two different answers to the same question.

■ **EXAMPLE 7.22 Borel-Kolmogorov Paradox**

Assume that X and Y have joint density

$$f(x, y) = 4xy, \quad 0 \le x \le 1,\ 0 \le y \le 1,$$

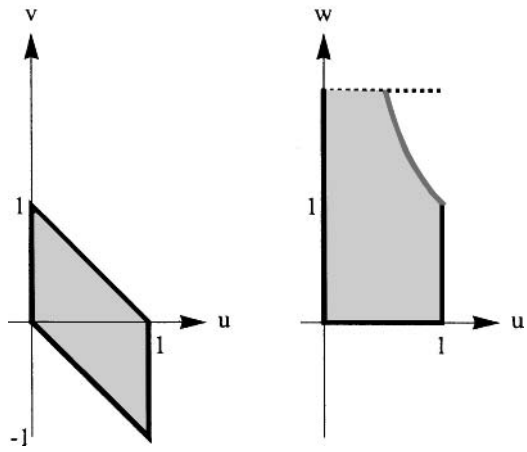


Figure 7.12 Supports of (U, V) and (U, W)

so X and Y are independent, with the same marginal densities

$$f_X(t) = f_Y(t) = 2t, \quad 0 \le t \le 1.$$

We will try to determine the conditional density of X given the event $X = Y$. As we will see, the answer will depend on the representation of event $X = Y$, as $Y - X = 0$ or as $Y/X = 1$.

SOLUTION 1. We introduce new variables, $U = X, V = Y - X$. Then our problem becomes equivalent to finding the density of U given $V = 0$. Thus we have to find the joint density $h(u, v)$ of (U, V) and determine the marginal density $h_V(v)$. Our answer will be

$$\varphi_X(x|X = Y) = \frac{h(u, v)}{h_V(v)}\bigg|_{v=0} = \frac{h(u, 0)}{h_V(0)}.$$

The transformation $u = x, v = y - x$, has the inverse $x = u, y = u + v$, so the Jacobian is

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1.$$

Consequently, the joint density of (U, V) is $h(u, v) = 4u(u + v)$ for $0 \le u \le 1$, $-u \le v \le 1 - u$ (see Figure 7.12). The marginal density of V is

$$h_V(v) = \int_0^{1-v} 4u(u + v)\, du = 4\left[\frac{(1 - v)^3}{3} + v\frac{(1 - v)^2}{2}\right] = \frac{2}{3}(1 - v)^2(2 + v),$$

for $0 \le v \le 1$, while for $-1 \le v \le 0$ we have

$$h_V(v) = \int_{-v}^{1} 4u(u + v)\, du = 4\left[\frac{1}{3} + \frac{v}{2} - \frac{v^3}{6}\right] = -\frac{2}{3}(v + 1)^2(v - 2).$$

Both formulas give $h_V(0) = 4/3$. Thus, as the conditional density of U at $V = 0$, we obtain

$$\frac{h(u,0)}{h_V(0)} = \frac{4u^2}{\frac{4}{3}} = 3u^2, \quad 0 \le u \le 1.$$

SOLUTION 2. We introduce the transformation $U = X, W = Y/X$. Then the conditional density of X, given $X = Y$, is the same as the conditional density of U, given $W = 1$.

The transformation $u = x, w = y/x$ has an inverse $x = u, y = uw$, and the Jacobian equals

$$J = \begin{vmatrix} 1 & 0 \\ w & u \end{vmatrix} = u.$$

Consequently, the joint density of (U, W) is $g(u,w) = 4u^3w$ for $0 \le u \le 1, 0 \le w \le 1/u$ (see Figure 7.12). The marginal density $g_W(w)$ is

$$g_W(w) = \int_0^1 4u^3w\,du = w \quad \text{for} \quad 0 \le w \le 1,$$

while for $w \ge 1$ we have

$$g_W(w) = \int_0^{1/w} 4u^3w\,du = \frac{1}{w^3}.$$

For $w = 1$ we have $g_W(1) = 1$, and therefore the conditional density equals

$$\frac{g(u,1)}{g_W(1)} = 4u^3, \quad 0 \le u \le 1.$$

Thus we obtained two different solutions, and we may ask: Which of them—if either—is correct?

The formal answer is that both solutions are correct, and we can choose, for the density of X given $X = Y$, any other function as well. This is simply because the conditional densities of a variable, given conditions with probability zero, can be modified arbitrarily on any set of conditions whose total probability is zero (in particular, on a single condition, or on finitely many conditions). This answer is, however, not quite satisfactory. We also want to understand why the techniques of transformation lead to different answers. To explain this phenomenon, let us assume that we need the density of X given $X = Y$, but we cannot observe exactly whether the conditions holds. Specifically, consider two situations: (1) we can observe the difference $X - Y$, within some small error ϵ; (2) we can observe the ratio the ratio Y/X, again within small error ϵ. In such cases the "natural" approach is to use limiting passage, with $\epsilon \to 0$. For any fixed ϵ the conditioning events $|X - Y| < \epsilon$ and $|Y/X - 1| < \epsilon$ are different (see Figure 7.13), and it is not surprising that the ratio Y/X favors larger values of u.

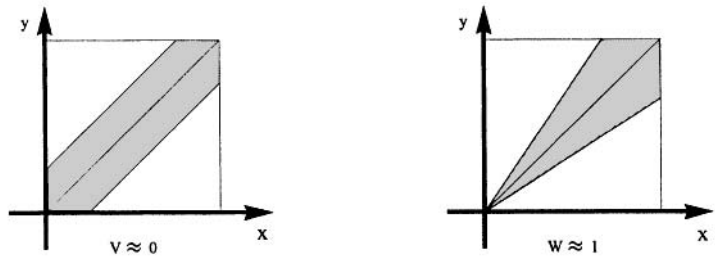


Figure 7.13 Approximations of two conditioning events

PROBLEMS

7.4.1 Let X, Y be independent, each with a standard normal distribution. Find the distribution of: **(i)** $V = X/Y$. **(ii)** $U = (X - Y)^2/2$. **(iii)** $W = X/|Y|$.

7.4.2 Let random variables X_1, X_2 be independent and have both EXP(λ) distribution. **(i)** Find the cdf and the density of $X_1 - X_2$. **(ii)** Find the cdf and the density of X_1/X_2. **(iii)** Show that $Z_1 = X_1 + X_2$ and $Z_2 = X_1/(X_1 + X_2)$ are independent.

7.4.3 Let X and Y be independent random variables with densities

$$f_X(x) = c_1 x^{\alpha-1} e^{-x}, \qquad f_Y(y) = c_2 y^{\beta-1} e^{-y}$$

for $x > 0, y > 0, \alpha > 0, \beta > 0$, and normalizing constants c_1, c_2. Find the density of $W = X/(X + Y)$.

7.4.4 Independent random variables X and Y are uniformly distributed over intervals $(-1, 1)$, and $(0, 1)$, respectively. Find the joint distribution of variables $U = XY$ and $V = Y$. Determine the support of the density function.

7.4.5 Random variables (X, Y) have a joint density

$$f(x, y) = \begin{cases} k(ax + by) & \text{if } 0 < x < 2,\ 0 < y < 2 \\ 0 & \text{otherwise,} \end{cases}$$

where $a > 0, b > 0$ are given constants. Find: **(i)** The value of k as a function of a and b. **(ii)** The density of random variable $Z = 1/(Y + 1)^2$. (*Hint:* Express $F(t) = P\{Z \leq t\}$ in terms of the cdf of Y.)

7.4.6 Random variables X and Y have joint density $f(x, y) = cx$ for $-x < y < x,\ 0 < x < 1$, and zero otherwise. Find the density and the cdf of $W = X - Y$.

7.4.7 Let f be the joint density of a pair (X, Y) of random variables, and let a and b be two constants. Find the densities of: **(i)** $aX + bY$. **(ii)** XY. **(iii)** X/Y. **(iv)** (U, V), where $U = aX + b, V = cY + d$, and $ac \neq 0$.

7.4.8 Let R be a nonnegative random variable with density $f(r)$. Let (X, Y) be a bivariate distribution obtained as follows: First randomly choose a value of R, and then chose a value of U according to its U(0, 1) distribution. Now put

$$X = R\cos(2\pi U), \qquad Y = R\sin(2\pi U). \tag{7.28}$$

Find: **(i)** The joint density of (X, Y). **(ii)** $P\{XY > 0\}$. **(iii)** $P\{X > 0\}$. **(iv)** $P\{X^2 + Y^2 \leq t\}$. [*Hint:* For (i) find the joint density of (R, U) first and then use transformation (7.28).]

7.4.9 A current of I amperes following through a resistance of R ohms varies according to the probability distribution with density

$$f(i) = \begin{cases} bi(1-i) & 0 < i < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the density of the power $W = I^2R$ watts, if the resistance R varies independently of the current according to probability distribution with the density $g(r) = 2r$ for $0 < r < 1$, and zero otherwise.

7.5 MULTIDIMENSIONAL DISTRIBUTIONS

We will now extend the concepts introduced in the case of bivariate distributions to the case of multivariate (or multidimensional) distributions. The motivation for these concepts lies in the frequency of practical situations when the analysis concerns many random variables simultaneously.

The examples here are easy to find. First, often a description of the phenomenon studied (sampled objects, some process, effects of treatment, etc.) uses several attributes at once. Formally, we have several random variables, $X_1, \ldots, X_n$, defined on a sample space, with X_i being the value of the ith attribute recorded for a given object. In such situations a natural choice of sample space is to take the population of the objects under consideration, with probability P being generated by a specific scheme of selecting elements from the population.

In another context, we may think of repeated measurements of the same attribute so that X_i is the result of the ith measurement. Now the probability P reflects possible dependence (or lack of it) in the measurement process.

Whatever the interpretation, formally we have a vector of random variables

$$\mathbf{X} = \mathbf{X}^{(n)} = (X_1, \ldots, X_n).$$

In the discrete case, the joint distribution of $\mathbf{X}$ consists of all probabilities of the form

$$P_{\mathbf{X}}(\mathbf{x}) = P\{\mathbf{X} = \mathbf{x}\} = P\{X_1 = x_1, \ldots, X_n = x_n\}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$. Clearly, we must have $\sum_{\mathbf{x}} P_{\mathbf{X}}(\mathbf{x}) = 1$, the summation extended over all possible values $\mathbf{x} = (x_1, \ldots, x_n)$ of vector $\mathbf{X}$.

In the continuous case, we have the joint density of the vector $\mathbf{X}$, in the form of a nonnegative function $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ of n variables such that

$$P\{\mathbf{X} \in Q\} = \underbrace{\int \cdots \int}_{\mathbf{x} \in Q} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

with

$$\underbrace{\int \cdots \int}_{R^n} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1.$$

The notions of marginal and conditional distributions and densities remain very much the same as in the bivariate case, except that the marginal distributions may now be themselves multivariate, and the same applies to conditional distribution, with the additional feature that the conditioning event may involve several random variables.

The simplicity of concepts can easily be obscured by confusing notation. While in any special case there is seldom any danger of confusion (we know the meaning of the variables and it is usually clear what is needed and what has to be done), the formulas covering the general cases may be confusing. We will use the following notation only presenting the theory; in examples we will try to simplify the notation, whenever possible.

First, in the case of marginal distribution, we need to specify the variables of interest. They form a subset of the variables $X_1, \ldots, X_n$. Thus we let $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, where $\mathbf{Y}$ are the variables of interest and $\mathbf{Z}$ are the remaining variables. The question of ordering is irrelevant. For instance, if $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$ and we are interested in the joint distribution of X_2 and X_5, then $\mathbf{Y} = (X_2, X_5)$ and $\mathbf{Z} = (X_1, X_3, X_4)$. We let $\mathbf{y}$ and $\mathbf{z}$ denote the values of vectors $\mathbf{Y}$ and $\mathbf{Z}$ so that $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ is the partitioning of $\mathbf{x}$ into the corresponding two subsets of coordinates. We now introduce the following definition:

Definition 7.5.1 In the discrete case, the *marginal distribution* of $\mathbf{Y}$ is given by

$$p_{\mathbf{Y}}(\mathbf{y}) = P\{\mathbf{Y} = \mathbf{y}\} = \sum_{\mathbf{z}} P\{\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}\}.$$

In the continuous case, the *marginal density* of $\mathbf{Y}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \int \cdots \int_{\mathbf{z}} f(\mathbf{y}, \mathbf{z})\, d\mathbf{z},$$

where the integrals represent the multiple integration over all variables X_i that are in vector $\mathbf{Z}$. □

Before considering examples, let us look at the corresponding definitions for the case of conditional distributions. As in the bivariate case, the discrete distributions present no difficulty, and we will simply state that all formulas can be deduced starting from the definition of conditional probability for events, namely $P(A|B) = P(A \cap B)/P(B)$ if $P(B) > 0$. In the continuous case, we have to partition $\mathbf{X}$ into three components So we write

$$\mathbf{X} = (\mathbf{Y}, \mathbf{Z}, \mathbf{W}),$$

where $\mathbf{Y}$ is the set of variables of interest, $\mathbf{Z}$ is the set of variables that will appear in the condition (whose values are assumed known), and $\mathbf{W}$ is the set (possibly empty) of variables that are neither in the condition nor of interest in the given instance. We need to define the conditional density of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$.

Definition 7.5.2 The conditional density of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ is defined as

$$g_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) = \frac{f_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})}{f_{\mathbf{Z}}(\mathbf{z})} = \frac{\int f(\mathbf{y},\mathbf{z},\mathbf{w})d\mathbf{w}}{\int\int f(\mathbf{y},\mathbf{z},\mathbf{w})d\mathbf{y}d\mathbf{w}},$$

where the integral symbols represent multiple integration over the variables in vectors w and y. □

We illustrate that by several examples.

■ **EXAMPLE 7.23 Trinomial Distribution**

Let us consider an experiment in which the outcome can be classified into one of three exclusive and exhaustive categories. We let these categories be A, B, and C, and let α, β, and γ be their respective probabilities. Outcomes A, B, and C form a partition of all outcomes so that $\alpha + \beta + \gamma = 1$.

In the case of experiment on treatment effectiveness A, B, and C could represent "improvement," "relapse," and "no change"; in cases of quality testing the categories may be "acceptable," "repairable, " "unacceptable and beyond repair," and so on.

We can perform the experiment n times assuming that the repetitions are independent. Let X_1, X_2 stand for the counts of A and B among n repetitions (the count of C is $n - X_1 - X_2$). We then have the following theorem:

Theorem 7.5.1 *The joint distribution of* $\mathbf{X} = (X_1, X_2)$ *defined above is given by*

$$P\{X_1 = x_1, X_2 = x_2\} = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!}\alpha^{x_1}\beta^{x_2}\gamma^{n-x_1-x_2}, \qquad (7.29)$$

where x_1, x_2 *are nonnegative integers such that* $x_1 + x_2 \leq n$.

Proof. The probability of x_1 outcomes A, x_2 outcomes B, and $n - x_1 - x_2$ outcomes C in a specific order equals

$$P(A)^{x_1}P(B)^{x_2}P(C)^{n-x_1-x_2} = \alpha^{x_1}\beta^{x_2}\gamma^{n-x_1-x_2},$$

by assumption of independence. Considering all possible

$$\binom{n}{x_1, x_2, n - x_1 - x_2} = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!}$$

orders (see Theorem 3.5.1), we obtain (7.5.1). □

■ **EXAMPLE 7.24**

Continuing Example 7.23, let us find the marginal distribution of X_2. Formally, for $x_2 = 0, 1, \ldots, n$, we have

$$p_{X_2}(x_2) = P\{X_2 = x_2\} = \sum_{x_1, x_3} P\{X_1 = x_1, X_2 = x_2, X_3 = x_3\}.$$

However, $x_3 = n - x_1 - x_2$, so

$$P\{X_1 = x_1, X_2 = x_2, X_3 = n - x_1 - x_2\} = P\{X_1 = x_1, X_2 = x_2\}.$$

Therefore, using (7.29), we write

$$\begin{aligned} p_{X_2}(x_2) &= \sum_{x_1=0}^{n-x_2} P\{X_1 = x_1, X_2 = x_2\} \\ &= \sum_{x_1=0}^{n-x_2} \frac{n!}{x_1!x_2!(n-x_1-x_2)!}\alpha^{x_1}\beta^{x_2}(1-\alpha-\beta)^{n-x_1-x_2} \\ &= \frac{n!}{x_2!(n-x_2)!}\beta^{x_2}(1-\beta)^{n-x_2} \\ &\quad \times \sum_{x_1=0}^{n-x_2} \frac{(n-x_2)!}{x_1![(n-x_2)-x_1]!} \times \frac{\alpha^{x_1}}{(1-\beta)^{x_1}} \left(1 - \frac{\alpha}{1-\beta}\right)^{(n-x_2)-x_1} \\ &= \binom{n}{x_2}\beta^{x_2}(1-\beta)^{n-x_2}, \end{aligned}$$

since the last sum above reduces to

$$\sum_{x_1=0}^{n-x_2} \binom{n-x_2}{x_1} \pi^{x_1}(1-\pi)^{n-x_2-x_1} = 1,$$

where $\pi = \alpha/(1-\beta)$.

Thus the marginal distribution of a variable in a trinomial distribution is binomial, as could have been expected. It is enough to interpret an occurrence of B as a success and the occurrence of anything else (A or C) as failure. So the probabilities of success and failure are β and $\alpha + \gamma = 1 - \beta$, respectively. The number of successes, X_2, has a binomial distribution, as it should. In the general case, suppose that the outcomes of each experiment are categorized into $m + 1$ classes, their probabilities being $\pi_1, \pi_2, \ldots, \pi_{m+1}$, with $\pi_1 + \pi_2 + \cdots + \pi_{m+1} = 1$. If the experiment is repeated independently n times, and X_i ($i = 1, \ldots, m$) is the number of outcomes of category i among n outcomes, then for nonnegative integers $x_1, \ldots, x_m$ with $x_1 + \cdots + x_m \leq n$ we have

$$P\{X_1 = x_1, \ldots, X_m = x_m\} = \frac{n!}{x_1! \cdots x_m! x_{m+1}!}\pi_1^{x_1} \cdots \pi_m^{x_m}\pi_{m+1}^{x_{m+1}},$$

where $x_{m+1} = n - (x_1 + \cdots + x_m)$ and $\pi_{m+1} = 1 - (\pi_1 + \cdots + \pi_m)$. This is the *multinomial* distribution.

Let now $\mathbf{Y} = (X_{i_1}, \ldots, X_{i_k})$ be a subset of variables $\mathbf{X} = (X_1, \ldots, X_m)$ with a multinomial distribution. Proceeding in the same way as in Example 7.24, we can show that for $\mathbf{y} = (y_{i_1}, \ldots, y_{i_k})$ with coordinates y_{i_j} being nonnegative integers with $y_{i_1} + \cdots + y_{i_k} \leq n$, we have

$$P\{Y_{i_1} = y_{i_1}, \ldots, Y_{i_k} = y_{i_k}\} = \frac{n!}{y_{i_1}! \cdots y_{i_k}! y_{i_{k+1}}!}\pi_{i_1}^{y_{i_1}} \cdots \pi_{i_k}^{y_{i_k}}\pi_{i_{k+1}}^{y_{i_{k+1}}},$$

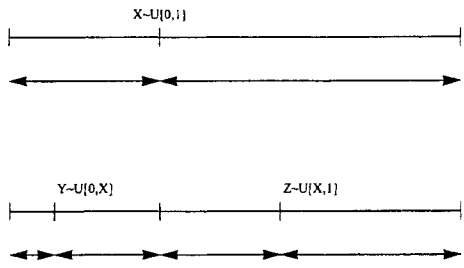


Figure 7.14 First two generations in the process of grinding

where $y_{i_{k+1}} = n - y_{i_1} - \cdots - y_{i_k}$ and $\pi_{i_{k+1}} = 1 - \pi_{i_1} - \cdots - \pi_{i_k}$. Thus a marginal distribution in a multinomial distribution is again multinomial with lower dimensionality.

■ EXAMPLE 7.25 Model of Grinding

The basic model suggested for the process of grinding rock begins with a piece of rock that is randomly divided into two parts. Each of these two parts is again divided randomly, and so on, until in the nth "generation" we have 2^n pieces of rock. Assume that this process continues for a large number of generations.

With the model one can obtain a distribution, as $n \to \infty$, of sizes of rock. Naturally, with each generation of grinding the actual sizes of rock will tend to zero, so that the sizes of 2^n pieces in the nth generation will have to be multiplied by the approximate scaling factor.

The results allow us to predict what fraction of the initial mass of rock will be ground into gravel with sizes contained between specific limits a and b, and so on.

Instead of the whole process, we will study only the first two divisions and the resulting sizes. For simplicity, we assume that the initial size is 1, represented as interval $[0, 1]$. In the first division, a point X—with uniform distribution—partitions the unit interval into two pieces of length X and $1-X$. In the next partition, length X is divided by point Y (with Y distributed uniformly between 0 and X), and the remainder is divided by point Z, distributed uniformly on $[X, 1]$. We therefore have four fragments (see Figure 7.14) of length

$$X_1 = Y, \quad X_2 = X - Y, \quad X_3 = Z - X, \quad X_4 = 1 - Z.$$

We will start by deriving the joint trivariate density of (X, Y, Z). By assumption, X has density

$$f_X(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Given $X = x$, variables Y and Z are independent with densities

$$g_{Y|X}(y|x) = \begin{cases} \frac{1}{x} & 0 \le y \le x \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad g_{Z|X}(z|x) = \begin{cases} \frac{1}{1-x} & x \le z \le 1 \\ 0 & \text{otherwise}. \end{cases}$$

Thus $g_{Y,Z|X}(y, z|x) = g_{Y|X}(y|x)g_{Z|X}(z|x)$, and the joint density is

$$f(x, y, z) = \begin{cases} \frac{1}{x(1-x)} & \text{for } 0 \le y \le x \le z \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{7.30}$$

The marginal joint density of Y and Z, for $0 \le y \le z \le 1$ will now be

$$f_{Y,Z}(y, z) = \int_y^z \frac{dx}{x(1-x)} = \int_y^z \left(\frac{1}{x} + \frac{1}{1-x}\right) dx = \log \frac{z}{1-z} + \log \frac{1-y}{y},$$

so the marginal density of Y will be

$$f_Y(y) = \int_y^1 \left(\log \frac{z}{1-z} + \log \frac{1-y}{y}\right) dz = -\log y$$

for $0 < y < 1$. Finally, we obtain the conditional density of X given $Y = y$ and $Z = z$:

$$g_{X|Y,Z}(x|y, z) = \frac{f(x, y, z)}{f_{YZ}(y, z)} = \frac{\frac{1}{x(1-x)}}{\log \frac{z}{1-z} + \log \frac{1-y}{y}}$$

for $0 < y < x < z < 1$.

To derive the density of (X_1, X_2, X_3, X_4), we need to derive only the density of (X_1, X_2, X_3), since $X_1 + X_2 + X_3 + X_4 = 1$. This can be done by applying the transformation

$$x_1 = y, \qquad x_2 = x - y, \qquad x_3 = z - x \tag{7.31}$$

to the trivariate density (7.30).

The inverse transformation to (7.31) is

$$y = x_1, \qquad x = x_1 + x_2, \qquad z = x_1 + x_2 + x_3. \tag{7.32}$$

with Jacobian

$$J = \begin{vmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \frac{\partial y}{\partial x_3} \\ \frac{\partial x}{\partial x_1} & \frac{\partial x}{\partial x_2} & \frac{\partial x}{\partial x_3} \\ \frac{\partial z}{\partial x_1} & \frac{\partial z}{\partial x_2} & \frac{\partial z}{\partial x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{vmatrix} = 1.$$

Thus the joint density $g(x_1, x_2, x_3)$ of (X_1, X_2, X_3) is obtained from (7.30) by substituting (7.32), that is,

$$g(x_1, x_2, x_3) = \begin{cases} \frac{1}{(x_1+x_2)(1-x_1-x_2)} & x_1 > 0,\ x_2 > 0,\ x_3 > 0, \\ & x_1 + x_2 + x_3 < 1. \\ 0 & \text{otherwise.} \end{cases}$$

PROBLEMS

7.5.1 Two cards are drawn without replacement from an ordinary deck. Let X be the number of aces, Y be the number of red cards, and Z be the number of hearts. Find: **(i)** The joint distribution of (X, Y, Z). **(ii)** $P\{Y = Z\}$. **(iii)** The conditional distribution of Z given $Y = y$.

7.5.2 Suppose that X_1, X_2, X_3 have joint density

$$f(x_1, x_2, x_3) = \begin{cases} x_1 x_2 x_3 & \text{for } 0 < x_i < c,\ i = 1, 2, 3 \\ 0 & \text{otherwise.} \end{cases}$$

Find: **(i)** c. **(ii)** The joint density of (Y_1, Y_2, Y_3), where $Y_1 = X_1, Y_2 = X_1 X_2$, and $Y_3 = X_1 X_2 X_3$.

7.5.3 Let random variables X, Y, and Z have joint density $f(x, y, z) = c(x+y+z)$ for $0 < x < y < z < 1$. Find: **(i)** c. **(ii)** $P(X + Z > 1)$. **(iii)** $P(X + Z > 1 | Y = 0.5)$.

7.5.4 A can of Three-Bean-Salad contains beans of varieties A, B, and C (plus other ingredients which are of no concern for the problem). Let X, Y, and Z denote the relative weights of varieties A, B, and C in a randomly selected can (so that $X+Y+Z = 1$). Moreover, let the joint distribution of (X, Y) have the density $f(x, y) = kx^2 y$ for $x > 0, y > 0, x + y < 1$, and $f(x, y) = 0$ otherwise. Find: **(i)** k. **(ii)** Probability $P(X > 0.5)$ that beans A take more than half of the total weight. **(iii)** $P(Z > 0.5)$. **(iv)** Probability that none of the three varieties of beans will take more than half of the weight. **(v)** The marginal density of (X, Z) and of Z.

CHAPTER 8

EXPECTATION

8.1 INTRODUCTION

The probabilistic concepts discussed in Chapters 6 and 7 could have also been developed without using random variables but in a clumsy and awkward way. Random variables were used there as a convenient tool of describing large classes of events. Indeed, once we considered events of the form $\{a < X \leq b\}$, it was quite natural to reduce the analysis to even simpler events $\{X \leq t\}$. Then the probability of such an event, regarded as a function of the argument t (i.e., the cdf of X) turned out to carry all information about the distribution of X. And once the notion of the cdf was introduced, it was natural to look for classes of cdf's that allow especially simple description (hence the definition of discrete and continuous random variables).

In this chapter we introduce the notion of *expectation*. Expectation, or the *expected value* of a random variable, cannot be formulated without the concept of random variables.

The intuitive content of the notion of expectation is as follows: Consider a random variable X, defined on some sample space $\mathcal{S}$. An experiment consists of a random selection of a point s of $\mathcal{S}$, and $X(s)$ is interpreted as the gain (loss, if $X(s) < 0$) of a hypothetical gambler.

If the experiment is repeated n times, the sample space becomes $\mathcal{S}^n = \mathcal{S} \times \cdots \times \mathcal{S}$ (n times). An outcome is $s^{(n)} = (s_1, \ldots, s_n)$, with $s_i \in \mathcal{S}, i = 1, \ldots, n$. The

gambler's accumulated gain becomes $X(s_1) + X(s_2) + \cdots + X(s_n)$, traditionally written as $X_1 + X_2 + \cdots + X_n$, with $X_i = X(s_i)$ being the outcome of the ith trial.

The average gain per gamble now becomes $(X_1 + \cdots + X_n)/n$. As n becomes larger, this average fluctuates less and less, tending to stabilize at some value, which we call the expectation of X and denote by $E(X)$.

It is not difficult to calculate this value in case of a discrete random variable. Indeed, if X assumes values $x_1, x_2, \ldots, x_m$ with probabilities $p_i = P\{X = x_i\}, i = 1, 2, \ldots m$—and if the experiment is to be repeated n times, then $\{X = x_i\}$ will occur n_i times $(i = 1, \ldots, m)$ and $n_1 + n_2 + \cdots + n_m = n$. According to the frequency interpretation of probability, the ratio n_i/n tends to $p_i = P\{X = x_i\}$ as the number n of repetitions of the experiments increases. The total gain $X_1 + \cdots + X_n$ equals $x_1 n_1 + x_2 n_2 + \cdots + x_m n_m$, and the average gain per experiment,

$$x_1 \frac{n_1}{n} + x_2 \frac{n_2}{n} + \cdots + x_m \frac{n_m}{n},$$

tends to

$$x_1 p_1 + x_2 p_2 + \cdots + x_m p_m.$$

We will use the last quantity, that is,

$$E(X) = x_1 P\{X = x_1\} + \cdots + x_m P\{X = x_m\} \tag{8.1}$$

as a formal definition of an expected value $E(X)$ of a discrete random variable X. The definition will later be extended to also cover random variables that are not discrete.

8.2 EXPECTED VALUE

For simplicity, the definition of an expected value of a random variable X is now formulated separately for the discrete and continuous random variables.

Definition 8.2.1 If X is a discrete random variable that assumes values $x_1, x_2, \ldots$ with probabilities $P\{X = x_i\}, i = 1, 2, \ldots$, then the *expected value* of X is defined as

$$E(X) = \sum_i x_i P\{X = x_i\}, \tag{8.2}$$

provided that

$$\sum_i |x_i| P\{X = x_i\} < \infty. \tag{8.3}$$

If X is a continuous random variable with the density $f(x)$, then the *expected value* of X is defined as

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx, \tag{8.4}$$

provided that

$$\int_{-\infty}^{+\infty} |x| f(x) dx < \infty. \tag{8.5}$$

□

Definition 8.2.1 covers formula (8.1) as a special case, since if X assumes only finitely many values, condition (8.3) is always satisfied. Conditions (8.3) and (8.5) are necessary to eliminate the situation where the value of an infinite sum depends on the order of summation (or integration). The absolute convergence (or absolute integrability) guarantees that the expected value is unambiguously defined.

The other symbols used for the expected value $E(X)$ in probabilistic and statistical literature may be EX, μ_X, m_X, or simply μ, m, if it is clear which random variable is being studied. In physics, the symbol used is $< X >$. Regarding terminology, the expected value is also called the expectation, or the mean. Moreover, the expected value is often associated not with X, but with its distribution. Consequently, one sometimes uses symbols such as μ_F or m_F, where F is a cumulative distribution function.

We begin with a series of examples of expectations for discrete random variables.

■ **EXAMPLE 8.1**

Let A be some event, and let $p = P(A)$. Imagine that you are to receive \$1 if A occurs and nothing if A does not occur (i.e., if A^c occurs). Let X be the random variable defined as "the amount you receive." In this case X assumes only two values, and its distribution is given by the array

Value	0	1
Probability	$1-p$	p

According to formula (8.2),

$$E(X) = 0 \times (1-p) + 1 \times p = p.$$

The result is quite intuitive: if you play such a game over and over, then your average outcome per game will equal the probability of winning, p.

Observe that X can also be described as "the number of occurrences of A in a single experiment." We will use this interpretation later.

■ **EXAMPLE 8.2**

Suppose now that you are to win \$1 if A occurs, but lose \$1 (i.e., "win" -1) if A does not occur. Then the distribution of Y, the random variable describing your winnings, is

Value	-1	1
Probability	$1-p$	p

and $E(Y) = -1 \times (1-p) + 1 \times p = 2p - 1$. Thus $E(Y) > 0$ if and only if $p > 0.5$, which agrees with our intuition: the game is favorable if the probability p of winning exceeds 0.5.

■ **EXAMPLE 8.3 Expected Value of Binomial Distribution**

The binomial random variable was defined in Example 6.8 as the total number of successes in n independent Bernoulli trials, where the probability of a success in any single trial is p. The possible values of the binomial random variable are $0, 1, \ldots, n$ with probabilities given by (6.21). Consequently,

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-1-(k-1))!}p^{k-1}(1-p)^{n-1-(k-1)} \\ &= np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j} = np, \end{aligned}$$

where in the last sum we introduced the new index of summation $j = k - 1$. The last sum turns out to be a Newtonian expansion of $[p + (1-p)]^{n-1} = 1$.

■ **EXAMPLE 8.4 Expectation of Geometric Distribution**

Let an event A ("success") have probability p. We keep repeating independently the experiments until A occurs for the first time, with X being the number of trials up to and including the first success. As already mentioned (see Example 6.9), X is a random variable with geometric distribution and $P\{X = k\} = (1-p)^{k-1}p$ for $k = 1, 2, \ldots$. Find $E(X)$.

SOLUTION. Letting $q = 1 - p$, we write

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} kP\{X = k\} = \sum_{k=1}^{\infty} kq^{k-1}p = p\sum_{k=1}^{\infty}\frac{d}{dq}(q^k) \\ &= p\frac{d}{dq}\sum_{k=1}^{\infty}q^k = p\frac{d}{dq}\left(\frac{q}{1-q}\right) = p \times \frac{1}{(1-q)^2} = \frac{1}{p}. \end{aligned}$$

The answer is intuitive. If, for example, the probability of success is $p = 0.2$, then on average, every fifth trial is a success, and therefore the average number of trials until the next success is 5, which agrees with formula (8.6).

The following theorem provides a very useful interpretation of the expected value of a random variable:

Theorem 8.2.1 *Let X be a random variable which assumes only nonnegative integer values. Then*

$$E(X) = \sum_{n=0}^{\infty} P\{X > n\}. \tag{8.6}$$

Proof. For $p_n = P\{X = n\}$ we have

$$\begin{aligned} E(X) &= p_1 + 2p_2 + 3p_3 + \cdots \\ &= p_1 + p_2 + p_3 + \cdots \\ &\quad + p_2 + p_3 + \cdots \\ &\quad + p_3 + \cdots \\ &= P\{X > 0\} + P\{X > 1\} + P\{X > 2\} + \cdots, \end{aligned}$$

where a change of order of summation is allowed, since all terms are nonnegative. □

■ **EXAMPLE 8.5**

Let us use formula (8.6) to find the expectation of the geometric distribution from Example 8.4. We have

$$P\{X > n\} = P\{\text{only failures in the first } n \text{ trials}\} = q^n,$$

or, if this reasoning appears suspiciously simple and one puts more trust in algebra,

$$P\{X > n\} = \sum_{k=n+1}^{\infty} P\{X = k\} = \sum_{k=n+1}^{\infty} q^{k-1}p = p\frac{q^n}{1-q} = q^n.$$

Consequently,

$$E(X) = \sum_{n=0}^{\infty} q^n = \frac{1}{1-q} = \frac{1}{p}.$$

The expected value in Example 8.5 was finite. Such situations are most common in practice, as otherwise the very concept of expected value would make no sense. However, there are cases of both practical and theoretical significance, when the condition (8.3) or (8.5) is not met.

To be more precise, one should distinguish between two cases. We will illustrate the possibilities of condition (8.3) involving the series; the situation with an integral is analogous. Let us consider separately sums $\sum x_i P\{X = x_i\}$ involving positive and negative terms, that is, $U^+ = \sum[\max(x_i, 0)] \times P\{X = x_i\}$ and $U^- = -\sum[\min(x_i, 0)] \times P\{X = x_i\}$. It is clear that if the series (8.3) diverges, then U^+, U^-, or both must be infinite, for otherwise the series of absolute values (8.3), equal $U^+ + U^-$, would be finite. If only U^+ is infinite, we can say that $E(X) = +\infty$; similarly, if only U^- is infinite, we can say that $E(X) = -\infty$. It is only when both U^+ and U^- are infinite that we have the expression of the form $\infty - \infty$, to which we cannot assign any value. In the last case we say that $E(X)$ *does not exist.*

In the following example random variable will have infinite expectation:

EXAMPLE 8.6 Petersburg Paradox

Suppose that you have the possibility of participating in the following game. A fair coin is tossed repeatedly until a head occurs. If the first head occurs at the kth toss, you win 2^k dollars. Clearly, this is a game in which you cannot lose, so the question is: What is the amount you should be willing to pay to participate in such game?

SOLUTION. The most common argument offered in such cases is that one should be willing to pay the fee as long as the expected winnings exceed or equal the fee for participation in the game. Then, "in the long run," one should come even or ahead.

Let X denote the winnings in the game (not counting the fee for participation in the game). Then the event $\{X = 2^k\}$ occurs if the first $k-1$ tosses are tails, and the kth toss is heads; chances of this event are $1/2^k$. Consequently, we have

$$E(X) = 2 \times \frac{1}{2} + 2^2 \times \frac{1}{2^2} + 2^3 \times \frac{1}{2^3} + \cdots = \infty.$$

Because the expected winnings in this game are infinite, one should be willing to pay an infinite fee for the right to play. However, in any particular game, the winnings will always be finite, so one is certain to lose.

This phenomenon, discovered in eighteen century, was named the Petersburg paradox. Attempts to solve it led to the introduction of the concept of utility. Without going into detail, it is postulated that the utility of money—the value a person attaches to a given amount of money—is not constant but depends on how much money the person already has. This appears intuitively acceptable: \$20 may seem to be a lot for someone who has nothing but does not seem like very much to a millionaire.

If $u(x)$ is the utility of $\$x$, then the above postulate implies that $u(x)$ is not proportional to x. The *expected utility* from participating in the game is

$$E[u(X)] = \sum_{k=1}^{\infty} u(2^k) \times \frac{1}{2^k},$$

a quantity that may be finite for some "plausible" choices of utility function $u(x)$. In particular, Bernoulli suggested $u(x) = c \log x$.

EXAMPLE 8.7 Bad Luck

This example is due to Feller (1971). Consider an event that can happen to Peter and some other (perhaps all) people—one may think of a specific illness, breakdown of a newly bought car, and so on. Let T_0 denote the time until this event happens to Peter. Occasionally Peter learns about other people who were in the situation analogous to his, and for whom the event in question already occured. Let their waiting times be $T_1, T_2, \ldots$. Assume symmetry and lack of interference here so that the random variables $T_0, T_1, \ldots$ have the same distribution and are independent. To avoid the necessity of considering ties, we assume that $T_i, i = 0, 1, \ldots$, have a continuous distribution. This means that an equality $T_i = T_j$ has probability zero and hence can be disregarded.

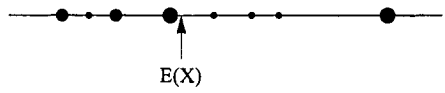


Figure 8.1 Interpretation of expected value of a discrete random variable

We can measure the "bad luck" of Peter in terms of a number of people, N, he has to encounter before finding one "worse off" (i.e., with $T_i < T_0$). Thus $N = n$ if $T_0 < T_1, T_0 < T_2, \ldots, T_0 < T_{n-1}, T_0 > T_n$, and $N > n$ if T_0 is the smallest among $T_0, T_1, \ldots, T_n$, that is,

$$P\{N > n\} = P\{T_0 = \min(T_0, \ldots, T_n)\}.$$

But by symmetry, T_0 is as likely to be the smallest of $n+1$ variables $T_0, \ldots, T_n$ as any other. Hence we must have

$$P\{N > n\} = \frac{1}{n+1},$$

and consequently,

$$P\{N = n\} = P\{N > n-1\} - P\{N > n\} = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}.$$

It follows that

$$E(N) = \sum_{n=1}^{\infty} n \times P(N = n) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty,$$

So the expected bad luck of Peter (and of everyone else!) is infinite.

Before going further, it may be worthwhile to give the following "mechanical" interpretation of $E(X)$. Consider a discrete random variable, with possible values x_i and the corresponding probabilities $p_i = P\{X = x_i\}, i = 1, 2, \ldots$.

Let us construct a mechanical device corresponding to the distribution of X. Consider an infinitely thin and infinitely rigid wire extended from $-\infty$ to ∞ (this is an abstraction that physicists use: an ideal void, a perfectly black body, etc.). At points with coordinates $x_1, x_2, \ldots$ we attach weights that are proportional to probabilities p_i (see Figure 8.1 where the sizes of the dots symbolize the variability in weight). Then $E(X)$ is the coordinate of the point at which the whole device would balance, or its center of gravity.

A similar interpretation also holds for continuous random variables. In this case we have to imagine that an infinitely rigid metal sheet of uniform weight is cut into the shape of density $f(x)$. Then $E(X)$ is the coordinate of the point at which the figure will balance if laid flat (see Figure 8.2).

We will now give some examples of expectations of continuous type random variables.

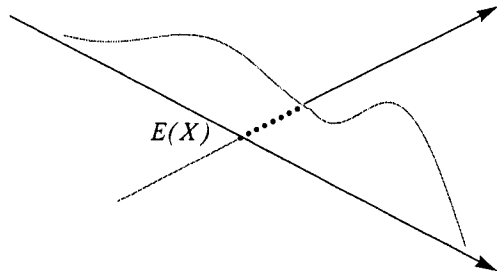


Figure 8.2 Interpretation of expected value of a continuous random variable

■ **EXAMPLE 8.8**

Let X have a distribution uniform on interval $[A, B]$ so that (see Example 6.15)

$$f(x) = \begin{cases} \frac{1}{B-A} & \text{for } A \le x \le B \\ 0 & \text{otherwise.} \end{cases}$$

According to the interpretation given above, the metal sheet figure is just a rectangle. Hence it will balance if supported in the middle between A and B. Thus we can expect that $E(X) = (A + B)/2$. Indeed the computations give

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_A^B \frac{x}{B-A}dx = \frac{B^2 - A^2}{2(B-A)} = \frac{A+B}{2}.$$

■ **EXAMPLE 8.9**

We will now find the expected value of random variable X with exponential distribution with density $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$ and 0 otherwise.

SOLUTION. Integrating by parts, we obtain

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x}dx = -\frac{1}{\lambda}e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}.$$

Thus, if the lifetime (in hours) of a certain type of electric bulb has a density $0.001e^{-0.001t}$ for $t > 0$, then the expected lifetime equals $(0.001)^{-1} = 1000$ hours.

■ **EXAMPLE 8.10**

Find the expectation of a normal random variable with the density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

SOLUTION. We substitute $z = (x - \mu)/\sigma, dx = \sigma dz$, obtaining

$$\begin{aligned} E(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\mu + z\sigma) e^{-\frac{z^2}{2}} dz \\ &= \mu \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z e^{-\frac{z^2}{2}} dz = \mu. \end{aligned}$$

The first integral in the last line equals μ (since the standard normal density integrates to 1), while the second integral is 0 because the integrand is an odd function. So the integrals over the positive and negative parts of the z-axis (each being finite) cancel each other. Thus μ is the expected value of the random variable with distribution $N(\mu, \sigma^2)$.

We will next derive an analogue of formula (8.6) for an expected value of a nonnegative continuous random variable.

Theorem 8.2.2 *Let T be a nonnegative random variable of continuous type, with density $f(t)$ and cdf $F(t)$, so that $f(t) = F(t) = 0$ for $t < 0$. Then*

$$E(T) = \int_0^{\infty} [1 - F(t)] dt.$$

Proof. Replacing t with $\int_0^t dx$ and changing the order of integration, we have

$$\begin{aligned} E(T) &= \int_0^{\infty} x f(x)\, dx = \int_0^{\infty} \left(\int_0^x dt \right) f(x)\, dx = \int_0^{\infty} \int_t^{\infty} f(x)\, dx\, dt \\ &= \int_0^{\infty} [1 - F(t)]\, dt. \end{aligned}$$

□

■ **EXAMPLE 8.11**

Let us compute again the expectation of a random variable with EXP(λ) distribution. We have here $F(x) = 1 - e^{-\lambda x}$ for $x > 0$. Using Theorem 8.2.2, we obtain $E(X) = \int_0^{\infty} e^{-\lambda x} dx = 1/\lambda$.

■ **EXAMPLE 8.12**

Finally, let us consider the Cauchy distribution with the density given by

$$f(x) = \frac{1}{\pi} \times \frac{1}{1 + x^2}.$$

It is easy to check that the function $f(x)$ is a density of a random variable:

$$\frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{dx}{1 + x^2} = \frac{1}{\pi} \arctan x \Big|_{-\infty}^{+\infty} = 1.$$

Regarding expectation, the positive part of the defining integral equals

$$\frac{1}{\pi}\int_0^{+\infty}\frac{xdx}{1+x^2}=\frac{1}{\pi}\times\frac{1}{2}\log(1+x^2)\Big|_0^{+\infty}=+\infty.$$

In a similar fashion we check that $\int_{-\infty}^{0} xf(x)dx=-\infty$. Hence the expected value of Cauchy distribution does not exist.

PROBLEMS

8.2.1 Variable X assumes values 0, 1, 2, and 3 with probabilities 0.3, a, 0.1, and b, respectively. Find a and b if: **(i)** $E(X)=1.5$. **(ii)** $E(X)=m$. First determine all possible values of m.

8.2.2 The density of the lifetime T of some part of electronic equipment is $f(t)=\lambda^2te^{-\lambda t}, t>0$. Find $E(T)$.

8.2.3 Suppose there are $k=10$ types of toys (plastic animals, etc.) to be found in boxes of some cereal. Assume that each type occurs with equal frequency. **(i)** Find $E(X)$, where X is the number of boxes you must buy until you collect three different types of toys. **(ii)** Suppose that your little sister is collecting the cereal toys and would be happy to get the whole set of 10 different types. Find $E(X)$, where X is the number of boxes you must buy until she gets a complete collection.

8.2.4 Find the expected value $E(|X|)$, where X is a normal random variable with parameters $\mu=0$ and σ^2 (the distribution of $|X|$ is sometimes called *folded normal* or *half normal*).

8.2.5 Let $X_1, X_2, \ldots, X_n$ be independent, each with the same distribution. Find $E[\min(X_1,\ldots,X_n)]$ if the distribution of variables is: **(i)** EXP(λ). **(ii)** U$[a,b]$. [*Hint:* First find the cdf (or density) of the random variable whose expectation is being computed.]

8.2.6 Let X be a nonnegative continuous random variable with hazard rate $h(t)=t$. Find $E(X)$.

8.2.7 Show that if the expectation of a continuous type random variable X exists, and the density $f(x)$ of X satisfies the condition $f(x)=f(2a-x)$ for all $x\geq 0$, then $E(X)=a$.

8.2.8 Let X be a continuous random variable with density f, and let Y be the area under f between $-\infty$ and X. Find $E(Y)$. (*Hint:* Start by determining the cdf and the density of Y.)

8.3 EXPECTATION AS AN INTEGRAL*

Definition 8.2.1 of expected value in the discrete and continuous cases covers most situations occurring in practice. Many textbooks do not provide the general definition, and some do not even mention the fact that a general definition exists.

One of the consequences of such an omission is that one could come to the conclusion that the symbol $E(X+Y)$ makes no sense if X is discrete and Y is continuous (since $X+Y$ is neither a continuous nor discrete random variable). In fact the expectation of a random variable is generally defined as an integral of a random variable, treated as a function on the sample space $\mathcal{S}$. We will now sketch this definition.

Riemann Integral

Let us start by briefly recalling the definition of an "ordinary" (i.e., Riemann) integral $\int_a^b g(x)dx$. For the moment assume that the function g is continuous on $[a,b]$. We first choose a sequence of partitions of the interval $[a,b]$. To simplify the presentation, suppose that the nth partition divides the interval $[a,b]$ into 2^n equal parts, so that the kth point of the nth partition is

$$x_{n,k} = a + \frac{k}{2^n}(b-a), \qquad k = 0, 1, \ldots, 2^n.$$

Let $g_{n,k}^{(-)}$ be the minimum of function g in the kth interval, that is,

$$g_{n,k}^{(-)} = \min_{x_{n,k-1} \le x \le x_{n,k}} g(x).$$

We now form the (lower) sum approximating the integral

$$\underline{S}_n = \sum_{k=1}^{2^n} g_{n,k}^{(-)}(x_{n,k} - x_{n,k-1}) = \frac{b-a}{2^n} \sum_{k=1}^{2^n} g_{n,k}^{(-)}. \tag{8.7}$$

The upper sum $\overline{S}_n$ is defined in the same way, with the minimum in (8.7) replaced by the maximum, say $g_{n,k}^{(+)}$.

It is not difficult to show that $\underline{S}_n \le \overline{S}_n$ for all n, and that the sequences $\{\underline{S}_n\}$ and $\{\overline{S}_n\}$ are monotone, the first increasing and the second decreasing. If they converge to the same limit, say S, then this limit is called the Riemann integral of g, and it is denoted by $\int_a^b g(x)dx$.

This is the essence of the definition; the details may vary in two respects. First, instead of dividing the interval $[a,b]$ into 2^n equal parts in the nth partition, we can take any points

$$a = x_{n,0} < x_{n,1} < \cdots < x_{n,k_n} = b$$

as long as

$$\lim_{n\to\infty} \max_{1 \le i \le k_n} (x_{n,i} - x_{n,i-1}) = 0$$

(this implies that we must have $k_n \to \infty$).

Second, instead of taking the minimum and the maximum of g over the subintervals of partitions (which need not exist if g is not continuous), we can take the value of function g at a selected point in each subinterval. We then have one partial sum (instead of lower and upper sums), and the requirement is that the limits of these partial sums exist and are the same regardless of the choice of the partitions and of the intermediate points.

The main results of the theory built on this definition are very well known: continuous functions are integrable on closed finite intervals and if g has antiderivative G (i.e., $G' = g$), then

$$\int_a^b g(x)dx = G(b) - G(a).$$

The Riemann integral over an infinite range, $\int_{-\infty}^{+\infty} g(x)dx$, is defined in the usual manner through a limiting passage; we omit the details, which can be found in most calculus texts.

We recall the definition of the Riemann integral in order to better stress the differences and also the analogy between the principles of definition of the Lebesgue and Riemann integrals. The difference between the two definitions seems small and, at first, not essential. Yet the concept of the Lebesgue integral is of tremendous consequence, allowing us at once to free the concept of integral of all inessential constraints, and to stress its most crucial and significant features.

Lebesque Integral

Again, assume at first that the function g, defined on the interval $[a, b]$, is continuous, and let $A < g(x) < B$. Instead of partitioning the interval $[a, b]$ on the x-axis, let us partition the interval $[A, B]$ on the y-axis. For simplicity, let us again take the partitions into 2^n equal parts,

$$A = y_{n,0} < y_{n,1} < \cdots < y_{n,2^n} = B,$$

where

$$y_{n,k} = A + \frac{k}{2^n}(B - A), \quad k = 0, 1, \ldots, 2^n.$$

The lower sum approximating the integral can now be written as

$$\underline{S}_n^* = \sum_{k=0}^{2^n-1} y_{n,k} \times l(C_{n,k}), \tag{8.8}$$

where l stands for length and $C_{n,k}$ is the set of points x where the function g lies between $y_{n,k}$ and $y_{n,k+1}$; more precisely,

$$C_{n,k} = \{x : y_{n,k} \leq g(x) < y_{n,k+1}\}, k = 0, 1, \ldots, 2^n - 1.$$

The upper sum $\overline{S}_n^*$ is defined in a similar way, with $y_{n,k}$ in (8.8) replaced by $y_{n,k+1}$.

The lower sums $\underline{S}_n$ and $\underline{S}_n^*$ for Riemann and Lebesgue integrals of the same functions are illustrated as parts (R) and (L) in Figure 8.3. The Lebesgue integral is now defined in the same way as the Riemann integral, namely as the common limit (if it exists) of the upper and lower sums. Again, we omit the details of the extension of the definition to an infinite range.

We may capture the difference in two definitions as follows: In the Riemann integral, when we partition the x-axis, we control the intervals on the x-axis, while the maxima and minima over these intervals depend on function g. In the Lebesgue integral, when we partition the y-axis, we control the values, but corresponding sets

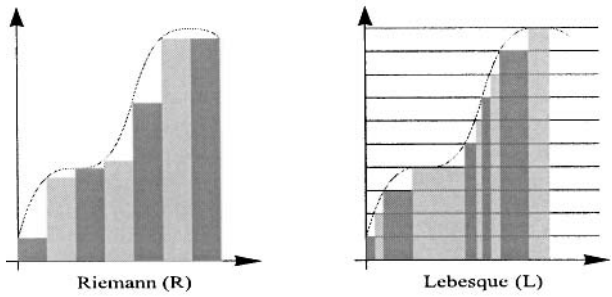


Figure 8.3 Approximating sums for Riemann and Lebesgue integrals

on the x-axis are determined by the function g. The distinction was summarized by Lebesgue, as a comparison of two cashiers who, at the end of the day, have to count the cash in their drawers. The "Riemann" cashier takes the money out in the order it came in. She counts: "A five-dollar bill, and a penny makes 5.01; then I have a quarter, which makes 5.26; then ...," and so on. On the other hand, the "Lebesgue" cashier proceeds systematically: "I have n_1 twenty-dollar bills, which makes $20n_1$ dollars. Then I have n_2 ten-dollar bills, which together makes $20n_1 + 10n_2$ dollars ..., " and so on.

One of the main results in the theory of the Lebesgue integral is that if a function is Riemann integrable, then it is also Lebesgue integrable, and both integrals are equal. Thus we may use standard calculus to compute Lebesgue integrals, whenever it is applicable. However, it is important to point out that the class of functions that are Lebesgue integrable is substantially wider than the class of functions that are Riemann integrable. To see that it is so, consider the following example:

■ **EXAMPLE 8.13**

Let g be a function continuous on a closed finite interval $[a, b]$ so that g is Riemann integrable. It is clear that if we modify g at a single point x_0 by defining

$$g^*(x) = \begin{cases} g(x) & \text{if } x \neq x_0 \\ c & \text{if } x = x_0, \end{cases}$$

where c is any number such that $c \neq g(x_0)$, then the function $g^*(x)$ is Riemann integrable and $\int_a^b g(x)dx = \int_a^b g^*(x)dx$. We can modify in this way the function g at any *finite* set of points, and we will still have a Riemann integrable function, with the same integral as the original function g. However, if we modify g in such a way at a *countable* set of points, then the resulting function may not be Riemann integrable. This is because for a modification at finitely many points, say N, the sums approximating an integral of g and an integral of g^* differ in, at most, N terms. The difference tends to zero, since the sum of lengths of intervals on which the functions g and g^* differ tends to

zero. But this is not necessarily the case where g and g^* differ at countably many points. Thus g^* may be not integrable in the Riemann sense.

However, g^* is integrable in Lebesgue sense, because two sets that differ by a countable set of points have the same length. So the sums approximating the Lebesgue integral are the same for g and g^*.

We will now present two generalizations of the concept of integral introduced thus far. The first generalization concerns both the Riemann and Lebesgue integrals, leading in a natural way to the corresponding concepts of Riemann-Stieltjes and Lebesgue-Stieltjes integrals. The second generalization, more important for our purposes, will concern specifically the Lebesgue integral.

Riemann-Stieltjes Integral

Regarding the first generalization, let us observe that in the approximating sums $\underline{S}_n$ in the Riemann integral given by (8.7), we can write the terms $g_{n,k}^{(-)}(x_{n,k} - x_{n,k-1})$ as $g_{n,k}^{(-)} l([x_{n,k-1}, x_{n,k}])$, where $l(\cdot)$ stands for the length of the set (in this case, of a single interval). This is similar to the sums (8.8) approximating the Lebesgue integral.

Now, instead of the length being the difference between the coordinates of endpoints, one can take a "length-like" function, defined as the differences between values of some function, F, of coordinates. Thus the terms of the approximating sums are now $g_{n,k}^{(-)}[F(x_{n,k}) - F(x_{n,k-1})]$. Naturally function F has to satisfy some conditions if such an extension is to lead to a meaningful concept of an integral.

Without striving for a general definition (which can be found in most books on advanced calculus), we will simply consider the case where F is a cdf, meaning a nondecreasing function, continuous on the right, satisfying the conditions $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

The common limit (if it exists) of the two sequences

$$\underline{S}_n = \sum g_{n,k}^{(-)}[F(x_{n,k}) - F(x_{n,k-1})]$$

and

$$\overline{S}_n = \sum g_{n,k}^{(+)}[F(x_{n,k}) - F(x_{n,k-1})]$$

will be denoted $\int_a^b g(x)dF(x)$, and called the Riemann–Stieltjes (R-S) integral of function g with respect to function F. Again, we omit the details of an extension of the concept to the improper integral

$$\int_{-\infty}^{+\infty} g(x)dF(x).$$

■ **EXAMPLE 8.14**

Let us consider a special case where F_X is a cdf of a discrete random variable with the set of possible values $x_1, x_2, \ldots$ and the corresponding probabilities

$P\{X = x_i\} = p_i$. As we know from Chapter 6, the cdf of X is a step function that is constant between points x_i, and its steps equal p_i at points x_i.

The approximating (lower) sum $\underline{S}_n$ equals

$$\underline{S}_n = \sum g_{n,k}^{(-)}[F_X(x_{n,k}) - F_X(x_{n,k-1})].$$

Here the difference $F_X(x_{n,k}) - F_X(x_{n,k-1})$ of values of F_X at two consecutive points of the nth partition is zero if the interval $(x_{n,k-1}, x_{n,k}]$ does not contain any point x_i of the increase of F_X. When the partition becomes finer as $n \to \infty$, the only terms in $\underline{S}_n$ that remain will be those corresponding to intervals covering the points x_i, and the differences $F_X(x_{n,k}) - F_X(x_{n,k-1})$ for intervals covering point x_i will converge to p_i. Under some regularity assumptions (e.g., continuity of g) the values $g_{n,k}^{(-)}$ correspond ing to nonzero terms will converge to corresponding values $g(x_i)$, and the limit will be

$$\int_{-\infty}^{+\infty} g(x)dF_X(x) = \sum_i g(x_i)p_i. \tag{8.9}$$

From formula (8.9) for the special case $g(x) = x$, we see that

$$\int_{-\infty}^{+\infty} xdF_X(x) = \sum x_i P\{X = x_i\},$$

provided that $\sum |x_i| P\{X = x_i\} < \infty$ (which turns out to be the condition for existence of the improper integral $\int_{-\infty}^{+\infty} xdF_X(x)$).

We have the following theorem:

Theorem 8.3.1 *If X is a discrete variable with cdf F_X and if $E(X)$ exists, then*

$$E(X) = \int_{-\infty}^{+\infty} xdF_X(x).$$

■ **EXAMPLE 8.15**

Consider the R-S integral in the case where the function F is a cdf of a continuous random variable so that $F'(x) = f(x)$, where f is the density of random variable X. In this case the lower approximating sum can be written, using the mean value theorem, as

$$\underline{S}_n = \sum_k g_{n,k}^{(-)}[F(x_{n,k}) - F(x_{n,k-1})] = \sum_k g_{n,k}^{(-)} f(u_{n,k})(x_{n,k} - x_{n,k-1}),$$

where $u_{n,k}$ is a point between $x_{n,k-1}$ and $x_{n,k}$. Again, if g is continuous, the limiting value of $\underline{S}_n$ (and also $\overline{S}_n$) will be the integral of $g(x)f(x)$ between appropriate finite or infinite limits. Thus in this case the R-S integral becomes

$$\int g(x)dF(x) = \int g(x)f(x)dx.$$

Again, taking $g(x) = x$, we obtain

Theorem 8.3.2 *If X is a continuous random variable with cdf F_X, and if $E(X)$ exists, then*

$$E(X) = \int_{-\infty}^{+\infty} x dF_X(x).$$

We therefore have a single expression for an expected value of a random variable, which reduces to formulas (8.2) and (8.4) in the case of discrete and continuous random variables.

Lebesque-Stieltjes Integral

The Lebesgue-Stieltjes integral is defined in very much the same way as the Lebesgue integral. If g is bounded (e.g., on interval $[a, b]$) so that $A \leq g(x) \leq B$ for some A and B, we use the sequence of partitions

$$A = y_{n,0} < y_{n,1} < \cdots < y_{n,2^n} = B$$

and take the (lower) approximating sum as

$$\underline{S}_n = \sum_{k=0}^{2^n} y_{n,k} l_F(C_{n,k}),$$

where $C_{n,k} = \{x : y_{n,k} \leq g(x) < y_{n,k+1}\}, k = 0, 1, \ldots, 2^n - 1, C_{n,2^n} = \{x : g(x) = B\}$, and l_F is the generalized length of a set. This length is induced by the cdf F, in the sense explained above (see the construction of the Riemann-Stieltjes integral). The upper sum $\overline{S}_n$ is defined similarly, and the common limit (if it exists) of these two sequences of sums is the Lebesgue-Stieltjes integral

$$\int_a^b g(x) dF(x).$$

Again, we omit the details of the extension to integrals over an infinite interval.

As before, if the function g is R-S integrable with respect to F, then it is also L-S integrable, and the integrals are equal. If F corresponds to a discrete distribution with masses p_i at points x_i, then

$$\int_{-\infty}^{+\infty} g(x) dF(x) = \sum_i g(x_i) p_i,$$

provided that $\sum |g(x_i)| p_i < \infty$. If F is a cdf of continuous distribution with density f, then

$$\int_{-\infty}^{+\infty} g(x) dF(x) = \int_{-\infty}^{+\infty} g(x) f(x) dx,$$

provided that $\int_{-\infty}^{+\infty} |g(x_i)| f(x) dx < \infty$.

Lebesque Integral: General Case

We now outline the second direction of generalization, applicable only to the case of the Lebesgue (or Lebesgue-Stieltjes) integral.

This generalization, by far the most important and profound extension of the concept of the Lebesgue integral, is based on the observation that approximating sums (8.8) make sense also for functions g that are defined on sets other than the real line. In fact we can consider real-valued functions g defined on an arbitrary set (in particular, the sample space $\mathcal{S}$). The sets $C_{n,k}$ are then subsets of sample space (hence events), and we let probability P play the role of length l.

In more familiar notation, let X be a random variable defined on sample space $\mathcal{S}$. Assume first that $X \geq 0$. Similarly to (8.8), we define the approximating sums as

$$S_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} P\left\{s : \frac{k}{2^n} \leq X(s) < \frac{k+1}{2^n}\right\} \tag{8.10}$$

Observe that as n increases, the partitions become finer and also their range increases. Observe also that according to the comment made in the footnote on the opening page of Chapter 6, the probabilities in the sum (8.10) are well defined (i.e., arguments of probabilities are events, if X is a random variable).

We will omit the details of construction, which can be found in any advanced text on probability. Roughly, two properties are shown:

1. For every random variable $X \geq 0$, the sums S_n converge to a finite or infinite limit.

2. This limit exists and is the same if instead of $y_{n,k} = k/2^n$ we take any other sequence of partitions, provided that these partitions become finer and eventually cover the whole set $[0, \infty)$.

Consequently, we define the integral as the limit of sums (8.10):

$$\int_S X dP = \lim S_n.$$

The extension to arbitrary random variables (not necessarily nonnegative) now consists in defining

$$X^+ = \max(X, 0) \quad \text{and} \quad X^- = \max(-X, 0)$$

so that $X = X^+ - X^-$ is represented as a difference of two nonnegative functions. One then puts

$$\int_S X dP = \int_S X^+ dP - \int_S X^- dP,$$

provided that at least one of the integrals on the right-hand side is finite (so that only indeterminate case $\infty - \infty$ is ruled out).

We now have, more formally, the following definition:

Definition 8.3.1 The *expectation* $E(X)$ of a random variable X is defined as the Lebesgue integral of X:

$$E(X) = \int_S X dP,$$

provided $E|X| = \int_S |X| dP < \infty$. If $\int_S X^+ dP = \infty$ while we have $\int_S X^- dP < \infty$, we define $E(X) = \infty$ (and similarly for $E(X) = -\infty$). If $\int_S X^+ dP = \int_S X^- dP = \infty$, we say that expectation of X does not exist. □

The expectation of a function of two (or more than two) random variables is defined in a similar way. To sketch the construction, consider a nonnegative function $g(X, Y)$ of a pair of random variables (X, Y). The sums approximating the integral $E[g(X, Y)]$ are of the following form, similar to (8.10),

$$S_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} P\left\{ s : \frac{k}{2^n} \le g(X(s), Y(s)) < \frac{k+1}{2^n} \right\}.$$

The rest of the construction is analogous to the case of a single random variable, leading to the definition of the expectation $E[g(X, Y)]$.

8.4 PROPERTIES OF EXPECTATION

The identification of the expected value of a random variable X with Lebesgue integral of X allows us to formulate the properties of the latter as the properties of the expectation. Again, for the proofs we refer the reader to any advanced textbook on probability.

Theorem 8.4.1 *The expectation of random variables has the following properties:*
(a) *Linearity. If $E(X)$ exists, then for all α, β,*

$$E(\alpha X + \beta) = \alpha E(X) + \beta. \tag{8.11}$$

(b) *Nonnegativity. If $X \ge 0$, then $E(X) \ge 0$.*

(c) *Modulus inequality. For any random variable X,*

$$|E(X)| \le E(|X|).$$

We list here two of the most important consequences of this theorem. By putting $\beta = 0$ in (8.11), we have $E(\alpha X) = \alpha E(X)$. This means, in particular, that if we change the units of measurement of X, then the expectation of X changes in the same way as X. On the other hand, by putting $\alpha = 0$, we obtain the property $E(\beta) = \beta$. So the expectation of a random variable equal to a constant β is equal to the same constant.

In the future, we will often use the following theorem, which gives a sufficient condition for the existence of an expected value.

Theorem 8.4.2 *If $|X| \le Y$, where $E(Y) < \infty$, then $E(X)$ exists and is finite.*

Thus, to prove that the expectation of a random variable exists, it suffices to find another random variable Y that dominates it, and whose expectation exists. The domination $|X| \leq Y$ means that $|X(s)| \leq Y(s)$ for every point s in sample space $\mathcal{S}$. Similarly, the symbol $\lim_{n\to\infty} X_n = X$ is understood as a pointwise limit of random variables as functions on $\mathcal{S}$, that is, $\lim_{n\to\infty} X_n(s) = X(s)$ for every $s \in \mathcal{S}$.

Next we give two principal theorems that connect expectation with convergence, allowing passing to the limit under the integral (expectation) sign.

Theorem 8.4.3 (Monotone Convergence Theorem) *If $X_1, X_2, \ldots$ is a sequence of random variables such that $0 \leq X_1 \leq X_2 \leq \cdots$ and*

$$\lim_{n\to\infty} X_n = X,$$

then

$$\lim_{n\to\infty} E(X_n) = E(X).$$

Theorem 8.4.4 (Dominated Convergence Theorem) *Let $X_1, X_2, \ldots$ be a sequence of random variables satisfying the condition*

$$\lim_{n\to\infty} X_n = X.$$

Moreover, assume that there exists a random variable Y with $E(Y) < \infty$ such that $|X| \leq Y$ and $|X_n| \leq Y$ for $n = 1, 2, \ldots$. Then

$$\lim_{n\to\infty} E(X_n) = E(X).$$

In both theorems the assertion is that $\lim E(X_n) = E(\lim X_n)$, which means that we may interchange the order of integration and passage to the limit.

Instead of presenting formal proofs, we will show the necessity of the assumptions with examples.

■ **EXAMPLE 8.16**

We will now show by an example that the monotonicity alone is not enough for the assertion in the monotone convergence theorem; nonnegativity is essential. We let $\mathcal{S} = (0, 1)$ and let s be chosen according to the uniform distribution. We let X_n be defined as

$$X_n(s) = \begin{cases} -1/s & \text{if } 0 < s < 1/n \\ 0 & \text{if } 1/n \leq s < 1. \end{cases}$$

For each fixed s, the sequence $\{X_n(s)\}$ is of the form

$$-1/s, -1/s, \ldots, -1/s, 0, 0, \ldots,$$

hence is monotone, and $\lim X_n \equiv 0$. However, $EX_n = \int_0^{1/n} (-1/s)ds = -\infty$, while the integral of the limit random variable $X \equiv 0$ is $EX = 0$.

■ **EXAMPLE 8.17**

We will show that the existence of the bound Y with finite expectation is essential in the dominated convergence theorem. Again, let $\mathcal{S} = (0, 1)$, with P being uniform on $\mathcal{S}$. Now

$$X_n(s) = \begin{cases} 4n^2 s & \text{for } 0 < s < 1/2n \\ 4n(1 - ns) & \text{for } 1/2n \leq s < 1/n \\ 0 & \text{for } 1/n \leq s < 1. \end{cases}$$

For every fixed s, the sequence of numbers $\{X_n(s)\}$ converges to 0. Again, the limiting random variable is $X \equiv 0$, and $E(X) = 0$. But it is easy to check that $E(X_n) = 1$ for every n so that the assertion of the theorem does not hold. This time each random variable X_n is bounded, but there is no integrable common bound for all of them.

It is now necessary to connect two definitions of expectation of a random variable X: as a Lebesgue integral of X, and as a Riemann–Stieltjes integral of function $g(x) \equiv x$ with respect to cdf of X.

One may raise a doubt: Starting with Chapter 1, we stress the fact that for the same phenomenon (for the same random variable), the sample space $\mathcal{S}$ can be chosen in a number of different ways. But if we choose different sample spaces for describing the same random variable X, how can we guarantee that the integral of X is the same, regardless of the choice of $\mathcal{S}$?

The answer lies in the next theorem, which provides methods of computing Lebesgue integrals by reducing them to Lebesgue–Stieltjes and Riemann–Stieltjes integrals.

Theorem 8.4.5 *Let X be a random variable defined on probability space $\mathcal{S}$, and let $F(x)$ be its cdf. If $\int_{\mathcal{S}} |X| dP < \infty$, then*

$$\int_{\mathcal{S}} X dP = \int_{-\infty}^{+\infty} x dF(x). \tag{8.12}$$

More generally, if g is a real function such that $\int_{\mathcal{S}} |g(X)| dP < \infty$, then

$$E[g(X)] = \int_{\mathcal{S}} g(X) dP = \int_{-\infty}^{+\infty} g(x) dF(x). \tag{8.13}$$

The right-hand sides of formulas (8.12) and (8.13) provide means of computing the expectations. In the case of discrete and continuous random variables, formula (8.12) reduces to (8.2) or (8.4). Similarly formula (8.13) reduces to

$$Eg(X) = \sum_i g(x_i) P\{X = x_i\} \tag{8.14}$$

and

$$Eg(X) = \int_{-\infty}^{+\infty} g(x) f(x) dx. \tag{8.15}$$

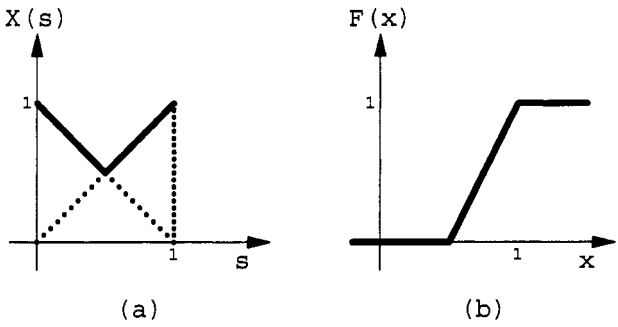


Figure 8.4 Graph of $X = \max(s, 1-s)$ and its cdf

■ **EXAMPLE 8.18**

Formulas (8.12) and (8.13) are sometimes referred to as "theorems of the unconscious statistician." The reason for the name is that in calculating expectations of some random variables, a statistician chooses either the left or right side of these formulas, often without being aware that using the other side may occasionally be simpler. We illustrate the situation by carrying out the calculations for both sides of (8.15).

Imagine that we have a stick of length 1, and we break it at a random point (i.e., the breaking point is chosen accordingly to the uniform distribution). What is the expected length of the longer of the two parts?

If we take the interval [0, 1] as the sample space, with measure P being uniform (i.e., probabilities are proportional to lengths), then the length of the longer part of the stick, if the break occurs at s, is

$$X = \max(s, 1-s).$$

The graph of X is presented at Figure 8.4(a). It is clear that $E(X) = \int_S X dP = \int_0^1 \max(s, 1-s) ds = 3/4$. This can be obtained by actually computing the integral or by observing that the area under the curve in Figure 8.4(a) is three quarters of the square of side 1.

We can also find the cdf of X and, realizing that X is a continuous random variable, use the right-hand side of (8.15). Since $1/2 \leq X \leq 1$, we can write

$$\begin{aligned} P\{X \leq x\} &= P\{\max(s, 1-s) \leq x\} = P\{s \leq x, 1-s \leq x\} \\ &= P\{1-x \leq s \leq x\} = x - (1-x) = 2x - 1. \end{aligned}$$

Consequently,

$$F(x) = \begin{cases} 0 & x \leq \frac{1}{2} \\ 2x-1 & \frac{1}{2} \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

[see Figure 8.4(b)], and

$$E(X) = \int_{-\infty}^{+\infty} x dF(x) = \int_{1/2}^{1} x dF(x) = \int_{1/2}^{1} x 2dx = \frac{3}{4}.$$

Let us now consider functions of two random variables—we will only discuss discrete and continuous bivariate distributions as the two most important cases. Let $g(x, y)$ be a real function such that $E|g(X, Y)| < \infty$. Suppose first that the distribution of (X, Y) is concentrated on $A \times B$, where $A = \{x_1, x_2, \ldots\}$ and $B = \{y_1, y_2, \ldots\}$. Let $p_{ij} = P\{X = x_i, Y = y_j\}$. Then

$$E[g(X, Y)] = \sum_{ij} g(x_i, y_j) p_{ij}. \tag{8.16}$$

Since $E|g(X, Y)| = \sum |g(x_i, y_j)| p_{ij} < \infty$, the sum (8.16) does not depend on the order of summation. So one can choose the order that leads to simpler calculations:

$$\begin{aligned} E[g(X, Y)] &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} g(x_i, y_j) p_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} g(x_i, y_j) p_{ij} \\ &= \sum_{k=1}^{\infty} \sum_{r=1}^{k} g(x_r, y_{k-r}) p_{r,k-r}, \end{aligned}$$

and so on. The choice depends on the function g and probabilities p_{ij}.

Let (X, Y) have the joint density $f(x, y)$. We then have

Theorem 8.4.6 *If* $E|g(X, Y)| < \infty$, *then*

$$\begin{aligned} E[g(X, Y)] &= \int_{R^2} \int g(x, y) f(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} g(x, y) f(x, y) dx \right] dy \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} g(x, y) f(x, y) dy \right] dx. \end{aligned} \tag{8.17}$$

The last two expressions provide the computational formulas that can be used in practice, since they reduce the computation of a double (two-dimensional) integral to an iteration of two single (one-dimensional) integrals.

As an illustration of applicability of Theorem 8.4.6, we give the proof (in the case of continuous random variables) of the following theorem:

Theorem 8.4.7 *Assume that* $E|X| < \infty$ *and* $E|Y| < \infty$. *Then*

$$E(X + Y) = E(X) + E(Y).$$

Proof. Let $f(x,y)$ be the density of (X,Y), and let f_X and f_Y be the marginal densities of X and Y, respectively. Letting $g(x,y) = x + y$ in (8.17), we have

$$\begin{aligned} E(X+Y) &= \int_{R^2}\int (x+y)f(x,y)dxdy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)dxdy + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yf(x,y)dxdy \\ &= \int_{-\infty}^{+\infty} x\left[\int_{-\infty}^{+\infty} f(x,y)dy\right]dx + \int_{-\infty}^{+\infty} y\left[\int_{-\infty}^{+\infty} f(x,y)dx\right]dy \\ &= \int_{-\infty}^{+\infty} xf_X(x)dx + \int_{-\infty}^{+\infty} yf_Y(y) = E(X) + E(Y). \end{aligned}$$

The order of integration can be changed because of the assumption that $\int_{-\infty}^{+\infty} |x| f_X(x)dx < \infty$, $\int_{-\infty}^{+\infty} |y| f_Y(y)dy < \infty$. □

The property of additivity extends immediately to any finite number of random variables with finite expectations:

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n),$$

or, combining it with Theorem 8.4.1(a), we have

$$E(\alpha_1 X_1 + \cdots + \alpha_n X_n + \beta) = \sum_{i=1}^{n} \alpha_i E(X_i) + \beta.$$

Finally, the nonnegativity property (b) in Theorem 8.4.1 implies that expectation is monotone: ff X and Z have finite expectations, then $Y \leq Z$ implies $E(Y) \leq E(Z)$. To see this, we write $Z = Y + (Z - Y)$, where now $Z - Y \geq 0$. We have

$$E(Z) = E(Y) + E(Z - Y) \geq E(Y).$$

We have also the following important theorem:

Theorem 8.4.8 *If X and Y are independent random variables such that $E|XY| < \infty$, then*

$$E(XY) = E(X)E(Y).$$

Proof. We give the proof only for the case of continuous random variables.

$$\begin{aligned} E(XY) &= \int_{R^2}\int xyf(x,y)dxdy = \int_{R^2}\int xyf_X(x)f_Y(y)dxdy \\ &= \left(\int_{-\infty}^{+\infty} xf_X(x)dx\right) \times \left(\int_{-\infty}^{+\infty} yf_Y(y)dy\right) = E(X)E(Y). \end{aligned}$$ □

Replacing the double integral by the iterated integrals, as well as the change of the order of integration (or summation), is very often taken for granted. Actually, the fact that iterated integrals are equal to one another is not a "law of nature," as some

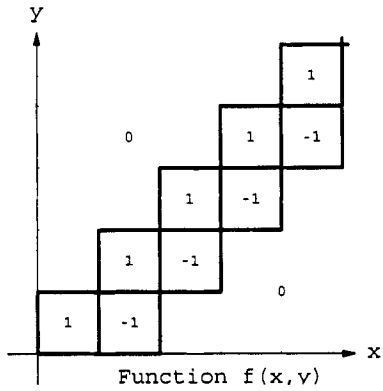


Figure 8.5 Nonintegrable function whose iterated integrals exist and are not equal

are inclined to believe. It is a fact that is true under specific assumptions, namely under the existence of the double integral.

We will not provide here the precise statement of the relevant theorem (see the Fubini theorem or Lebesgue integration in any advanced textbook on probability). Instead, we will give a simple example that shows that the iterated integrals need not be equal.

■ **EXAMPLE 8.19**

Consider $f(x, y)$ described in Figure 8.5. Obviously here $\int_0^\infty f(x, y)dx = 0$ for every y, so $\int_0^\infty \int_0^\infty f(x, y)dxdy = 0$. On the other hand, $\int_0^\infty f(x, y)dy$ is 1 for $0 \leq x \leq 1$ and 0 otherwise, so $\int_0^\infty \int_0^\infty f(x, y)dxdy = 1$.

It ought to be clear that the reason why there is a difference between the values of two iterated integrals is that the double integral (the sum of infinitely many "volumes" + 1 and infinitely many "volumes" -1) does not exist.

■ **EXAMPLE 8.20**

Consider now a function $h(x, y)$ similar to that of Example 8.19 given in Figure 8.6. In this case we have $\int_0^\infty h(x, y)dx = \int_0^\infty h(x, y)dy = 0$, so both iterated integrals exist and are equal zero. Still the double integral does not exist, for the same reason as in Example 8.19.

Examples 8.19 and 8.20 show that if the double integral does not exist, we cannot say anything about equality of iterated integrals.

PROBLEMS

8.4.1 Show that $E(X - E(X)) = 0$.

8.4.2 Assume that X has density $f(x) = ax + bx^3$ for $0 \leq x \leq 2$ and $f(x) = 0$ otherwise. Find a and b if $E(X^2) = 2.5$.

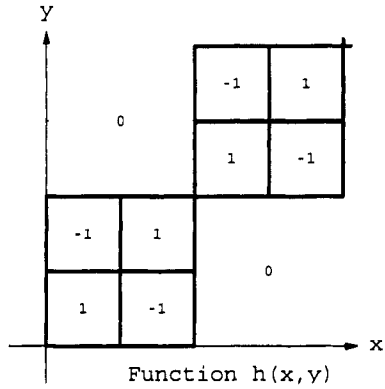


Figure 8.6 Nonintegrable function whose iterated integrals exist and are equal

8.4.3 Find $E(X)$, $E(1/X)$, and $E(2^X)$ if $X = 1, 2, \ldots, 8$ with equal probabilities 1/8.

8.4.4 Let X be a random variable with density $f(x) = 1/2$ for $-1 \leq x \leq 1$ and $f(x) = 0$ otherwise. Find: **(i)** $E(X)$. **(ii)** $E(X^2)$. **(iii)** $E(2X - 3)^2$.

8.4.5 Let X have the density

$$f(x) = \begin{cases} cx & 0 \leq x \leq 1 \\ c(2-x) & 1 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

Find: **(i)** c. **(ii)** $E(X)$. **(iii)** $E(2 - X)^3$. **(iv)** $E[1/(2 - X)]$.

8.4.6 An urn contains w white and r red balls. We draw $n \leq r$ balls from the urn without replacement, and we let X be the number of red balls drawn. Find $E(X)$, by defining indicator variables: **(i)** $X_1, \ldots, X_n$ such that $X = X_1 + \cdots + X_n$. **(ii)** $Y_1, \cdots, Y_r$ such that $X = Y_1 + \cdots + Y_r$.

8.4.7 A cereal company puts a plastic bear in each box of cereal. Every fifth bear is red. If you have three red bears, you get a free box of the cereal. If you decide to keep buying this cereal until you get one box free, how many boxes would you expect to buy before getting a free one? [*Hint:* Represent the answer as $E(X_1 + X_2 + X_3)$ where X_i is the number of boxes you buy after getting the $(i - 1)$st red bear and until getting ith red bear.]

8.4.8 Show that if X is such that $P(a \leq X \leq b) = 1$, then $E(X)$ exists and $a \leq E(X) \leq b$.

8.4.9 We say that X is *stochastically smaller* than Y ($X \leq_{st} Y$) if $P\{\leq t\} \geq P\{Y \leq t\}$ for all t. Show that if X and Y have finite expectations and $X \leq_{st} Y$, then $E(X) \leq E(Y)$. (*Hint:* Start with nonnegative X and $Y$and use Theorem 8.2.2. Then use the decomposition into a positive and negative part.) Show also that the converse assertion is false: there exist random variables X and Y such that $E(X) < E(Y)$ and X is not stochastically smaller than Y.

8.5 MOMENTS

We begin with the following definition.

Definition 8.5.1 For any random variable X, the expectation of X^n, if it exists, will be called the nth *ordinary moment* (or *the moment of the order* n) of X and denoted

$$m_n = E(X^n). \qquad \square$$

The nth moment, m_n, exists if $E|X|^n < \infty$. The moment of the order 0 always exists and equals 1, while m_1 is simply the expectation of X. If the nth moment exists, it may be computed from the formula

$$m_n = E(X^n) = \int_{-\infty}^{+\infty} x^n dF(x),$$

where F is the cdf of X; this formula is obtained by substituting $g(x) = x^n$ in (8.13). Observe that any bounded random variable X, that is, a random variable such that $P(|X| \leq M) = 1$ for some $M < \infty$, has finite moments of any order. This follows at once from Theorem 8.4.2, and in this case we have $m_n \leq M^n$.

■ **EXAMPLE 8.21**

Let X have POI(λ) distribution. Then

$$m_1 = E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

On the other hand,

$$\begin{aligned} m_2 = E(X^2) &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} [k(k-1) + k] \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda^2 + \lambda, \end{aligned}$$

since the second sum equals $m_1 = \lambda$.

■ **EXAMPLE 8.22**

Let X have the U[a, b] distribution. Then its density is $f(x) = 1/(b-a)$ for $a \leq x \leq b$ and $f(x) = 0$ otherwise so that

$$m_n = \int_a^b x^n \times \frac{dx}{b-a} = \frac{1}{n+1} \times \frac{b^{n+1} - a^{n+1}}{b-a}.$$

When the distribution is symmetric around 0, then $a = -b$, and we have $m_{2k} = b^{2k}/(2k+1)$ while $m_{2k+1} = 0$.

Before introducing the concepts which will illustrate the usefulness of the notion of moments, let us introduce some definitions pertaining to a variety of types of moments.

Definition 8.5.2 An *absolute* moment of order n is defined as

$$\beta_n = E(|X|^n). \qquad \square$$

The existence of an absolute moment of a given order implies the existence of an "ordinary" moment m_n of the same order. Note that the order of an absolute moments need not be an integer. The same applies to ordinary moments of positive random variables.

Definition 8.5.3 Ordinary moments of the random variable $Y = X - E(X)$, where $E(X) = m_1$, are called *central moments* of X so that

$$\gamma_n = E[(X - m_1)^n]. \qquad \square$$

Clearly, the first central moment γ_1 is always equal to 0.

Definition 8.5.4 A *factorial* moment of X of order n *is defined as*

$$\pi_n = E[X(X-1)\cdots(X-n+1)]. \qquad \square$$

We will now prove the following:

Theorem 8.5.1 *If an absolute moment of order $\alpha > 0$ exists, then all moments (ordinary, absolute, central, and factorial) of orders $r \leq \alpha$ exist.*

Proof. We will prove first that if $E(|X|^\alpha) < \infty$, then $E(|X|^r) < \infty$ for all $r < \alpha$. Clearly, if $|X| \geq 1$, then $|X|^r \leq |X|^\alpha$, while if $|X| < 1$, then $|X|^r \leq 1$. Consequently,

$$|X|^r \leq \max(1, |X|^\alpha), \tag{8.18}$$

and (see Theorem 8.4.2) it remains to be proved that the right-hand side of (8.18) has finite expectation.

Now, if $E|X|^\alpha < \infty$, then

$$E(|X|^\alpha) = \int_{-1}^{1} |x|^\alpha dF(x) + \int_{|x|\geq 1} |x|^\alpha dF(x),$$

and it follows that $\int_{|x|\geq 1} |x|^\alpha dF(x) < \infty$ as a difference of two finite quantities. Thus we can write

$$E(|X|^r) \leq E\{\max(1, |X|^\alpha)\} = \int_{-1}^{1} dF(x) + \int_{|x|\geq 1} |x|^\alpha dF(x) < \infty.$$

It remains to prove that if $E|X|^\alpha < \infty$, then all other types of moments of order α exist. This is true for ordinary moments (by definition). Regarding central and factorial moments of order n, they are linear combinations of ordinary moments of order $k \leq n$, which proves the theorem. $\square$

The following theorem will be given without proof:

Theorem 8.5.2 (Liapunov Inequality) *If* $0 < \alpha < \beta < \infty$, *then*

$$\{E(|X|^{\alpha})\}^{1/\alpha} \leq \{E(|X|^{\beta})\}^{1/\beta}.$$

Note that the first part of the proof of Theorem 8.5.1 is an immediate consequence of the Liapunov inequality.

We will now introduce a function that will be a valuable tool in analyzing random variables and limiting behavior of their sums.

Definition 8.5.5 The function of real variable t defined as

$$m_X(t) = E(e^{tX})$$

is called the *moment generating function* (mgf) of a random variable X. □

For any random variable, its mgf exists for $t = 0$. If X is a positive random variable, and the mgf of X exists for some t_0, then $m_X(t)$ exists for all $t \leq t_0$. This fact follows from Theorem 8.4.2.

■ **EXAMPLE 8.23**

If X has the BIN(n, p) distribution, then

$$m_X(t) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k q^{n-k} = (pe^t + q)^n$$

so that $m_X(t)$ exists for all t.

■ **EXAMPLE 8.24**

If X has EXP(λ) distribution, then we can write

$$m_X(t) = \int_0^{\infty} e^{tx} \times \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx = \frac{\lambda}{\lambda - t}, \tag{8.19}$$

provided that $t < \lambda$. For $t \geq \lambda$ the integral in (8.19) diverges.

Let us explore some of the properties of moment generating functions. First, observe that the concept of the mgf is connected with a distribution rather than with a random variable: two different random variables with the same distribution will have the same mgf.

The name "moment generating function" is related to the following theorem:

Theorem 8.5.3 *Let* X *be a random variable with the mgf* $m_X(t)$, *assumed to exist in some neighborhood of* $t = 0$. *If* $E(|X|^n) < \infty$, *then for* $k = 1, \ldots, n$, *the* k*th moment of* X *is given by the formula*

$$m_n = \frac{d^n}{dt^n} m_X(t)\bigg|_{t=0}.$$

Proof. Differentiating formally the expression for the mgf under the sign of expectation (i.e., under the integral or summation sign), we obtain

$$m_X'(t) = E(Xe^{tX}),\quad m_X''(t) = E(X^2e^{tX}),\ldots,\frac{d^k}{dt^k}m_X(t) = E(X^ke^{tX}),$$

for $k = 1, 2, \ldots n$. Substitution of $t = 0$ gives the required result.

The validity of this argument depends crucially on whether or not formal differentiation under the integral sign is allowed. The answer is positive if the corresponding derivatives are absolutely integrable for $t = 0$, which is ensured by the existence of the nth absolute moment. □

■ **EXAMPLE 8.25**

Since the mgf of random variable with BIN(n, p) distribution is $m_X(t) = (pe^t + q)^n$, we have $m_X'(t) = n(pe^t + q)^{n-1}pe^t$ while $m_X''(t) = n(n-1)(pe^t + q)^{n-2}(pe^t)^2 + n(e^t + q)^{n-1}pe^t$. Thus $E(X) = m_X'(0) = np$ and $E(X^2) = m_X''(0) = n(n-1)p^2 + np$.

■ **EXAMPLE 8.26**

The moment generating function of the U$[a, b]$ distribution is

$$m_X(t) = \int_a^b e^{tx} \times \frac{1}{b-a}dx = \frac{e^{bt} - e^{at}}{t(b-a)}. \tag{8.20}$$

Determination of the values of derivatives of $m_X(t)$ at $t = 0$ requires repeated usage of the de L'Hospital rule. We can, however, expand exponentials into power series, and after some algebra we obtain

$$m_X(t) = 1 + \frac{1}{2!}\frac{b^2 - a^2}{b-a}t + \frac{1}{3!}\frac{b^3 - a^3}{b-a}t^2 + \cdots.$$

This is a Taylor expansion of $m_X(t)$ about $t = 0$. So we must have

$$\begin{aligned} m_n &= m_X^{(n)}(0) = \frac{n!}{(n+1)!}\frac{b^{n+1} - a^{n+1}}{b-a} \\ &= \frac{1}{n+1}[b^n + b^{n-1}a + b^{n-2}a^2 + \cdots + a^n]. \end{aligned}$$

Next we will prove

Theorem 8.5.4 *If X is a random variable with mgf $m_X(t)$, then the random variable $Y = \alpha X + \beta$ has the mgf*

$$m_Y(t) = e^{\beta t}m_X(\alpha t). \tag{8.21}$$

Proof. From the properties of the expectation,

$$m_Y(t) = E(e^{tY}) = E(e^{t(\alpha X+\beta)}) = E(e^{(\alpha t)X} \times e^{\beta t}) = e^{\beta t}m_X(\alpha t).$$

■ **EXAMPLE 8.27**

Let us find the mgf's of random variables U and V, with distributions U[0, 1] and U[−1, 1], respectively.

If X has a U[a, b] distribution, and hence has the mgf (8.20), then $U = (X - a)/(b - a)$ has a distribution uniform on [0, 1]. Using formulas (8.20) and (8.21) for $\alpha = 1/(b - a)$ and $\beta = -a/(b - a)$, we obtain

$$m_U(t) = e^{-at/(b-a)} m_X\left(\frac{t}{b-a}\right) = \frac{e^t - 1}{t}. \tag{8.22}$$

Next, if V has a U[−1, 1] distribution, then $V = 2U - 1$. Taking $\alpha = 2$ and $\beta = -1$ and using (8.22), we get

$$m_V(t) = e^{-t} \times \frac{e^{2t} - 1}{2t} = \frac{e^t - e^{-t}}{2t} = \frac{\sinh t}{t}.$$

Obviously the expressions for these mgf's can be obtained directly from the definition.

■ **EXAMPLE 8.28**

The moment generating function of a N(μ, σ^2) distribution can be found by first finding the mgf of a standard normal random variable Z ∼ N(0, 1) as

$$m_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-\frac{z^2}{2}} dz = \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1/2)(z-t)^2} dz = e^{t^2/2}.$$

Then, for any $X \sim \mathrm{N}(\mu, \sigma^2)$, $X = \mu + \sigma Z$, and based on Theorem 8.5.4 we obtain

$$m_X(t) = e^{\mu t + \sigma^2 t^2/2}. \tag{8.23}$$

We now prove

Theorem 8.5.5 *If X and Y are independent random variables with moment generating functions $m_X(t)$ and $m_Y(t)$, respectively, then the mgf of $X + Y$ is*

$$m_{X+Y}(t) = m_X(t) m_Y(t).$$

Proof. Observe that random variables e^{tX} and e^{tY} are independent for each t, so by Theorem 8.4.8 we have

$$m_{X+Y}(t) = (e^{t(X+Y)}) = E(e^{tX}) E(e^{tY}) = m_X(t) m_Y(t). \qquad \square$$

Before proceeding with examples, we state one more important theorem. Its proof is beyond the scope of this book.

Theorem 8.5.6 *If X and Y are two random variables such that their moment generating functions $m_X(t)$ and $m_Y(t)$ coincide in some neighborhood of the point $t = 0$, then X and Y have the same distribution.*

Theorem 8.5.6 asserts that if two mgf's agree in some neighborhood of $t = 0$, then they agree for all t (for which they are defined), and that an mgf determines uniquely the distribution. In other words, random variables with different distributions must have different mgf's.

■ **EXAMPLE 8.29**

Let X and Y be independent and have Poisson distributions with means λ_1 and λ_2, respectively. Let us first determine the mgf's of X and of Y. We have

$$m_X(t) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda_1^k}{k!} e^{-\lambda_1} = e^{-\lambda_1} \sum_{k=0}^{\infty} \frac{(\lambda_1 e^t)^k}{k!} = e^{\lambda_1(e^t-1)}.$$

Similarly $m_Y(t) = e^{\lambda_2(e^t-1)}$, and by Theorem 8.5.6,

$$m_{X+Y}(t) = m_X(t)m_Y(t) = e^{(\lambda_1+\lambda_2)(e^t-1)}. \tag{8.24}$$

We recognize (8.24) as the mgf of the Poisson distribution. In view of Theorem 8.5.6, $X + Y$ has a Poisson distribution with mean $\lambda_1 + \lambda_2$.

The main disadvantage of moment generating functions is that they may not exist for any $t \neq 0$. There exist random variables X such that the random variable e^{tX} has no expectation for any $t \neq 0$. This restricts the usefulness of mgf's as a tool (to be explored in Chapter 10) for obtaining limit theorems. For proofs using mgf's to be valid, they have to be confined to classes of random variables for which the mgf's exist in some neighborhood of $t = 0$. The corresponding theorems, however, are typically valid without this assumption, and the proofs usually require nothing more than replacing the mgf's with so-called *characteristic functions*. Although we will not use characteristic functions in this book, it is worthwhile to provide their definition and simplest properties. The concepts below require a rudimentary knowledge of complex numbers.

Definition 8.5.6 Let X be a random variable with the cdf F. The function φ_X of real argument t ($-\infty < t < +\infty$) defined as

$$\varphi_X(t) = E(e^{itX}) = \int e^{itx} dF(x) \tag{8.25}$$

is called the *characteristic function* (chf) of X (or of the cdf F). □

In the definition above i is the imaginary unit (i.e., $i^2 = -1$). From the formula $e^{i\xi} = \cos\xi + i\sin\xi$, we obtain

$$\varphi_X(t) = E\{\cos(tX)\} + iE\{\sin(tX)\}, \tag{8.26}$$

so $\varphi_X(t)$ is a complex-valued function of a real argument. Since $|e^{itX}| = \cos^2(tX) + \sin^2(tX) = 1$, the expectation in (8.25) exists for all t, so the characteristic functions always exist.

Below we list some basic properties of characteristic functions; the proofs are left as exercises. In particular, Theorems 8.5.3, 8.5.4, and 8.5.5 carry over to the case of characteristic functions almost without any change. In the following theorems, $\varphi_X, \varphi_Y, \ldots$ are chf's of random variables $X, Y, \ldots$:

Theorem 8.5.7 *For every random variable X, its chf $\varphi_X(t)$ is uniformly continuous and satisfies the conditions*

$$\varphi_X(0) = 1, \quad |\varphi_X(t)| \leq 1, \tag{8.27}$$

for every real t. Moreover, for any real a, b,

$$\varphi_{aX+b}(t) = \varphi_X(at)e^{ibt}. \tag{8.28}$$

In particular, for $a = -1, b = 0$, using (8.26,) we obtain

$$\varphi_{-X}(t) = \varphi(-t) = \overline{\varphi(t)}, \tag{8.29}$$

where $\overline{z}$ stands for the complex conjugate of z.

■ **EXAMPLE 8.30**

If $X \sim \text{BIN}(n, p)$, then

$$\varphi_X(t) = E(e^{itX}) = \sum_{k=0}^{n} e^{itk}\binom{n}{k}p^k q^n = (pe^{it} + q)^n = [(p\cos t + q) + ip\sin t]^n.$$

In this case $\varphi_X(t)$ is a periodic function.

■ **EXAMPLE 8.31**

If X is uniform on $[0, 1]$, then

$$\varphi_X(t) = \int_0^1 e^{itx}dx = \frac{1}{it}(e^{it} - 1).$$

If $Y = 2X - 1$, then Y has uniform distribution on $[-1, 1]$ and (8.28) gives

$$\varphi_Y(t) = \varphi_{2X-1}(t) = \varphi_X(2t)e^{-it} = \frac{1}{2it}(e^{2it} - 1)e^{-it} = \frac{\sin t}{t}.$$

■ **EXAMPLE 8.32**

If $X \sim \text{N}(0, 1)$, then

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{itx}e^{-x^2/2}dx = \frac{e^{-t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-(1/2)(x-it)^2}dx = e^{-t^2/2}.$$

The addition of independent random variables corresponds to multiplications of their chf's.

Theorem 8.5.8 *If random variables X and Y are independent, then*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

As in the case of mgf's, characteristic functions uniquely determine the distributions:

Theorem 8.5.9 *If two cdfs F and G have the same characteristic function, then $F \equiv G$.*

Finally, the relationship between moments of X and behavior of characteristic function in the neighborhood of $t = 0$ is given by the following theorem:

Theorem 8.5.10 *If X is a random variable such that $E(|X|^k) < \infty$ for some integer $k \geq 1$, then*

$$\varphi_X(t) = \sum_{j=0}^{k} \frac{i^j}{j!} m_j t^j + o(|t|^k)$$

in some neighborhood of $t = 0$. Here $m_j = E(X^j)$, and $o(x)$ is such that $\lim_{x \downarrow 0} o(x)/x = 0$.

PROBLEMS

8.5.1 Find the moment generating function of a discrete random variable X with distribution $P\{X = k\} = 1/n, k = 0, 1, \ldots, n-1$.

8.5.2 Let X be a nonnegative integer-valued random variable. The function $g_X(s) = Es^X$, defined for $|s| \leq 1$, is called a *probability generating function*, or simply a generating function, of X. Find $g_X(s)$ for random variables with: **(i)** Geometric distribution. **(ii)** Binomial distribution. **(iii)** Poisson distribution.

8.5.3 Find the fourth factorial moment of the random variable X with a POI(λ) distribution.

8.5.4 Find the moment generating function for a random variable with a density: **(i)** $f(x) = xe^{-x}$ for $x > 0$. **(ii)** $f(x; \sigma^2) = \sqrt{2/\pi} e^{-x^2/2\sigma^2}$ for $x > 0$.

8.5.5 Let $X_1, \ldots, X_n$ be independent random variables with the common distribution N(μ, σ). Find constants α_n and β_n such that $U = (X_1 + \cdots + X_n - \alpha_n)/\beta_n$ and X_1 have the same distribution.

8.5.6 A continuous random variable X is called *symmetric about* c if its density f satisfies the condition $f(c-x) = f(c+x)$ for all x. Show that: **(i)** If X is symmetric about c and $E(X)$ exists, then $E(X) = c$. **(ii)** If X is symmetric about 0, then all moments of odd order (if they exist) are equal 0.

8.5.7 Let X be a random variable with $E(X) = \mu, \mathrm{Var}(X) = \sigma^2$ and such that the third central moment $\gamma_3 = E(X - \mu)^3$ exists. The ratio γ_3/σ^3 is called the *coefficient of skewness*. Find skewness of the following distributions: **(i)** U[0, 1]. **(ii)** $f(x) = \alpha x^{\alpha-1}$ for $0 \leq x \leq 1$ and $f(x) = 0$ otherwise ($\alpha > 1$). **(iii)** BIN($1, p$). **(iv)** POI(λ).

8.5.8 Let X be a random variable with $E(X) = \mu, \mathrm{Var}(X) = \sigma^2$ and such that $\gamma_4 = E(X - \mu)^4$ exists. Then γ_4/σ^4 is called the *coefficient of kurtosis*. Find kurtosis of the following distributions: **(i)** N(0, 1). **(ii)** N(μ, σ^2). **(iii)** BIN($1, p$). **(iv)** POI(λ).

8.5.9 Find the characteristic function of the following distributions: **(i)** POI(λ). **(ii)** GEO(p). **(iii)** EXP(λ). **(iv)** N(μ, σ^2).

8.5.10 A family $\mathcal{G}$ of distributions is said to be *closed under convolution*, if whenever independent random variables X and Y have distributions in $\mathcal{G}$; the same is true for the random variable $X + Y$. Show closeness under convolution in families of: **(i)** Poisson distributions. **(ii)** Normal distributions.

8.5.11 Show that if $\varphi(t)$ is a chf, then $|\varphi(t)|^2$ is also a chf.

8.5.12 Show that the distribution of X is symmetric around 0 if and only if $\varphi_X(t)$ is real.

8.6 VARIANCE

We will now introduce a concept that plays an extremely important role in statistical analysis.

Definition 8.6.1 If $E(X^2) < \infty$, then the second central moment of X,

$$\text{Var}(X) = E[(X - m_1)^2],$$

is called the *variance* of X. Its positive square root is called the *standard deviation* of X. □

The other symbols used for the variance are $\text{V}(X)$ and σ_X^2.

We begin by listing some basic properties of the variance and providing preliminary examples. First, observe that

$$\begin{aligned} \text{Var}(X) &= E(X^2 - 2m_1X + m_1^2) = E(X^2) - 2m_1E(X) + m_1^2 \\ &= E(X^2) - [E(X)]^2. \end{aligned} \tag{8.30}$$

Formula (8.30) gives an alternative way of calculating the variance. Note that variance is always nonnegative, as the expectation of a nonnegative random variable $(X - m_1)^2$. Consequently $E(X^2) \geq [E(X)]^2$; that is, $E(X^2)^{1/2} \geq |E(X)|$, which is a special case of the Liapunov inequality given in Theorem 8.5.2.

Theorem 8.6.1 *If* $\text{Var}(X)$ *exists, then*

$$\text{Var}(aX + b) = a^2\text{Var}(X). \tag{8.31}$$

Proof. Using the fact that $E(aX + b) = aE(X) + b$, we write

$$\begin{aligned} \text{Var}(aX + b) &= E[(aX + b)^2] - [aE(X) + b]^2 \\ &= E[a^2X^2 + 2abX + b^2] - a^2[E(X)]^2 - 2abE(X) - b^2 \\ &= a^2[E(X^2) - (E(X))^2] = a^2\text{Var}(X). \end{aligned}$$ □

■ **EXAMPLE 8.33**

Let X have the Poisson distribution with mean λ. Then the mgf of X is $m_X(t) = e^{\lambda(e^t-1)}$ (see Example 8.29). After some algebra we obtain

$$m_X''(t) = \lambda(\lambda e^t + 1)e^{\lambda(e^t-1)+t}$$

so that $E(X^2) = m_X''(0) = \lambda^2 + \lambda$. Since $E(X) = \lambda$, we obtain $\text{Var}(X) = (\lambda^2 + \lambda) - \lambda^2 = \lambda$. Thus for the Poisson distribution the mean and variance coincide.

■ **EXAMPLE 8.34**

From Example 8.25 we know that if X has BIN(n, p) distribution, then $E(X) = np$, $E(X^2) = n(n-1)p^2 + np$. Consequently, using (8.30), we obtain

$$\text{Var}(X) = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq. \tag{8.32}$$

■ **EXAMPLE 8.35**

Let us find the variance of random variable X with a U$[a, b]$ distribution. We have $f(x) = 1/(b-a)$ for $a \le x \le b$, $E(X) = (a+b)/2$, and

$$E(X^2) = \frac{1}{3} \times \frac{b^3 - a^3}{b-a} = \frac{1}{3}[b^2 + ab + a^2],$$

so $\text{Var}(X) = (b-a)^2/12$.

■ **EXAMPLE 8.36**

In Example 8.24 we found that the mgf of the EXP(λ) distribution is $m(t) = \lambda(\lambda - t)^{-1}$. Consequently $m'(t) = \lambda(\lambda - t)^{-2}$, $m''(t) = 2\lambda(\lambda - t)^{-3}$. Then $E(X) = 1/\lambda$, $E(X^2) = 2/\lambda^2$ and $\text{Var}(X) = 1/\lambda^2$, so in the exponential distribution the mean and standard deviation coincide.

To interpret the variance, let us now consider the case of a discrete random variable. The variance equals

$$\text{Var}(X) = \sum_i (x_i - m_1)^2 P(X = x_i). \tag{8.33}$$

For the variance to be small, all terms of the sum (8.33) must be small. Hence the values x_i with the large difference $|x_i - m_i|$ must have a low probability. Qualitatively speaking, a small variance means that the values of X are concentrated closely to the mean of X, so large deviations are unlikely. The following theorem provides another interpretation of the variance:

Theorem 8.6.2 *The mean square deviation from ξ, namely $E(X-\xi)^2$, is minimized at $\xi = E(X)$, and the minimal value equals* $\text{Var}(X)$.

Proof. Since $f(\xi) = E(X^2) - 2\xi E(X) + \xi^2$ represents a parabola with branches directed upward, the minimum occurs at the point ξ^* at which $f'(\xi^*) = 0$, so $\xi^* = E(X)$. □

The function $f(\xi)$ corresponds to the case where, qualitatively speaking, "small errors are almost negligible, large errors are very serious." The function $g(\xi)$ in Example 8.37 treats the seriousness of an error as proportional to its size.

■ **EXAMPLE 8.37**

A natural question that arises from problem posed in Theorem 8.6.2 is to determine ξ that minimizes

$$g(\xi) = E|X - \xi|.$$

The minimization of $g(\xi)$ is not as simple as that of $f(\xi)$. We have to turn here to the following result:

Theorem 8.6.3 *The mean absolute deviation from ξ, namely $E|X-\xi|$, is minimized at $\xi = m$, where m is the median of X.*

Proof. We will present the proof in the case of the continuous random variable X with density f. It suffices to show that $E(|X - m|) \leq E(|X - a|)$ for every a. Assume now that $m < a$. Then

$$\begin{aligned} E(|X-m|) - E(|X-a|) &= \int_{-\infty}^{m} [(m-x) - (a-x)] f(x)dx \\ &\quad + \int_{m}^{a} [(x-m) - (a-x)] f(x)dx \\ &\quad + \int_{a}^{\infty} [(x-m) - (x-a)] f(x)dx, \end{aligned}$$

which simplifies to

$$\int_{-\infty}^{m} (m-a) f(x)dx + \int_{m}^{a} (2x - m - a) f(x)dx + \int_{a}^{\infty} (a-m) f(x)dx.$$

Since $2x - m - a \leq 2a - m - a = a - m$ for $m \leq x \leq a$, combining the second and third integral, we obtain the inequality

$$E(|X-m|) - E(|X-a|) \leq (m-a)[P(X \leq m) - P(X \geq m)].$$

Since $m - a < 0$ and $P(X \leq m) - P(X \geq m) \geq 1/2 - 1/2 = 0$, we have $E(|X - m|) - E(|X - a|) \leq 0$. The proof for $a < m$ is analogous. □

We now explore the behavior of the variance of sums of random variables. This will naturally lead us to certain new important concepts.

To make the notation more readable, we let $m_X = E(X)$ and $m_Y = E(Y)$. Then

$$\begin{aligned}\operatorname{Var}(X+Y) &= E(X+Y)^2 - [E(X+Y)]^2 \\ &= E(X^2+2XY+Y^2) - (m_X+m_Y)^2 \\ &= E(X^2) - m_X^2 + E(Y^2) - m_Y^2 + 2[E(XY) - m_X m_Y] \\ &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2[E(XY) - m_X m_Y].\end{aligned}$$

The last quantity appears sufficiently often to deserve a separate name.

Definition 8.6.2 The quantity

$$\operatorname{Cov}(X,Y) = E(XY) - E(X)E(Y) = E[(X-m_X)(Y-m_Y)] \tag{8.34}$$

is called the *covariance* of random variables X and Y. □

We shall show that $\operatorname{Cov}(X,Y)$ exists whenever X and Y have finite second ordinary moments. Indeed, $0 \le (X-Y)^2 = X^2 - 2XY + Y^2$, and this gives $2XY \le X^2 + Y^2$. Consequently

$$|XY| \le \frac{X^2}{2} + \frac{Y^2}{2}.$$

Thus, if $E(X^2) < \infty$ and $E(Y^2) < \infty$, then expectations are finite, and $E(|XY|) < \infty$. So $\operatorname{Cov}(X,Y)$ exists.

We have therefore

Theorem 8.6.4 *If $E(X_1^2)$ and $E(X_2^2)$ exist, then*

$$\operatorname{Var}(X_1+X_2) = \operatorname{Var}(X_1) + \operatorname{Var}(X_2) + 2\operatorname{Cov}(X_1,X_2),$$

and more generally, if $E(X_i^2) < \infty$ for $i = 1,\dots,n$ then

$$\operatorname{Var}(X_1+\cdots+X_n) = \sum_{j=1}^{n} \operatorname{Var}(X_j) + 2\sum_{i<j} \operatorname{Cov}(X_i,X_j).$$

Definition 8.6.3 The random variables X, Y for which $\operatorname{Cov}(X,Y) = 0$ are called *uncorrelated* or *orthogonal.* □

Observe that in view of Theorem 8.4.8, independent random variables with finite variances are uncorrelated. Consequently, we have

Theorem 8.6.5 *If random variables $X_1,\dots,X_n$ are pairwise uncorrelated, then*

$$\operatorname{Var}(X_1+\cdots+X_n) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n). \tag{8.35}$$

In particular, (8.35) *holds if $X_1,\dots,X_n$ are independent.*

Let us find the variance of a linear combination of random variables, that is, a variance of the sum

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n.$$

Using Theorem 8.6.4, we have

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(a_i X_i) + \sum_{i<j} \mathrm{Cov}(a_i X_i, a_j X_j).$$

By Theorem 8.6.1, $\mathrm{Var}(a_i X_i) = a_i^2 \mathrm{Var}(X_i)$, while

$$\begin{aligned} \mathrm{Cov}(a_i X_i, a_j X_j) &= E(a_i X_i \times a_j X_j) - E(a_i X_i) E(a_j X_j) \\ &= a_i a_j [E(X_i X_j) - E(X_i) E(X_j)] = a_i a_j \mathrm{Cov}(X_i, X_j). \end{aligned}$$

Consequently,

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2 \sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j).$$

In particular, if $X_1, \ldots, X_n$ are uncorrelated, then

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i). \tag{8.36}$$

■ **EXAMPLE 8.38**

For the variance of a difference of two random variables, $X_1 - X_2$, we take $a_1 = 1, a_2 = -1$, and obtain

$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) - 2\mathrm{Cov}(X_1, X_2);$$

hence for uncorrelated random variables

$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2). \tag{8.37}$$

■ **EXAMPLE 8.39 Averaging**

If $X_1, \ldots, X_n$ are independent, with the same distribution, then in statistics we call them a *random sample* and their average

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

is referred to as a sample mean. If $\mathrm{Var}(X_i) = \sigma^2$, then using $a_i = 1/n, i = 1, \ldots, n$ in formula (8.36), we obtain

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

Thus averaging decreases the variability (as measured by the variance) by the factor $1/n$. The standard deviation of the sample mean is therefore decreased (as compared to the standard deviation of a single observation) by the factor $1/\sqrt{n}$.

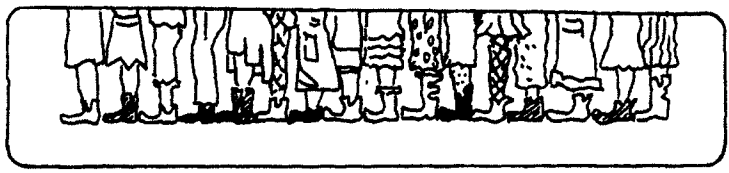


Figure 8.7 Length of 16 feet (Drawing by S. Niewiadomski)

■ **EXAMPLE 8.40 A Foot**

The effect of averaging on variability, computed in Example 8.39, appears to have been understood long before the beginning of probability and statistics. This is illustrated by the following law from the Middle Ages that defined the length of one foot (before you read any further, think for awhile: How could a measure of length be defined centuries ago so as to be—as much as possible—uniform throughout a country?).

The law specified the following procedure (apparently, the standard of one foot was necessary only on market days—in this case on Sundays after Mass). The shoes of the first 16 men leaving the church (this was an attempt to get a random sample!) lined up, toe to heel, gave the "right and lawful rood" (e.g., see Stigler, 1996). Then 1/16 of it was to be used as measure of 1 foot. The number 16 was clearly chosen because it was easy to divide a string into 16 equal parts, by folding it four times into halves.

This procedure cuts down the variability of the length of feet, as measured by the standard deviation, by the factor of 4.

■ **EXAMPLE 8.41 Problem of Design**

Suppose that you have a scale and a set of weights at your disposal. The weight of two objects, A and B, has to be determined as precisely as possible, with the scale used only twice.

SOLUTION. If you put an object on a scale and balance it with weights, you obtain a measurement of the true weight of the object, w, with an error. One of the possible assumptions here, met in many practical situations, is that what one observes is a value of a random variable, X, such that $E(X) = w$ and $\text{Var}(X) = \sigma^2$, with different measurements (even of the same object) being independent. In other words, we have $X = w + \epsilon$, where $E(\epsilon) = 0, \text{Var}(\epsilon) = \sigma^2$, with σ being the standard deviation of the error of measurement.

In our situation, it would seem that all one has to do is to put object A on the scale, balance it with weights, and then do the same with object B, observing two random variables, X and Y, with $E(X) = w_A, E(Y) = w_B$, and $\text{Var}(X) = \text{Var}(Y) = \sigma^2$.

One can, however, proceed differently. Suppose that on the first weighting, one puts both objects A and B on one scale, and then balances the scales, observing random variable $X_{A+B} = w_A + w_B + \epsilon_1$, where ϵ_1 is the measure-

Figure 8.8 Two weightings of A and B

ment error. On the second weighting, one puts A and B on opposite sides of the scale and adds weights as needed for balance (see Figure 8.8). Thus on the second measurement we observe $X_{A-B} = w_A - w_B + \epsilon_2$, where ϵ_2 is independent of ϵ_1, with $E(\epsilon_1) = E(\epsilon_2) = 0, \text{Var}(\epsilon_1) = \text{Var}(\epsilon_2) = \sigma^2$. We easily find that

$$\frac{X_{A+B} + X_{A-B}}{2} = w_A + \frac{\epsilon_1 + \epsilon_2}{2}$$

and

$$\frac{X_{A+B} - X_{A-B}}{2} = w_B + \frac{\epsilon_1 - \epsilon_2}{2}.$$

We have now $\text{Var}[(\epsilon_1 + \epsilon_2)/2] = (1/4)\text{Var}(\epsilon_1 + \epsilon_2) = \sigma^2/2$, and $\text{Var}[(\epsilon_1 - \epsilon_2)/2] = \sigma^2/2$. Using the scale twice, we obtained the measurements of w_A and w_B, each with standard deviation of the error equal $\sigma/\sqrt{2} = 0.707\sigma$. This means an error reduction by about 30% obtained *at no additional cost* (i.e., with the same number of observations). This is one of the simplest examples of the effects of choosing a proper *design of experiment*.

Let us now investigate the concept of covariance more closely. First, observe that according to formula (8.34),

$$\text{Cov}(X, X) = E[(X - m_X)^2] = \text{Var}(X). \tag{8.38}$$

So, in some sense, covariance is a generalization of variance.

Next we have the identity describing the behavior of covariance for a linear combination of random variables.

Theorem 8.6.6 *If $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two sets (not necessarily disjoint) of random variables with finite second moments, then for any constants $a_1, \ldots, a_n, b_1, \ldots, b_m$ and c, d,*

$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i + c, \sum_{i=1}^{m} b_j Y_j + d\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, X_j). \tag{8.39}$$

The proof involves straightforward checking and will be omitted.

The fact that $\text{Cov}(X, Y) = 0$, if X and Y are independent, suggests using the covariance as a measure of dependence. However, to achieve comparability across various measurements it needs to be standardized. Accordingly we introduce the following definition:

Definition 8.6.4 Let X and Y be two random variables with finite second moments. Then

$$\rho = \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

is called the coefficient of *correlation* between X and Y.

■ **EXAMPLE 8.42**

Let U, V, W be independent random variables such that $\mathrm{Var}(U) = \sigma_U^2$, $\mathrm{Var}(V) = \sigma_V^2$, and $\mathrm{Var}(W) = \sigma_W^2$. Compute the coefficient of correlation between $X = U + W$ and $Y = V + W$.

SOLUTION. We have $\sigma_X^2 = \sigma_U^2 + \sigma_W^2$ and $\sigma_Y^2 = \sigma_V^2 + \sigma_W^2$, whereby $\mathrm{Cov}(X,Y) = \mathrm{Cov}(U+W, V+W) = \mathrm{Cov}(U,V)+\mathrm{Cov}(U,W)+\mathrm{Cov}(W,V)+\mathrm{Cov}(W,W) = \sigma_W^2$ by (8.38). Consequently

$$\rho_{XY} = \frac{\sigma_W^2}{\sqrt{(\sigma_U^2 + \sigma_W^2)(\sigma_V^2 + \sigma_W^2)}}. \tag{8.40}$$

Situations like this are quite common; variables often are related because they are influenced by a common other variable (W). One can think about the prices of two items X and Y that depend on the price W of some components, which are used in both X and Y. To see the effect of σ_W^2, let us write (8.40) as

$$\rho_{XY} = \frac{1}{\sqrt{\left(\frac{\sigma_U^2}{\sigma_W^2} + 1\right)\left(\frac{\sigma_V^2}{\sigma_W^2} + 1\right)}}.$$

If σ_W^2 is large compared with both σ_U^2 and σ_V^2, then $\rho_{X,Y}$ is close to 1. On the other hand, if σ_W^2 is small compared to one (or both) of σ_U^2 and σ_V^2, then $\rho_{X,Y}$ is close to 0. These results are expected of $\rho_{X,Y}$ as a measure of dependence of random variables X and Y.

■ **EXAMPLE 8.43**

Let X, Y have a joint density uniform on the square with vertices $(-1,0)$, $(0,1)$, $(1,0)$ and $(0,-1)$ (see Figure 8.9).

Clearly, $E(X) = E(Y) = 0$ by the symmetry of marginal distributions. Also $E(XY) = 0$, since in the integral $\int\int xyf(x,y)\,dx\,dy$ the contribution arising from domains with $xy > 0$ and $xy < 0$ cancel one another. Thus $\mathrm{Cov}(X,Y) = 0$, and therefore $\rho_{X,Y} = 0$. Yet the random variables are dependent. This can be seen immediately from Theorem 7.2.2, as well as from the fact that if one knows, for example, that X is close to 1, then one can infer that Y must be close to zero.

This example shows that correlation, as a measure of dependence, is not perfect. There are dependent random variables for which $\rho_{X,Y} = 0$. To analyze the properties of the correlation coefficient, we need to introduce the following inequality:

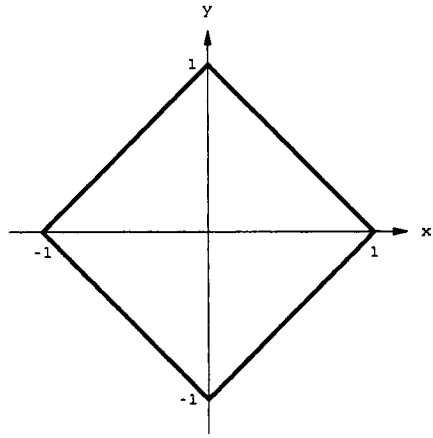


Figure 8.9 Dependent but uncorrelated random variables

Theorem 8.6.7 (Schwarz Inequality) *For any variables X and Y,*

$$[E(XY)]^2 \leq E(X^2) \times E(Y^2). \tag{8.41}$$

Proof. If $E(X^2)$ or $E(Y^2)$ is infinite, the inequality (8.41) holds. On the other hand, if $E(X^2) = 0$, then $P(X = 0) = 1$, $P(XY = 0) = 1$, and $E(XY) = 0$. Hence, again, inequality (8.41) holds. The same argument applies if $E(Y^2) = 0$. So we can assume that $0 < E(X^2) < \infty$ and $0 < E(Y^2) < \infty$.

Consider now the random variable $Z_t = tX + Y$. For every t we have

$$\begin{aligned} 0 \leq E(Z_t^2) &= E(t^2X^2 + 2tXY + Y^2) \\ &= t^2E(X^2) + 2tE(XY) + E(Y^2). \end{aligned} \tag{8.42}$$

The right-hand side is a quadratic function of t, which is nonnegative for all t. Thus its discriminant must satisfy the condition

$$4[E(XY)]^2 - 4E(X^2)E(Y^2) \leq 0$$

which is the same as (8.41). □

We can now prove the following theorem, asserting the basic properties of the correlation coefficient:

Theorem 8.6.8 *The coefficient of correlation ρ_{XY} satisfies the inequality*

$$-1 \leq \rho_{X,Y} \leq 1 \tag{8.43}$$

with $|\rho_{X,Y}| = 1$ if, and only if, there exist constants a ($a \neq 0$) and b such that $P\{Y = aX + b\} = 1$.

Moreover, $|\rho|$ is invariant under linear transformation of random variables; more precisely, for any a, b, c, d with $ac \neq 0$,

$$\rho_{aX+b,cY+d} = \epsilon\rho_{X,Y} \tag{8.44}$$

with $\epsilon = +1$ *if* $ac > 0$ *and* $\epsilon = -1$ *if* $ac < 0$.

Proof. By formula (8.39), we have $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$. On the other hand, $\text{Var}(aX + b) = a^2\text{Var}(X)$, $\text{Var}(cY + d) = c^2\text{Var}(Y)$, so that

$$\rho_{aX+b,cY+d} = \frac{ac\text{Cov}(X,Y)}{\sqrt{a^2\text{Var}(X) \times c^2\text{Var}(Y)}} = \frac{ac}{|ac|} \times \rho_{X,Y},$$

which proves (8.44). We can now prove (8.43). By what is already shown in (8.39), we can assume that $EX = EY = 0$, so

$$\rho_{X,Y} = E(XY)/\sqrt{E(X^2)E(Y^2)}.$$

The condition $|\rho_{X,Y}| \leq 1$ is equivalent to the Schwarz inequality (8.41).

It remains to prove that $\rho^2 = 1$ is equivalent to the existence of a linear relationship between X and Y. Again, we can assume that $E(X) = E(Y) = 0$. If $Y = aX$ with $a \neq 0$, then $E(XY) = E[X(aX)] = aE(X^2)$ and $E(X)E(Y) = a[E(X)]^2$. Consequently, $\text{Cov}(X, Y) = a\{E(X^2) - [E(X)]^2\} = a\text{Var}(X)$, and since $\text{Var}(Y) = a^2\text{Var}(X)$, we get

$$\rho_{X,Y} = \frac{a\text{Var}(X)}{\sqrt{\text{Var}(X) \times a^2\text{Var}(X)}} = \frac{a}{|a|} = \pm 1.$$

Conversely, assume that $\rho^2_{X,Y} = 1$. Then (assuming $E(X) = E(Y) = 0$) we have $[E(XY)]^2 = E(X^2)E(Y^2)$, meaning X and Y such that we have equality in (8.42). The proof of Theorem 8.6.7 shows that this occurs if the discriminant of the right-hand side of (8.42) is 0, hence if there exists t^* such that $E(t^*X + Y)^2 = 0$. But the expectation of a nonnegative random variable is 0 if, and only if, this variable assumes only the value zero, so that $P(t^*X + Y = 0) = 1$. If we had $t^* = 0$, then we would have $P(Y = 0) = 1$, and hence $E(Y^2) = 0$, a case that we eliminated. This shows that $t^* \neq 0$, and hence that there is a linear relationship between X and Y. □

■ **EXAMPLE 8.44**

Consider the situation where the value of some random variable Z has to be predicted based on the values of two other variables X and Y.

The optimal solution to this problem is well known. Adapting Theorem 8.6.2 to the case of conditional prediction, we see that given $X = x$ and $Y = y$, one should predict the conditional expectation $\xi = E(Z|X = x, Y = y)$. However, the practical implementation of this solution presupposes the knowledge of the joint distribution of (X, Y, Z), and—which may be analytically difficult—the conditional expectation of Z given (X, Y).

Quite often we simply do not have this information. If appropriate, one can then use here the best *linear* prediction, which requires the knowledge of the first two moments of the joint distribution of (X, Y, Z) only, that is, expectations, variances, and covariances (or equivalently, correlations).

SOLUTION. Without loss of generality, we can assume that $E(X) = E(Y) =$

$E(Z) = 0$ (since if we know the mean, predicting the random variable is equivalent to predicting its deviation from the mean). We will be looking for a predictor of the form $\xi = \alpha X + \beta Y$ that minimizes the expected square error

$$E(Z - \xi)^2 = E(Z - \alpha X - \beta Y)^2. \tag{8.45}$$

Condition (8.45) can be expanded to

$$\begin{aligned} E(Z - \alpha X - \beta Y)^2 &= E(Z^2) + \alpha^2 E(X^2) + \beta^2 E(Y^2) \\ &\quad -2\alpha E(XZ) - 2\beta E(YZ) + 2\alpha\beta E(XY) \\ &= \sigma_Z^2 + \alpha^2\sigma_X^2 + \beta^2\sigma_Y^2 \\ &\quad -2\alpha\rho_{XZ}\sigma_Z\sigma_X - 2\beta\rho_{YZ}\sigma_Z\sigma_Y + 2\alpha\beta\rho_{XY}\sigma_X\sigma_Y. \end{aligned}$$

Differentiating with respect to α and β and setting the derivatives equal to zero, we obtain the system of linear equations:

$$\begin{aligned} \alpha\sigma_X + \beta\rho_{XY}\sigma_Y &= \rho_{XZ}\sigma_Z, \\ \alpha\rho_{XY}\sigma_X + \beta\sigma_Y &= \rho_{YZ}\sigma_Z. \end{aligned}$$

The solution exists if $\rho_{XY}^2 \neq 1$:

$$\alpha = \frac{\rho_{XZ} - \rho_{XY}\rho_{YZ}}{1 - \rho_{XY}^2} \times \frac{\sigma_Z}{\sigma_X} \quad \text{and} \quad \beta = \frac{\rho_{YZ} - \rho_{XY}\rho_{XZ}}{1 - \rho_{XY}^2} \times \frac{\sigma_Z}{\sigma_Y}$$

(if $\rho_{XY}^2 = 1$, then $X = Y$ or $X = -Y$, so X and Y provide the same information). This solution is not very important by itself. More important is the fact that this method can easily be applied to determine the best linear predictor of one or more random variables. All that we need are the means and second order moments of all variables in question—those to be predicted and those serving as predictors. The next example will illustrate the class of situations in which such a prediction method can be usefully applied.

■ **EXAMPLE 8.45 Moving Averages**

Consider the process $\{X_t\}$ of prices of some commodity. In many cases the process $\{X_t\}$ is subject to seasonal variations. For example, if t is measured in months, and X_t represents the (average) monthly price of tomatoes, then X_t varies with the period of 12. One of the methods of detecting trends (or prediction), which takes such periodic seasonal variation into account, is based on taking the averages of X_t over the most recent complete period. Thus we can define

$$Y_t = \frac{X_t + X_{t-1} + \cdots + X_{t-11}}{12}.$$

Here a possible assumption about X_t may be $X_t = f_t + \xi_t$, where f_t is some nonrandom function with period 12 (i.e., $f_t = f_{t-12}$ for every t) and ξ_t's are independent random variables with the same distribution (sometimes referred to as "shocks") representing random effects. Under these assumptions

$$Y_t = C + \frac{\xi_t + \xi_{t-1} + \cdots + \xi_{t-11}}{12},$$

where $C = \frac{1}{12}(f_t + f_{t-1} + \cdots + f_{t-11})$, which is a constant independent of t. Clearly, we can assume, without loss of generality, that $E(\xi_t) = 0$ for all t. Let $\mathrm{Var}(\xi_t) = E(\xi_t^2) = \sigma^2 > 0$. Then $E(Y_t) = C$, $\mathrm{Var}(Y_t) = \sigma^2/2$ and

$$\begin{aligned}\mathrm{Cov}(Y_t, Y_{t-m}) &= \mathrm{Cov}\left(\frac{\xi_t + \xi_{t-1} + \cdots + \xi_{t-11}}{12}, \frac{\xi_{t-m} + \cdots + \xi_{t-m-11}}{12}\right) \\ &= \frac{1}{144}\mathrm{Cov}(\xi_t + \cdots + \xi_{t-11}, \xi_{t-m} + \cdots + \xi_{t-m-11}). \\ &= K \times \frac{\sigma^2}{144}\end{aligned}$$

where K is the number of overlapping terms in sums $\xi_t + \xi_{t-1} + \cdots + \xi_{t-11}$ and $\xi_{t-m} + \cdots + \xi_{t-m-11}$. Obviously the number of such terms is 0 if $m \geq 12$ and equals $12 - m$ otherwise.

Consequently, we have

$$\mathrm{Cov}(Y_t, Y_{t-m}) = \max\left(0, \frac{12-m}{144}\sigma^2\right)$$

so that

$$\rho_{Y_t, Y_{t-m}} = \begin{cases} 1 - \frac{m}{12} & \text{for } m = 0, 1, \ldots, 12 \\ 0 & \text{for } m \geq 12. \end{cases} \tag{8.46}$$

PROBLEMS

8.6.1 The random variable X has binomial distribution with mean 5 and standard deviation 2. Find $P\{X = 6\}$.

8.6.2 Find the variance of X if its first ordinary moment is 3 and the second factorial moment is 52.

8.6.3 For random variables X and Y such that $E(X) = 2, E(Y) = 1, E(X^2) = 10, E(Y^2) = 3$ and $E(XY) = c$, find: **(i)** $\mathrm{Var}(3X - 5Y)$. **(ii)** $\rho_{X,Y}$. **(iii)** The range of values of c for which the assumptions of the problem are consistent.

8.6.4 Find the variance of a random variable X with a cdf

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \sqrt{x} & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1. \end{cases}$$

8.6.5 Let variables X and Y be such that $E(X) = E(Y) = 0, \mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, and $\rho_{X,Y} = \rho$. Find $E(W), \mathrm{Var}(W)$, and $\rho_{W,Y}$ if $W = X - \rho Y$.

8.6.6 Find the correlation of random variables X and Y jointly uniformly distributed on: **(i)** The triangle with vertices $(0,0), (1,0), (1,2)$. **(ii)** The quadrangle with vertices $(0,0), (a,0), (a,2)$, and $(2a,2)$, where $a > 0$.

8.6.7 Let X, Y be independent, with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. Show that $\mathrm{Var}(XY) = \sigma_X^2\sigma_Y^2 + \sigma_X^2\mu_Y^2 + \sigma_Y^2\mu_X^2$.

8.6.8 Let $X_1, \ldots X_n$ be independent random variables having the same distribution with a mean μ and variance σ^2. Let $\overline{X} = (X_1 + \cdots + X_n)/n$. Show that $E\{\sum_{i=1}^n (X_i - \overline{X})^2\} = (n-1)\sigma^2$. [*Hint:* Since $X_i - \overline{X} = (X_i - \mu) - (\overline{X} - \mu)$, we have $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\overline{X} - \mu)^2$.]

8.7 CONDITIONAL EXPECTATION

The concept of conditional probability was introduced in Chapter 4 and then extended to the concept of conditional distribution of a random variable in Chapters 6 and 7. Now we will introduce the expectation of this distribution as a number, $E(X|Y = y)$, determined for every particular value $Y = y$, provided that it exists.

Definition 8.7.1 For two random variables X and Y, the *conditional expectation* $E(X|Y)$ is defined as the random variable that assumes the value $E(X|Y = y)$ when $Y = y$. □

■ **EXAMPLE 8.46**

Let Y have POI(λ) distribution, and given $Y = n$, let X have BIN(n, p) distribution. We can think here (recall Example 7.14) of Y being the number of eggs laid by a bird and of X as the number of those eggs that hatch (assuming that the eggs hatch independently, each with probability p). This means that the expected number of eggs that hatch, if $Y = n$, is $E(X|Y = n) = np$.

■ **EXAMPLE 8.47**

Let (X, Y) have a distribution uniform on the triangle with vertices (0, 0), (0, 1) and (1, 1). Given Y, the distribution of X is uniform on the intersection of the line $Y = y$ and the support of the density. Since the expectation of the uniform distribution on an interval is its midpoint, we have $E(X|Y = y) = y/2$, and similarly $E(Y|X = x) = (1+x)/2$, where $0 \le X \le 1, 0 \le Y \le 1$.

■ **EXAMPLE 8.48**

Consider now a case where one variable is discrete and the other is continuous. Let $P\{Y = 1\} = \alpha$ and $P\{Y = 2\} = 1 - \alpha$, and given Y, let X have the density

$$f(x|Y = y) = yx^{y-1}, \quad 0 \le x \le 1.$$

In this case, if $Y = 1$, then X is uniform on [0, 1]; hence $E(X|Y = 1) = 1/2$. If $Y = 2$, then X has density $f(x|Y = 2) = 2x$ on [0, 1]; hence $E(X|Y = 2) = 2/3$. On the other hand, given $X = x$, we have

$$\begin{aligned} P\{Y = 1|X = x\} &= \frac{\alpha}{\alpha + 2x(1-\alpha)}, \\ P\{Y = 2|X = x\} &= \frac{2x(1-\alpha)}{\alpha + 2x(1-\alpha)}. \end{aligned}$$

This is a consequence of Bayes' formula, interpreting the values of the density $f(x|Y)$ as infinitesimal probability $P\{X = x|Y\}dx$. Thus

$$\begin{aligned} E(Y|X=x) &= 1 \times P(Y=1|X=x) + 2 \times P(Y=2|X=x) \\ &= \frac{\alpha + 4x(1-\alpha)}{\alpha + 2x(1-\alpha)}. \end{aligned}$$

We will now prove

Theorem 8.7.1 *For any random variables X, Y, we have*

$$E[E(X|Y)] = E(X), \tag{8.47}$$

provided that $E(X)$ exists.

Proof. We will present the proof only in the case of continuous random variables. Letting $f_1(x)$ and $f_2(y)$ denote the marginal densities of the joint density $f(x, y)$, we have

$$E(X|Y=y) = \int x g(x|Y=y)dx = \int x \frac{f(x,y)}{f_2(y)}dx.$$

Hence, taking expectation with respect to $Y = y$, we obtain

$$\begin{aligned} E[E(X|Y)] &= \int \left[\int \frac{x f(x,y)}{f_2(y)}dx\right] f_2(y)dy = \int x \left[\int f(x,y)dy\right] dx \\ &= \int x f_1(x)dx = E(X). \end{aligned}$$

The interchange of the order of integration is legitimate in view of the assumption that $E(X)$ exists. □

■ **EXAMPLE 8.49**

In Example 8.46 we had $E(X|Y) = Yp$. Since Y has Poisson distribution with mean λ, we obtain $E(X) = E(Y) \times p = \lambda p$. This result also follows from the fact (see Example 7.14) that X has Poisson distribution with mean λp.

■ **EXAMPLE 8.50**

In Example 8.47 we have $E(X) = E[E(X|Y)] = E(Y/2)$ and $E(Y) = E[E(Y|X)] = E[(1+X)/2]$, which gives $E(X) = E(Y)/2, E(Y) = 1/2 + E(X)/2$. This system of two equations can easily be solved, leading to $\mu_X = 1/3$ and $\mu_Y = 2/3$.

■ **EXAMPLE 8.51**

Finally, in Example 8.48 we have

$$E(X) = E[E(X|Y)] = \alpha E(X|Y=1) + (1-\alpha)E(X|Y=2) = \frac{1}{2}\alpha + \frac{2}{3}(1-\alpha).$$

On the other hand, since $f_1(x) = \alpha + 2x(1-\alpha)$, we have

$$\begin{aligned} E(Y) &= E[E(Y|X)] = \int_0^1 \frac{\alpha + 4x(1-\alpha)}{\alpha + 2x(1-\alpha)} f_1(x)dx \\ &= \int_0^1 [\alpha + 4x(1-\alpha)]dx = 2 - \alpha. \end{aligned}$$

Theorem 8.7.1 can be very helpful in determining the expectation of a random variable X by conditioning it with respect to some other variable (Y). The choice of Y is crucial here if a simplification is to be achieved.

The following theorem is an analogue of Theorem 8.7.1:

Theorem 8.7.2 *For any random variables X, Y, if $E(X^2)$ exists, then*

$$\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}[E(X|Y)]. \tag{8.48}$$

Proof. Let $\mu_X = E(X)$. We have

$$\begin{aligned} \mathrm{Var}(X) &= E(X-\mu_X)^2 = E\{[X - E(X|Y)] + [E(X|Y) - \mu_X]\}^2 \\ &= E\{[X - E(X|Y)]^2\} + 2E\{[X - E(X|Y)] \times [E(X|Y) - \mu_X]\} \\ &\quad + E\{[E(X|Y) - \mu_X]^2\} \\ &= A + B + C. \end{aligned}$$

Using Theorem 8.7.1, we obtain

$$\begin{aligned} A &= E\{E[X - E(X|Y)]^2|Y\} = E[\mathrm{Var}(X|Y)], \\ C &= E[E(X|Y) - \mu_X]^2 = \mathrm{Var}[E(X|Y)], \end{aligned}$$

where the last equality follows from the fact that $\mu_X = E[E(X|Y)]$ (see (8.47)). It remains to prove that $B = 0$.

As before, we have

$$\begin{aligned} B &= E\{[X - E(X|Y)][E(X|Y) - \mu_X]\} \\ &= E(E\{[X - E(X|Y)] \times [E(X|Y) - \mu_X]|Y\}) \\ &= E(E[U(X,Y) \times V(Y)|Y]), \end{aligned}$$

where $U(X,Y) = X - E(X|Y)$ and $V(Y) = E(X|Y) - \mu_X$ is constant for every fixed value of Y. Consequently, $V(Y)$ can be factored out, leaving

$$B = E\{V(Y)E[U(X,Y)|Y]\}.$$

But $E[U(X,Y)|Y] = E[X - E(X|Y)|Y] = E(X|Y) - E(X|Y) = 0$. Thus $B = E[V(Y) \times 0] = 0$, which completes the proof. □

■ **EXAMPLE 8.52**

For an application of Theorem 8.7.2, let us return again to the Example 8.46. We have there $X \sim \text{BIN}(Y, p)$, so $E(X|Y) = Yp$, and $\text{Var}(X|Y) = Ypq$. Consequently, since Y has POI(λ) distribution with $E(Y) = \text{Var}(Y) = \lambda$, we obtain

$$\begin{aligned} \text{Var}(X) &= E\text{Var}(X|Y) + \text{Var}[E(X|Y)] = E(Ypq) + \text{Var}(Yp) \\ &= pq\lambda + p^2\lambda = \lambda p, \end{aligned}$$

again, in agreement with our finding from Example 7.14 that the marginal distribution of X is Poisson with a parameter λp.

Finally, as another example of an application, we will prove the theorem due to Hotelling and Solomons (1932) that connects the mean, median, and standard deviation of any random variable. The present proof is a slight modification of the proof by O'Cinneide (1990).

Theorem 8.7.3 *Let X be a random variable with $E(X^2) < \infty$, and let μ and σ be its mean and standard deviation. Moreover, let m be any median of X, that is, a number such that $P(X \geq m) \geq 0.5$ and $P(X \leq m) \geq 0.5$. Then for any random variable X,*

$$|\mu - m| \leq \sigma;$$

that is, the mean is within one standard deviation of any median.

Proof. Let $\pi^- = P(X < m), \pi^+ = P(X > m)$, and $\pi^0 = P(X = m)$ so that $\pi^- + \pi^0 \geq 0.5$ and $\pi^0 + \pi^+ \geq 0.5$. Let Y be the random variable defined as follows:
(a) If $\pi^0 = 0$, then

$$Y = \begin{cases} 1 & \text{if } X < m \\ 2 & \text{if } X > m. \end{cases} \tag{8.49}$$

(b) If $\pi^0 > 0$, then in addition to (8.49), we let

$$P\{Y = 1|X = m\} = 0.5 - \pi^- \quad \text{and} \quad P\{Y = 2|X = m\} = 0.5 - \pi^+.$$

Clearly, $P\{Y = 1\} = P\{Y = 2\} = 0.5$.

For simplicity, put $\mu_i = E(X|Y = i)$, for $i = 1, 2$. Then $\mu = E(X) = E[E(X|Y)] = 0.5E(X|Y = 1) + 0.5E(X|Y = 2)$, that is,

$$\mu = 0.5\mu_1 + 0.5\mu_2. \tag{8.50}$$

Assume that $m \leq \mu$ (the argument in the case of opposite inequality is analogous). Clearly, we have $\mu_1 \leq m$, which in view of (8.50) gives the inequality

$$\mu - m \leq \mu - \mu_1 = 0.5(\mu_2 + \mu_1) - \mu_1 = 0.5(\mu_2 - \mu_1). \tag{8.51}$$

Using Theorem 8.7.2 we can now write

$$\begin{aligned} \sigma^2 = \text{Var}(X) &= E[\text{Var}(X|Y)] + \text{Var}[E(X|Y)] \geq \text{Var}[E(X|Y)] \\ &= \left(\frac{\mu_1 - \mu_2}{2}\right)^2 \geq (\mu - m)^2. \end{aligned}$$

So $\sigma \geq |\mu - m|$, as asserted.

PROBLEMS

8.7.1 Variables X and Y are jointly distributed with the density $f(x, y) = Cx(3x + 2y)$ for $0 < x < y < 1$, and $f(x, y) = 0$ otherwise. Find: **(i)** C. **(ii)** $E(Y|X = x)$.

8.7.2 Let X and Y have the joint density uniform on the triangle with vertices $(0, 0), (2, 0)$ and $(3, 1)$. Find: **(i)** $E(X|Y)$ and $E(Y|X)$. **(ii)** $\mathrm{Var}(X|Y)$ and $\mathrm{Var}(Y|X)$. **(iii)** The expectations and variances of X and Y using formulas (8.47) and (8.48).

8.7.3 Let X, Y be continuous random variables with a joint density f. Assume that $E(Y|X = x) = \mu$ for all x. Show that

$$\mathrm{Var}(Y) = \int \mathrm{Var}(Y|X = x) f_X(x)\, dx.$$

8.7.4 The number of traffic accidents that occur in a certain city in a week is a random variable with mean μ and variance σ^2. The numbers of people injured in an accident are independent random variables, each with mean m and variance k^2. Find the mean and variance of the number of people injured in a week.

8.8 INEQUALITIES

In this section we will introduce three important inequalities, all involving expectations that will be used in the following chapters.

Theorem 8.8.1 *If V is a random variable such that $E(V) = 0$ and $\mathrm{Var}(V) = 1$, then for every $t > 0$,*

$$P\{|V| \geq t\} \leq 1/t^2. \tag{8.52}$$

Proof. We will give the proof in the case of continuous random variables; the proof in the discrete case (as well as in the general case) will require only notational changes. Letting f denote the density of V, we have

$$\begin{aligned} 1 &= \mathrm{Var}(V) = E(V^2) = \int_{-\infty}^{+\infty} v^2 f(v) dv \geq \int_{|v| \geq t} v^2 f(v) dv \\ &\geq t^2 \int_{|v| \geq t} f(v) dv = t^2 P\{|V| \geq t\}, \end{aligned}$$

which was to be shown. □

Clearly, inequality (8.52) is not informative for $t \leq 1$.

If X is a random variable with finite positive variance $\mathrm{Var}(X) = \sigma^2 > 0$ and $\mu = E(X)$, we can always *standardize* X by defining

$$V = \frac{X - \mu}{\sigma}. \tag{8.53}$$

Obviously $E(V) = 0$ and $\mathrm{Var}(V) = 1$; transformation of X into V amounts to introducing a new scale of expressing values of X, with the origin of the scale located at $E(X)$, and the unit of measurement being the standard deviation σ. Applying formula (8.52) to random variable (8.53), we obtain the Chebyshev inequality:

Theorem 8.8.2 (Chebyshev Inequality) *If* $\mathrm{Var}(X) = \sigma^2 < \infty$, *then for every* $t > 0$,

$$P\{|X - \mu| \geq t\sigma\} \leq \frac{1}{t^2} \tag{8.54}$$

or for $\epsilon = t\sigma$,

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}. \tag{8.55}$$

Both (8.55) and (8.54) show the role of the standard deviation and variance. We obtain the bounds on probabilities of random variable X deviating from its mean μ by more than a certain amount, which can be expressed as a multiple of σ in (8.54) or in original units in (8.55).

■ **EXAMPLE 8.53**

For any random variable, the probability that it will deviate from its mean by more than three standard deviations is, at most, $1/9$. This probability, however, can be much less than $1/9$ for some random variables. For instance it equals 0.0026 for a normal distribution; see (9.59).

Consequently, the *three-sigma rule* (which says that we can practically disregard the probability that a random variable will deviate from its mean by more than three standard deviations) should be used with caution. It may safely be used for random variables with either a normal distribution, or close to normal, but in the general case, the probability $1/9$ can hardly be disregarded.

We will now find the bounds given by the Chebyshev inequality in the case of few selected distributions.

■ **EXAMPLE 8.54**

Suppose that we toss a coin 20 times. Let us estimate the probability that the number of heads will deviate from 10 by 3 or more (then we will have at most 7 or at least 13 heads). Here $\mu = np = 10, \sigma^2 = npq = 5$ (see Examples 8.3 and 8.34). We have

$$P\{|X - 10| \geq 3\} \leq \frac{\sigma^2}{3^2} = \frac{5}{9} = 0.555.$$

The exact probability (see Table A.1) equals 0.2632.

■ **EXAMPLE 8.55**

Let X have exponential distribution with density $f(x) = \lambda e^{-\lambda x}, x > 0$. We have here $\mu = \sigma = 1/\lambda$ (see Example 8.36). Consequently

$$P\left\{\left|X - \frac{1}{\lambda}\right| > \frac{t}{\lambda}\right\} \leq \frac{1}{t^2}.$$

For instance, if $t = 1$ we obtain a noninformative bound 1, while the probability that X will deviate from $1/\lambda$ by more than $1/\lambda$ is (remember that exponential random variable can assume only positive values)

$$P\left\{X > \frac{2}{\lambda}\right\} = e^{-\lambda(2/\lambda)} = e^{-2} = 0.13534.$$

As can be seen from these examples, the quality of bounds given by the Chebyshev inequality is not impressive. However, the most important thing here is that the Chebyshev inequality gives a universal bound, valid for *all* random variables with finite variance. In fact the bound as such cannot be improved.

Among the most important consequences of the Chebyshev inequality are the so-called laws of large numbers. We explore this topic in some detail in Chapter 10. Here we illustrate the situation by the following example:

■ **EXAMPLE 8.56 Binomial Distribution**

Consider the binomial random variable S_n = number of successes in n trials. We have then $E(S_n) = np$ and $Var(S_n) = np(1-p)$ when p is the probability of success. Consequently, $E(S_n/n) = p, Var(S_n/n) = pq/n$, and the Chebyshev inequality gives

$$P\{|\frac{S_n}{n} - p| \geq \epsilon\} \leq \frac{pq}{n\epsilon^2}.$$

Letting $n \to \infty$ we obtain the following theorem:

Theorem 8.8.3 *If S_n has binomial distribution with parameter p, then for every $\epsilon > 0$*

$$\lim_{n\to\infty} P\{|\frac{S_n}{n} - p| \geq \epsilon\} = 0.$$

This theorem appears to explain why the empirical frequency of an event, namely S_n/n, approaches, as the number of trials n increases, the probability p of the event. It tells us that for any positive number ϵ it becomes increasingly unlikely that the empirical frequency will deviate from theoretical probability by more than ϵ. The Chebyshev inequality assumes that the random variable has finite second moment. One occasionally needs a bound for a tail of the distribution without this assumption. In such cases we have the following theorem:

Theorem 8.8.4 (Markov Inequality) *If X is a nonnegative random variable with $E(X) < \infty$, then for every $t > 0$,*

$$P\{X > t\} \leq \frac{E(X)}{t}. \tag{8.56}$$

Proof. We give the proof for the discrete case. Let $x_1, x_2, \ldots$ be possible values of X. Then

$$\begin{aligned} E(X) &= \sum_i x_i P(X = x_i) \geq \sum_{x_i > t} x_i P(X = x_i) \\ &\geq t \sum_{x_i > t} P(X = x_i) = tP(X > t). \end{aligned}$$

□

The last inequality will be given without proof:

Theorem 8.8.5 (Kolmogorov Inequality) *Let independent random variables $X_1, X_2, \ldots$ have $E(X_i) = 0$ and finite variances $\sigma_j^2, j = 1, 2, \ldots$ (in the case of nonzero means we can always consider new variables $X_j' = X_j - E(X_j)$ with means that equal zero). If $S_j = X_1 + X_2 + \cdots + X_j$, then for every $t > 0$,*

$$P\{\max_{1 \leq j \leq n} |S_j| \geq t\} \leq \frac{\mathrm{Var}(S_n)}{t^2}. \tag{8.57}$$

PROBLEMS

8.8.1 Show that $E(X) \geq 0.2$ if $P(X \geq 0) = 1$ and $P(X \geq 2) = 0.1$.

8.8.2 Assume that $E(X) = 12, P(X \geq 14) = 0.12$, and $P(X \leq 10) = 0.18$. Show that the standard deviation of X is at least 1.2.

8.8.3 Prove the Markov inequality when X is a continuous random variable.

8.8.4 Derive the Chebyshev inequality from the Markov inequality.

8.8.5 Show that if X has a mgf bounded by the mgf of exponential distribution (i.e., $\lambda/(\lambda - t)$ for $t < \lambda$), then for $\lambda\epsilon > 1$ we have

$$P\{X > \epsilon\} \leq \lambda\epsilon e^{-(\lambda\epsilon - 1)}.$$

(*Hint:* Use the mgf of exponential distribution to obtain the bound for $P\{X > \epsilon\}$, then determine its minimum.)

8.8.6 Let X have the Poisson distribution with mean λ. Show that

$$P\left\{X \leq \frac{\lambda}{2}\right\} \leq \frac{4}{\lambda} \quad \text{and} \quad P\{X \geq 2\lambda\} \leq \frac{1}{\lambda}.$$

8.8.7 Let X be a random variable such that a mgf $m_X(t)$ exists for all t. Use the same argument as in the proof of the Chebyshev inequality to show that $P\{X \geq y\} \leq e^{-ty} m_X(t), t \geq 0$.

8.8.8 Show that if X has the Poisson distribution with mean λ, then $P(X \leq \lambda/2) \leq (2/e)^{\lambda/2}$ and $P(X \geq 2\lambda) \leq (e/4)^{\lambda}$. (*Hint:* Use the inequality in Problem 8.8.7, and find minimum of the right-hand sides for t.)

This Page Intentionally Left Blank

CHAPTER 9

SELECTED FAMILIES OF DISTRIBUTIONS

In this chapter we will review the most commonly used distributions. Almost all of them were introduced in the preceding chapters, mostly as examples. We will now present them in a systematic way, adding new information while also providing references to the examples in the preceding chapters.

9.1 BERNOULLI TRIALS AND RELATED DISTRIBUTIONS

Bernoulli trials refer to independent repetitions of some experiment in which we are interested in an event A that occurs in each trial with the same probability p. We refer to event A as "success," and the event A^c as a "failure." The decision about which of the two events of interest, A and A^c, is labeled success is arbitrary and usually implies nothing about the nature of event A in any practical applications.

The Bernoulli random variable, a building block of the theory, is just a count of the number of successes in a single trial. Thus X is 1 or 0, depending on whether A or A^c occurred, and consequently the distribution of X is given by

$$P\{X = 0\} = 1 - p, \qquad P\{X = 1\} = p. \tag{9.1}$$

It will be convenient to let $q = 1 - p$. The expectation and variance are then

$$E(X) = p, \qquad \mathrm{Var}(X) = p(1 - p) = pq.$$

Moreover, since $0^n = 0$ and $1^n = 1$, the Bernoulli random variable is the only (nondegenerate) random variable X that satisfies the relations $X^n = X$ for all $n \geq 1$. Consequently the moments of X are

$$m_n = E(X^n) = E(X) = p$$

for all $n \geq 1$. Then, by the fact that $\frac{d^n}{dt^n} m_X(t)|_{t=0} = p$ for all $n \geq 1$, the Taylor expansion of the moment generating function of X is

$$m_X(t) = m_0 + \sum_{n=1}^{\infty} m_n \frac{t^n}{n!} = 1 + p \sum_{n=1}^{\infty} \frac{t^n}{n!} = 1 + p(e^t - 1) = pe^t + q.$$

This result, of course, could have been obtained in a much simpler way using the distribution (9.1) of X.

Binomial Distribution

The binomial distribution plays the central role in probability theory and statistics, as a model of a total count of the total number of successes in n Bernoulli trials. In this chapter we let S_n denote the binomial random variable, and we use the representation

$$S_n = X_1 + \cdots + X_n, \tag{9.2}$$

where $X_1, \ldots, X_n$ are the Bernoulli random variables describing the outcomes of successive trials (i.e., $X_i = 1$ or 0, depending on whether the ith trial results in success or in failure).

We have encountered binomial random variables in the preceding chapters (e.g., in Examples 6.8, 7.14, 8.3, and 8.34). We also know that

$$P\{S_n = k\} = \binom{n}{k} p^k (1-p)^{n-k} = b(k; n, p) \tag{9.3}$$

and from (8.32), that

$$E(S_n) = np, \qquad \mathrm{Var}(S_n) = np(1-p).$$

The symbol $\mathrm{BIN}(n, p)$ denotes binomial a distribution with parameters n and p, so we can say that S_n has distribution $\mathrm{BIN}(n, p)$ or simply $S_n \sim \mathrm{BIN}(n, p)$. The Bernoulli random variable has the distribution $\mathrm{BIN}(1, p)$.

To find the most likely number of successes (the mode), we proceed as follows: We have

$$\frac{b(k; n, p)}{b(k-1; n, p)} = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\binom{n}{k-1} p^{k-1} (1-p)^{n-k+1}} = \frac{n-k+1}{k} \times \frac{p}{q}.$$

Since $\frac{n-k+1}{k} \times \frac{p}{q} > 1$ if and only if $k < (n+1)p$, the probabilities $b(k; n, p)$ initially increase and then decrease. If $(n+1)p$ is not an integer, the unique maximum of probabilities $b(k; n, p)$ is attained at $k^* = [(n+1)p]$ (an integer part of $(n+1)p$), while if $(n+1)p$ is an integer, the maximum is attained at two values, $(n+1)p$ and $(n+1)p - 1$.

The moment generating function of binomial distribution was obtained in Example 8.23:

$$m_{S_n}(t) = E(e^{tS_n}) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k q^{n-k} = (pe^t + q)^n.$$

The same result can be obtained using Theorem 8.5.5 and formula (9.2) by observing that $m_{S_n}(t) = [m_X(t)]^n$, where $m_X(t)$ is the moment generating function of the Bernoulli random variable.

Calculating the numerical values of binomial probabilities is simple, in principle, but can be cumbersome, especially when—as is often the case—we need to know probabilities of the form $P\{a \leq S_n \leq b\}$, which require calculating the individual probabilities $P\{S_n = k\}$ for all k between a and b.

For small and moderate n, the situation is somewhat remedied by tables (see Table A.1 in the Appendix). This table, as most of the binomial tables found in the literature, gives the cdf values of S_n, that is, of probabilities $P\{S_n \leq k\}$ for selected n, k, and p. Then, $P\{a \leq S_n \leq b\} = P\{S_n \leq b\} - P\{S_n \leq a-1\}$, and the required probabilities are obtained by subtraction of the two terms (rather than adding $b - a + 1$ terms).

■ **EXAMPLE 9.1**

Assume that it is known (from past experience, research surveys, etc.) that 40% of buyers of FATCOW butter buy unsalted butter, and the remaining 60% buy salted butter. The store expects to sell no more than 20 packages of FATCOW butter per day, so they put on the shelf 8 packs of unsalted, and 12 packs of salted FATCOW butter. It happened that only 15 persons bought FATCOW butter on a given day, each person buying one package. What is the probability that all buyers found the kind of butter they wanted?

SOLUTION. We assume that the choices made by different people are independent (which may be a reasonable assumption, but we advise the reader to think of a situation where assumption of independence is not valid). If we let S_{15} denote the number of persons (out of 15 buyers) who bought unsalted FATCOW butter, then S_{15} has binomial distribution BIN$(15, 0.4)$. If all customers are to find the kind of butter they want, we must have $S_{15} \leq 8$ and also $15 - S_{15} \leq 12$, that is $S_{15} \geq 3$. Thus we need $P\{3 \leq S_{15} \leq 8\}$. Rather than computing the sum $\sum_{k=3}^{8} \binom{15}{k}(0.4)^k(0.6)^{15-k}$, it is more convenient to get the required probability as $P\{S_{15} \leq 8\} - P\{S_{15} \leq 2\}$, and read the two values from Table A.1 for $n = 15$ and $p = 0.4$.

Table A.1 gives the binomial probabilities for selected values of p only up to $p = 0.5$. To use binomial tables when $p > 0.5$, it suffices to observe the following: *if* S_n *has distribution* BIN(n, p), *then* $S_n' = n - S_n$ *has the distribution* BIN$(n, 1-p)$. To put it simply, the role of "successes" and "failures" need to be reversed.

■ **EXAMPLE 9.2 A Warning**

The probability that a Montesuma rose will blossom during the first year after planting is 80%. Mrs. Smith bought 20 bushes of Montesuma roses and planted them in her garden. What is the probability that less than 75% of her roses will blossom the first year?

SOLUTION. We regard the number S_{20} of rose bushes that will blossom in the first year as a random variable with distribution BIN$(20, 0.8)$. The probability $P\{S_{20} < 15\} = P\{S_{20} \leq 14\}$ cannot be obtained directly from the Table A.1, since $p = 0.8$ exceeds 0.5. Therefore we introduce $S'_{20} = 20 - S_{20}$—the number of roses that do not blossom. S'_{20} has binomial distribution BIN$(20, 0.2)$, and $P\{S_{20} < 15\} = P\{20 - S'_{20} < 15\} = P\{S'_{20} > 5\} = 1 - P\{S'_{20} \leq 4\} = 1 - 0.630 = 0.370$, the latter value obtained from Table A.1.

The truth is, however, that the *real* answer to this problem simply cannot be computed. Indeed, the probability 0.8 of blossoming in the first year after planting presumably represents some kind of overall success average, obtained from data for various years, soil conditions, gardening techniques, and so on. The roses of Mrs. Smith are likely to be subjected to the specific conditions, such as same type of care. Consequently, the probability of success for Mrs. Smith garden need not be 0.8, and—more important—blossoming of her roses is unlikely to be independent of one another.

A more realistic assumption here might be that probability of blossoming p is random. Keeping the assumption of independence for every p, this will make the blossoming of Mrs. Smith's roses exchangeable (but not independent) events (see Chapter 4). Thus the solution above is obtained only at the cost of accepting the assumption of independence and $p = 0.8$.

This example is placed here to make the reader aware that modeling real situations requires accepting some assumptions. Justification of these assumptions can sometimes be a rather delicate issue.

A difficulty with tabulating binomial distribution lies in the fact that it is necessary to have a separate table for each pair n and p, for $0 < p \leq 0.5$. Moreover, the tables become increasingly cumbersome when n increases. As we will show in subsequent sections, the situation can be remedied substantially for large n. As a preparation for the approximation introduced later in this chapter, let us consider an example.

■ **EXAMPLE 9.3**

Assume that about one birth in 80 is a twin birth. What is the probability that there will be no twin births among the next 200 births in the maternity ward of a given hospital?

SOLUTION. Clearly, we have here a binomial situation. If S_{200} denotes the number of twin births among the next 200 births, we need $P\{S_{200} = 0\}$,

where $S_{200} \sim \text{BIN}(200, 1/80)$. Thus, remembering that $\lim_{n\to\infty}(1 - \frac{c}{n})^n = e^{-c}$, we write:

$$\begin{aligned} P\{S_{200} = 0\} &= \binom{200}{0}\left(\frac{1}{80}\right)^0\left(\frac{79}{80}\right)^{200} = \left(1 - \frac{1}{80}\right)^{200} \\ &= \left(1 - \frac{200/80}{200}\right)^{200} \approx e^{-200/80} = 0.0821. \end{aligned} \tag{9.4}$$

The approximation works well if the probability of success p is small and number of trials n is large. Specifically, the error, equal to $|P\{S_n = 0\} - e^{-np}|$, depends on the value of the product np. As we will see, (9.4) is an example of the Poisson approximation to the binomial distribution, which we will discuss later.

Geometric Distribution

Another random variable related to Bernoulli trials is the number X of failures preceding the first success. We earlier encountered this variable in Examples 6.9 and 8.4. Clearly,

$$P\{X = k\} = q^k p, \quad k = 0, 1, 2, \ldots. \tag{9.5}$$

Since the tails of the distribution of X, that is,

$$P\{X \geq k\} = q^k p + q^{k+1}p + \cdots = p\frac{q^k}{1-q} = q^k, \tag{9.6}$$

are obtained by summing the geometric series, the distribution (9.5) is often called *geometric*. Based on formula (9.6), one immediately obtains the cdf formula, since

$$F_X(k) = P(X \leq k) = 1 - q^{k+1}.$$

A distribution closely associated with (9.5) is that of the number Y of trials up to and including the first success, which we write as

$$P\{Y = n\} = P\{X = n - 1\} = q^{n-1}p, \quad n = 1, 2, \ldots. \tag{9.7}$$

Actually, both distributions, (9.5) and (9.7), are called geometric in the literature. Since $Y = X + 1$, these distributions are closely related, and most results are valid under either definition. In this book we will use both definitions, specifying (in cases where it makes a difference) whether (9.5) or (9.7) is used. We use both definitions deliberately, not in order to confuse readers but to make them flexible in using terminology that is not well established. We will also use the same symbol for both distributions, (9.5) and (9.7). In other words, we will use the notation GEO (p) to denote the geometric distribution with probability of success being p.

The expected value of the geometric distribution was found in Example 8.2.4. For the distribution (9.7), we have

$$E(Y) = \frac{1}{p}, \tag{9.8}$$

so for the distribution (9.5), we obtain

$$E(X) = E(Y - 1) = \frac{1}{p} - 1 = \frac{q}{p}. \tag{9.9}$$

The variance of the geometric distribution is the same for both X and Y (see Theorem 8.6.1) and is easily found to be

$$\mathrm{Var}(X) = \mathrm{Var}(Y) = \frac{q}{p^2}. \tag{9.10}$$

Finally, the mgf of the geometric distribution equals

$$m_X(t) = \frac{p}{1 - qe^t}$$

for $qe^t < 1$; hence for $t < -\log q$. Consequently, by Theorem 8.5.4,

$$m_Y(t) = m_{X+1}(t) = \frac{pe^t}{1 - qe^t}.$$

Let us observe that for all $m, n = 0, 1, 2, \ldots$ we have

$$P\{X \geq m + n | X \geq m\} = P\{X \geq n\}. \tag{9.11}$$

Indeed, since $X \geq m + n$ implies that $X \geq m$, we can write

$$P\{X \geq m + n | X \geq m\} = \frac{P\{X \geq m + n\}}{P\{X \geq m\}} = \frac{q^{m+n}}{q^m} = q^n = P\{X \geq n\}.$$

Formula (9.11) is said to express the *memoryless property* of geometric distribution: if waiting time for the first success is at least m, then the probability that it will be at least $m + n$ is the same as the probability that the waiting time for the first success will be at least n.

Formula (9.11) is valid for random variable X. For the random variable Y (number of trials up to and including the first success) we have a formula with the strict inequalities

$$P\{Y > m + n | Y > m\} = P\{Y > n\}. \tag{9.12}$$

Formula (9.11) characterizes geometric distribution (9.5); equivalently, (9.12) characterizes distribution (9.7). We will prove

Theorem 9.1.1 *If Y is a discrete random variable that may assume values $1, 2, \ldots$ and satisfies* (9.12)*, then Y has the distribution* (9.7) *for some p and $q = 1 - p$.*

Proof. Let $\pi_k = P\{Y = k\}$ and $\eta_k = \pi_{k+1} + \pi_{k+2} + \cdots = P\{Y > k\}$. Then (9.12) means that we must have $\eta_{m+n}/\eta_m = \eta_n$ for all $m, n = 1, 2, \ldots$. In particular, $\eta_2 = \eta_1^2$, and by induction, $\eta_k = \eta_1^k$. Consequently, $\pi_k = \eta_k - \eta_{k+1} = \eta_1^k - \eta_1^{k+1} = \eta_1^k(1 - \eta_1)$ for all k, which means that Y has geometric distribution with $q = \eta_1 = P\{Y > 1\}$. □

It is sufficient to require that (9.12) holds for $m = 1$ and all n or for $n = 1$ and for all m. Also the Theorem 9.1.1 holds if m is replaced by a random variable (see Srivastava, 1981).

Theorem 9.1.2 *Let U be a random variable which assumes only positive integer values. If*

$$P\{Y > U + n | Y > U\} = P\{Y > n\}, \tag{9.13}$$

then Y has geometric distribution (9.7) *for some p and $q = 1 - p$.*

■ **EXAMPLE 9.4**

In Chapter 5 we discussed Markov chains, these being processes that describe the evolution of a system that may at any time be in one of a specified set of states and may change its state at times $t = 1, 2, \ldots$. The main assumption is that if at some time t the system is in state i, then the probability that it will pass to state j at time $t + 1$ is p_{ij}, regardless of the history of transitions prior to time t. This means that given the present (state at time t), the future and the past are independent. Suppose now that at some time t the system enters state i, and let T_i be the duration of the current visit to state i. More precisely, we define $T_i = k$ if the state at times $t, t + 1, t + 2, \ldots, t + k$ is i, but the state at time $t + k + 1$ is different from i. We have then

$$P\{T_i = k\} = p_{ii}^k(1 - p_{ii}), \quad k = 1, 2, \ldots,$$

and we see that the duration of a visit in state i has geometric distribution with $q = p_{ii}$ and $p = 1 - p_{ii}$.

The memoryless property of geometric distribution is in fact the Markov property. Indeed, suppose that the system stayed in state i for m units of time prior to time t. Then the probability of staying there for at least n additional units is the same as the probability of staying in state i for at least $m + n$ units of time. Given the present state, the future (in particular, the duration of remaining stay in the present state) is independent of the past (in particular, of how long the system has already stayed in its present state).

Theorem 9.1.2 asserts a seemingly identical property, but a moment of reflection shows that the condition is now much stronger: the "present" is not some fixed time, but it is a time that may be random. In particular, this randomness may be affected by the process itself.

To give an example, let $X = T_i$, the time of remaining in the state i just entered, and let U be the longest visit in state i recorded thus far (so that U depends on the history of the process up to the present time). Theorem 9.1.2 asserts, in particular, the following: given that the duration of the present state in state i will break the record ($T_i > U$), the probability of breaking the record by at least 3 units (i.e., $T_i > U + 3$) is the same as the probability of a visit lasting longer than 3 units ($T_i > 3$). This property stands in contrast with what one observes in sport records: the consecutive improvements of the world record in any discipline tend to be smaller.

The property asserted in Theorem 9.1.2 is generally called *strong* Markov property. It means that given the present, the future is independent of the past, even if the "present" is selected at random, as long as this randomness depends only on the past, and not on the future.

Theorems 9.1.1 and 9.1.2 are examples of *characterizations* of probability distributions. This kind of theorem singles out a certain property of a type of distribution and shows that this property is valid only for distributions of this type (in the case above, the memoryless property for the geometric distribution).

The characterization theorems are of a great practical value. As we will see in the later chapters, if one is interested in more than merely summarizing and presenting

the statistical data, it is useful to regard the data (results of observations, experiments, surveys, etc.) as values of some random variables. Roughly speaking, the more we know about the distributions of these random variables, the better is our understanding of the phenomenon studied, as well as the better are our possibilities of prediction and control.

To fix the idea, imagine that a statistician's data can be regarded as independent observations of some random variable X. It happens quite often that identification of the distribution of X is accomplished in two steps. The first step consists of identifying a class of distributions that contains the distribution of X. The second step is the identification of the particular distribution in this class.

Now, sometimes the first step is easy. The statistician often knows the type of distribution of X because he *makes* it belong to a given type by appropriate sampling. For instance, in independent sampling, each element sampled is classified to one of two categories: defective/nondefective, treatment successful/unsuccessful, and so on. The total number X of elements of a given type is binomial, as long as the sample is really random. Out of two parameters, n is controlled by the experimenter, and p (the fraction of elements of a given kind in the population) is unknown and is to be estimated.

In many cases, however, determining the class of distributions that contains distribution of X is not so simple. In such situations one can sometimes use a characterization theorem: if one knows that the distribution has some property that turns out to be characteristic for a given class, then the distribution must be in that class (e.g., if one knows that the distribution of X is memoryless, then it must be geometric).

We will now give one more example connected with the geometric distribution.

■ **EXAMPLE 9.5 Family Planning**

Assume that the sexes of consecutive children born to the same parents are independent, with the probability of a child being a girl equal to π. We will disregard twin births, and consider family sizes under various plans.

Plan 1. The couple decides to stop having children as soon as their first girl is born. Let ξ_1 be the number of children according to this plan. Then ξ_1 is a geometric random variable with $P\{\xi_1 = n\} = (1-\pi)^{n-1}\pi$ so that $E(\xi_1) = 1/\pi$. If the chances of a child being a boy are the same as those of being a girl, we have $\pi = 0.5$ and $E(\xi_1) = 2$.

Plan 2. The couple decides to have children until they have a boy and a girl and then stop. Let ξ_2 be the number of children under this plan. To determine $P(\xi_2 = n)$, let G and B denote the event "first child is a girl (boy)." Then, by the total probability formula (4.9), for $n \geq 2$ we have

$$\begin{aligned} P(\xi_2 = n) &= \pi P(\xi_2 = n|G) + (1-\pi)P(\xi_2 = n|B) \\ &= \pi P\{Y^B = n-1\} + (1-\pi)P\{Y^G = n-1\}. \end{aligned}$$

Here Y^B and Y^G are the numbers of children (excluding the first) that a family will have until the first boy (girl) is born. Clearly, Y^B and Y^G have the distribution given by (9.6), with probability of success being $1-\pi$ for Y^B

and π for Y^G. Thus, for $n = 2, 3, \ldots$,

$$\begin{aligned} P\{\xi_2 = n\} &= \pi[\pi^{n-2}(1-\pi)] + (1-\pi)[(1-\pi)^{n-2}\pi] \\ &= \pi^{n-1}(1-\pi) + (1-\pi)^{n-1}\pi. \end{aligned}$$

To find the expected number of children under plan 2, we write

$$E(\xi_2) = \pi \sum_{n=2}^{\infty} nP\{Y^B = n-1\} + (1-\pi) \sum_{n=2}^{\infty} nP\{Y^G = n-1\},$$

and, putting $k = n - 1$, obtain

$$\begin{aligned} \sum_{n=2}^{\infty} nP\{Y^B = n-1\} &= \sum_{n=2}^{\infty} [(n-1)+1]P\{Y^B = n-1\} \\ &= \sum_{k=1}^{\infty} kP\{Y^B = k\} + \sum_{k=1}^{\infty} P\{Y^B = k\} \\ &= \frac{1}{1-\pi} + 1. \end{aligned}$$

By symmetry, $\sum_{n=2}^{\infty} nP\{Y^G = n-1\} = 1/\pi + 1$, so

$$\begin{aligned} E(\xi_2) &= \pi\left(\frac{1}{1-\pi} + 1\right) + (1-\pi)\left(\frac{1}{\pi} + 1\right) = \frac{\pi}{1-\pi} + \frac{1-\pi}{\pi} + 1 \\ &= \frac{1}{\pi(1-\pi)} - 1. \end{aligned}$$

Again, if $\pi = 1/2$, then $E(\xi_2) = 3$.

Negative Binomial Distribution

The geometric distribution allows us to make an immediate and natural extension. Rather than to consider the number of Bernoulli trials up to the first success, we can consider the number Y of Bernoulli trials up to and including the rth success. In analogy with the geometric distribution, we will also consider the random variable X defined as the number of failures preceding the rth success.

We will start by deriving the probability distribution of the random variables X and Y. Clearly, the possible values of Y are integers $r, r+1, r+2, \ldots$. The event $\{Y = n\}$ occurs if:

1. The nth trial results in success.

2. The first $n-1$ trials give exactly $r-1$ successes (and $n-r$ failures).

Indeed, the conjunction of 1 and 2 ensures that the rth success occurs at trial n. The events 1 and 2 are independent (since their occurrence is determined by disjoint sets of trials), and their probabilities are p and $\binom{n-1}{r-1}p^{r-1}q^{n-r}$, respectively. Consequently,

$$P\{Y = n\} = \binom{n-1}{r-1} p^r q^{n-r} \tag{9.14}$$

for $n = r, r+1, r+2, \ldots$. The number X of failures preceding the rth success is such that $X + r = Y$. So for $k = 0, 1, 2, \ldots$, we have

$$P\{X = k\} = P\{Y = k + r\} = \binom{k+r-1}{r-1} p^r q^k. \tag{9.15}$$

Both random variables X and Y are referred to as having a *negative binomial* distributions. In either case, the distribution depends on two parameters, r and p. We will use the symbol NBIN(r, p) to denote the negative binomial distribution, whether (9.14) or (9.15).

Let us verify that (9.15) and (9.14) are probability distributions. We must show that

$$\sum_{k=0}^{\infty} \binom{k+r-1}{r-1} p^r q^k = 1. \tag{9.16}$$

Let us observe first that by formulas (3.14) and (3.26),

$$\begin{aligned}
\binom{k+r-1}{r-1} &= \binom{r-1+k}{k} \\
&= \frac{(r-1+k)(r-1+k-1)\cdots(r-1+k-k+1)}{k!} \\
&= \frac{r(r+1)\cdots(r+k-1)}{k!} \\
&= \frac{(-1)^k(-r)(-r-1)\cdots(-r-k+1)}{k!} = (-1)^k \binom{-r}{k}.
\end{aligned}$$

Thus, using Newton's formula (3.28), we have

$$\sum_{k=0}^{\infty} \binom{k+r-1}{r-1} p^r q^k = p^r \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} q^k = p^r (1-q)^{-r} = 1.$$

Obviously the proof that

$$\sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r q^{n-r} = 1$$

is similar, and we can omit the details.

To determine the mean and variance of a random variable with negative binomial distribution, one can proceed in several ways, of which we will demonstrate two. The first uses direct calculations while the second uses a representation as a sum of simpler random variables.

Observe that since $Y + r = X$, we have $E(Y) + r = E(X)$ and $\mathrm{Var}(Y) = \mathrm{Var}(X)$ (the latter property being a consequence of Theorem 8.6.1). Thus it suffices to study one of the random variables X and Y.

We begin by finding the mgf of X. Proceeding as in the proof of (9.16), we have

$$\begin{aligned}
m_X(t) &= Ee^{Xt} = \sum_{k=0}^{\infty} e^{kt} \binom{k+r-1}{r-1} p^r q^k = p^r \sum_{k=0}^{\infty} \binom{k+r-1}{r-1} (qe^t)^k \\
&= p^r \sum_{k=0}^{\infty} \binom{-r}{k} (-qe^t)^k = \frac{p^r}{(1-qe^t)^r},
\end{aligned}$$

provided that $|-qe^t| < 1$, or equivalently $t < \log(1/q)$.

Consequently, $E(X) = m_X'(0) = rq/p$, and

$$E(Y) = r + E(X) = \frac{r}{p}. \tag{9.17}$$

After some algebra, we also obtain

$$\text{Var}(X) = \text{Var}(Y) = m_X''(0) - [E(X)]^2 = \frac{rq}{p^2}. \tag{9.18}$$

■ **EXAMPLE 9.6**

A salesman calls prospective buyers to make a sales pitch. Assume that the outcomes of consecutive calls are independent, and that on each call he has 15% chance of making a sale. His daily goal is to make 3 sales, and he can make only 20 calls in a day. What is the probability that he achieves his goal in 18 trials? What is the probability that he does not achieve his daily goal?

SOLUTION. The "strategy" of solving a problem like this is to identify the type of distribution to be analyzed. Assuming that we identify the situation as Bernoulli trials (i.e., repeated independent trials with the same probability p of success), the problem most typically concerns either a binomial or negative binomial distribution. The crucial question here is: Is the number of trials fixed (and then the number of successes is random), or is the number of successes fixed (and the number of trials is random)? In the first case we have the binomial distribution, and in the second case, the negative binomial distribution.

For our salesman, we want the probability that his third sale (success) come at trial 18. Thus the number of successes $r = 3$ is fixed (this is the salesman's goal), and it is the number of calls that is random. We have $p = 0.15$ and we ask for $P\{X = 18\}$, where X is the number of trials up to and including the third success. Substitution to formula (9.14) gives

$$P\{X = 18\} = \binom{17}{2}(0.15)^3(0.85)^{15} = 0.0401.$$

The second question is about the probability of the salesman not attaining his daily goal. Here the number of trials is fixed ($n = 20$,) and we can treat the problem as involving the binomial distribution. Thus, if S_{20} is the number of successes in 20 trials, the salesman does not achieve his goal if $S_{20} \leq 2$. The answer is therefore

$$\begin{aligned} P\{S_{20} \leq 2\} &= P\{S_{20} = 0\} + P\{S_{20} = 1\} + P\{S_{20} = 2\} \\ &= \sum_{k=0}^{2} \binom{20}{k}(0.15)^k(0.85)^{20-k} = 0.4049. \end{aligned}$$

We can also use the negative binomial distribution here. If X is the number of trials up to and including the third success, then the salesman does not attain

his goal when $X > 20$. So the answer is

$$P\{X > 20\} = \sum_{n=21}^{\infty} P\{X = n\} = \sum_{n=21}^{\infty} \binom{n-1}{2} (0.15)^3 (0.85)^{n-3},$$

a quantity that is much harder to evaluate numerically than $P\{S_{20} \leq 2\}$.

The example above suggests that the probabilities in binomial and negative binomial distributions with the same probability p of success are related. The following theorem provides a useful identity:

Theorem 9.1.3 *Let $S_n \sim \text{BIN}(n, p)$ (the number of successes in n trials) and let $Y^{(r)} \sim \text{NB}(r, p)$ (the number of trials up to and including the rth success). Then, for every $k = 0, 1, \ldots,$*

$$P\{Y^{(r)} > k\} = P\{S_k < r\}. \tag{9.19}$$

Proof. Observe that both sides of (9.19) refer to the probability of the same event: the waiting time for the rth success exceeds k if and only if fewer than r successes occur in the first k trials. □

A word of warning: Since $Y^{(r)}$ and S_k are both discrete random variables, it matters whether or not the inequalities are strict. Thus (9.19) can be written in any of the forms, such as

$$P\{Y^{(r)} > k\} = P\{S_k \leq r - 1\}, \quad P\{Y^{(r)} < k + 1\} = P\{S_k \geq r\}.$$

An inspection of the proof of the formula (9.16) shows that the fact that r was an integer was never used. Formally, a negative binomial distribution is defined for any $r > 0$ and $0 < p < 1$ (although the interpretation of the probabilities in terms of Bernoulli trials is no longer valid).

Let us also note that the mgf of the negative binomial distribution is the rth power of the mgf of the geometric distribution. For integer values of r this fact shows that the random variable $X^{(r)}$ (respectively, $Y^{(r)}$) is a sum of r independent geometric random variables (of the form X or Y, depending on whether we represent $X^{(r)}$ or $Y^{(r)}$). This representation (e.g., in the case of $X^{(r)}$) means that

$$X^{(r)} = X_1 + \cdots + X_r,$$

where X_i is interpreted as the number of failures falling between the $(i-1)$st and ith success (e.g., in case of $r = 3$, if the consecutive trials are FFFSSFFFFS, then $X^{(3)}$= 7 = number of failures preceding the third success, with $X_1 = 3, X_2 = 0, X_3 = 4$ being the numbers of failures between the three consecutive successes).

Consequently,

$$E(X^{(r)}) = E(X_1) + \cdots + E(X_r)$$

and (because of independence)

$$\text{Var}(X^{(r)}) = \text{Var}(X_1) + \cdots + \text{Var}(X_r),$$

which in view of (9.8) and (9.10) gives formulas (9.17) and (9.18). The representation of random variables $Y^{(r)}$ is analogous and will be omitted.

PROBLEMS

9.1.1 Label statements below as true or false:
(i) Suppose that 6% of all cars in a given city are Toyotas. Then the probability that there are 4 Toyotas in a row of 12 cars parked in the municipal parking is $\binom{12}{4}(0.06)^4(0.94)^8$.
(ii) Suppose that 6% of all Europeans are French. Then the probability that in a random sample of 12 inhabitants of a major European capital there are 4 Frenchmen is $\binom{12}{4}(0.06)^4(0.94)^8$.

9.1.2 Assume that we score $Y = 1$ for a success and $Y = -1$ for a failure. Express Y as a function of the number X of successes in a single Bernoulli trial, and find moments $E(Y^n)$, $n = 1, 2, \ldots$.

9.1.3 Suppose that random variables $X_1, \ldots, X_n$ are independent, each with the same Bernoulli distribution. Given that $\sum_{i=1}^{n} X_i = r$, find: (i) The probability that $X_1 = 1$. **(ii)** The covariance between X_i and X_j, $1 \leq i < j \leq n$.

9.1.4 Assume that X_1 and X_2 are independent random variables, with binomial distributions with parameters n_1, p and n_2, p respectively. Find a correlation coefficient between X_1 and $X_1 + X_2$.

9.1.5 An experiment consists of tossing a fair coin 13 times. Such an experiment is repeated 17 times. Find the probability that in a majority of repetitions of the experiment the tails will be in minority.

9.1.6 Two players (or two teams) are negotiating the rules for determining the championship. The two possibilities are "best of five" or "best of seven." This means that whoever wins three (respectively four) games is the champion. Assume that games are independent, and that the probability of winning a game by the first player (there are no ties) is p. For what values of p, should the first player favor the scheme "best out of five"?

9.1.7 Show that if S_n is a binomial random variable, then for $k = 1, 2, \ldots, n$,

$$P\{S_n = k\} = \frac{(n-k+1)p}{k(1-p)} P\{S_n = k-1\}.$$

9.1.8 Show that for the binomial distribution we have

$$P\{S_n \leq k\} = (n-k)\binom{n}{k} \int_0^{1-p} x^{n-k-1}(1-x)^k dx.$$

9.1.9 Six dice are tossed simultaneously until, for the first time, all of them show the same face. Find $E(U)$ and $\mathrm{Var}(U)$, where U is the number of tosses until this happens.

9.1.10 Assume that the probability of twins being identical is β, and that the sexes of children are determined independently, with probability of a boy being b (possibly $b \neq 1/2$). Find the expected number of twin births recorded in the hospital before the first pair of: **(i)** Boys. **(ii)** Girls. **(iii)** Different genders. Note that identical twins must be of the same sex.

9.1.11 Assume that the probability that a birth is a multiple one (twins, triplets, etc.) is π. Given that a birth is a multiple one, probabilities $\alpha_2, \alpha_3, \ldots$ of twins, triplets, ...satisfy the condition $\alpha_{k+1} = \gamma\alpha_k$. Find π, α_2 and γ if it is known that the expected number of children born in 100 births is $100 + c$, and the expected number of single births observed before recording a multiple birth is M (assume that M and c given).

9.1.12 A hospital needs 20 volunteers for the control group in testing the efficiency of some treatment. The candidates are subject to psychological and medical screening, and on average, only 1 in 15 candidates is found acceptable for the experiment. The cost of screening, whether or not a candidate is found acceptable, is \$50 per person. The granting agency argues that one needs about 50×15 dollars to find one acceptable candidate, and therefore allows $20 \times 50 \times 15 = 15,000$ dollars for the cost of screening. **(i)** Write the formula for the probability that the allocated sum will be enough to find 20 acceptable candidates. **(ii)** Use the Chebyshev inequality to find the sum that gives at least a 90% chance of finding 20 acceptable candidates before the testing money runs out.

9.1.13 Assume that in a tennis match between A and B, the probability of winning a tennis set by player A is p, and that the results of sets are independent. Let T be the number of sets played in a match. Find the distribution of T and $E(T)$ as a function of p, assuming that the match is played by: **(i)** Men. **(ii)** Women (Note that men play "best out of five," while women play "best out of three" sets.)

9.1.14 In the flowchart in Figure 9.1, $m > 0$ is an integer and $0 < p < 1$. The block "sample U" means that a value of a random variable U is sampled from the U[0, 1] distribution, with consecutive samplings being independent. Find the distribution of X, and then calculate its mean and variance.

9.1.15 Show that if X has a negative binomial distribution NBIN(r, p), then

$$E[(r-1)/(r+X-1)] = p.$$

9.2 HYPERGEOMETRIC DISTRIBUTION

The distributions discussed in the previous section were based on the notion of a Bernoulli trial—an event that results in a "success" or a "failure." One of the necessary assumptions of Bernoulli trials is that the probabilities of a "success" (or equivalently "failure") are the same in each trial. That requires either sampling from an *infinite* (practically very large) population or returning an element after it was selected (sampling with replacement). There are, however, situations where neither the population is large nor the replacing can be done. We then have sampling without replacement, with the probabilities changing after every selection and each time

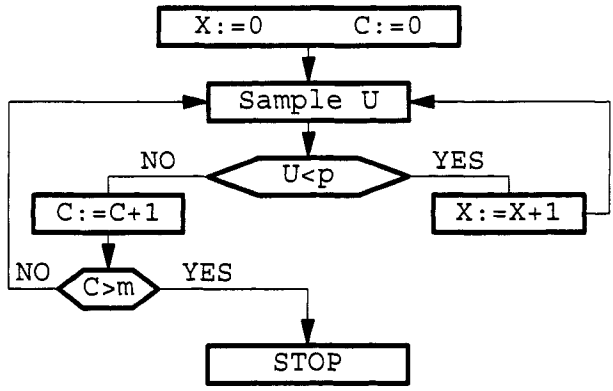


Figure 9.1 Flowchart

depending on the type of the element recovery as often used for statistical modeling (e.g., in zoology to study small animal or plant populations). We have already encountered hypergeometric distribution in Examples 3.6 and 3.7.

To proceed systematically, we assume that initially the population has N elements of which a are of one kind ("successes") and $b = N - a$ are of another kind ("failures"). We sample n elements without replacement, and we let X denote the number of successes in the sample. Let us first determine the range of the random variable X. Clearly, $0 \le X \le a$, but we must also have a similar inequality for the number of failures, namely $0 \le n - X \le N - a$. This yields

$$\max(0, n - (N - a)) \le X \le \min(n, a). \tag{9.20}$$

If k satisfies the constraint (9.20), then

$$P\{X = k\} = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}. \tag{9.21}$$

To check that the probabilities in (9.21) add up to 1, we will use a trick. Let us consider the identity

$$(1 + x)^N = (1 + x)^a (1 + x)^{N-a}$$

and compare the coefficients of x^n on both sides. On the left-hand side the coefficient is $\binom{N}{n}$. On the right-hand side the coefficient equals the sum of all possible terms of the form $\binom{a}{k}\binom{N-a}{n-k}$, where k must satisfy the constraint (9.20). This shows that the sum of all terms (9.21) is 1, as asserted.

■ **EXAMPLE 9.7**

A class consists of 10 boys and 12 girls. The teacher selects 6 children at random for some task. Is it more likely that she chooses 3 boys and 3 girls or that she chooses 2 boys and 4 girls?

SOLUTION. If X is the number of boys in the selected set of children, then

$$P\{X=3\}=\frac{\binom{10}{3}\binom{12}{3}}{\binom{22}{6}}=0.3538, \qquad P\{X=2\}=\frac{\binom{10}{2}\binom{12}{4}}{\binom{22}{6}}=0.2985.$$

So the first probability is higher than the second.

As expected, for large populations it does not matter too much whether we sample with or without replacement. We shall now formulate the corresponding approximation theorem:

Theorem 9.2.1 *Let $N \to \infty$ and $a \to \infty$ in such a way that $a/N \to p$, where $0 < p < 1$. Then for every fixed n and $k = 0, 1, \ldots n$ we have*

$$\frac{P\{X=k\}}{b(k;n,p)} \to 1 \tag{9.22}$$

as $N \to \infty$. Here $P\{X=k\}$ is the probability (9.21) for hypergeometric distribution, while $b(k;n,p)$ is the binomial probability given by (9.3).

Proof. Letting J_N denote the ratio in (9.22), we write

$$\begin{aligned} J_N &= \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}\binom{n}{k}p^k(1-p)^{n-k}} \\ &= \frac{a!(N-a)!(N-n)!}{(a-k)!(N-a-n+k)!N!p^k(1-p)^{n-k}}, \end{aligned}$$

which cancels to

$$\frac{[a(a-1)\cdots(a-k+1)][(N-a)(N-a-1)\cdots(N-a-n+k+1)]}{N(N-1)\cdots(N-n+1)p^k(1-p)^{n-k}}.$$

Multiplying and dividing by N^n, we obtain $J_N = A_N B_N / C_N$, where

$$\begin{aligned} A_N &= \frac{a}{N}\left(\frac{a}{N}-\frac{1}{N}\right)\cdots\left(\frac{a}{N}-\frac{k-1}{N}\right), \\ B_N &= \left(1-\frac{a}{N}\right)\left(1-\frac{a}{N}-\frac{1}{N}\right)\cdots\left(1-\frac{a}{N}-\frac{n-k-1}{N}\right), \\ C_N &= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)p^k(1-p)^{n-k}. \end{aligned}$$

Since n and k are fixed and $a/N \to p$, $A_N \to p^k$, $B_N \to (1-p)^{n-k}$, and $(1-1/N)\cdots(1-(n-1)/N) \to 1$, we have $J_N \to 1$. This proves the theorem. □

The practical use of Theorem 9.2.1 consists mostly of replacing probabilities

$$P\{X=k\}=\frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$$

by their approximations $\binom{n}{k}\left(\frac{a}{N}\right)^k\left(1-\frac{a}{N}\right)^{n-k}$, which are much easier to compute. The relative error of such approximation depends on all four factors, N, n, a, and k.

Let us now find the expectation and variance of the hypergeometric random variable X with distribution (9.21). We will use the method that we already used in the case of binomial and negative binomial random variables, by representing X as a sum of simpler random variables. We let

$$X = \xi_1 + \xi_2 + \cdots + \xi_n, \tag{9.23}$$

where ξ_i equals 1 or 0, depending on whether the ith draw results in success or failure.

Observe that $\xi_1, \cdots, \xi_n$ are dependent random variables. This fact will affect our calculation of the variance of X. Generally, we will use the formulas

$$\begin{aligned} E(X) &= E(\xi_1) + \cdots + E(\xi_n), \\ \mathrm{Var}(X) &= \mathrm{Var}(\xi_1) + \cdots + \mathrm{Var}(\xi_n) + 2\sum_{i<j}\mathrm{Cov}(\xi_i, \xi_j). \end{aligned}$$

Note that ξ_i's are Bernoulli random variables (i.e., they assume only values 0 and 1). So letting $p_i = P\{\xi_i = 1\}$, we have

$$E(\xi_i) = p_i, \mathrm{Var}(\xi_i) = p_i(1-p_i). \tag{9.24}$$

Observe that $\xi_i\xi_j$ is also a Bernoulli random variable. So letting $p_{ij} = P\{\xi_i = 1, \xi_j = 1\}$ we have

$$\mathrm{Cov}(\xi_i, \xi_j) = E(\xi_i\xi_j) - E(\xi_i)E(\xi_j) = P\{\xi_i\xi_j = 1\} - p_ip_j = p_{ij} - p_ip_j. \tag{9.25}$$

To determine the probabilities p_{ij}, we need the joint distribution of (ξ_i, ξ_j). Generally, one of the ways of visualizing the distribution of the $(\xi_1, \ldots, \xi_n)$ is as follows: The elements of the population (of a successes and $N-a$ failures) are ordered at random. This gives the sample space of $N!$ permutations, all of them being equally likely. The random vector $(\xi_1, \ldots, \xi_n)$ is then defined at any sampled point (permutation) as the initial n elements of this permutation.

It is now clear that the joint distribution of (ξ_i, ξ_j) does not depend on (i, j), meaning it is the same as the joint distribution of (ξ_1, ξ_2). Indeed, the probability that $(\xi_i = x, \xi_j = y)$ depends on the number of permutations in the sample space that have specific elements x and y at places i and j. By symmetry, this number is the same as the number of permutations that have elements x and y in the first two places (or in any other designated pair of places).

Consequently, the marginal distributions of ξ_i are the same, and for all i we must have

$$p_i = P\{\xi_i = 1\} = \frac{a}{N}.$$

Similarly

$$p_{ij} = P\{\xi_1 = 1, \xi_2 = 1\} = P\{\xi_1 = 1\} \times P\{\xi_2 = 1 | \xi_1 = 1\} = \frac{a}{N} \times \frac{a-1}{N-1}.$$

From (9.24) and (9.25) we get $E(\xi_i) = a/N, \mathrm{Var}(\xi_i) = (a/N)(1-a/N)$. Therefore

$$\begin{aligned}\mathrm{Cov}(\xi_i, \xi_j) &= E(\xi_i\xi_j) - E(\xi_i)E(\xi_j) = p_{ij} - E(\xi_i)E(\xi_j) \\ &= \frac{a}{N} \times \frac{a-1}{N-1} - \left(\frac{a}{N}\right)^2 = -\frac{1}{N} \times \frac{a(N-a)}{N(N-1)} = -\frac{\mathrm{Var}(\xi_i)}{N-1}.\end{aligned}$$

Thus $E(X) = n(a/N)$. For variance, observe that the number of pairs (i, j) with $i < j$ is $\binom{n}{2} = n(n-1)/2$. Consequently

$$\begin{aligned}\mathrm{Var}(X) &= n \times \frac{a}{N} \times \frac{N-a}{N} - 2 \times \frac{n(n-1)}{2} \times \frac{1}{N} \times \frac{a(N-a)}{N(N-1)} \\ &= n \times \frac{a}{N} \times \frac{N-a}{N}\left(1 - \frac{n-1}{N-1}\right).\end{aligned}$$

Letting $a/N = p$ denote the probability of success prior to selecting the first element, we have the following theorem:

Theorem 9.2.2 *If X has the hypergeometric distribution* (9.21)*, then*

$$\begin{aligned}E(X) &= n \times \frac{a}{N} = np, \\ \mathrm{Var}(X) &= n \times \frac{a}{N} \times \left(1 - \frac{a}{N}\right)\frac{N-n}{N-1} = np(1-p)\frac{N-n}{N-1}.\end{aligned}$$ □

Thus the expected number of successes in a sample of size n is np regardless of whether we sample with or without replacement. The variance in the case of sampling without replacement is smaller than the corresponding binomial variance by the factor $(N-n)/(N-1)$ (sometimes referred to as *finite population correction factor*).

The behavior of the variances of binomial and hypergeometric distributions is different as the sample size, n, increases. In a binomial distribution, each new element of the sample contributes the same amount to variance, so the latter grows linearly with n. In sampling without replacement, variance changes in proportion to the product $n(N-n)$; hence it initially grows to reach a maximum when sampling exhausts half of the population, and then declines to zero when $n = N$ (at $n = N$ we exhaust the whole population, so there is no longer variability in the sample). Variance for $n = 1$ is the same as for $n = N - 1$. This is clear, since variability involved in sampling one element is the same as variability involved with leaving just one element unsampled.

The next theorem connects the binomial and hypergeometric distributions in the situation when the successes come from two sources, each following the binomial distribution, and the total number of successes is fixed. Thus we will prove

Theorem 9.2.3 *Let X and Y be independent random variables with binomial distributions $X \sim \mathrm{BIN}(m, p)$ and $Y \sim \mathrm{BIN}(n, p)$. Then*

$$P\{X = k | X + Y = r\} = \frac{\binom{m}{k}\binom{n}{r-k}}{\binom{m+n}{r}}. \tag{9.26}$$

Proof. Since $X + Y \sim \text{BIN}(m + n, p)$, we have

$$\begin{aligned} P\{X = k | X + Y = r\} &= \frac{P\{X = k, X + Y = r\}}{P\{X + Y = r\}} \\ &= \frac{P\{X = k\}P\{Y = r - k\}}{P\{X + Y = r\}}, \end{aligned}$$

and substitution of binomial probabilities gives (9.26). □

The right-hand side of (9.26) is the hypergeometric probability of k successes in sampling r elements from a population with a total of $m + n$ elements, of which m are successes.

The scheme of sampling without replacement can be generalized as follows: Assume that the population consists initially of a elements of one kind (successes) and $b = N - a$ elements of the second kind (failures). Each time an element is sampled, it is returned, and c elements of the same kind as just sampled are added to the urn. This process continues for n samplings. Let X be the number of successes in the sample. This random variable X is said to have *Pólya distribution*, and the sampling described above is called the *Pólya scheme.*

The reason for designing this scheme was as follows. Adding elements of the same kind to the population increases the probability of selecting elements of the kind most recently sampled. Such effect is known to occur when one samples from a population to determine the fraction of persons infected with a disease. Typically, if one finds one person with the disease, then the chances of finding others with the same disease increase. Pólya introduced this scheme of sampling to model such effects.

Observe that we can formally put $c = -1$ (elements are simply not returned). Thus the special case $c = -1$ yields random variable X with a hypergeometric distribution. To find $P\{X = k\}$ in the general case, let us first find the probability of sampling the elements in a specific order, for example, first k successes and then $n - k$ failures. This probability, by the chain rule (4.5), is

$$\frac{a}{N}\frac{a+c}{N+c}\cdots\frac{a+(k-1)c}{N+(k-1)c}\frac{b}{N+kc}\frac{b+c}{N+(k+1)c}\cdots\frac{b+(n-k-1)c}{N+(n-1)c}. \tag{9.27}$$

Let us observe that the probability (9.27) remains the same regardless of the order of k successes and $n - k$ failures. Each time the number of elements in the population increases by c, the product of all denominators is $N(N+1)\cdots(N+(n-1)c)$. Similarly the numerators corresponding to sampling successes are $a, a+c, a+2c, \ldots, a+(k-1)c$ regardless of when the successes are to occur, and the same holds for failures. Consequently, we have

Theorem 9.2.4 *In Pólya scheme, for* $k = 0, 1, \ldots, n$

$$P\{X = k\} = \frac{n!p(p+\gamma)\cdots(p+(k-1)\gamma)q(q+\gamma)\cdots(q+(n-k-1)\gamma)}{k!(n-k)!(1+\gamma)(1+2\gamma)\cdots(1+(n-1)\gamma)}$$

where $p = a/N, q = b/N = 1 - a/N$ *and* $\gamma = c/N$.

At the end, it is worth mentioning that the Pólya scheme can be modified in the following way: Again, we have an urn, containing initially a balls of one color ("successes") and b balls of another color ("failures"). Balls are drawn successively. After a ball is drawn, it is returned. If it was a success, c_1 balls representing *failure* are added; if it was a failure, c_2 balls representing *success* are added. Thus a success causes an increase of probability of failure in the next trial, and conversely, a failure causes an increase of probability of success in the next trial. This scheme has been used to model accidents at work. Here the situation is such that an accident causes an increase in observing safety measures, awareness of danger, and so on. On the other hand, the longer the time without an accident, the more laxity in observing safety regulations, and so on. Under the appropriate choice of a, b, c_1, and c_2, it is possible to model the "after-effects of accidents" as described above, thus modeling the distribution of times between accidents, number of accidents in a given period, and so on.

Unfortunately, the formulas are much more complicated than those for the Pólya scheme, and we will not pursue this topic here.

PROBLEMS

9.2.1 An urn contains nine chips, five of them red and four blue. Three chips are drawn without replacement. Find the distribution of X = number of red chips drawn.

9.2.2 An urn contains six chips, three red and three green. Four chips are selected without replacement. Find $E(X)$ and $\mathrm{Var}(X)$ where X = number of red chips in the sample.

9.2.3 Instead of (9.23), write $X = \eta_1 + \eta_2 + \cdots + \eta_a$, where $\eta_j = 1$ or 0 depending on whether or not the jth element representing success was selected. Use this representation to derive formulas for the mean and variance of X as given in Theorem 9.2.2.

9.3 POISSON DISTRIBUTION AND POISSON PROCESS

We start from the following definition:

Definition 9.3.1 A random variable X is said to have a Poisson distribution, POI(λ), if for some $\lambda > 0$,

$$P\{X = n\} = \frac{\lambda^n}{n!} e^{-\lambda}, \qquad n = 0, 1, \ldots. \tag{9.28}$$

□

First we want to check that the terms in (9.28) add to 1. We have

$$\sum_{n=0}^{\infty} P\{X = n\} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1.$$

To determine the moments of the Poisson distribution, let us compute the mgf:

$$\begin{aligned} m_X(t) = Ee^{tX} &= \sum_{n=0}^{\infty} e^{nt} P\{X=n\} = \sum_{n=0}^{\infty} e^{nt} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}. \end{aligned}$$

Thus a moment generating function of the Poisson distribution is defined for all t and is differentiable an arbitrary number of times. We have

$$E(X) = m_X'(t)|_{t=0} = \lambda e^t e^{\lambda(e^t-1)}|_{t=0} = \lambda.$$

An easy differentiation yields $E(X^2) = m_X''(t)|_{t=0} = \lambda^2 + \lambda$, and therefore

$$\mathrm{Var}(X) = \lambda.$$

The next theorem shows that the family of Poisson distributions is closed under addition of independent random variables:

Theorem 9.3.1 *If X and Y are independent, with distributions* POI(λ_1) *and* POI(λ_2), *respectively, then $X+Y$ has a* POI($\lambda_1+\lambda_2$) *distribution.*

Proof. We will present a direct proof, to show the kind of calculations involved in evaluating the distribution of the sum. The proof using moment generating functions can be found in Example 8.29.

$$\begin{aligned} P\{X+Y=n\} &= \sum_{j=0}^{n} P\{X=j, Y=n-j\} = \sum_{j=0}^{n} P\{X=j\}\, P\{Y=n-j\} \\ &= \sum_{j=0}^{n} \frac{\lambda_1^j e^{-\lambda_1}}{j!} \times \frac{\lambda_2^{n-j} e^{-\lambda_2}}{(n-j)!} = \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{j=0}^{n} \binom{n}{j} \lambda_1^j \lambda_2^{n-j} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!} \quad \text{(by Newton's formula)}, \end{aligned}$$

which completes the proof. □

Finally, we introduce an analogue of Theorem 9.2.3.

Theorem 9.3.2 *If X and Y are independent and $X \sim$* POI(λ_1), *$Y \sim$* POI(λ_2), *then for $k = 0, 1, \ldots, n$,*

$$P\{X=k|X+Y=n\} = \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}. \tag{9.29}$$

Thus the conditional distribution of X given $X+Y$ is binomial, with the number of trials $X+Y$ and the probability of success $p = \lambda_1/(\lambda_1+\lambda_2)$.

Proof. Using Theorem 9.3.1, we have

$$\begin{aligned} P\{X=k|X+Y=n\} &= \frac{P\{X=k, X+Y=n\}}{P\{X+Y=n\}} \\ &= \frac{P\{X=k\}\,P\{Y=n-k\}}{P\{X+Y=n\}} \\ &= \frac{\frac{\lambda_1^k}{n!}e^{-\lambda_1}\frac{\lambda_2)^{n-k}}{(n-k)!}e^{-\lambda_2}}{\frac{(\lambda_1+\lambda_2)^k}{n!}e^{-(\lambda_1+\lambda_2)}}, \end{aligned}$$

which reduces to the right-hand side of (9.29). □

The following theorem explains a very important application of the Poisson distribution:

Theorem 9.3.3 *If* $p \to 0$ *and* $n \to \infty$ *in such a way that* $\lim np = \lambda > 0$, *then for* $k = 0, 1, \ldots,$

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{9.30}$$

Proof. We will prove (9.30) under simplifying assumption $np = \lambda$ for all n. The proof in the general case is based on the same idea but obscured by some technical points. Replacing p by λ/n, we have

$$\begin{aligned} \binom{n}{k} p^k (1-p)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{k}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{k!\left(1-\frac{\lambda}{n}\right)^k}. \end{aligned}$$

The factor $(1-\lambda/n)^n$ converges to $e^{-\lambda}$, while each of the remaining factors involving n converges to 1. Since the number of such factors does not depend on n, their product also tends to 1, which proves the theorem. □

To see the applicability of Theorem 9.3.3, observe that it can be used as an approximation of binomial probability of k successes in n trials, valid for small p and large n:

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{(np)^k}{k!} e^{-np}. \tag{9.31}$$

■ **EXAMPLE 9.8**

Continuing Example 9.3, we assume that on average, one birth in 80 is a multiple birth. What is the probability that among 400 births that occurred in the maternity ward of a given hospital during the first three months of a year there were fewer than 4 multiple births?

SOLUTION. If X stands for the number of multiple births during the period analyzed, then $X \sim \text{BIN}(400, 1/80)$. We have

$$P\{X < 4\} = \sum_{k=0}^{3} P\{X = k\} = \sum_{k=0}^{3} \binom{400}{k} \left(\frac{1}{80}\right)^k \left(\frac{79}{80}\right)^{400-k}.$$

This sum can be computed directly, and the value is

$$0.0065 + 0.0331 + 0.0835 + 0.1402 = 0.2633.$$

We have here $np = 400 \times \frac{1}{80} = 5$, so the approximation (9.31) by Poisson distribution gives

$$P\{X < 4\} \approx \sum_{k=0}^{3} \frac{5^k}{k!} e^{-5} = 0.0067 + 0.0337 + 0.0842 + 0.1404 = 0.2650.$$

The relative errors of consecutive approximating terms are, respectively, 3.08%, 1.31%, 0.34% and 0.14%, while the relative error of the final answer is 0.55%. Whether or not this may be regarded as a good approximation depends on the goal of finding the probability in question. For most purposes, a relative error below 1% is quite acceptable. One can imagine, however, situations where it need not be so. For instance, an insurance company that is to cover the cost of delivery and hospital care for multiple births might conceivably want to know the probability of fewer than 4 multiple births with precision better than second decimal in order to decide on the premium.

■ **EXAMPLE 9.9**

Suppose that on average, one in every 100 passengers does not show up for a flight. An airline sold 250 tickets for a flight serviced by an airplane that has 247 seats. What is the probability that every person who shows up for flight will get a seat?

SOLUTION. Let X be the number of passengers who do not show up for the flight in question, and let us treat X as a binomial random variable with $n = 250$, $p = 0.01$, so that $np = 2.5$. The probability of the event $X \geq 3$, using Poisson approximation, is

$$\begin{aligned} P\{X \geq 3\} &= 1 - P\{X < 3\} = 1 - P\{X = 0\} - P\{X = 1\} - P\{X = 2\} \\ &\approx 1 - e^{-2.5} - \frac{2.5}{1!} e^{-2.5} - \frac{(2.5)^2}{2!} e^{-2.5} = 0.4562, \end{aligned}$$

while directly from the binomial distribution we have

$$P\{X \geq 3\} = 1 - \sum_{j=0}^{2} \binom{250}{j} (0.01)^j (0.99)^{250-j} = 0.4568. \tag{9.32}$$

This time the approximation by a Poisson distribution to binomial probabilities (9.32) has a relative error of 1.3%.

It has to be pointed out that whereas in Example 9.8 the claim that X has a binomial distribution was fully justified (whether or not a birth is a multiple birth is independent on the multiplicity of other births in the same period), the situation is not so clear in case of passengers missing an airline flight. The point is that people often fly together (typically in families or other groups). In these cases the fact that one person misses the flight may affect the chances of some other persons missing the same flight. Consequently X is at best *approximately* binomial. Our calculations therefore give a relative error of an approximation to a number that is already an approximation (to the actual probability).

We will now try to capture features responsible for the fact that the number of occurrences of some event in a given interval of time follows a Poisson distribution.

We consider the class of situations in which a certain event occurs at random points in time. Examples are quite common: arrivals of customers at service stations, twin births in a hospital, earthquakes of specified intensity occurring in a given region, fire alarms in a given town, and so on. To increase practical applicability, the theory focuses only on the times of their occurrence, disregarding other specific features of the events under consideration. The random variable one needs to analyze here is the number $N_{[t_1,t_2]}$ of events that occur between times t_1 and t_2. The theory built for analyzing such processes in the most general case is called the *theory of point processes*. We will analyze only a special case of point processes, the Poisson process.

The assumptions underlying the Poisson processes attempt to capture the intuitive notion of "complete randomness." In particular, in the Poisson process knowledge of the past provides no clue in regard to the future.

To express the properties that will imply that a given stream of events is a Poisson process, it will be convenient to introduce a mathematical notation, which will later be useful also in other contexts: The symbol $o(x)$ denotes any function $f(x)$ such that

$$\lim_{x \to 0} \frac{f(x)}{x} = 0.$$

■ **EXAMPLE 9.10**

A power function x^a is $o(x)$ if $a > 1$, and so is every function of the form $x^a h(x)$ if $a > 1$ and h is continuous at 0 (hence bounded in the neighborhood of 0). For instance, if S_n is a binomial random variable, then $P\{S_n = k\} = o(p)$ for $k > 1$. Indeed, $P\{S_n = k\}/p = \binom{n}{k} p^{k-1}(1-p)^{n-k}$, which converges to 0 when $p \to 0$ if $k > 1$.

We will often use the following facts:

1. *If* $\lim_{x \to 0} \frac{h(x)}{x} = c \neq 0$, *then* $h(x) = cx + o(x)$.
2. *If the functions* $f_1, f_2, \ldots, f_N$ *are* $o(x)$, *then* $f_1 + \cdots + f_N$ *is also* $o(x)$.

We can now formulate the postulates of the Poisson process.

Postulate 1. *The numbers of events occurring in two nonoverlapping time intervals are independent.*
Postulate 2. *The probability of at least one event occurring in an interval of length* Δt *is* $\lambda \Delta t + o(\Delta t)$ *for some constant* $\lambda > 0$.
Postulate 3. *The probability of two or more events occurring in an interval of length* Δt *is* $o(\Delta t)$.

The first postulate is the one that asserts that knowledge of the past is of no help in predicting the future. The second postulate asserts stationarity, in the sense that probability of an event occurring in a short time interval is (roughly) proportional to the length of this interval, but does not depend on the location of this interval. Finally, the third postulate asserts that events occur one at a time. That is chances of two events occurring within an interval of length Δt become negligible as $\Delta t \to 0$.

Let us now fix the zero on time scale, and let $P_n(t)$ denote the probability of exactly n events prior to t, so that $P_n(t) = P\{N_{[0,t)} = n\}$. We will prove

Theorem 9.3.4 *Under all three postulates*

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \qquad n = 0, 1, \ldots \tag{9.33}$$

Proof. By postulates 2 and 3, for every t and $\Delta t > 0$,

$$\begin{aligned} P\{N_{[t,t+\Delta t)} = 1\} &= \lambda \Delta t + o_1(\Delta t), \\ P\{N_{[t,t+\Delta t)} = 0\} &= 1 - \lambda \Delta t + o_2(\Delta t). \end{aligned} \tag{9.34}$$

For $n = 0$ we write, using postulate 1 and (9.34),

$$P_0(t + \Delta t) = P\{N_{[0,t)} = 0, N_{[t,t+\Delta t)} = 0\} = P_0(t)[1 - \lambda \Delta t + o_2(\Delta t)]$$

which gives the difference ratio

$$\frac{P_0(t + t + \Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t) + \frac{1}{\Delta t} o_2(\Delta t) P_0(t). \tag{9.35}$$

Passing to the limit with $\Delta t \to 0$, we obtain[14]

$$P_0'(t) = -\lambda P_0(t). \tag{9.36}$$

The initial condition is $P_0(0) = 1$, and the relation of (9.36) gives

$$P_0(t) = e^{-\lambda t}. \tag{9.37}$$

[14]The limit of the right-hand side exists, and is equal to the right derivative of P_0 at t. To justify the existence of the left derivative in (9.35), observe that one can replace t by $t - \Delta t$ in (9.35). Since $o_2(\Delta t)$ does not depend on t, we see that P_0 is continuous, and that the left derivative is equal to the right derivative.

Now, if $n \geq 1$, we write

$$\begin{aligned} P_n(t+\Delta t) &= \sum_{j=0}^{n} P\{N_{[0,t)} = n-j, N_{[t,t+\Delta t)} = j\} \\ &= \sum_{j=0}^{n} P_{n-j}(t) P\{N_{[t,t+\Delta t)} = j\} \\ &= P_n(t)[1 - \lambda\Delta t + o_2(\Delta t)] + P_{n-1}(t)[\lambda\Delta t + o_1(t)] + o_3(t), \end{aligned}$$

where $o_3(t)$ is the term obtained from combining together all terms involving two or more events occurring between t and $t + \Delta t$.

Forming the difference ratio and passing to the limit with $\Delta t \to 0$, we obtain the equations, valid for $n = 1, 2, \ldots,$

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t), \tag{9.38}$$

which can be solved recursively using (9.37), with the initial conditions now being $P_n(0) = 0, n = 1, 2, \ldots$. Alternatively, we could use induction to check that probabilities (9.33) satisfy (9.37) and (9.38). □

We will now discuss some examples of Poisson processes.

■ **EXAMPLE 9.11**

The maternity ward in a certain hospital has, on average, 30 births per week. Given that there were 6 births on a specific day, find the probability of: (a) Three births on each of the following two days. (b) A total of six births during the following two days. (c) The expected number of days with exactly one birth during the month of May.

SOLUTION. We assume here that the births in the maternity ward in question form a Poisson process. Consequently, the number of births on a given day does not affect the number of births in future intervals. To answer all these questions, we must first choose the unit of time. This is a totally arbitrary choice, but the important point is that once this choice is made, we must express the parameter λ in the chosen units. Then λ is the expected number of events in the unit of time.

Let a time unit equal one day. Then $\lambda = 30/\text{week} = 4.286/\text{day}$. Consequently, the probability of three births in a given day is $(\lambda^3/3!)e^{-\lambda} = 0.1806$, and probability of such event on two consecutive days is $[(\lambda^3/3!)e^{-\lambda}]^2 = 0.0326$. As regards (b), the probability of 6 births in 2 days is $[(2\lambda)^6/6!]e^{-2\lambda} = 0.1043$. For (c), the number of days in May when there is exactly one birth is has binomial distribution with $n = 31$ and $p = \lambda e^{-\lambda} = 0.0590$, so the expectation equals $31p = 1.83$.

As mentioned, the postulates of Poisson process attempt to capture the idea of "complete randomness." The theorems below indicate to which extent this attempt is successful.

To facilitate formulation of the theorems, we let $X_t = N_{(0,t)}$ denote the number of events occurring in $(0, t)$, and also let $T_1, T_2, \ldots$ denote the times of occurrence of successive events. Thus

$$T_k = \inf\{t : X_t \geq k\}.$$

In the analogy with (9.19), we have the following identity:

$$T_k \leq t \quad \text{if and only if} \quad X_t \geq k$$

and consequently $P\{T_k \leq t\} = P\{X_t \geq k\}$. Since X_t has Poisson distribution with parameter λt, we have

$$P\{X_t \geq k\} = \sum_{j=k}^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t} = 1 - \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} e^{-\lambda t}.$$

Then the cdf of T_k is

$$P\{T_k \leq t\} = P\{X_t \geq k\} = 1 - \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} e^{-\lambda t}. \tag{9.39}$$

Now let $U_1 = T_1, U_2 = T_2 - T_1, U_3 = T_3 - T_2, \ldots$ be the time to the first event (U_1) and consecutive times between events ($U_2, U_3, \ldots$). From (9.39) for $k = 1$we have $P\{U_1 \leq t\} = P\{T_1 \leq t\} = 1 - e^{-\lambda t}$, so U_1 has an exponential distribution with mean $1/\lambda$. Next

$$\begin{aligned} P\{U_2 > t | T_1 = \tau\} &= P\{U_2 > t | U_1 = \tau\} \\ &= P\{\text{ no events in } (\tau, t+\tau) | T_1 = \tau\} \\ &= P\{\text{ no events in } (\tau, t+\tau)\} = e^{-\lambda t}, \end{aligned}$$

which means that U_2 also has exponential distribution with mean $1/\lambda$ and is independent of U_1. Since the argument can be repeated for all other U_j's, we have proved the following theorem:

Theorem 9.3.5 *In a Poisson process the time U_1 until the first event and the times $U_2, U_3, \ldots$ between subsequent events are independent random variables, each with the same exponential distribution with mean $1/\lambda$.*

Since the origin of time scale $t = 0$ was chosen arbitrarily, this theorem asserts that if we start observing a Poisson process at an arbitrarily selected time, fixed or randomly chosen,[15] then the waiting time for the first event has the same distribution as the times between subsequent events. This property is connected closely with the memoryless property of geometric distribution, specified in Theorem 9.1.2.

[15]The phrase "randomly chosen" ought to be qualified here. Suppose that the "random choice" is to start observing Poisson process 5 minutes *before* the next event. Technically such a choice gives a random moment of beginning of observation (since the time of event is random), and for this choice $U_1 = 5$ minutes. Here the qualification of "random choice" is that the decision depends on the past but *not* on the future of the process.

Theorem 9.3.6 *If X is a random variable with exponential distribution, then for all $s, t > 0$,*

$$P\{X > s + t | X > s\} = P\{X > t\}. \tag{9.40}$$

Conversely, if X may assume only positive values and satisfies (9.40) for all $s, t > 0$, then X has exponential distribution.

Proof. Let X have exponential distribution. The left-hand side of (9.40) is the ratio

$$\frac{P\{X > s + t\}}{P\{X > s\}} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P\{X > t\}.$$

Conversely, (9.40) implies that the tail of the cdf of X, that is, the function $\phi(x) = P\{X > x\}$, satisfies the equation $\phi(s + t) = \phi(s)\phi(t)$. One then shows (e.g., see Feller, 1968) that any bounded solution of this equation must be of the form $\phi(t) = e^{-\lambda t}$ for some $\lambda > 0$. □

The property of a Poisson process discussed above is one of the arguments for the claim that assumptions of a Poisson process capture "maximal randomness." While in the Poisson process, knowledge of the past does not give a clue to the future; it is not so for other streams of events. For instance, if one arrives at a bus stop just after a bus has left (knowledge of the past), one may expect a longer wait for the next bus.

The following theorem shows that in Poisson processes, knowledge of the number of events in the past gives us, in a sense, no additional information not only about the future but also about the past. We have the following:

Theorem 9.3.7 *The events in Poisson process satisfy the following property: for every $0 < u < t$,*

$$P\{T_1 < u | X_t = 1\} = \frac{u}{t}. \tag{9.41}$$

Proof. We have

$$P\{T_1 < u | X_t = 1\} = \frac{P\{T_1 < u, X_t = 1\}}{P\{X_t = 1\}} = \frac{P\{T_1 < u, X_t = 1\}}{\lambda t e^{-\lambda t}}. \tag{9.42}$$

Conditioning on T_1, the time of occurrence of the first event, and using the fact that T_1 has exponential distribution, the numerator in (9.42) becomes

$$P\{T_1 < u, X_t = 1\} = \int_0^u P\{X_t = 1 | T_1 = z\} \lambda e^{-\lambda z} dz.$$

If $T_1 = z$, the event $X_t = 1$ occurs if the time U_2 between the first and second event exceeds $t - z$. So

$$P\{X_t = 1 | T_1 = z\} = P\{U_2 > t - z\} = e^{-\lambda(t-z)}.$$

Substituting (9.40), we obtain formula (9.41). □

This theorem asserts that if we know that only one event occurred between 0 and t, then the conditional distribution of the time of occurrence of this event is uniform on $(0, t)$. In a sense, then, we have no information as to when the event occurred.

Finally, let us observe that we can obtain the unconditional density of T_k by differentiating the cdf given by formula (9.39).

Theorem 9.3.8 *The density $f_{T_k}(t)$ of the time T_k of the kth event in Poisson process is*

$$f_{T_k}(t) = \frac{\lambda^k}{(k-1)!} t^{k-1} e^{-\lambda t}, \quad t > 0. \tag{9.43}$$

The distribution of T_k is sometimes called the Erlang distribution. As we will see later, (9.43) is a special case of the gamma density.

■ **EXAMPLE 9.12**

Fires in a certain town occur according to a Poisson process. If there were 10 fires in a given week, what is the probability that at least one of them occurred on Friday?

SOLUTION. What is of interest here is that we do not need to know the intensity λ of the Poisson process in question. We know that given that the number of fires was 10, their times of occurrence fall within a week according to the uniform distribution. The probability that a single fire does not fall on Friday is $6/7$; hence chances of at least one of 10 fires falling on Friday is $1 - (6/7)^{10} = 0.7859$.

■ **EXAMPLE 9.13**

Suppose that traffic accidents on a given intersection occur according to a Poisson process, with the rate on Saturdays being twice the rate on weekdays and the rate on Sundays being double the rate on Saturdays. The total rate is about five accidents per week. What is more likely: two accidents on each of two consecutive weekends (Saturday + Sunday), or a total of four accidents on weekdays in a given week?

SOLUTION. If λ is the average number of accidents on a weekday, then it is 2λ on a Saturday and 4λ on a Sunday. Consequently, we have $5\lambda + 2\lambda + 4\lambda = 5$, which gives $\lambda = 5/11$. The number of accidents on a weekend is the sum of the numbers of accidents on Saturday and Sunday. These are independent Poisson random variables; hence their sum (see Theorem 9.3.1) also has Poisson distribution, with the mean $2\lambda + 4\lambda = 30/11$. Consequently, the first probability is $P\{X = 2\}^2$, where $X \sim \text{POI}(30/11)$, or $[(30/11)^2 e^{-30/11}/2!]^2 = (0.2432)^2 = 0.0592$. The second probability is $P\{Y = 4\}$, where now $Y \sim \text{POI}(5\lambda) = \text{POI}(25/11)$, which equals $(25/11)^4 e^{-25/11}/4! = 0.1145$.

To continue, suppose that on a Friday (which happened to be Friday the 13th) the number of accidents was as high as the number of accidents on the following weekend. Is such an event unusual?

Let us evaluate the chances of such an occurrence without reference to any magic connected with Friday the 13th, that is, probability $P\{X = Y\}$, where X and Y are independent, $X \sim \text{POI}(\lambda)$ and $Y \sim \text{POI}(6\lambda)$ for $\lambda = 5/11$.

$$\begin{aligned} P\{X=Y\} &= \sum_{j=0}^{\infty} P\{X=j\}P\{Y=j\} \\ &= \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}e^{-\lambda}\frac{(6\lambda)^j}{j!}e^{-6\lambda} = \sum_{j=0}^{\infty}\frac{(6\lambda^2)^j}{(j!)^2}e^{-7\lambda} \\ &= 0.0415 + 0.0515 + 0.0159 + 0.0022 + \cdots = 0.1111. \end{aligned}$$

As we see, the chances here are slightly over 10%; hence such an event need not necessarily be regarded as highly unusual. However, as one may note, almost all of this probability is due to the "unattractive" possibility that $X = Y = 0$ or $X = Y = 1$.

Similarly the probability that there are more accidents on Friday the 13th than on the whole following weekend is still not negligible:

$$P(X > Y) = \sum_{k=1}^{\infty} P\{X=k\}\sum_{j=0}^{k-1} P\{Y=j\} = 0.0406.$$

This time most of the probability is contributed by the terms $P\{X = 1\} \times P\{Y = 0\}$ and $P\{X = 2\} \times [P\{Y = 0\} + P\{Y = 1\}]$.

The concept of a Poisson process allows a number of generalizations. First of all, note that the symbol t need not be interpreted as time: it may be some other attribute interpretable as a linear dimension. Thus one can regard faults on a magnetic tape or misprints in a text (regarded as a continuous string of letters) as Poisson processes, provided that one can reasonably expect that the postulates of the Poisson process hold.

One of generalizations of the Poisson process concerns the extension to a higher dimension. Instead of events occurring in time (i.e., random points on a line), one can consider the case of points allocated at random on a plane or in space. The postulates of the Poisson process in such cases are analogous to the postulates in one dimension. The basic random variable is $X(A)$ = number of points falling into A, where A is a set on the plane or in space. The main postulate asserts that the numbers of points falling into disjoint sets are independent. The second postulate asserts that the probability that $X(A) = 1$ depends on the size of the set A, not on its location, and equals $\lambda|A| + o(|A|)$, where $|A|$ stands for the area or volume of the set A. Finally, the third postulate asserts that the probability that $X(A) \geq 2$ is of the order $o(|A|)$. Under these postulates one can show that

$$P\{X(A)=k\} = \frac{(\lambda|A|)^k}{k!}e^{-\lambda|A|}, \qquad k = 0, 1, \ldots. \tag{9.44}$$

As in a one-dimensional case, λ is the expected number of points falling into a region of unit size.

■ **EXAMPLE 9.14**

The data below are taken from Feller (1968) who lists them among examples of phenomena fitting the Poisson distribution. The observations concerned points of hits in south London by flying bombs during World War II. The entire area under study was divided into $N = 576$ small areas of 0.25 square kilometer each, and the numbers N_k of areas that were hit k times were counted. The total number of hits is $T = \sum kN_k = 537$, so the average number of hits per area is $\lambda = T/N = 0.9323$. The fit of the Poisson distribution is excellent, as can be seen from comparison of the actual numbers N_k and expected numbers $NP\{X = k\} = N\frac{\lambda^k}{k!}e^{-\lambda}$ for $\lambda = 0.9323$.

k	0	1	2	3	4	5 or more
N_k	229	211	93	35	7	1
$NP\{X = k\}$	226.74	211.39	98.54	30.62	7.14	1.57

The chi-square goodness-of-fit criterion (to be discussed in further chapters) shows that in about 88% of cases one should expect worse agreement.

This example has become sort of a "classic" in the sense of being reproduced in numerous textbooks on statistics, invariably without any comments (other than remarks that the fit is very good). The readers may therefore get the impression that statisticians have somewhat ghoulish interests. In fact, however, the fit to the Poisson distribution was a piece of information of considerable value as military intelligence: it showed the state of German technology in regard to the precision of their aiming devices. Perfect randomness of hits in a given large area indicated that it was not possible to select any specific target within this area.

Feller writes: "It is interesting to note that most people believed in a tendency of the points of impact to cluster. If this were true, there would be a higher frequency of areas with either many hits or no hit, and a deficiency in the intermediate classes. The data indicates perfect randomness and homogeneity of the area; we have here an instructive illustration of the established fact that, to the untrained eye, randomness appears as regularity or tendency to cluster." It appears that Feller fully knew the reason for collecting and analyzing the data in question but could only make a veiled allusion: his book was first published in 1950, just five years after the end of World War II, when many things were still secret.

At the end, we will give an example involving the Poisson process in three dimensions.

■ **EXAMPLE 9.15**

The Poisson Bakery makes a special kind of cookies, called Four-Raisin cookies. Raisins (10,000) are added to the dough for 2500 cookies, and after thorough mixing, the dough is divided into equal parts of which 2500 cookies are formed and baked. What is the proportion of Four-Raisin cookies that have no raisins at all? What proportion will have exactly four raisins?

SOLUTION. Here we have a spatial Poisson process, with raisins playing the role of points located randomly in space. If we take a cookie as a unit of volume, then $\lambda = 4$, as there are, on average, four raisins per cookie. If we let X denote the number of raisins in a randomly selected cookie, then $P\{X = 0\} = e^{-4} = 0.0183$, so slightly below 2% of all Four-Raisin cookies are raisin-less. On the other hand, $P\{X = 4\} = (4^4/4!)e^{-4} = 0.1954$.

Now suppose that you buy a box of 200 Four-Raisin cookies. What is the probability that no more than two of them have no raisins? We have here in a situation of a Poisson approximation to the binomial distribution, with "success" being a cookie with no raisins so that $np = 200e^{-4} = 3.66$. The number Y of raisinless cookies in the box has a binomial distribution BIN(200, 0.0183), which is approximated by the Poisson distribution with $\lambda = np = 3.66$. Thus $P\{X \leq 2\} = P\{Y = 0\} + P\{Y = 1\} + P\{Y = 2\} \approx (1 + 3.66 + 3.66^2/2)e^{-3.66} = 0.29$; hence $P\{Y \geq 3\}$ is about 0.71, which means that 71% of all boxes will contain three or more Four-Raisin cookies with no raisins at all.

PROBLEMS

9.3.1 Let X have the POI(λ) distribution. Find: (i) The mode of X (i.e., the most likely value of X). **(ii)** $P(X$ is even). (*Hint.* Write the Taylor expansions for e^{λ} and $e^{-\lambda}$. Any ideas?)

9.3.2 A book with 500 pages contains, on average, three misprints per ten pages. What is the probability that there will be more than one page containing at least three misprints?

9.3.3 Accidents in a given plant occur at a rate of 1.5 per month. The numbers of accidents in different months are independent and follow the Poisson distribution. Find the probability of: **(i)** Five accidents in a period of five consecutive months. **(ii)** One accident in each of five consecutive months.

9.3.4 Suppose that the daily numbers of ships arriving to a certain port are independent, each with POI(3) distribution. Find: **(i)** The expected number of days in April when there are no arrivals. **(ii)** The expected number and variance of days during the summer months (June, July, August) with the number of arrivals equal to the mean daily arrival rate.

9.3.5 A certain store makes, on average, two sales per hour between 9:00 a.m. and 2:00 p.m., and three sales per hour between 2:00 p.m. and 9:00 p.m. The numbers of sales in different time periods are independent and have a Poisson distribution. Find: **(i)** The probability of more than three sales between 10:00 a.m. and noon, and also between 1:00 p.m. and 3:00 p.m. **(ii)** The probability that the number of sales between 10:00 a.m. and 11:00 a.m. will be the same as number of sales between 6:00 p.m. and 7:00 p.m.

9.3.6 Weekly numbers of traffic accidents at intersections A, B, and C are independent, each with a Poisson distribution. It is known that, on the average, the number

of accidents at intersection A is the same as the number of accidents at intersections B and C combined, while the average number of accidents at intersection B is half of that at intersection C. **(i)** If there were, in a given week, 16 accidents at intersections A, B, and C, what is the probability that exactly four of them were at intersection C? **(ii)** What is the probability that there were more accidents at intersection C than at intersection A?

9.3.7 Find the approximate probability that in 1000 randomly chosen persons there are exactly: **(i)** Two born on New Year and two born on Christmas. **(ii)** Four born on either Christmas or New Year.

9.3.8 (Does Nature Prefer Even Numbers?) Generalizing Problem ??, let X be an integer-valued random variable such that $X = X_1 + X_2$, where X_1, X_2 are independent, identically distributed integer-valued random variables. Show that $P\{X \text{ is even}\} \geq 0.5$ (this property has been pointed out to us by Steve MacEachern, personal communication).

9.3.9 Let X be the number of failures preceding the rth success in a sequence of Bernoulli trials with probability of success p. Show that if $q \to 0, r \to \infty$ in such a way that $rq = \lambda > 0$, then

$$P\{X = k\} \to \frac{\lambda^k}{k!} e^{-\lambda}$$

for every $k = 0, 1, 2, \ldots$. This shows that the negative binomial distribution can be, for large r and small q, approximated by a Poisson distribution. (*Hint:* Use an argument similar to that in the proof of Theorem 9.3.3.)

9.3.10 Suppose that the number of eggs X laid by a bird has a Poisson distribution. Each egg hatches with probability p, independently of what happens to other eggs. Let V_1 and V_2, $V_1 + V_2 = X$, denote the numbers of eggs that hatch, and the number of eggs that do not hatch, respectively. Show that V_1 and V_2 are independent.

9.3.11 Traffic accidents at a given intersection occur following a Poisson process. **(i)** Given that 10 accidents occurred in June, what is the probability that the seventh accident occurred before June 10? **(ii)** If it is known that n accidents occurred in April, what is the expected number of accidents that occurred during the second week of that month?

9.3.12 Consider two independent Poisson processes with the same parameter λ. Let $N_i(t)$, i=1, 2 be the number of events in i-th process which occurred up to time t, and let U_T be the set of all those times t with $0 \leq t \leq T$ at which $N_1(t) = N_2(t)$. Find $E(U_T)$ given that: **(i)** $N_1(T) = N_2(T) = 2$. **(ii)** $N_1(T) = 2, N_2(T) = 3$. (*Hint.* Use the fact that the sum of two independent Poisson processes is a Poisson process and Theorem 9.8.5.)

9.3.13 Assume that chocolate chips are distributed within a cake according to a Poisson process with parameter λ. A cake is divided into two parts of equal volume (disregard the possibility of cutting through a chocolate chip). Show that the

probability that each part of the cake has the same number of chips is

$$e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda/2)^{2k}}{(k!)^2}.$$

9.4 EXPONENTIAL, GAMMA AND RELATED DISTRIBUTIONS

We have encountered the exponential distribution on a number of occasions, notably, as the distribution of times between events in Poisson process in one dimension. To repeat the definition, a random variable X has exponential distribution EXP(λ) if for some $\lambda > 0$ and $x \geq 0$,

$$F(x) = P\{X \leq x\} = 1 - e^{-\lambda x}, \quad x \geq 0,$$

so that for $x \geq 0$ the density of X is $f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$. We know that

$$E(X) = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2},$$

and the mgf of X is

$$m_X(s) = \frac{\lambda}{\lambda - s} \text{ for } s < \lambda.$$

The hazard function (see Section 6.5) of exponential random variable is constant:

$$h(x) = \frac{f(x)}{1 - F(x)} = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \lambda,$$

a property closely related to the memoryless property of an exponential distribution asserted in Theorem 9.3.6.

Readers should be aware of the fact that the phrase "exponential distribution with parameter λ" is ambiguous, since an exponential distribution is sometimes introduced in the form $f(x) = (1/\lambda)e^{-x/\lambda}$, $x > 0$. In this notation $E(X) = \lambda$ and $\mathrm{Var}(X) = \lambda^2$. Consequently, unless it is clear whether the parameter appears in the numerator or denominator of the exponent, one could use phrases that convey the information about the mean. Thus "exponential distribution with mean θ" will have density $(1/\theta)e^{-x/\theta}$, and so on.

Next, the sums of independent and exponentially distributed random variables also appeared in the Poisson process, as times $T_1, T_2, \ldots$ of consecutive events. The cdf's of these random variables were obtained using the identity

$$P\{T_r > t\} = P\{X_t < r\} = \sum_{k=0}^{r-1}\frac{(\lambda t)^k}{k!}e^{-\lambda t}.$$

The density of T_r follows now by the differentiation,

$$f_r(t) = \frac{\lambda^n}{(r-1)!}t^{r-1}e^{-\lambda t}, \quad t > 0. \tag{9.45}$$

Since T_r is the sum of r independent waiting times, each with the same exponential distribution, the mgf of T_r exists for $s < \lambda$ and equals

$$m_{T_r}(s) = \left(\frac{\lambda}{\lambda - s}\right)^r.$$

The definition of distribution of T_r as the sum of waiting times involves using an integer value of r. We have a complete analogy with the binomial distribution (count of successes in fixed number of trials) and negative binomial distribution (count of number of trials up to a fixed number of successes), on one hand, and the Poisson distribution (number of events until fixed time) and distribution (9.45) (time till fixed number of events occurs), on the other hand. However, unlikely as in the case of a negative binomial distribution, r need not be an integer, and we can define the class of distributions comprising densities (9.45) as a special case. To this end let us introduce the following definition:

Definition 9.4.1 For $t \geq 0$ we define the *gamma function* as

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx. \tag{9.46}$$

□

We can show that this function is well defined for all $t \geq 0$. Integration by parts gives

$$\Gamma(t) = (t-1)\Gamma(t-1). \tag{9.47}$$

Since $\Gamma(1) = 1$, by induction, for any integer $n \geq 1$, we obtain

$$\Gamma(n) = (n-1)!$$

Consequently, a gamma function can be seen as an extension of the factorial function, which it "fills in" for noninteger values of n.

Let us now show that $\Gamma(\frac{1}{2}) = \sqrt{2}$. Indeed,

$$\begin{aligned}\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty x^{1/2-1}e^{-x}dx = \int_0^\infty x^{-1/2}e^{-x}dx = \int_0^\infty \frac{\sqrt{2}}{y}e^{-y^2/2}ydy \\ &= 2\sqrt{\pi}\int_0^\infty \frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy = 2\sqrt{\pi}\frac{1}{2} = \sqrt{2},\end{aligned} \tag{9.48}$$

where we used substitution $x = y^2/2$ and the fact that the last integral equals $1/2$ as a half of the integral of the standard normal density.

We can now introduce the following definition:

Definition 9.4.2 A random variable with density of the form

$$f(x) = \begin{cases} Cx^{\alpha-1}e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x < 0 \end{cases}$$

for some $\alpha > 0$ and $\lambda > 0$ is said to have a *gamma distribution* with *shape* parameter α and *scale* parameter λ, GAM(α, λ). C is the normalizing constant. □

Since

$$C\int_0^\infty x^{\alpha-1}e^{-\lambda x}dx = 1,$$

by substituting $\lambda x = z$, we easily obtain from (9.46) that

$$C = \frac{\lambda^\alpha}{\Gamma(\alpha)}. \tag{9.49}$$

We can now compute the moments of gamma distribution. That is,

$$\begin{aligned} E(X) &= \int_0^\infty xf(x)dx = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^\alpha e^{-\lambda x}dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} = \frac{\alpha}{\lambda} \end{aligned} \tag{9.50}$$

in view of (9.47). Similarly

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 f(x)dx = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha+1}e^{-\lambda}dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha+2)}{\lambda^{(\alpha+2)}} = \frac{\alpha(\alpha+1)}{\lambda^2}, \end{aligned}$$

so

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{\alpha}{\lambda^2}. \tag{9.51}$$

The moment generating function of gamma distribution can be evaluated as follows:

$$\begin{aligned} m(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty e^{tx}x^{\alpha-1}e^{-\lambda x}dx = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-(\lambda-t)x}dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha)}{(\lambda-t)^\alpha} = \frac{1}{(1-t/\lambda)^\alpha}, \end{aligned}$$

provided that $t < \lambda$. It is easy to show that the kth ordinary moment is

$$m_k = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\lambda^k}. \tag{9.52}$$

The following closure property of gamma distributions is a consequence of Theorem 8.5.2:

Theorem 9.4.1 *If X and Y are independent with $X \sim \mathrm{GAM}(\alpha_1, \lambda)$, $Y \sim \mathrm{GAM}(\alpha_2, \lambda)$, then $X + Y \sim \mathrm{GAM}(\alpha_1 + \alpha_2, \lambda)$.*

Since the waiting time T_k for the kth event in a Poisson process has a distribution $\mathrm{GAM}(k, \lambda)$, Theorem 9.4.1 (in the case of integers α_1 and α_2) expresses the simple fact that the waiting time $T_{\alpha_1+\alpha_2}$ is the sum of two independent waiting times, T_{α_1} and T_{α_2}.

Let us note that exponential distribution with parameter λ is the same as GAM(1, λ). Also let us recall that in Example 6.28 we found the density of the square of standard normal variable [i.e., of $Y = Z^2$, where $Z \sim \mathrm{N}(0,1)$]. This density equals

$f(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}$, which we recognize as GAM(1/2, 1/2). The sums of squares of independent standard normal variables appear so often in statistics, that the distribution of such sum bears its own name:

Definition 9.4.3 For integer ν, the distribution GAM $(\nu/2, 1/2)$ is called the *chi-square* distribution with ν *degrees of freedom.* A random variable with such distribution is typically denoted by χ^2_ν. □

The following theorem will be very important in statistical inference; its proof is an immediate consequence of Theorem 9.4.1.

Theorem 9.4.2 *If $Z_1, \ldots, Z_n$ are independent, each with standard normal distribution, then*

$$X = Z_1^2 + \cdots + Z_n^2$$

has chi-square distribution with n degrees of freedom.

Moreover, if $X_1, \ldots, X_k$ are independent, chi-square distributed random variables, $X_i \sim \chi^2_{\nu_i}$, then $Y = X_1 + \cdots + X_k$ has chi-square distribution with ν degrees of freedom, where $\nu = \nu_1 + \cdots + \nu_k$.

Another property that will be useful in further chapters devoted to statistical inference is provided by the next theorem:

Theorem 9.4.3 *If random variable X has* GAM(α, λ) *distribution, then $Y = 2\lambda X$ has* GAM$(\alpha, 1/2)$ *distribution. If 2α additionally is a positive integer, then Y has $\chi^2_{2\alpha}$ distribution.*

Proof. Let $Y = aX$, where $a > 0$. Then $x = y/a = \psi(y)$ and $\psi'(y) = 1/a$. Consequently, based on (9.4.2),

$$\begin{aligned} f_Y(y) &= f_X(\psi(y)) \times |\psi'(y)| = \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{y}{a}\right)^{\alpha-1} e^{-\lambda y/a} \times \frac{1}{a} \\ &= \frac{(\lambda/a)^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-(\lambda/a)y}. \end{aligned} \tag{9.53}$$

The density $f_Y(y)$ has a GAM$(\alpha, \lambda/a)$ distribution, which is the same as $\chi^2_{2\alpha}$, when $a = 2\lambda$ and 2α is a positive integer. □

With wide availability of statistical packages today, the values of cdf as well as percentiles of chi-square distributions for various numbers of degrees of freedom can be easily obtained. For given n and x, we can obtain $P\{\chi^2_n \leq x\}$, as well as for given n and p, we can obtain x such that $P\{\chi^2_n \leq x\} = p$. Selected quantiles and upper percentiles can also be found in Table A.4. at the end of the book.

At the end of this section we will introduce two families of distributions somewhat related to the exponential family.

Definition 9.4.4 The distribution with the density function

$$f(x; \lambda) = \frac{1}{2} \lambda e^{-\lambda|x|} \tag{9.54}$$

for $-\infty < x < \infty, \lambda > 0$, is called *Laplace* or a *double exponential.* □

The Laplace distribution is symmetric, and it has higher values of a kurtosis than the normal distributions. The double exponential distribution can be generalized in many ways. For example, the distribution with density

$$f(x;\lambda_1,\lambda_2)=\begin{cases}\frac{1}{2}\lambda_1 e^{-\lambda_1|x|} & \text{for } x<0\\ \frac{1}{2}\lambda_2 e^{-\lambda_2|x|} & \text{for } x>0\end{cases}$$

will be skewed if $\lambda_1 \neq \lambda_2$. The Laplace distribution is very useful for modeling in the biological sciences, economics, and finance. Interested readers can find more information in Kotz et al. (2001).

The moment generating function of the Laplace distribution is

$$\begin{aligned}E(e^{tX}) &= \int_{-\infty}^{+\infty} e^{tx}\frac{1}{2}\lambda e^{-\lambda|x|}dx = \frac{\lambda}{2}\int_{-\infty}^{0} e^{x(t+\lambda)}dx + \frac{\lambda}{2}\int_{0}^{+\infty} e^{x(t-\lambda)}dx\\ &= \frac{\lambda e^{x(t+\lambda)}}{2(t+\lambda)}\Big|_{-\infty}^{0} + \frac{\lambda e^{x(t-\lambda)}}{2(t-\lambda)}\Big|_{0}^{+\infty}\\ &= \frac{\lambda}{2}\left(\frac{1}{t+\lambda}-\frac{1}{t-\lambda}\right) = \frac{\lambda^2}{\lambda^2-t^2}.\end{aligned} \tag{9.55}$$

The density (9.54) of the Laplace distribution is symmetric around 0; therefore its expected value is 0. The computation of other moments will be left as an exercise.

The last family of distributions to be introduced in this section is named after Waloddi Weibull, a Swedish engineer, scientist, and mathematician. The family of distributions that he introduced has a lot of flexibility and, as such, is widely used in industrial and engineering applications, such as reliability analysis, determination of wind spread distribution, or to model the dispersion of the received signals in radar systems.

Definition 9.4.5 The random variable with density of the form

$$f(x;k,\theta)=k\theta\,(x\theta)^{k-1}e^{-(x\theta)^k} \tag{9.56}$$

for $x>0, \theta>0, k>0$ is said to have a *Weibull* distribution, denoted WEI(k,θ); k and θ are called its *shape* and *scale* parameters, respectively. □

The cdf of a Weibull distribution is

$$F(x)=1-e^{-(x\theta)^k},$$

and consequently its hazard function equals $k\theta\,(x\theta)^{k-1}$. For $k<1$, the hazard function is decreasing, and it is increasing for $k>1$. If $k=1$, the hazard function is constant and WEI$(1,\theta)$ distribution becomes exponential.

It can be shown (we leave this for the readers as an exercise) that ordinary moments, m_n of Weibull distribution equal

$$m_n=\theta^{-n}\Gamma\left(1+\frac{n}{k}\right), \tag{9.57}$$

and that the mean and the variance are $\theta^{-1}\Gamma(1+\frac{1}{k})$ and $\theta^{-2}[\Gamma(1+\frac{2}{k})-\Gamma^2(1+\frac{1}{k})]$, respectively.

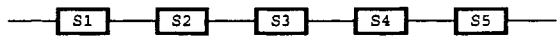


Figure 9.2 Series system

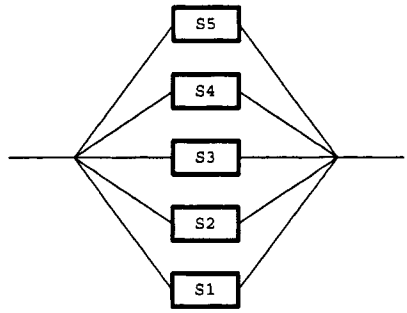


Figure 9.3 Parallel system

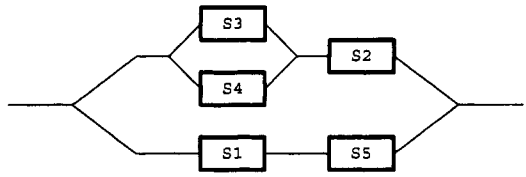


Figure 9.4 Series-parallel system

PROBLEMS

9.4.1 Show that if variable X has a GAM$(n, 1)$ distribution, where n is a positive integer, then its cdf is given by the following formula:

$$F_X(x) = \frac{1}{\Gamma(n)} \int_0^x t^{n-1} e^{-t} dt = 1 - e^{-x} \sum_{j=0}^{n-1} \frac{x^j}{j!}.$$

(*Hint:* Integrate by parts and use induction.)

9.4.2 A system consists of five components. Suppose that the lifetimes of the components are independent, with exponential distributions EXP(λ_1), ..., EXP(λ_5). Find the cdf and density of variable T = time to failure of the system if the components are connected: **(i)** In series (see Figure 9.2), so that the system fails as soon as one of its components fails. **(ii)** In parallel (see Figure 9.3), so that the system works as long as at least one component is operating. **(iii)** As in Figure 9.4.

9.4.3 In the flowchart of Figure 9.5, the block denoted "sample U" means that the computer samples a value of random variable U with a distribution uniform on (0, 1), the samplings being independent each time the program executes this instruction.

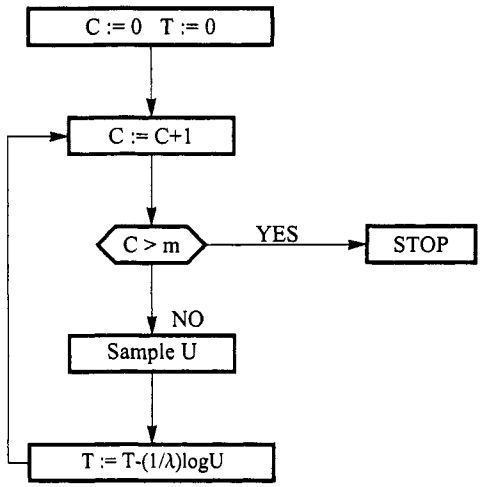


Figure 9.5 Flowchart

Assume that m is a positive integer and $\lambda > 0$. Find: **(i)** $P\{T \leq 2\}$ if $\lambda = 2$ and $m = 3$. **(ii)** The cdf and density of T, $E(T)$, and Var(T) in the general case.

9.4.4 For a random variable X that has a Laplace distribution with $\lambda = 1$, find: **(i)** Survival and hazard functions. **(ii)** Variance. **(iii)** Kurtosis.

9.4.5 Find the distribution of $X = \theta(-\log U)^{1/k}$, if $U \sim \mathrm{U}(0, 1)$.

9.4.6 **(i)** Find the median and the mode of the Weibull distribution with a density (9.56). **(ii)** Prove the formula (9.57).

9.4.7 It was found that the survival time (in years) in a group of patients who had a certain medical treatment and are in the similar risk group follows WEI$(2, 1/3)$ distribution. Find: **(i)** The median survival time for such patients. **(ii)** The probability that a randomly selected patient will live at least five more years if he already survived one year after the treatment.

9.5 NORMAL DISTRIBUTION

We have already encountered the normal distribution in Chapter 6. Let us recall that the univariate normal distribution with parameters μ and σ^2, denoted N(μ, σ^2), has density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty.$$

It has been shown in Example 6.18 that f is indeed a density. Moreover, if X has distribution N(μ, σ^2), then $E(X) = \mu$, Var$(X) = \sigma^2$, which gives a direct interpretation of the parameters.

The moment generating function of the normal distribution was obtained in Example 8.28 as

$$m(t) = e^{\mu t + \sigma^2 t^2/2}.$$

We have the following closure properties of the normal distribution:

Theorem 9.5.1 *A linear transformation of a normally distributed random variable again has a normal distribution. In particular, if random variable X has a normal distribution* $\mathrm{N}(\mu, \sigma^2)$, *then the random variable*

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

has the standard normal distribution $\mathrm{N}(0, 1)$.

Proof. We have, using (8.21),

$$m_Z(t) = m_{1/\sigma X - (\mu/\sigma)}(t) = m_X(t/\sigma)\, e^{-\mu/\sigma t} = e^{t^2/2}.$$

The right-hand side is the mgf of a standard normal random variable. □

The property asserted in Theorem 9.5.1 has important practical consequences. It shows that to determine probabilities for *any* normally distributed random variable, it suffices to have access to the probabilities for a standard normal random variable.

This property is especially important in view of the fact that the cdf of a normal distribution

$$\Phi(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt$$

cannot be integrated in closed form. Thus it is necessary to use the tables, and Theorem 9.5.1 implies that one table is sufficient to calculate probabilities for all normal distributions. We will let Z denote the standard normal random variable, with density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and cdf

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx. \tag{9.58}$$

Theorem 9.5.1 asserts, in effect, that given the tables of function (9.58), we have, for any $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$\begin{aligned} P\{a \le X \le b\} &= P\left\{\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right\} \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

Since $\phi(x)$ is symmetric around 0, we have $P\{Z \le z\} = P\{Z \ge -z\}$; hence the cdf of standard normal distribution satisfies the relation

$$\Phi(z) = 1 - \Phi(-z).$$

Consequently, many statistical tables give the values of $\Phi(z)$ only for $z \geq 0$.

An inspection of Table A.2. shows that $\Phi(3) = 0.9987$. Since for $z > 0$ we have

$$\begin{aligned} P\{|Z| > z\} &= P\{Z > z\} + P\{Z < -z\} \\ &= 1 - \Phi(z) + \Phi(-z) = 2\Phi(-z) = 2[1 - \Phi(z)], \end{aligned}$$

we see that $P\{|Z| > 3\} = 0.0026$. Using Theorem 9.5.1, we see that if $X \sim N(\mu, \sigma^2)$, then

$$P\{|X - \mu| > 3\sigma\} = P\{|Z| > 3\} = 0.0026. \tag{9.59}$$

This explains the origin of the *three-sigma rule*, according to which one is allowed to disregard the possibility of a random variable deviating from its mean more than three standard deviations.

Most tables of standard normal distribution do not give the values of $\Phi(z)$ for $z > 3$. If such values are needed, we have

Theorem 9.5.2 *For $z > 0$ the function $\Phi(z)$ satisfies the inequality*

$$\left(\frac{1}{z} - \frac{1}{z^3}\right)\phi(z) \leq 1 - \Phi(z) \leq \frac{1}{z}\phi(z) \tag{9.60}$$

and consequently, as $z \to \infty$,

$$1 - \Phi(z) \approx \frac{1}{z}\phi(z). \tag{9.61}$$

Proof. We have here the obvious inequality

$$\left(1 - \frac{3}{x^4}\right)\phi(x) \leq \phi(x) \leq \left(1 + \frac{1}{x^2}\right)\phi(x).$$

A simple check shows that this inequality is equivalent to

$$-\frac{d}{dx}\left[\left(\frac{1}{x} - \frac{1}{x^2}\right)\phi(x)\right] \leq -\frac{d}{dx}[1 - \Phi(x)] \leq -\frac{d}{dx}\left[\frac{1}{x}\phi(x)\right]. \tag{9.62}$$

Integrating (9.62) between z and ∞, we obtain (9.60). □

Numerically the relation (9.61) gives $1 - \Phi(4) \approx 3.36 \times 10^{-5}, 1 - \Phi(5) \approx 2.97 \times 10^{-7}, 1 - \Phi(6) \approx 1.01 \times 10^{-9}$. Such small probabilities could be of interest in estimating the chances of, say, an accident in a nuclear power plant.

From a practical viewpoint, the knowledge of these probabilities allows us to assume the normality of distribution in many cases where "logically" the distribution cannot possibly be normal. To illustrate the point, it is often assumed that an attribute as, say, height in human population is normally distributed (e.g., among men, we have mean μ about 70 inches and standard deviation σ of about 2 inches). But since a normally distributed random variable may *always* assume both positive and negative values, one could argue that height—which cannot be negative—cannot have a normal distribution. The chances of a random variable X with normal distribution

N(70, 4) being negative are of the order $\Phi(-35) = 1 - \Phi(35) \approx 0.0069 \times e^{-612}$ So events with such a probability can be safely disregarded, and it turns out that it is much easier to work with the assumption of normality than invent a distribution for height that does not allow negative values.

We will now prove a very useful and important property of normal distribution and illustrate its use.

Theorem 9.5.3 *Let X and Y be independent random variables with distributions $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Then the random variable $U = \alpha X + \beta Y$ has the distribution* $\mathrm{N}(\alpha\mu_1 + \beta\mu_2, \alpha^2\sigma_1^2 + \beta^2\sigma_2^2)$.

Proof. Since $m_X(t) = exp\{\mu_1 t + \sigma_1^2 \frac{t^2}{2}\}$, $m_Y(t) = exp\{\mu_2 t + \sigma_2^2 \frac{t^2}{2}\}$, and αX, βY are independent, we have

$$\begin{aligned} m_U(t) &= m_{\alpha X+\beta Y}(t) = m_{\alpha X}(t)\, m_{\beta Y}(t) = m_X(\alpha t)\, m_Y(\beta t) \\ &= \exp\left\{\mu_1 \alpha t + \sigma_1^2 \frac{\alpha^2 t^2}{2}\right\} \exp\left\{\mu_2 \alpha t + \sigma_2^2 \frac{\beta^2 t^2}{2}\right\} \\ &= \exp\left\{(\mu_1\alpha + \mu_2\beta)t + (\sigma_1^2\alpha^2 + \sigma_2^2\beta^2)\frac{t^2}{2}\right\}. \end{aligned}$$

We recognize the right-hand side as an mgf of $\mathrm{N}(\mu_1\alpha + \mu_2\beta, \sigma_1^2\alpha^2 + \sigma_2^2\beta^2)$ distribution. □

■ **EXAMPLE 9.16**

Assume that the height of men in a certain population is normal with mean $\mu_M = 70$ inches and standard deviation $\sigma_M = 2$ inches. The height of women is also normal, with mean $\mu_W = 68$ inches and $\sigma_W = 1.5$ inches.

One man and one woman are selected at random. What is the probability that the woman selected is taller than the man selected?

SOLUTION. Let X and Y be the heights of the randomly selected man (X) and woman (Y). We need $P\{Y > X\}$.

Without Theorem 9.5.3, we could proceed as follows (this solution is applicable to *any* distribution of X and Y, and is therefore of some general interest): If F and G are cdf's for X and Y, respectively, and f and g are their densities, then conditioning on values of Y, we have

$$\begin{aligned} P\{Y > X\} &= \int P\{Y > X | Y = y\}\, g(y)\, dy = \int P\{X < y\}\, g(y)\, dy \\ &= \int F(y)\, g(y)\, dy. \end{aligned} \tag{9.63}$$

We also condition on values of X, obtaining

$$\begin{aligned} P\{Y > X\} &= \int P\{Y > X | X = x\}\, f(x)\, dx = \int P\{Y > x\}\, f(x)\, dx \\ &= \int [1 - G(x)]\, f(x)\, dx. \end{aligned}$$

Using now the assumption of normality of X and Y, based on (9.63), we have

$$\begin{aligned} P\{Y > X\} &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{y} \frac{1}{\sigma_W \sqrt{2\pi}} e^{-(x-\mu_W)^2/2\sigma_W^2} dx \right] \\ &\quad \times \frac{1}{\sigma_M \sqrt{2\pi}} e^{-(y-\mu_M)^2/2\sigma_M^2} dy \\ &= \frac{1}{2(3/2)2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{y} e^{-(x-70)^2/8-(y-68)^2/(18/4)} dx dy. \end{aligned}$$

However, by Theorem 9.5.3, we have

$$P\{Y > X\} = P\{Y - X > 0\} = P\{U > 0\}$$

where $U = Y - X$ has normal distribution with mean $\mu = \mu_W - \mu_M = 68 - 70 = -2$ and $\sigma_U^2 = \sigma_W^2 + \sigma_M^2 = (3/2)^2 + 2^2 = 25/4$, hence $\sigma_U = 2.5$ inches (we use here Theorem 9.5.3 with $\alpha = -1, \beta = 1$). Consequently, since U is normally distributed,

$$\begin{aligned} P\{Y > X\} &= P\{Z > \frac{0-(-2)}{2.5}\} = P\{Z > 0.8\} \\ &= 1 - \Phi(0.8) = 1 - 0.7881 = 0.2119. \end{aligned}$$

Before we discuss the multivariate normal distribution we will introduce two distributions that are related to normal: *lognormal* and *folded normal* distributions.

Definition 9.5.1 Random variable X is said to have a *lognormal distribution* if the variable $Y = \log X$ has a normal distribution. The density of a random variable X is then

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{ -\frac{[\log(x) - \mu]^2}{2\sigma^2} \right\},$$

where $N(\mu, \sigma^2)$ is the distribution of X. □

The mean and the variance of the lognormal distribution are

$$E(X) = e^{\mu + \sigma^2/2} \quad \text{and} \quad V(X) = e^{2\mu+\sigma^2} \times (e^{\sigma^2} - 1),$$

and the general formula for m_k, the kth ordinary moment, which we give without proof, is

$$m_k = \exp\left\{ k\mu + \frac{k^2\sigma^2}{2} \right\}.$$

The lognormal distribution has one very unique property: Even though it has finite moments of all orders, the moment generating function is infinite at any positive number.

The lognormal distribution is widely applicable in modeling, for example, where there is a multiplicative product of many small independent factors. A typical example is the long-term return rate on a stock investment that can be considered as the product of the daily return rates.

In many experiments, measurements such as dimensions (as well as time and angles) are often recorded with respect to certain expected standards, where the magnitude of the difference, not the direction, is important. Consequently, the actual distribution of the measurement, X, is replaced by a distribution of absolute measurements, $Y = |X|$. When the underlying distribution is normal, the resulting distribution is called the *folded normal*, or *half-normal*.

Any normal distribution $N(\mu, \sigma^2)$ can be "folded" at zero, but we will discuss here only the simplest case with $\mu = 0$. So for $X \sim N(0, \sigma^2)$, the density of $Y = |X|$ is

$$f(y, \sigma) = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-y^2/2\sigma^2}$$

for $x > 0$, and 0 otherwise. We will now find the expected value $E(Y)$ and the variance $\mathrm{Var}(Y)$. Notice that σ^2 is no longer is a variance of the distribution.

$$E(Y) = \int_0^{+\infty} y \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-y^2/2\sigma^2} dy = \sigma \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-t} dt = \sigma \sqrt{\frac{2}{\pi}},$$

where we used substitution $t = y^2/2\sigma^2$. For $\mathrm{Var}(Y)$ we need

$$\begin{aligned} E(Y^2) &= \int_0^{+\infty} y^2 \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-y^2/2\sigma^2} dy = \frac{1}{2} \int_{-\infty}^{+\infty} x^2 \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} e^{-x^2/2\sigma^2} dy \\ &= \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dy = E(X^2) = \sigma^2, \end{aligned}$$

since $X \sim N(0, \sigma^2)$. Finally, we obtain

$$\mathrm{Var}(Y) = \sigma^2 - \left[\sigma \frac{2}{\pi}\right]^2 = \sigma^2 \left(1 - \frac{2}{\pi}\right).$$

Derivation of the moment generating function will be left to the reader as an exercise.

Let us now consider the multivariate normal distribution. We will start from the case of two dimensions. Some situations with more than two dimensions will be considered in later chapters.

Definition 9.5.2 The pair (X, Y) of random variables is said to have a *bivariate normal distribution*, $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, if the joint density is of the form

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{x-\mu_1}{\sigma_1} \times \frac{y-\mu_2}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]}, \quad (9.64)$$

where $-\infty < \mu_1 < \infty$, $-\infty < \mu_2 < +\infty$, $\sigma_1 > 0$, $\sigma_2 > 0$, and $|\rho| < 1$. □

To find marginal distributions of X and Y and determine the interpretation of parameters, let us change the variables into

$$U = \frac{X - \mu_1}{\sigma_1}, \qquad V = \frac{Y - \mu_2}{\sigma_2}$$

This amounts to substitution

$$x = \sigma_1 u + \mu_1, \qquad y = \sigma_2 v + \mu_2$$

with the Jacobian $J = \sigma_1\sigma_2$. Consequently, the joint density of (U, V) is

$$g(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[u^2 - 2\rho uv + v^2]\}.$$

Since $u^2 - 2\rho uv + v^2 = (u-\rho v)^2 + (1-\rho^2)\,v^2$, we write

$$g(u,v) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\left(\frac{u-\rho v}{\sqrt{1-\rho^2}}\right)^2\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{v^2}{2}\right\},$$

where

$$\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\left(\frac{u-\rho v}{\sqrt{1-\rho^2}}\right)^2\right\} du = 1$$

for every v. Hence the second factor is the marginal density of V, while the first factor is the conditional density of U given $V = v$. Thus V is standard normal, while $U|V \sim \mathrm{N}(\rho V, 1-\rho^2)$. By symmetry, U must also be standard normal and $V|U \sim \mathrm{N}(\rho U, 1-\rho^2)$. Therefore we have proved

Theorem 9.5.4 *If (X, Y) have bivariate normal distribution given by (9.64), then both X and Y have normal distributions, $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$, respectively. Moreover, the conditional distributions are also normal:*

$$X|Y \sim N\left(\rho\frac{\sigma_1}{\sigma_2}(Y-\mu_2) + \mu_1, \sigma_1^2(1-\rho^2)\right) \tag{9.65}$$

and

$$Y|X \sim N\left(\rho\frac{\sigma_2}{\sigma_1}(Y-\mu_1) + \mu_2, \sigma_2^2(1-\rho^2)\right). \tag{9.66}$$

Proof. Only the last two statements require some proof. The normality of the conditional distribution is obvious in view of the normality of conditional distributions of U given V. We have

$$\begin{aligned} E(X|Y) &= E(\sigma_1 U + \mu_1 | \sigma_2 V + \mu_2) = \sigma_1 E(U|\sigma_2 V + \mu_2) + \mu_1 \\ &= \sigma_1 E(U|V) + \mu_1 = \sigma_1 \rho V + \mu_1 = \sigma_1 \rho \frac{Y-\mu_2}{\sigma_2} + \mu_1. \end{aligned}$$

Similarly

$$\mathrm{Var}(X|Y) = \mathrm{Var}(\sigma_1 U + \mu_1 | \sigma_2 V + \mu_2) = \sigma_1^2 \mathrm{Var}(U|V) = \sigma_1^2(1-\rho^2),$$

which proves (9.65). The proof of (9.66) is analogous. □

Theorem 9.5.5 *If X, Y have the bivariate normal distribution* (9.64), *then ρ is the coefficient of the correlation between X and Y.*

Proof. Clearly, $\rho_{X,Y} = \text{Cov}(U, V) = E(UV)$, and we may write

$$\begin{aligned} E(UV) &= \int\int uvg(u,v)\,du\,dv \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}}\int\int uv\exp\{-\frac{1}{2(1-\rho^2)}[u^2-2\rho uv+v^2]\}du\,dv \\ &= \int \frac{ve^{-v^2/2}}{\sqrt{2\pi}}\left[\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\int u\exp\left\{-\frac{1}{2}\left(\frac{u-\rho v}{\sqrt{1-\rho^2}}\right)^2\right\}du\right]dv \\ &= \frac{1}{\sqrt{2\pi}}\int ve^{-\frac{v^2}{2}}[\rho v]dv = \rho\frac{1}{\sqrt{2\pi}}\int v^2\exp\{-\frac{v^2}{2}\} = \rho. \end{aligned}$$

□

■ **EXAMPLE 9.17**

Assume, as in Example 9.16, that the height of men X and the height of women Y in a population have N(70, 4) and N(68, 2.25) distributions, respectively. Assume also that the heights of siblings are correlated with $\rho = 0.6$. If we sample a brother and a sister, what is the probability that the sister is taller than her brother?

SOLUTION. Proceeding as in Example 9.16, $P\{Y > X\} = P\{Y - X > 0\}$, and we know that $Y - X$ is normal, with $E(Y - X) = E(Y) - E(X) = 68 - 70 = -2$ inches. For the variance we have

$$\begin{aligned} \text{Var}(Y-X) &= \text{Var}(Y) - 2\text{Cov}(X,Y) + \text{Var}(X) = \sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2 \\ &= (1.5)^2 - 2\times 0.6\times 2\times 1.5 + 2^2 = 2.65. \end{aligned}$$

Consequently,

$$P\{Y > X\} = P\left\{Z > \frac{0-(-2)}{\sqrt{2.65}}\right\} = 1 - \Phi(1.23) = 0.1093.$$

To develop the intuition concerning the bivariate normal distribution and correlation coefficient, let us consider some possible schemes that lead to the appearance of bivariate normal distribution with correlated variables.

■ **EXAMPLE 9.18 Sequential Formation**

It is possible that one of the values, X, is formed at random, following the normal distribution $\text{N}(\mu_1, \sigma_1^2)$. This means that some random process leads to the creation of an element of the population with a specific value x of an attribute X. Subsequently the value of attribute Y is formed by some other random process that generates Y according to the normal distribution with mean $\rho(\sigma_2/\sigma_1)(x - \mu_1) + \mu_2$ and standard deviation $\sigma_2\sqrt{1-\rho^2}$.

Examples of such "sequential" generation of attributes are quite common: We could think here of X and Y being the temperatures at noon at a specific place today

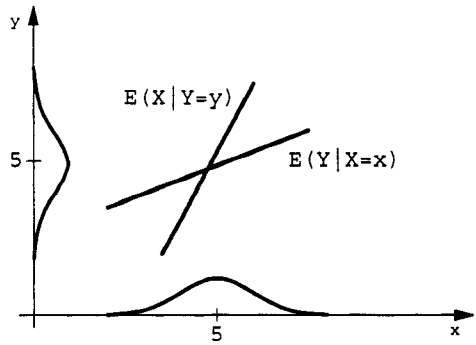


Figure 9.6 Two regression lines

and tomorrow, water levels on the same river at two specific times, or in two places, one downstream from another, the height of a father and a son, and so on. In a sense, we have a "natural" ordering, with X being the first variable, whose value affects the second variable Y.

To enable a clear geometrical representation, let $\sigma_1 = \sigma_2 = 1$ and let μ_1, μ_2 be removed sufficiently far from the origin, to have clear plots of marginals, say $\mu_1 = \mu_2 = 5$ (see Figure 9.6). The conditional expectation of Y, given X (the *regression* of Y on X, discussed in detail in Chapter 14), is now, by formula (9.66),

$$E(Y|X) = \rho(X - 5) + 5,$$

which is the line $y = \rho x + 5(1-\rho)$ with slope ρ, passing through the point $(\mu_1, \mu_2) = (5, 5)$. Observe that regression of X on Y is the line $E(X|Y) = \rho(Y - 5) + 5$; that is, $y = (1\rho)x - 5[(1 - \rho)/\rho]$, which also passes through the point (5, 5) but has the slope $1/\rho$.

After the value of $X = x$ is sampled, the value of Y is sampled from the normal distribution centered at the appropriate point on the regression line, and having variance $1 - \rho^2$. Remembering that the total variance of Y is 1, we have the decomposition of $\mathrm{Var}(Y)$ into a sum of the form $\rho^2 + (1 - \rho^2)$, the second term being the variance of the deviation $Y - [\rho X - 5(1 - \rho)]$ of Y from the regression line. The first term, ρ^2, is therefore the contribution to the variability of Y coming from the variability of X. In generally accepted terminology, $100\rho^2$ is the "percentage of variance of Y *explained* by the variability of X."

A glance at Figure 9.6 reveals that if ρ were larger, the line would be more steep, and this would increase the contribution of X to the variability of Y. (Remember that X and Y are standardized, so their variances remain equal to 1. This explains why the regression line $E(Y|X)$—in the case of standardized variables—cannot be steeper than the diagonal. So, when ρ increases, to keep $\mathrm{Var}(Y) = 1$, one has to decrease the variance of deviations from the regression line.)

■ EXAMPLE 9.19

Another interpretation of the correlation coefficient ρ, not related to any temporal or casual ordering of X and Y, is connected with the following situa-

tion: Suppose that we have a population of objects of some kind, and that an attribute ξ of objects in this population has a normal distribution $N(\mu, \sigma^2)$. Elements of this population are sampled and their attribute ξ is measured twice. The measurements are subject to error, and the errors ϵ_1 and ϵ_2 of the two measurements are independent of ξ and from one another, with the same normal distribution $\epsilon_1 \sim N(0, \sigma_\epsilon^2)$, $\epsilon_2 \sim N(0, \sigma_\epsilon^2)$. The observed results of measurement are $X = \xi + \epsilon_1, Y = \xi + \epsilon_2$. In this case X and Y are normal with the same means and variances $\mu_1 = \mu_2 = \mu, \sigma_1^2 = \sigma_2^2 = \sigma^2 + \sigma_\epsilon^2$. The covariance between X and Y equals $E(XY) - \mu^2 = E(\xi + \epsilon_1)(\xi + \epsilon_2) - \mu^2 = \sigma^2$, in view of the assumed independence and $E(\epsilon_i) = 0$, we have

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\sigma_1^2 \sigma_2^2}} = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\epsilon^2}.$$

Here $\rho > 0$, which means that the results of measurements of the same (random) quantity, subject to independent errors, are *always positively correlated.*

It should be pointed out here that—as opposed to Example 9.18—ρ (not ρ^2) represents the fraction of variance of X (or Y) "explained" by the variability of ξ. The situation is different from that in Example 9.18, since now we "explain" the variance of one variable (X), not through the variability of the second variable of the pair (Y), but through the variability of some other variable (ξ) that affects both X and Y.

PROBLEMS

In all problems of this section, Z stands for a standard normal variable, and Φ is its cdf.

9.5.1 Use the tables of normal distribution to determine the probabilities: **(i)** $P(0 \le Z \le 1.34)$. **(ii)** $P(0.14 \le Z \le 2.01)$. **(iii)** $P(-0.21 \le Z \le -0.04)$. **(iv)** $P(-0.87 \le Z \le 1.14)$. **(v)** $P(|Z| \ge 1.02)$. **(vi)** $P(Z \ge 1.11)$.

9.5.2 Determine x in the following cases (interpolate, if necessary): **(i)** $\Phi(x) = 0.62$. **(ii)** $\Phi(x) = 0.45$. **(iii)** $P(|Z| \le x) = 0.98$. **(iv)** $P(1.4 \le Z \le x) = 0.12$.

9.5.3 Find $P(|X - 2| \le 0.5)$ if $X \sim N(1, 4)$.

9.5.4 Let random variable X have a $N(\mu, \sigma^2)$ distribution. Find: **(i)** μ if $\sigma^2 = 2$ and $P(X \le 12) = 0.72$. **(ii)** σ^2 if $\mu = 2$ and $P(X \ge 5) = 0.39$.

9.5.5 A "100-year water," or flood, is the water level that is exceeded once in a hundred years (on average). Suppose that the threatening water levels occur once a year and have a normal distribution. Suppose also that at some location the 100-year water means the level of 30 feet above average. What is the 10,000-year water level?

9.5.6 Find γ_3—the skewness of a lognormal distribution with parameters μ and σ^2.

9.5.7 Assume that X_1 and X_2 are independent, with $N(3, 6)$ and $N(-1, 2)$ distributions, respectively. Find: **(i)** $P(3X_1 - 2X_2 \ge 14)$. **(ii)** $P(X_1 < X_2)$.

9.5.8 Assume that X_1 and X_2 have a bivariate normal distribution with $E(X_1) = 3, E(X_2) = 2, \mathrm{Var}(X_1) = 4, \mathrm{Var}(X_2) = 1$ and $\rho = -0.6$. Find: **(i)** $P\{X_1 \le 4 | X_2 = 3\}$. **(ii)** $P\{|X_2 - 1| \ge 1.5 | X_1 = 2\}$.

9.6 BETA DISTRIBUTION

The family of *beta distributions* defined below is known for its usefulness in modeling the researcher's uncertainty about the unknown probability p of some event.

Definition 9.6.1 A random variable X has *beta distribution* with parameters $\alpha > 0$ and $\beta > 0$, $X \sim \text{BETA}(\alpha, \beta)$, if the density of X equals

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{9.67}$$

for $0 \le x \le 1$ and $f(x) = 0$ outside interval $[0, 1]$. Here Γ is the function defined by (9.46). □

First, we will check that (9.67) is indeed a density:

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \left(\int_0^\infty u^{\alpha-1}e^{-u}du\right)\left(\int_0^\infty v^{\beta-1}e^{-v}dv\right) \\ &= \int_0^\infty\int_0^\infty u^{\alpha-1}v^{\beta-1}e^{-(u+v)}dudv. \end{aligned}\tag{9.68}$$

Introducing

$$z = u + v \quad \text{and} \quad x = \frac{u}{u+v},$$

we have $0 \le x \le 1, 0 \le z \le \infty$. Since $u = xz$ and $v = z(1-x)$, the Jacobian equals z. Consequently, after substitution to (9.68) the variables separate, and we obtain

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty z^{\alpha-1}\, z^{\beta-1}e^{-z}\, z\, dz \int_0^1 x^{\alpha-1}\,(1-x)^{\beta-1}dx \\ &= \Gamma(\alpha+\beta)\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx,\end{aligned}$$

as was to be shown.

Once we know that

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

we can easily compute the moments of the beta distribution. Indeed, if $X \sim$ BETA (α, β), then

$$\begin{aligned}E(X^k) &= \frac{\Gamma(\alpha+\beta}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{k+\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(k+\alpha)\Gamma(\beta)}{\Gamma(k+\alpha+\beta)} = \frac{\Gamma(\alpha+\beta)\Gamma(k+\alpha)}{\Gamma(\alpha)\Gamma(k+\alpha+\beta)}.\end{aligned}$$

In particular, using formula (9.47), for $k = 1$ and $k = 2$, we obtain

$$E(X) = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad E(X^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)},$$

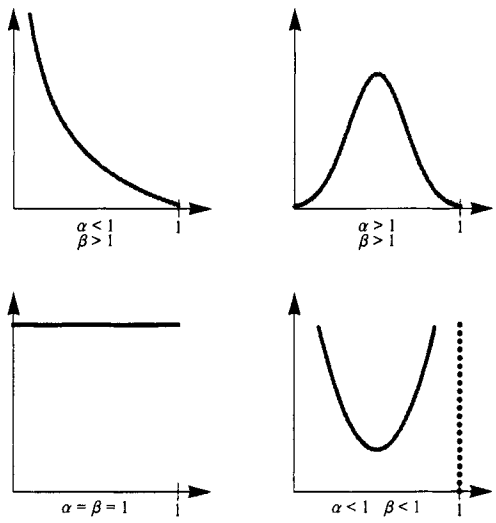


Figure 9.7 Shapes of beta distributions

respectively, so that

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{9.69}$$

The various shapes of the beta distribution are illustrated in Figure 9.7. If $\alpha > 1, \beta > 1$, the distribution is bell-shaped, with the peak becoming more narrow for larger α and/or β. For $\alpha = \beta = 1$, the distribution is uniform. If $\alpha < 1$ and $\beta < 1$, the distribution is U-shaped, whereas if $\alpha < 1, \beta \geq 1$ or $\alpha \geq 1, \beta < 1$, the distribution is J-shaped.

As mentioned, the most typical application of a beta distribution is when we consider a binomial distribution with an unknown p, and represent knowledge (or uncertainty) about p by assuming that p is random with a beta distribution. The examples below illustrate this approach.

■ **EXAMPLE 9.20**

One of the questions posed by Laplace is: What is the probability that the sun will rise tomorrow?

SOLUTION. The question may sound silly in the phrasing above. But in a slight reformulation, it is: Suppose that some event A occurred in all of a large number N of trials. What is the probability that it will occur again on the next trial?

We assume the Bernoulli model of independent trials, and let A_n and $A^{(n)}$ stand for the events "A occurs on n consecutive trials" and " A occurs on nth

trial." Following Laplace, we take the uniform distribution, which is BETA(1, 1), as the distribution of p (presumably, such a choice suggests total impartiality).

We want the expected value of the conditional probability $P(A_{N+1}|A_N)$, the expectation taken with respect to the conditional distribution of p given the data. According to Theorem 4.4.2, on combining evidence, we can proceed in either of two ways:

1. Find the expected probability of joint occurrence of evidence so far and the event in question (hence $A \cap A_N = A_{N+1}$) with respect to the unconditional distribution of p.
2. Find the expected probability of the event in question ($A^{(N+1)}$), with respect to the conditional probability of p given the evidence (A_N).

Using the first approach, we have the probability of the conjunction $A \cap A_N$ equal p^{N+1}, so the answer is, assuming $p \sim$ BETA(1, 1),

$$E(p^{N+1}) = \int_0^1 p^{N+1} dp = \frac{N+1}{N+2}.$$

A rather conservative estimate of the time when we *know* that the sun was rising every day is about 7000 years. Much as we may be convinced that it is not the case, it is possible that before the first written records, the sun did not rise but was (say) switched on suddenly in the sky or operated under some other principles (after all, why would Laplace bother to calculate the probability of the sun rising tomorrow if he did not allow for the possibility of some other mechanism?). Thus, letting $N = 365 \times 7000$, we have $(N + 1)/(N + 2) \approx 1 - 4 \times 10^{-7}$.

■ **EXAMPLE 9.21**

Continuing the Laplace problem in a more realistic setup, we observe Bernoulli trials where we know that p is random with distribution BETA(α, β). So far we know that there were 3 successes in 4 trials. What is the probability of 2 successes in the next 3 trials?

Reasoning in exactly the same way as in the preceding example, the combined data and event in question is "3 successes in 4 trials and 2 successes in the next 3 trials" (which is *not* the same as 5 successes in 7 trials. Why?). Conditioning on p, we obtain the probability

$$\binom{4}{3} p^3(1-p) \binom{3}{2} p^2(1-p) = 12p^5(1-p)^2.$$

Taking expectation with respect to $p \sim$ BETA(α, β), we have, for the expected probability E(2, 3) of 2 successes in 3 trials

$$\begin{aligned} E(2,3) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 12p^5(1-p)^2 p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= 12 \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(5+\alpha)\Gamma(2+\beta)}{\Gamma(7+\alpha+\beta)}. \end{aligned}$$

Using formula (9.47), which asserts that $\Gamma(t) = (t-1)\Gamma(t-1)$, we obtain, after some algebra,

$$E(2,3) = 12 \times \frac{\alpha(\alpha+1)\cdots(\alpha+4)\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+6)}.$$

■ **EXAMPLE 9.22**

To make the situation still more realistic, consider a client who comes to a statistical consulting center for help. It turns out that she is currently negotiating the purchase of a large shipment of some merchandise. Some of the items conform to the specification, and some do not (let us call them defective). Suppose that testing the quality of all items in the lot is not possible because of the prohibitive cost of such an operation, or because the process of testing is destructive. Some tests have been made, though, and the data are such that among 25 tested items 3 were found to be defective. What can one reasonably say about the fraction of defective items in the whole lot?

SOLUTION. Suppose that the lot in question is not the first lot purchased by the client. It turns out that she has been buying the same kinds of merchandise from the same company (or from other companies). In her experience, as judged by tests, data on returns, customer complaints, and so on, she estimates that on average, the percentages of items below specification (defectives) in various lots purchased were about 12%, with variability of about 2% in either direction.

The problem of the client at the statistical consulting center can be formulated as follows: First we determine the chances that the fraction p of defective items in the lot whose purchase is presently being negotiated is below 10%. Then we evaluate $P\{p < 0.1\}$ in light of the tests of the lot under consideration (3 out of 25 below specification), using the experience accumulated in the past. If we accept the assumption that the fractions in various lots follow a beta distribution, we have for the mean

$$\frac{\alpha}{\alpha+\beta} = 0.12. \tag{9.70}$$

Next the relatively small variability ($\pm 2\%$) suggests that we have $\alpha > 1$ and $\beta > 1$. Consequently the density of beta distribution is bell-shaped. Using the three-sigma rule, we take $3\sigma \approx 0.02$, and using (9.69), we obtain another equation for α and β:

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \left(\frac{0.02}{3}\right)^2. \tag{9.71}$$

The solution of (9.70) and (9.71) is $\alpha^* = 285, \beta^* = 2090$.

We have assumed here that the fraction p of defective items is a random variable whose distribution is BETA(α^*, β^*). If no testing of the currently negotiated lot was made, the answer would be

$$P\{p < 0.1\} = \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \int_0^{0.1} x^{\alpha^*-1}(1-x)^{\beta^*-1}dx.$$

However, we have additional information, namely that event A = "three defective in a sample of 25" occurred. Consequently, we should integrate the conditional density of p given A. We have

$$P(A|p) = \binom{25}{3} p^3(1-p)^{22}.$$

The conditional density is obtained from Bayes' formula, with the summation replaced by integration. Schematically

$$\varphi_p(x|A) = \frac{P(A|x)\varphi_p(x)}{\int P(A|u)\varphi_p(u)du}, \tag{9.72}$$

where $\varphi_p(x)$ and $\varphi_p(x|A)$ are the unconditional and conditional densities of p at x.

Typical reasoning is that the denominator in (9.72) is independent of x, and both sides represent densities in x. Collecting the terms that depend on x, and representing all other factors as a constant, the right-hand side of (9.72) is of the form

$$Cx^3(1-x)^{22}x^{\alpha^*-1}(1-x)^{\beta^*-1}, \tag{9.73}$$

so it is a beta density with parameters $\alpha^* + 3$ and $\beta^* + 22$. This gives

$$C = \frac{\Gamma(\alpha^* + \beta^* + 25)}{\Gamma(\alpha^* + 3)\Gamma(\beta^* + 22)},$$

and we get the answer to the original question by integrating (9.73) between 0 and 0.1.

Observe that the analysis above has some interesting consequences. We know that if the distribution of p is BETA(α, β) and we observe k successes in n trials (event A), then the conditional distribution of p given A is BETA$(\alpha+k, \beta+n-k)$—the family of beta distributions that constitutes the *conjugate priors* for binomial data. We will return to those problems in Chapter 12.

PROBLEMS

9.6.1 Let X have a distribution BETA(α, β). Find: **(i)** The distribution of $Y = 1 - X$. **(ii)** $E\{X^k(1-X)^m\}$.

9.6.2 Let $X_1, \ldots, X_n$ be independent variables with a U$[0, 1]$ distribution. The joint density of the $X = \min(X_1, \ldots, X_n)$ and $Y = \max(X_1, \ldots, X_n)$ has density $f(x, y) = n(n-1)(y-x)^{n-2}$ for $0 \le x \le y \le 1$ and $f(x, y) = 0$ otherwise. Find: **(i)** $E(X^m)$, $E(Y^m)$, and $\rho(X, Y)$. **(ii)** $E(Y|X)$.

9.6.3 Let X have a symmetric beta distribution. Find α and β if the coefficient of variation (ratio of standard deviation and the mean) is k. Does a solution exist for all k?

9.6.4 Let $X_1, X_2, \ldots$ be iid with an exponential distribution. For any positive m and n find the distribution of the ratio

$$T_{m,n} = \frac{X_1 + \cdots + X_m}{X_1 + \cdots + X_{m+n}}.$$

CHAPTER 10

RANDOM SAMPLES

10.1 STATISTICS AND THEIR DISTRIBUTIONS

In the previous chapters we referred to independent random variables $X_1, \ldots, X_n$ having the same distribution, as iid (independent, identically distributed) variables. In statistics we say that such variables constitute a random sample, since they can be thought of as observations (measurements) independently selected from the same population, or resulting from analogous statistical experiments. A random sample provides an important link between the observed data and the distribution in the population from which it has been selected. Most techniques of statistical inference that will be discussed in later chapters will be based on certain functions of random samples. Such functions are called *statistics* if they depend on observations $X_1, \ldots, X_n$, but not on parameter(s) of the distribution. We will discuss the properties of the distributions of statistics (called *sampling distributions*).

Before we start, it should be mentioned that there are several important issues related to the properties of a selection process so that it does yield a random sample. There are also other, and sometimes more appropriate, sampling schemes besides the random sample. That will be discussed in more detail in Chapter 11.

Among various statistics meaningful for the purpose of statistical inference is a sample mean

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

an arithmetic average of all measurements, or equivalently their sum

$$T_n = X_1 + \cdots + X_n.$$

Among other important statistic are: sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{10.1}$$

sample standard deviation $S = \sqrt{S^2}$, $\min(X_1, \ldots, X_n)$ and $\max(X_1, \ldots, X_n)$—the smallest and the largest values in the data, respectively.

The following example illustrates an interesting distributional property:

■ **EXAMPLE 10.1**

We will determine the distribution of $T_n = X_1 + \cdots + X_n$ and of $\overline{X} = (1/n)T_n$ in a random sample $X_1, \ldots, X_n$ selected from an EXP(λ) distribution.

SOLUTION. Based on Theorems 8.5.4 and 8.5.5, the moment generating function for the sample mean $\overline{X}$ is

$$m_{\overline{X}}(t) = m_{T_n}\left(\frac{t}{n}\right) = \left[m_X\left(\frac{t}{n}\right)\right]^n = \left(1 - \frac{\lambda t}{n}\right)^{-n}, \tag{10.2}$$

From Theorem 8.5.6 it is clear that distributions of T_n and $\overline{X}$ are GAM(n, λ) and GAM($n, \lambda/n$), respectively.

A similar property for the family of normal distributions will be shown in Section 10.2.

Regardless of the distribution in the sample, one can obtain expectations of some sampling distributions as functions of the moments of original distributions as given in the following theorem:

Theorem 10.1.1 *Let $X_1, \ldots, X_n$ be a random sample from a distribution with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then*

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}, \tag{10.3}$$

and similarly

$$E(X_1 + \cdots + X_n) = n\mu, \quad \mathrm{Var}(X_1 + \cdots + X_n) = n\sigma^2. \tag{10.4}$$

Moreover, $E(S^2) = \sigma^2$.

Proof. Formulas (10.3) and (10.4) can be found in Sections 8.4 and 8.7. The proof that $E(S^2) = \sigma^2$ is very simple. Since $E(W^2) = \mathrm{Var}(W) + [E(W)]^2$ for any random variable W, and $\sum(X_i - \overline{X})^2 = \sum X_i^2 - n(\overline{X})^2$, we have

$$\begin{aligned} E\left(\sum X_i^2\right) - nE(\overline{X})^2 &= n[E(X_i^2) - E(\overline{X})^2] \\ &= n[\sigma^2 + \mu^2 - (\sigma^2/n + \mu^2)] = (n-1)\sigma^2. \end{aligned}$$

□

PROBLEMS

10.1.1 Statistic G_k, defined for $k = 1, 2$ as

$$G_k = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|^k,$$

was proposed as a measure of variation by Jordan (1869). Show that $G_2 = 2S^2$, where S^2 is given by formula (10.1).

10.1.2 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two random samples from distributions with means μ_1 and μ_2, respectively, and the same variance σ^2. **(i)** Find $E(\overline{X} - \overline{Y})$ and $\mathrm{Var}(\overline{X} - \overline{Y})$. **(ii)** Assuming that $\mu_1 = \mu_2$ and both samples are of equal size, find $n = m$ such that $P(|\overline{X} - \overline{Y}| > \sigma/4) \leq 0.05$.

10.1.3 Let $\overline{X}_1, \overline{X}_2, \overline{X}_3$ be sample means in three independent samples of sizes n_1, n_2, n_3, respectively. Each sample was obtained from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Find the distribution of $V_1 = (1/3)(\overline{X}_1 + \overline{X}_2 + \overline{X}_3)$ and $V_2 = w_1\overline{X}_1 + w_2\overline{X}_2 + w_3\overline{X}_3$, where $w_i = n_i(n_1 + n_2 + n_3)$.

10.2 DISTRIBUTIONS RELATED TO NORMAL

The distributions that will be introduced in this section are of special importance in statistical inference. Their applications as sampling distributions of statistics obtained in random samples selected from the normal distribution will be discussed in later chapters. We start from the following theorem:

Theorem 10.2.1 *Let $X_1, \ldots, X_n$ be a random sample from the normal distribution* $\mathrm{N}(\mu, \sigma^2)$. *Then*

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \tag{10.5}$$

and

$$U = \frac{n-1}{\sigma^2} S^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{10.6}$$

are independent random variables with distributions $\mathrm{N}(\mu, \sigma^2/n)$ *and* χ^2_{n-1}, *respectively.*

Proof. There exist several ways of proving this theorem. We will use a multidimensional version of the method of finding the density of a function of random variables.

Let us observe that

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \overline{x} + \overline{x} - \mu)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 + n(\overline{x} - \mu)^2,$$

since $2\sum_{i=1}^{n}(x_i-\overline{x})(\overline{x}-\mu)=2(\overline{x}-\mu)\sum_{i=1}^{n}(x_i-\overline{x})=0$. Consequently, the joint density of $(X_1,\ldots,X_n)$ can be written as

$$\begin{aligned} f(x_1,\ldots,x_n) &= \frac{1}{\sigma^n(2\pi)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right\} \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}}\exp\left\{-\frac{u}{2\sigma^2}\right\}\exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right\}. \end{aligned} \tag{10.7}$$

Using transformation

$$x_j=\overline{x}+v_j\sqrt{u}, \qquad j=1,2,\ldots,n, \tag{10.8}$$

we obtain

$$\sum_{j=1}^{n}v_j=0, \qquad \sum_{j=1}^{n}v_j^2=n, \tag{10.9}$$

which means that two of the v_j's are functions of the remaining ones. We therefore solve (10.9) for v_{n-1} and v_n, obtaining two solutions:

$$v_{n-1}=\frac{A-B}{2}, \qquad v_n=\frac{A+B}{2},$$

or

$$v_{n-1}=\frac{A+B}{2}, \qquad v_n=\frac{A-B}{2},$$

where

$$A=-\sum_{k=1}^{n-2}v_k, \qquad B=\left(2n-2\sum_{k=1}^{n-2}v_k^2-\sum_{k=1}^{n-2}\sum_{l=1}^{n-2}v_kv_j\right)^{1/2}. \tag{10.10}$$

Thus, to each vector $(v_1,\ldots,v_{n-2},\overline{x},u)$ with $u>0$, there are two corresponding systems $(x_1,\ldots,x_n)$:

$$\begin{aligned} x_j &= \overline{x}+v_j\sqrt{u}, \qquad j=1,\ldots,n-2 \\ x_{n-1} &= \overline{x}+\frac{A-B}{2}\sqrt{u}, \qquad x_n=\overline{x}+\frac{A+B}{2}\sqrt{u} \end{aligned} \tag{10.11}$$

and

$$\begin{aligned} x_j &= \overline{x}+v_j\sqrt{u}, \qquad j=1,\ldots,n-2 \\ x_{n-1} &= \overline{x}+\frac{A+B}{2}\sqrt{u}, \qquad x_n=\overline{x}+\frac{A-B}{2}\sqrt{u}. \end{aligned} \tag{10.12}$$

The Jacobian of transformation (10.11) is

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial \overline{x}} & \cdots & \frac{\partial x_n}{\partial \overline{x}} \\ \frac{\partial x_1}{\partial u} & \cdots & \frac{\partial x_n}{\partial u} \\ \frac{\partial x_1}{\partial v_1} & \cdots & \frac{\partial x_n}{\partial v_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial x_1}{\partial v_{n-2}} & \cdots & \frac{\partial x_n}{\partial v_{n-2}} \end{vmatrix}$$

$$= \begin{vmatrix} 1 & 1 & \ldots & 1 & 1 & 1 \\ \frac{v_1}{2\sqrt{u}} & \frac{v_2}{2\sqrt{u}} & \cdots & \frac{v_{n-2}}{2\sqrt{u}} & \frac{A-B}{2} \times \frac{1}{2\sqrt{u}} & \frac{A+B}{2} \times \frac{1}{2\sqrt{u}} \\ \sqrt{u} & 0 & \ldots & 0 & \frac{\sqrt{u}}{2}\left(\frac{\partial A}{\partial v_1} - \frac{\partial B}{\partial v_1}\right) & \frac{\sqrt{u}}{2}\left(\frac{\partial A}{\partial v_1} + \frac{\partial B}{\partial v_1}\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & \sqrt{u} & \frac{\sqrt{u}}{2}\left(\frac{\partial A}{\partial v_{n-2}} - \frac{\partial B}{\partial v_{n-2}}\right) & \frac{\sqrt{u}}{2}\left(\frac{\partial A}{\partial v_{n-2}} + \frac{\partial B}{\partial v_{n-2}}\right) \end{vmatrix}.$$

Since for transformation (10.12) the Jacobian is the same, but with the last two columns interchanged, the absolute value of the Jacobian for transformation (10.12) is the same as for transformation (10.11). We have here

$$|J| = u^{(n-3)/2} h(v_1, \ldots, v_{n-2}), \tag{10.13}$$

where h is some function whose exact form is not needed for our purposes.

Substitution of new variables into (10.7), and multiplication by (10.13) gives the same joint density of vector $(\overline{X}, U, V_1, \ldots, V_{n-2})$ for transformations (10.11) and (10.11), namely

$$g(\overline{x}, u, v_1, \ldots, v_{n-2}) = \frac{1}{(\sigma\sqrt{2\pi})^n} u^{(n-3)/2} e^{-u/2\sigma^2} e^{-(n/2\sigma^2)(\overline{x}-\mu)^2} h(v_1 \ldots, v_{n-2}). \tag{10.14}$$

Since the density (10.14) can be written as the product of densities

$$c_1 u^{(n-1)/2-1} e^{-u/2\sigma^2} \times c_2 e^{-(n/2\sigma^2)(\overline{x}-\mu)^2} \times c_3 h(v_1, \ldots, v_{n-2}),$$

it is now clear that the random variables $U, \overline{X}$ and $(V_1, \ldots, V_{n-2})$ are independent. An inspection of terms involving $\overline{x}$ and u shows that after we adjust the constants, $\overline{X}$ has a $N(\mu, \sigma^2/n)$ distribution and U/σ^2 has a chi-square distribution with $n-1$ degrees of freedom. □

The most remarkable fact here is that $\overline{X}$ and U are independent, even though $\overline{X}$ appears explicitly in the definition of U. This independence is characteristic for a normal distribution. We have the following theorem, which we leave without a proof:

Theorem 10.2.2 *If in a random sample $X_1, \ldots, X_n$ variables $\overline{X}$ and U are independent, then the distribution of X_i, $i = 1, \ldots, n$, is normal.*

We will now introduce the following definition:

Definition 10.2.1 Let Z be a standard normal random variable, and let U be a chi-square distributed random variable independent of Z, with ν degrees of freedom. Then the random variable

$$X = \frac{Z}{\sqrt{U/\nu}} \tag{10.15}$$

is said to have *Student's t distribution with ν degrees of freedom.* □

The density of X will be derived following the steps outlined in Chapter 7. Starting from the joint density of Z and U, we add a "companion" variable Y to X, find the Jacobian of the transformation $(z, u) \to (x, y)$, and finally, find the density of X as the marginal in the joint density of X and Y.

By the assumption of independence, the joint density of Z and U is

$$f(z,u) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \times \frac{1}{2^{\nu/2}\Gamma(\nu/2)} u^{\nu/2-1} e^{-u/2}, \qquad u > 0. \tag{10.16}$$

Selecting $Y = U$ as a companion variable to X, we have $z = x\sqrt{y/m}$ and $u = y$. The Jacobian equals $J = \sqrt{y/m}$; hence the joint density of X and Y is

$$\varphi(x,y) = Q y^{(\nu-1)/2} e^{-(1/2)(1+x^2/\nu)y}, \tag{10.17}$$

where

$$Q = \frac{1}{\sqrt{\pi\nu}\, 2^{(m+1)/2}\Gamma(\nu/2)}. \tag{10.18}$$

The density (10.17), except for a constant, is a gamma density with parameters $(m+1)/2$ and $(1/2)(1+x^2/m)$. Thus, the marginal density of X is obtained by integration as

$$\begin{aligned} f_\nu(x) &= \int_0^\infty Q y^{(\nu-1)/2} e^{-(1/2)(1+x^2/\nu)y} dy \qquad (10.19) \\ &= \frac{Q\Gamma((\nu+1)/2)}{[(1/2)(1+x^2/\nu)]^{(\nu+1)/2}} = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \times \left(1 + x^2/\nu\right)^{-(\nu+1)/2}. \end{aligned}$$

It is evident that the density (10.19) of the t_ν distribution is symmetric around $x = 0$. When $\nu = 1$, we have $\varphi(x) = K(1+x^2)^{-1}$, which is the density of the Cauchy distribution (see Example 8.12). Observe that for $\nu = 1$, the Student's t distribution has no mean (hence no higher moments either). For $\nu > 1$ and $k < \nu$, the kth ordinary moment of t_ν can be obtained as (see Problem 10.2.6):

$$E(X^k) = E(Z^k) \left(\frac{\nu}{2}\right)^{k/2} \frac{\Gamma\left(\frac{\nu-k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{10.20}$$

where Z has N(0, 1) distribution. Consequently, we have $E(X) = 0$ and $\mathrm{Var}(X) = \nu/(\nu-2)$ for variable X with t_ν distribution ($\nu > 2$).

When the number of degrees of freedom, ν, increases, t_ν distribution approaches the standard normal distribution.

Theorem 10.2.3 *Student's t_n distribution approaches the standard normal distribution* N(0, 1) *when the number of degrees of freedom n increases.*

Proof. The density of a t distribution with ν degrees of freedom is proportional to the product

$$\left[\left(1+\frac{x^2}{\nu}\right)^{\nu}\right]^{-1/2} \times \left[\left(1+\frac{x^2}{\nu}\right)\right]^{-1/2}.$$

When $\nu \to \infty$, the first term converges to $(e^{x^2})^{-1/2} = e^{-x^2/2}$ while the second term converges to 1. To preserve the integrability to 1, the constants must converge to $1/\sqrt{2\pi}$, and the limiting density is standard normal. □

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution. By Theorem 10.2.1, $\overline{X}$ and $\sum_{i=1}^n (X_i - \overline{X})^2$ are independent, with $\sqrt{n}(\overline{X}-\mu)/\sigma$ and $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2$ having standard normal and chi-square distribution with $n-1$ degrees of freedom, respectively. Consequently, the ratio

$$\frac{(\overline{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2/(n-1)\sigma^2}} = \frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \tag{10.21}$$

has a Student's t distribution with $n-1$ degrees of freedom. As we will see in Chapters 12 and 13, the random variable (10.21) plays an important role in building schemes for inference on μ.

Table A3. gives upper quantiles for t distribution for selected numbers of degrees of freedom, where the upper quantile $t_{\alpha,\nu}$ is defined by the relation $P\{X \geq t_{\alpha,\nu}\} = \alpha$, with X having the Student's t distribution with ν degrees of freedom (so that the upper α-quantile is a $(1-\alpha)$-quantile).

The following definition introduces yet another distribution important in statistical inference:

Definition 10.2.2 If U and V are independent random variables with distributions $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$, respectively, then the random variable

$$X = \frac{U/\nu_1}{V/\nu_2}, \tag{10.22}$$

is said to have *Snedecor's F distribution with ν_1 and ν_2 degrees of freedom*, to be denoted F_{ν_1,ν_2}. □

Let us derive the density of the random variable X. The joint density of U and V, in view of their independence, is

$$\varphi(u,v) = \frac{1}{2^{\nu_1/2}\Gamma(\nu_1/2)} u^{\nu_1/2-1} e^{-u/2} \times \frac{1}{2^{\nu_2/2}\Gamma(\nu_2/2)} v^{\nu_2/2-1} e^{-v/2} \tag{10.23}$$

for $u > 0, v > 0$. With $Y = V$ as a companion variable, the inverse transformation $(x,y) \to (u,v)$ is $u = (\nu_1/\nu_2)xy$, and $v = y$ so that the Jacobian equals $J = (\nu_1/\nu_2)y$. Thus, after substitution to (10.23), the joint density of X and Y is

$$\psi(x,y) = \frac{\nu_1^{\nu_1/2}}{\nu_2^{\nu_1/2} 2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} x^{\nu_1/2-1} y^{(\nu_1+\nu_2)/2-1} e^{-(1/2)[\nu_1 x/\nu_2+1]y}.$$

The marginal density of X equals, again using the fact that we can separate a gamma-type integral for y,

$$\begin{aligned}\varphi_{\nu_1,\nu_2}(x) &= \int_0^\infty \psi(x,y)dy = \frac{\nu_1^{\nu_1/2}x^{\nu_1/2-1}}{\nu_2^{\nu_1/2}2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \\ &\quad \times \int_0^\infty y^{(\nu_1+\nu_2)/2-1}e^{-(1/2)[\nu_1 x/\nu_2+1]y}dy \\ &= \frac{\nu_1^{\nu_1/2}x^{\nu_1/2-1}\Gamma((\nu_1+\nu_2)/2)2^{(\nu_1+\nu_2)/2}}{\nu_2^{\nu_1/2}2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)[\nu_1 x/\nu_2+1)^{(\nu_1+\nu_2)/2}} \\ &= \frac{\Gamma((\nu_1+\nu_2)/2))\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \times \frac{x^{\nu_1/2-1}}{(\nu_1 x+\nu_2)^{(\nu_1+\nu_2)/2}}.\end{aligned}$$

The upper quantiles of the F distribution can be found in Table A6. Since the distribution depends on a pair (ν_1, ν_2) of degrees of freedom (see the margins of the table), each entry gives just one quantile for the corresponding numbers of degrees of freedom—the 5% upper quantile in Table A6a. and the 1% quantile in Table A6b. Quantiles of order 1% and 5% can also be obtained from these tables (see Problem 10.2.2).

PROBLEMS

10.2.1 Let $X \sim t_\nu$. Show that X^2 has $F_{1,\nu}$ distribution.

10.2.2 Show that if x_α is the α-quantile of a random variable X with an F_{ν_1,ν_2} distribution, then $1/x_\alpha$ is the $(1-\alpha)$-quantile of a random variable with an F_{ν_2,ν_1} distribution.

10.2.3 Let X, Y, W be independent random variables such that $X \sim$N(0, 1), $Y \sim$N(1, 1), and $W \sim$N(2, 4), respectively. Use Table A6. to find k such that

$$P\left\{\frac{X^2+(Y-1)^2}{X^2+(Y-1)^2+(W-2)^2/4} > k\right\} = 0.05.$$

10.2.4 Let $X_1 \sim \text{GAM}(1,\lambda)$, $X_2 \sim \text{GAM}(2,\lambda)$, $X_3 \sim \text{GAM}(3,\lambda)$ be independent random variables. Find constant a such that variable $Y = aX_1/(X_2+X_3)$ has an F distribution.

10.2.5 Show that variable $aX/(1+aX)$, where $X \sim \text{F}(\nu_1,\nu_2)$ and $a = \nu_1/\nu_2$, has a BETA$(\nu_1/2, \nu_2/2)$ distribution.

10.2.6 Derive formula (10.20). [*Hint:* Use formula (9.52) for the kth moment of a GAM(α,λ) distribution and the fact that variables X and U in (10.15) are independent.]

10.2.7 Derive the formula for $E(X^k)$, where $X \sim \text{F}(\nu_1,\nu_2)$. (*Hint:* Use similar approach as in Problem 10.2.6).

10.3 ORDER STATISTICS

Let $X_1, X_2, \ldots, X_n$ be a random sample from continuous distribution with a cdf F and density f. We will consider random variables $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$, where $X_{i:n}$ is the ith in magnitude among $X_1, X_2, \ldots, X_n$. Thus

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}. \tag{10.24}$$

Since variables X_i's are continuous, the probability of two observations being equal is zero, and therefore we can assume that all inequalities in (10.24) are strict.

Definition 10.3.1 The random variables (10.24) are called *order statistics* of the sample $X_1, X_2, \ldots, X_n$, and for $k = 1, \ldots, n$, $X_{k:n}$ will be referred to as the kth order statistic. □

In particular, the first order statistic, $X_{1:n}$, is the minimum of the sample, while the nth order statistic, $X_{n:n}$, is the maximum. We will derive the joint distribution of all n order statistics, of two selected order statistics, and also the marginal distribution of any order statistic. We will start from the latter. Let G_k be the cdf of order statistic $X_{k:n}$. Then $X_{k:n} \leq t$ if at least k variables in the sample $X_1, X_2, \ldots, X_n$ satisfy the inequality $X_i \leq t$. The number of observations in the sample that satisfy $X_i \leq t$ has a binomial distribution with probability of success $p = F(t) = P(X_i \leq t)$, and therefore

$$G_k(t) = P\{X_{k:n} \leq t\} = \sum_{r=k}^{n} \binom{n}{r} [F(t)]^r [1 - F(t)]^{n-r}. \tag{10.25}$$

Two particular cases deserve attention here. If $k = 1$, we obtain

$$G_1(t) = P\{X_{1:n} \leq t\} = 1 - P\{X_1 > t, \ldots, X_n > t\} = 1 - [1 - F(t)]^n,$$

and if $k = n$, then

$$G_n(t) = P\{X_{n:n} \leq t\} = P\{X_1 \leq t, \ldots, X_n \leq t\} = [F(t)]^n.$$

Differentiating with respect to t, we obtain densities

$$g_1(t) = n[1 - F(t)]^{n-1} f(t) \quad \text{and} \quad g_n(t) = n[F(t)]^{n-1} f(t) \tag{10.26}$$

To find the density $g_k(t)$, we differentiate (10.25) obtaining

$$\begin{aligned} g_k(t) &= \sum_{r=k}^{n} \binom{n}{r} r[F(t)]^{r-1} f(t)[1 - F(t)]^{n-r} \\ &\quad - \sum_{r=k}^{n-1} \binom{n}{r} [F(t)]^r (n-r)[1 - F(t)]^{n-r-1} f(t) \\ &= \sum_{r=k}^{n} \frac{n!}{(r-1)!(n-r)!} [F(t)]^{r-1} [1 - F(t)]^{n-r} f(t) \\ &\quad - \sum_{r=k+1}^{n} \frac{n!}{(r-1)!(n-r)!} [F(t)]^{r-1} [1 - F(t)]^{n-r} f(t). \end{aligned}$$

All terms except one cancel in the last two sums.

Theorem 10.3.1 *The density of the kth order statistic is given by*

$$g_k(t) = \frac{n!}{(k-1)!(n-k)!}[F(t)]^{k-1}[1-F(t)]^{n-k}f(t). \tag{10.27}$$

To find distributions of more complex functions of order statistics, one should start from the joint distribution of all order statistics given in the following theorem:

Theorem 10.3.2 *The joint density of $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ is given by*

$$\phi(y_1, \ldots, y_n) = \begin{cases} n!f(y_1)\cdots f(y_n) & \text{if } y_1 < \cdots < y_n \\ 0 & \text{otherwise.} \end{cases} \tag{10.28}$$

The proof is very simple and will be omitted. The reader is urged nevertheless to consider the case $n = 3$, draw a picture, and determine the partition of the three-dimensional space that accounts for the appearance of the factor $3! = 6$ in the density. If $n = 3$ is too hard, we suggest to start with the case $n = 2$.

The next theorem gives the density of a joint distribution of any two order statistics.

Theorem 10.3.3 *Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density f and cdf F. For any $k < l$ and $s < t$, the joint distribution of $X_{k:n}$ and $X_{l:n}$ is given by the formula*

$$\begin{aligned} g_{k,l}(s,t) &= \frac{n!}{(k-1)!(l-k-1)!(n-l)!}[F(s)]^{k-1}[F(t)-F(s)]^{l-k-1} \\ &\quad \times[1-F(t)]^{n-l}f(s)f(t) \end{aligned} \tag{10.29}$$

and $g_{k,l}(s,t) = 0$ otherwise.

Proof. The density (10.29) is obtained by direct integration of the joint density (10.28) with respect to all y_i such that $i < k$, $k < i < l$, and $i > l$. □

■ EXAMPLE 10.2 Distribution of the Range

We present two different ways of deriving the distribution of a range $R = X_{n:n} - X_{1:n}$ of a random sample $X_1, \ldots, X_n$.

SOLUTION 1. The first solution consists of finding the joint distribution of $X_{1:n}$ and $X_{n:n}$, and then determining the distribution of $X_{n:n} - X_{1:n}$ using techniques for transformations of random variables discussed in Chapter 7.

For $s < t$, the direct application of (10.29) gives

$$\begin{aligned} g_{1,n}(s,t) &= \frac{n!}{(0)!(n-2)!(0)!}[F(t)-F(s)]^{n-2}f(s)f(t) \\ &= n(n-1)[F(t)-F(s)]^{n-2}f(s)f(t). \end{aligned}$$

Since $R = X_{n:n} - X_{1:n}$, we can take the transformation $r = t - s$ and its "companion" $w = t$. Solving for t and s, we obtain $s = w - r$ and $t = w$ with

$|J| = 1$. Thus the density of R is the marginal density for $r > 0$,

$$\begin{aligned} h_R(r) &= \int_{-\infty}^{\infty} g_{1,n}(w - r, w)dw \qquad (10.30) \\ &= \int_{-\infty}^{+\infty} n(n-1)[F(w) - F(w-r)]^{n-2} f(w-r)f(w)\, dw. \end{aligned}$$

SOLUTION 2. A more direct solution is to condition on one of the variables, say $X_{1:n}$. If $X_{1:n} = s$, then the probability that the range is less than r is $\{[F(s+r) - F(s)]/[1 - F(s)]\}^{n-1}$, since all remaining observations must fall between s and $s+r$. The density of $X_{1:n}$ is given by (10.26) and therefore

$$\begin{aligned} P\{R \le r\} &= \int_{-\infty}^{+\infty} \left\{ \frac{[F(s+r) - F(s)]^{n-1}}{[1 - F(s]^{n-1}} \right\}^{n-1} \times n[1 - F(s)]^{n-1} f(s)ds \\ &= \int_{-\infty}^{+\infty} n[F(s+r) - F(s)]^{n-1} f(s)ds. \end{aligned}$$

Differentiation with respect to r gives the density of R, in a somewhat different form.

It is clear that unlike variables $X_1, \ldots, X_n$ in a random sample, order statistics $X_{1:n}, \ldots, X_{n:n}$ are dependent variables. For example, $X_{i:n} \le X_{j:n}$ for any $1 \le i < j \le n$.

■ **EXAMPLE 10.3**

We will obtain correlation ρ of two extreme order statistics, $X_{1:n}$ and $X_{n:n}$, in a random sample from U[0, θ] distribution.

SOLUTION. To find ρ, we need the first and second moments of both variables $X_{1:n}$ and $X_{n:n}$, and their covariance. Based on Example 6.15 and densities (10.26) we have

$$\begin{aligned} E(X_{1:n}) &= \frac{n}{\theta} \int_0^{\theta} t \left(1 - \frac{t}{\theta}\right)^{n-1} dt = n\theta \int_0^1 (1-w)w^{n-1} dw \\ &= \frac{1}{n+1}\theta \qquad (10.31) \end{aligned}$$

and

$$E(X_{n:n}) = \frac{n}{\theta} \int_0^{\theta} t \left(\frac{t}{\theta}\right)^{n-1} dt = n\theta \int_0^1 u^n du = \frac{n}{n+1}\theta. \qquad (10.32)$$

Similarly the second moments are

$$\begin{aligned} E(X_{1:n}^2) &= \int_0^{\theta} t^2 g_1(t)dt = \frac{n}{\theta} \int_0^{\theta} t^2 \left(1 - \frac{t}{\theta}\right)^{n-1} dt \\ &= n\theta^2 \int_0^1 (1-w)^2 w^{n-1} dw = \frac{2\theta^2}{(n+1)(n+2)} \qquad (10.33) \end{aligned}$$

and

$$E(X_{n:n}^2) = \int_0^\theta t^2 g_n(t)dt = \frac{n}{\theta}\int_0^\theta t^2\left(\frac{t}{\theta}\right)^{n-1} dt = \frac{n\theta^2}{n+2}. \tag{10.34}$$

Consequently,

$$\mathrm{Var}(X_{1:n}) = \frac{2\theta^2}{(n+1)(n+2)} - \frac{\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)^2(n+2)}$$

and

$$\mathrm{Var}(X_{n:n}) = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)^2(n+2)} = \mathrm{Var}(X_{1:n}).$$

The equality of variances should not be surprising. Since $g_1(t) = g_n(\theta - t)$ for $0 \le t \le \theta$, the variability in distributions of $X_{1:n}$ and $X_{n:n}$ is the same. To find $\mathrm{Cov}(X_{1:n}, X_{n:n})$, we first obtain $E(X_{1:n}X_{n:n})$ using the joint density

$$g_{1,n}(s,t) = \frac{n(n-1)}{\theta^n}(t-s)^{n-2} \quad \text{for} \quad 0 \le s < t \le \theta.$$

Based on formula (10.29),

$$E(X_{1:n}X_{n:n}) = \int_0^\theta \int_0^t s\,t\,\frac{n(n-1)}{\theta^n}(t-s)^{n-2} ds\,dt,$$

and consequently

$$\mathrm{Cov}(X_{1:n}, X_{n:n}) = \frac{\theta^2}{n+2} - \frac{\theta}{(n+1)} \times \frac{n\theta}{(n+1)} = \frac{\theta^2}{(n+1)^2(n+2)}. \tag{10.35}$$

Finally,

$$\begin{aligned} \rho &= \frac{\mathrm{Cov}(X_{1:n}, X_{n:n})}{\sqrt{\mathrm{Var}(X_{1:n})}\sqrt{\mathrm{Var}(X_{n:n})}} \\ &= \frac{\theta^2}{(n+1)^2(n+2)} \times \frac{(n+1)^2(n+2)}{n\theta^2} = \frac{1}{n}. \end{aligned} \tag{10.36}$$

The obtained result is intuitive. The correlation of $X_{1:n}$ and $X_{n:n}$ decreases when the number of observations in the sample (sample size) increases.

■ EXAMPLE 10.4 Theory of Outliers

As another illustration of the use of order statistics, we will present basic definitions and results of the theory of outliers proposed by Neyman and Scott (1971).

An outlier is an observation that is sufficiently far away from the remaining observations to justify the suspicion that it results from an observational or recording error, or perhaps from a sudden undetected change of conditions

that made this particular observation obey different probability distributions than those obeyed by the remaining observations. What we present here is a theory of right outliers. The results for left outliers may be obtained by obvious modification. Thus we have a sample, $x_1, \ldots, x_n$, with values of order statistics $x_{1:n}, \ldots, x_{n:n}$ where $x_{n:n}$ is, in some sense, "too far away" from the rest of observations. The intuition of "being too far away" has been formalized in a number of ways in various approaches to the detection of outliers, typically in terms of relation of $x_{n:n}$ to the average and standard deviation of the sample.

Neyman and Scott (1971) proposed the definition based on order statistics. In what follows, we will always assume that $n \geq 3$ and $r > 0$.

Definition 10.3.2 The sample $x_1, \ldots, x_n$ contains an (r, n) *right outlier*, if

$$x_{n:n} - x_{n-1:n} > r(x_{n-1:n} - x_{1:n}). \tag{10.37}$$

We will simply use the term *outlier* for the right outlier. □

It is clear that if only $x_{n:n} \neq x_{n-1:n}$, the sample contains an (r, n) outlier for all r satisfying the condition

$$r < \frac{x_{n:n} - x_{n-1:n}}{x_{n-1:n} - x_{1:n}}.$$

Intuitively the *outlier* in the usual sense is an (r, n) outlier for r large enough.

Theorem 10.3.4 *A random sample $X_1, \ldots, X_n$ from a continuous distribution with cdf F and density f will contain an (r, n) outlier with probability*

$$\pi(r, n; F) = n(n-1) \int_{-\infty}^{\infty} \int_{x}^{\infty} \left[F\left(\frac{y + rx}{r+1} \right) - F(x) \right]^{n-2} f(y)\, f(x)\, dy\, dx. \tag{10.38}$$

Proof. Indeed, if we condition on $x_{1:n} = x$ and $x_{n:n} = y > x$, then the sample will contain an (r, n) outlier if (10.37) holds, hence if

$$x < X_{n-1:n} < \frac{y + rx}{r+1}.$$

This event, in turn, occurs if $n - 2$ remaining observations lie between x and $(y + rx)/(r+1)$, which occurs with probability $[F((y + rx)/(r + 1)) - F(x)]^{n-2}$. Multiplying by the joint density of $(X_{1:n}, X_{n:n})$ and integrating, we obtain (10.38). □

Formula (10.38) allows one to compute the probability that an (r, n) outlier will appear for a given F and n. One of the essential features that characterize situations in statistics is that the distribution F that governs the selection of the sample is not known. Typically we know only that the true F belongs to a certain family $\mathcal{F}$ of distributions. For example, the experimenter may know that the sample she observes was obtained from a normal distribution but not know the value(s) of the parameter(s).

Accordingly we introduce the following definitions:

Definition 10.3.3 The family $\mathcal{F}$ of distributions is (r, n) *outlier resistant* if

$$\sup_{F \in \mathcal{F}} \pi(r, n; F) < 1 \tag{10.39}$$

and is (r, n) *outlier prone* if

$$\sup_{F \in \mathcal{F}} \pi(r, n; F) = 1. \tag{10.40}$$

Moreover, a family $\mathcal{F}$ of distributions is *totally outlier resistant* if condition (10.39) holds for all $r > 0$ and $n \geq 3$; it is called *totally outlier prone*, if condition (10.40) holds for all $r > 0$ and $n \geq 3$. □

Among other results proved by Neyman and Scott (Neyman and Scott, 1971), we mention two theorems. One of them asserts that *the family of all normal distributions is totally outlier resistant*. The other asserts that *the family of all gamma distributions is totally outlier prone.* [16] The practical consequences of these two theorems are as follows: On the one hand, if we have a sample $x_1, \ldots, x_n$ from a normal distribution, we can find the largest r for which this sample contains an (r, n) outlier. This r is equal to the ratio $r^* = (x_{n:n} - x_{n-1:n})/(x_{n-1:n} - x_{1:n})$. We can then find the quantity (10.38) for $\mathcal{F}$ being a family of the normal distribution and $r = r^*$. If this quantity is sufficiently small, we have good reason to suspect that the element $x_{n:n}$ in the sample is a genuine outlier, in the sense of representing observations from a distribution other than that of the rest of the sample. This gives a practical procedure for the rejection of outliers.

On the other hand, if we have a sample $x_1, \ldots, x_n$ from a gamma distribution, we can *never* reject the largest element as an outlier if our decision is to be based on the observed values only. To see this, suppose that we have data such as $x_{1:4} = 0.5, x_{2:4} = 0.55, x_{3:4} = 1$, and $x_{4:4} = 1,000,001$. We cannot reject the observation $1,000,001$ as an outlier, since the probability of a configuration with the ratio

$$\frac{x_{4:4} - x_{3:4}}{x_{3:4} - x_{1:4}} = 2,000,000$$

or more has a probability arbitrarily close to 1 for *some* gamma distribution. More precisely, for every $\epsilon > 0$ there is a gamma distribution such that the probability of the configuration above exceeds $1 - \epsilon$.

Our conclusion that in practice one should never reject any sample element as an outlier if the sample is known to come from some gamma distribution is actually counterintuitive. Such a conclusion does unexpectedly provide a powerful argument for the Bayesian approach to statistics, but the point is that it seldom, if ever, happens that the statistician knows *only* that the sample comes from a gamma distribution. He usually has some idea about the parameters, based on his experience, imagination, understanding of the situation, and so on. Such knowledge, vague as it may be, allows the statistician to regard some gamma distributions in the family $\mathcal{F}$ as "more plausible" than others, and perhaps even eliminate some members of $\mathcal{F}$ as impossible in the given situation. Now, if we restrict the family $\mathcal{F}$ to *some* gamma

[16] For further development of the suggested theory of outliers, see Green (1974, 1976).

distributions only (by putting a bound on parameter α), then $\mathcal{F}$ is no longer totally outlier prone, and rejection of outliers becomes justifiable.

PROBLEMS

10.3.1 Determine the distribution of $X_{1:n}$ in a random sample selected from the EXP(λ) distribution.

10.3.2 Let $X_1, \ldots, X_n$ be a random sample selected from the U[0, 1] distribution. **(i)** Show that the distribution of $X_{j:n}$ is BETA($j, n - j + 1$) for $j = 1, \ldots, n$. **(ii)** Find the distribution of a sample median $X_{k+1:n}$, and obtain its variance if $n = 3, 7, 15, 2k + 1$. Do you see any trend?

10.3.3 Let X_1, X_2 be a random sample of size 2 from a continuous distribution with a median θ. **(i)** Find $P(X_{1:2} < \theta < X_{2:2})$. **(ii)** Generalize part (i) finding $P(X_{1:n} < \theta < X_{n:n})$ in a random sample of size n.

10.3.4 Use results from Example 10.3 to determine Var($X_{l:n} - X_{k:n}$, $l > k$,) in a random sample $X_1, \ldots, X_n$ from the U[0, θ] distribution.

10.3.5 Let $X_1, \ldots, X_n$ be a random sample from the U[0, 1] distribution. Find sample size n such that $E(R) = 0.75$, where $R = X_{n:n} - X_{1:n}$.

10.3.6 Let $X_1, \ldots, X_5$ be a random sample from the BETA(2, 1) distribution. Find: **(i)** The density of a joint distribution of $X_{1:5}, X_{2:5}, X_{4:5}$. **(ii)** $E(X_{2:5}|X_{4:5})$. **(iii)** The distribution of $Y = X_{2:5}/X_{1:5}$.

10.3.7 Let $X_1, \ldots, X_n$ and $Y_i, \ldots, Y_n$ be two independent random samples from the same continuous distribution with a density f. Show that $P\{X_{i:n} \leq t\} \geq P\{Y_{j:n} \leq t\}$ for every t if and only if $i \leq j$.

10.3.8 Find $\pi(r, n; F_\theta)$ when F_θ is the U[0, θ] distribution. Find $\pi(r, n; \mathcal{F})$ for the family $\mathcal{F} = \{\mathcal{F}_\theta\}, \theta > 0$.

10.4 GENERATING RANDOM SAMPLES

In various quantitative problems that are too complex to be studied mathematically, one can often use simulations to find an approximate solution. Such techniques of statistical sampling, known today as Monte Carlo methods,[17] have been used in a very simple form (e.g., tossing dice) for centuries, but became a widely applicable formal methodology only with the invention of electronic computers in the middle of the 20th century.

[17] Stanisław Ulam, a Polish born mathematician working with John von Neuman and Nicholas Metropolis on the Manhattan Project during World War II, developed computer algorithms for statistical sampling. He also worked on transforming deterministic problems into random ones that could be solved by simulations. In his spare time Ulam investigated probabilities of winning a card game of solitaire, and that is how the new methodology got named after a famous casino.

Since then, with increasing computing power and easy access to technology, simulations have become an important and powerful tool of modern statistics. Today complex statistical problems are often investigated based on generated random samples from specific distributions that are easy to obtain even using available statistical software packages. We now introduce related concepts and explain the theory behind the Monte Carlo simulation process.

To generate a random sample from any distribution, we first need to generate independent observations from the U[0, 1] distribution. Over the last 40 years, several algorithms have been proposed and extensively tested, so the ones that we use today are of really good quality. However, "random number generators" are not really generators of random numbers. They produce numbers that can be periodical but with periods much larger than we are able to detect, and therefore not affecting the quality of our simulation process. The interested readers should consult some additional literature, such as Ross (2006).

We will now discuss the random sample generation process, assuming that a good random generator of a U[0, 1] distribution is available.

The following example will show how one can generate a random sample from a simple discrete distribution:

■ **EXAMPLE 10.5**

Let X be a Bernoulli random variable such that

$$P(X = 1) = 1 - P(X = 0) = p = 0.2.$$

To generate a single observation from this distribution we must first generate U, a single observation from the U[0, 1] distribution, and then transform it into X observation in a following way:

$$X = \begin{cases} 1 & \text{if } U \leq 0.2 \\ 0 & \text{otherwise.} \end{cases} \tag{10.41}$$

To generate a random sample of size n, we simply need to generate a random sample $U_1, \ldots, U_n$ from the U[0, 1] distribution and then transform each U_i into X_i in the way explained above. The transformation given by (10.41) is not the only one possible here. Other possibilities could, for example, be

$$X = \begin{cases} 1 & \text{if } U > 0.8 \\ 0 & \text{otherwise} \end{cases} \quad \text{or} \quad X = \begin{cases} 1 & \text{if } 0.3 < U \leq 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

as long as $P(X = 1) = 0.2$.

To generate a random sample $Y_1, \ldots, Y_n$ from the binomial distribution BIN(k, p), we will need $k \times n$ observations from the U[0, 1] distribution. Then we will transform each of them into the BIN(1, p) according to (10.41) and obtain $Y_1, \ldots, Y_n$ as

$$Y_i = \sum_{j=(i-1)\times n+1}^{i\times n} X_j$$

for $i = 1, \ldots, k$. Random samples from any distribution related to Bernoulli trials can be generated in a similar way.

The next example shows how to generate values from other discrete distributions.

■ **EXAMPLE 10.6**

Let X be a random variable with a POI(λ) distribution, in which $P(X = x) = e^{-\lambda}\lambda^x/x!$. After a value of random variable U is generated from U[0, 1], the value x of X is determined in a following way: x equals 0 when $U \leq e^{-\lambda}$, otherwise x is the only positive integer for which the inequality

$$\sum_{i=0}^{x-1} \frac{e^{-\lambda}\lambda^i}{i!} < U \leq \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$$

holds.

To generate observations from continuous distributions with a cdf F such that F^{-1} has a closed form, we can apply Theorem 6.4.2, as illustrated by the following example:

■ **EXAMPLE 10.7**

To obtain a random sample $X_1, \ldots, X_n$ from the EXP(λ) distribution, we have to generate a random sample $U_1, \ldots, U_n$ from U[0, 1] and for any $i = 1, \ldots, n$, take

$$X_i = -\frac{1}{\lambda}\log(1 - U),$$

since $F_X^{-1}(u) = -\frac{1}{\lambda}\log(1 - u)$ is an inverse of the cdf $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.

The generation that we just presented can also serve as a first step in generating some distributions from a gamma family. By Theorem 9.4.1, variable $Y = X_1 + \cdots + X_k$, where variables X_i's are independent and have the same EXP(λ) distribution, will have a GAM(k, λ) distribution. Therefore, to generate $Y_1, \ldots, Y_n$, we could start by generating a random sample $U_i, \ldots, U_{n\times k}$ from U[0, 1] and then transform it into a sample $X_1, \ldots, X_{n\times k}$ to finally obtain a random sample $Y_1, \ldots, Y_n$, where for any $i = 1, \ldots, n$,

$$Y_i = \sum_{j=(i-1)\times k+1}^{i\times k} X_j.$$

Unfortunately, this approach can be used only if the shape parameter in a gamma distribution is a positive integer.

In general, there exist methods that can be used to generate all probability distributions with an invertible cdf that is not in a closed form. One of these techniques, called the Accept-Reject algorithm, is presented below.

The idea here is that the two observations are generated independently. One observation, U, is from the U[0, 1] distribution, and the other, Y, is from a selected continuous distribution, which is easy to generate. If a certain condition is satisfied, the Y observation is "accepted" as an X observation and included in a random sample $X_1, \ldots, X_n$. Otherwise, it is "rejected," and the next pair (U, Y) is generated. Pairs (U, Y) continue to be generated until there are n observations $X_1, \ldots, X_n$ in our random sample.

To introduce the condition that makes some of the Y observations become X observations and to explain why the process is really valid, we start with a constant C, such that

$$f(x) \leq Cg(y), \tag{10.42}$$

where f is the density of a distribution we are interested in (a "target" distribution), and g is a density of a "companion" distribution. The generated Y observation is "accepted" as an X observation if

$$U \leq \frac{f(Y)}{Cg(Y)}; \tag{10.43}$$

otherwise, it is "rejected." To validate the algorithm, we have to show that the distribution of X is the same as a conditional distribution of Y under the condition (10.43) so that

$$F_X(y) = F_Y\left(y \,|\, U \leq \frac{f(Y)}{Cg(Y)}\right).$$

Since

$$\begin{aligned} P\left(U \leq \frac{f(Y)}{Cg(Y)}\right) &= \int_{-\infty}^{+\infty} P\left(U \leq \frac{f(y)}{Cg(y)}\right) g(y)\, dy \\ &= \int_{-\infty}^{+\infty} \frac{f(y)}{Cg(y)} g(y)\, dy = \frac{1}{C}, \end{aligned} \tag{10.44}$$

we have

$$\begin{aligned} P\left(Y \leq y \,|\, U \leq \frac{f(Y)}{Cg(Y)}\right) &= \frac{P(Y \leq y, U \leq f(Y)/Cg(Y))}{P(U \leq f(Y)/Cg(Y))} \\ &= C \int_{-\infty}^{y} g(t) \int_{0}^{f(t)/Cg(t)} du\, dt \\ &= C \int_{-\infty}^{y} g(t) \frac{f(t)}{Cg(t)} dt = \int_{-\infty}^{y} f(t) dt = F_X(y) \end{aligned}$$

as was to be shown.

Two questions still need to be answered here: Does it matter how the constant C is selected? And how many (U, Y) pairs will have to be generated before we get $X_1, \ldots, X_n$? From (10.44) it is clear that although any $C < \infty$ that satisfies (10.42) can be used, the optimal choice is $C = \sup_y f_Y(y)/g(y)$. Unfortunately, such $C < \infty$ cannot be found for all distributions. In these cases one could use more advanced methods. The interested reader is referred to Ross (2006).

PROBLEMS

10.4.1 Obtain a sample of size 6 from a POI(2) distribution based on following six independent observations from U[0, 1] distribution: 0.090907, 0.185040, 0.124341, 0.299086, 0.428996, 0.927245.

10.4.2 Obtain a random sample of size 4 from a Pareto distribution with a density $f(x) = (1 + x)^{-2}$ for $x > 0$ and 0 otherwise. Use the following random sample from U[0, 1]: 0.187724, 0.386997, 0.182338, 0.028113.

10.4.3 The double exponential (or Laplace) distribution has a density given by the formula $f(x) = (\lambda/2)e^{-\lambda|x|}$ for $-\infty < x < \infty, \lambda > 0$. Obtain a random sample from the Laplace distribution with $\lambda = 2$ based on a random sample 0.744921, 0.464001 from the U[0, 1] distribution.

10.4.4 A generalized Laplace distribution has a density given by the formula

$$f(x) = \begin{cases} p\lambda_1 e^{-\lambda_1 x} & \text{if } x \geq 0 \\ (1-p)\lambda_2 e^{\lambda_2 x} & \text{otherwise,} \end{cases} \tag{10.45}$$

where $\lambda_1 > 0, \lambda_2 > 0$. Generate two independent observations from a generalized Laplace distribution with $p = 1/4, \lambda_1 = 3$ and $\lambda_2 = 1/2$, based on a random sample 0.647921, 0.049055 from U[0, 1] distribution.

10.4.5 Generate a random sample from the Gompertz distribution with survival function $S(t) = \exp\{1 - \exp(2t)\}$ using the following random sample from the U[0,1] distribution: 0.289365, 0.228349, 0.732889.

10.4.6 Apply an Accept/Reject method to the Laplace distribution with density $(1.5)e^{-3|x|}$ to generate observations from a standard normal distribution. List the obtained values and specify how many of them you were able to obtain using the random sample of size 5 from U[0, 1]: 0.222795, 0.516174, 0.847152, 0.466449, 0.914370. Use the optimal choice of C.

10.4.7 The Box-Muller transformation of two independent, uniform variables into two independent standard normal variables was presented in Theorem 7.4.1. Another algorithm, proposed by Marsaglia and Bray (1964), is to generate U_1 and U_2 as two independent observations from U[$-1, 1$]. If $V = U_1^2 + U_2^2 \leq 1$, then define

$$Z_1 = U_1\sqrt{\frac{-2\log V}{V}} \quad \text{and} \quad Z_2 = U_2\sqrt{\frac{-2\log V}{V}};$$

otherwise, start the procedure again. Show that Z_1 and Z_2 are independent and have an N$(0, 1)$ distribution.

10.4.8 Kennedy and Gentle (1980) provide the following algorithm for generating a beta distribution: Generate U_1 and U_2—two independent observations from the U[0, 1] distribution. For $\alpha > 0$ and $\beta > 0$ denote $V_1 = U_1^{\alpha}$ and $V_2 = U_2^{\beta}$. According to the Accept/Reject algorithm, let $X = V_1/(V_1 + V_2)$ if $V_1 + V_2 \leq 1$; otherwise, start the procedure again. **(i)** Determine the distribution of a random variable X. **(ii)** Use this algorithm to generate a random sample of size 3 from BETA(0.732, 1.281).

10.4.9 Genest (1987) provides the following algorithm for generating random samples from the so-called Frank family of bivariate distributions: (a) Generate two independent observations U_1 and U_2 from U[0, 1]. (b) Obtain $T = \alpha^{U_1} + (\alpha - \alpha^{U_1})U_2$. (c) Let $X = U_1$ and $Y = \log_\alpha[T/(T + (1-\alpha)U_2)]$, where $\alpha > 0, \alpha \neq 1$. **(i)** Show that the bivariate cdf of Frank's distribution has the following form:

$$H_\alpha(x, y) = P(X \leq x, Y \leq y) = \log_\alpha \left\{ 1 + \frac{(\alpha^x - 1)(\alpha^y - 1)}{\alpha - 1} \right\}.$$

(ii) Generate one observation of (X, Y) based on two independent observations from U[0, 1]: 0.548291 and 0.179112. Use $\alpha = 4$.

10.5 CONVERGENCE

Several limit theorems were already encountered in past chapters. Before proceeding with a systematic exploration of the topic, we recall them briefly. First, on several occasions we looked at the differences between sampling with and without replacement, and noted that these two schemes of sampling become closer one to another "as the population becomes larger." Second, we proved the Poisson approximation theorem, which asserts that as n becomes larger and p becomes smaller, the binomial and Poisson probabilities of the same events become close. Finally, in Chapter 5 we showed that (under some conditions) as the number of transitions of a Markov chain increases, the probability of finding the system in a given state approaches a limiting value.

The common feature of these theorems was that in each case the probabilities of certain events—or, more generally, distributions of certain random variables—approached some limits with the appropriate change of one or more parameters. Typically the parameter is an index of some sequence, such as sample size n, but the case of Poisson approximation shows that it can also be a simultaneous change of two parameters that drives the probabilities to their limiting values.

In addition to the limit theorem mentioned above, we have encountered a different kind of limit theorem, exemplified by the law of large numbers (Example 8.56). There we had convergence not only of the distributions but of the random variables themselves.

To grasp the difference between those two classes of situations, observe that one can study convergence of a sequence of distributions $\{F_n\}$ without considering any random variables. On the other hand, if $X_1, X_2, \ldots$ is a sequence of independent random variables with the same distribution F, then the sequence $\{F_n\}$ of their distributions clearly converges to F, but we cannot expect any regularity in behavior of the sequence $\{X_n\}$.

In what follows, we consider a sequence $\xi_1, \xi_2, \ldots$ of random variables defined on the same sample space $\mathcal{S}$ with σ-field $\mathcal{F}$ of events, and probability measure P on $\mathcal{S}$. Our first objective will be to distinguish various possible modes in which sequence $\{\xi_n\}$ can converge, and discuss their interrelationships. To connect the analysis with subsequent statistical concepts, it is worthwhile to start from some examples of sequences $\{\xi_n\}$ that arise in statistical research and practice.

■ **EXAMPLE 10.8**

One of the most common situations in statistics is when we have simple random samples. This means that we observe the beginning of a sequence of independent and identically distributed (iid) random variables $X_1, X_2, \ldots$. Depending on the goal of analysis, given the observations $X_1, \ldots, X_n$, the statistician computes the value of some statistic $\xi_n = H(X_1, \ldots, X_n)$ and uses ξ_n as means of inference. The behavior of the sequence $\{\xi_n\}$ as n increases tells the statistician to which extent it would be worthwhile to increase the sample size n.

We begin with the definition that captures the type of convergence encountered in the law of large numbers (Theorem 8.8.3).

Definition 10.5.1 The sequence $\{\xi_n\}$ *converges in probability to a constant* c if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\{|\xi_n - c| \geq \epsilon\} = 0. \tag{10.46}$$

More generally, we say that $\{\xi_n\}$ converges in probability *to a random variable* ξ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\{|\xi_n - \xi| \geq \epsilon\} = 0. \tag{10.47}$$

Convergence in probability, especially to a constant, is often called a *stochastic* convergence. □

The meaning of (10.46) is that as n increases, it becomes less and less likely that ξ_n will deviate from c by more than ϵ. In (10.47) the interpretation is the same, except that the constant c is replaced by a random quantity. So written more explicity, (10.47) reads

$$\lim_{n\to\infty} P\{s : |\xi_n(s) - \xi(s)| \geq \epsilon\} = 0.$$

We will use the symbol $\xrightarrow{P}$ to denote convergence in probability.

■ **EXAMPLE 10.9 Laws of Large Numbers**

The law of large numbers proved in Section 8.8 asserts convergence in probability of empirical frequencies of an event. Such theorems (asserting convergence in probability of averages of random variables) are called *weak* laws of large numbers (WLLNs), to distinguish them from *strong* laws of large numbers, discussed later.

For instance, in the case of a Bernoulli distribution we have the following rephrasing of Theorem 8.8.3:

Theorem 10.5.1 *If* S_n *has binomial distribution* BIN(n, p), *then*

$$\frac{S_n}{n} \xrightarrow{P} p.$$

A stronger type of convergence of sequences of random variables is given in the following definition:

Definition 10.5.2 Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables defined on some probability space $(\mathcal{S}, \mathcal{F}, \mathcal{P})$. If $\lim \xi_n(s) = \xi(s)$ exists for all points $s \in U$ where $P(U) = 1$, then we say that ξ_n converges to ξ almost *everywhere* (a.e.), *almost surely* (a.s.), or *with probability* 1. □

The following theorem is given without proof:

Theorem 10.5.2 *If* $\xi_n \to \xi$ *a.s., then* $\xi_n \xrightarrow{P} \xi$.

The converse to Theorem 10.5.2 is not true, as illustrated by the next example.

■ **EXAMPLE 10.10**

Let $\xi_1, \xi_2, \ldots$ be independent random variables, such that $P\{\xi_n = 1\} = 1/n, P\{\xi_n = 0\} = 1 - 1/n$. Thus, if $0 < \epsilon < 1$, then $P\{|\xi_n| \geq \epsilon\} = P\{\xi_n = 1\} = 1/n \to 0$, which shows that the sequence $\{\xi_n\}$ converges to 0 in probability. For any sample point, the sequence $\{\xi_n(s)\}$ is simply a sequence of 0's and 1's, and in order to converge to 0, there must be only a finite number of terms equal 1(i.e., all terms must be 0, starting from some N). But letting $A_n = \{\xi_n = 1\}$, we have $\sum P(A_n) = \sum 1/n = \infty$, and by the second Borel-Cantelli lemma (Theorem 4.5.5) we know that with probability 1 infinitely many events A_n will occur. Thus $P\{\xi_n \text{ converges}\} = 0$, which shows that convergence in probability does not imply a.s. convergence.

Theorem 10.5.3 *The sequence* $\{\xi_n\}$ *of random variables* ξ *converges a.s. to a random variable* ξ *if and only if for every* $k = 1, 2, \ldots,$

$$\lim_{N \to \infty} P\{|\xi_n - \xi| \geq \frac{1}{k} \text{ for some } n \geq N\} = 0, \tag{10.48}$$

or equivalently,

$$\lim_{N \to \infty} P\{\sup_{n \geq N} |\xi_n - \xi| \geq \frac{1}{k}\} = 0.$$

We will now use (10.48), together with the Kolmogorov inequality (Theorem 8.8.5), to prove the sufficiency part of Kolmogorov's three series theorem. This theorem provides conditions for the a.s. convergence of the series $\sum_{j=1}^{\infty} X_j$ of independent random variables $X_1, X_2, \ldots$.

We will also introduce the method of truncation, a powerful technique of handling limits of sequences of random variables. If X is a random variable and $c > 0$, we define the *truncation* of X at c as a random variable $Y = Y^c$, defined by

$$Y = \begin{cases} X & \text{if } |X| \leq c \\ 0 & \text{if } |X| > c. \end{cases}$$

Observe that Y is a bounded random variable, so $E(Y)$ and (Y) both exist. We have

Theorem 10.5.4 (Kolmogorov Three Series Theorem) *Let $X_1, X_2, \ldots$ be a sequence of independent random variables, and let Y_n be the truncation of X_n at level $c > 0$. Then $\sum_{n=1}^{\infty} X_n$ converges a.s. if and only if for some $c > 0$,*

(a) $\sum_{n=1}^{\infty} P\{|X_n| > c\} < \infty$.

(b) $\sum_{n=1}^{\infty} E(Y_n) < \infty$.

(c) $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty$.

Proof. As mentioned, only the sufficiency of conditions (a)–(c) will be shown; the proof of necessity is beyond the scope of this book.

Let us fix N, k, and $n > N$, and consider the sums

$$\sum_{j=N}^{r} (Y_j - E(Y_j))$$

for $r = N, \ldots, n$. By the Kolmogorov inequality (8.57), we have

$$P\left\{ \max_{N \le r \le n} \left| \sum_{j=N}^{r} (Y_j - E(Y_j)) \right| \ge \frac{1}{k} \right\} \le \frac{\sum_{j=N}^{n} \mathrm{Var}(Y_j)}{(1/k)^2}.$$

Letting $n \to \infty$, we have

$$P\left\{ \sup_{N \le r} \left| \sum_{j=N}^{r} (Y_j - E(Y_j)) \right| \ge \frac{1}{k} \right\} \le k^2 \sum_{n=N}^{\infty} \mathrm{Var}(Y_n).$$

In view of (c), the right-hand side converges to 0 as $N \to \infty$ for every fixed k. By Theorem 10.5.3, the sequence $\sum_{i=1}^{n} (Y_j - E(Y_j)), n = 1, 2, \ldots$, converges a.s. Since $\sum_{i=1}^{r} E(Y_i)$ converges (condition b), we infer that $\sum_{i=1}^{\infty} Y_i$ converges a.s. To complete the proof, observe that $P\{|X_n| > c\} = P\{X_n \ne Y_n\}$. In view of condition (a) and the first Borel-Cantelli lemma (Theorem 2.6.2), with probability 1 only finitely many terms Y_n will differ from terms X_n. Consequently, $\sum X_n$ and $\sum Y_n$ will a.s. differ only by a finite quantity, and since $\sum Y_n$ converges, so does $\sum X_n$. □

■ **EXAMPLE 10.11**

For the harmonic series we have $1 + \frac{1}{2} + \frac{1}{3} + \cdots = \infty$. With alternating signs, we have $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \frac{\pi}{4}$. What if the signs are allocated at random, by a flip of a fair coin? In other words, we ask about convergence of the series

$$X_1 + X_2 + X_3 + \cdots,$$

where X_n assumes values $\pm 1/n$ with probability $1/2$ and X_n's are independent. Taking $c = 2$, say, we have $Y_n = X_n$ for all n, and all terms of the series (a) in Theorem 10.5.4 are zero. Next we have $E(Y_n) = 0$ and $\mathrm{Var}(Y_n) = E(Y_n^2) = 1/n^2$. Thus all three series (a)–(c) converge; hence

$\sum X_n$ converges a.s. In other words, the probability is zero that random signs $+$ and $-$ in the harmonic series will come out so unbalanced as to make the series diverge.

Observe that this in not only the question of "balancing" the numbers of positive and negative signs. Indeed, the series $\sum 1/\sqrt{n}$ diverges, and the series $\sum \epsilon_n/\sqrt{n}$, where ϵ_n's are independent and $\epsilon_n = \pm 1$ with probability $1/2$, also diverges a.s.—since in this case the series (c) of variances diverges.

The third important type of convergence is the following:

Definition 10.5.3 Let $\xi_0, \xi_1, \xi_2, \ldots$ be a sequence of random variables, and let $F_n(t) = P\{\xi_n \leq t\}, n = 0, 1, 2, \ldots$, be their cdf's. The sequence $\{\xi_n\}$ *converges in distribution* to ξ_0 if

$$\lim_{n\to\infty} F_n(t) = F_0(t)$$

for every t at which $F_0(t)$ is continuous. In this case we can write $\xi_n \xrightarrow{d} \xi_0$. Alternatively, we will use the symbol $F_n \Longrightarrow F_0$. □

Before presenting the theorems characterizing convergence in distribution, we will make some comments, that can help clarify the motivation and intention of this concept.

First, observe that convergence in a distribution does not imply anything about the behavior of random variables. For instance, if the variables $\xi_0, \xi_1, \ldots$ are independent and identically distributed (iid), then $F_n(t) \equiv F_0(t)$, so $\xi_n \xrightarrow{d} \xi_0$, but we cannot expect any regularity in behavior of the observed values of random variables. Since we require only convergence of the distribution functions, we do not need to have any specific random variables ξ_n in mind. This is the reason behind the dual notation in Definition 10.5.3.

The second question concerns the reasons for requiring the convergence of $F_n(t)$ to $F_0(t)$ only at points of continuity of F_0 (and not at all points). The explanation has to do with the special role played by discontinuities of cdf's. Consider a sequence of degenerate random variables (i.e., constants), defined as $\xi_n = 1/n, n = 1, 2, \ldots$, and $\xi_0 = 0$. Obviously here $\lim \xi_n = \xi_0$ in the "usual" sense (of convergence of observed values of ξ_n to ξ_0), and we can choose to cover this case by the definition of convergence in the distribution.

We have here

$$F_n(t) = P\{\xi_n \leq t\} = \begin{cases} 0 & \text{if } t < 1/n \\ 1 & \text{if } t \geq 1/n \end{cases}$$

and

$$F_0(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0. \end{cases}$$

For any $t \neq 0$ (where F_0 is continuous), we have $\lim F_n(t) = F_0(t)$. However, at $t = 0$ we have $F_0(0) = 1$, while $F_n(0) = 0$ for all n. Observe that if we put $\xi_n = -1/n$, then $F_n(t)$ will converge to $F_0(t)$ at all points t, including $t = 0$.

We now present some theorems that connect the three types of convergence introduced above, and also some criteria for convergence in a distribution.

Theorem 10.5.5 (Slutsky) *If ξ_n and η_n are the sequences of random variables such that $\xi_n - \eta_n \xrightarrow{P} 0$ and $\eta_n \xrightarrow{d} \xi$, then $\xi_n \xrightarrow{d} \xi$.*

Proof. Let F be the cdf of ξ, and let F_n be the cdf of ξ_n. Let t be a continuity point of F, and let $\epsilon > 0$ be such that $t + \epsilon$ and $t - \epsilon$ are also continuity points of F. We write

$$\begin{aligned} F_n(t) &= P\{\xi_n \le t\} = P\{\xi_n \le t, |\xi_n - \eta_n| < \epsilon\} + P\{\xi_n \le t, |\xi_n - \eta_n| \ge \epsilon\} \\ &\le P\{\eta_n \le t + \epsilon\} + P\{|\xi_n - \eta_n| \ge \epsilon\}. \end{aligned}$$

By similar reasoning we obtain

$$P\{\eta_n \le t - \epsilon\} \le F_n(t) + P\{|\xi_n - \eta_n| \ge \epsilon\}.$$

As $n \to \infty, P\{\eta_n \le t \pm \epsilon\} \to F(t \pm \epsilon)$, and $P\{|\xi_n - \eta_n| \ge \epsilon\} \to 0$ by the assumption of the theorem. Consequently,

$$F(t - \epsilon) \le \liminf F_n(t) \le \limsup F_n(t) \le F(t + \epsilon).$$

Since $\epsilon > 0$ is arbitrary (subject only to the condition that F is continuous at $\pm\epsilon$), we must have

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

as was to be shown. □

Taking $\eta_n = \xi, n = 1, 2, \ldots$, we obtain the following:

Theorem 10.5.6 *If $\xi_n \xrightarrow{P} \xi$, then $\xi_n \xrightarrow{d} \xi$.*

Since we already know that convergence a.s. implies convergence in probability, we also have

Theorem 10.5.7 *If $\xi_n \to \xi$ a.s., then $\xi_n \xrightarrow{d} \xi$.*

Although, as already explained, convergence in distribution does not imply convergence in probability, such implication holds in the case where the convergence is to a constant. We have

Theorem 10.5.8 *If $\xi_n \xrightarrow{d} c$, then $\xi_n \xrightarrow{P} c$.*

Proof. The condition $\xi_n \xrightarrow{d} c$ means that $\{\xi_n\}$ converges in distribution to a random variable ξ such that $P\{\xi = c\} = 1$. Consequently, letting again $F_n(t) = P\{\xi_n \le t\}$, we have $F_n(x) \to 0$ for all $x < c$ and $F_n(x) \to 1$ for all $x > c$. But this means that $P\{|\xi_n - c| > \epsilon\} = P\{\xi_n < c - \epsilon\} + P\{\xi_n > c + \epsilon\} = F_n(c - \epsilon - 0) + [1 - F_n(c + \epsilon)] \to 0$ for every $\epsilon > 0$, which is what was to be proved. □

We shall state now a theorem that shows the extent to which one is allowed to carry the algebraic manipulations on sequences converging in distribution and in probability.

Theorem 10.5.9 *If ξ_n, α_n, and β_n are sequences of random variables such that $\xi_n \xrightarrow{d} \xi, \alpha_n \xrightarrow{P} \alpha, \beta_n \xrightarrow{P} \beta$, where α and β are finite constants, then*

$$\alpha_n \xi_n + \beta_n \xrightarrow{d} \alpha\xi + \beta.$$

In particular, we have here the following corollary:

Corollary 10.5.10 *If $\xi_n \xrightarrow{d} \xi$, and a, b, a_n, b_n are constants such that $a_n \to a, b_n \to b$, then*

$$a_n \xi_n + b_n \xrightarrow{d} a\xi + b.$$

The following theorem, which we state without proof, will serve as one of the main tools in proving convergence in distribution to normally distributed random variables:

Theorem 10.5.11 *Let $\xi_n, n = 1, 2, \ldots$, and ξ be random variables, such that their moment generating functions $m_n(t)$ and $m(t)$ exist. Then $\xi_n \xrightarrow{d} \xi$ if and only if $m_n(t) \to m(t)$ for every t in some interval around the point $t = 0$.*

The actual use of Theorem 10.5.11 will be based, in most cases, on the assumption of the existence of moments and the corresponding Taylor expansion of the mgf's.

Theorem 10.5.12 *Let the random variable X have moment generating function $m(t)$. Assume that $E(|X|^k) < \infty$ for some k, and let $m_j = E(X^j)$ for $j = 0, 1, \ldots, k$. Then*

$$m(t) = \sum_{j=0}^{k} \frac{m_j t^j}{j!} + o(|t|^k), \tag{10.49}$$

where $o(|t|^k)$ is some function such that $o(|t|^k)/|t|^k \to 0$ as $t \to 0$.

Proof. The proof consists of using the well-known theorem from calculus on Taylor expansions. Recalling from Theorem 8.5.3 that

$$m_j = E(X^j) = \left. \frac{d^{(j)} m(t)}{dt^j} \right|_{t=0}.$$

□

The following theorem states another fact known from calculus:

Theorem 10.5.13 *If $\{c_n\}$ is a sequence of numbers such that $\lim_{n\to\infty} c_n = c$, then*

$$\lim_{n\to\infty} \left(1 + \frac{c_n}{n}\right)^n = e^c.$$

As an illustration of applicability of Theorem 10.5.11, we sketch a proof of Theorem 8.8.3 (law of large numbers in the Bernoulli case). We know from (8.23) that the mgf of the binomial random variable S_n is $m_{S_n}(t) = Ee^{tS_n} = (q + pe^t)^n$. Using the fact that $m_{S_n/n}(t) = m_{S_n}(t/n)$, we have

$$m_{S_n/n}(t) = \left[q + p\left(1 + \frac{t}{n} + \frac{t^2}{2n^2} + \cdots\right)\right]^n = \left(1 + \frac{pt}{n} + \frac{pt^2}{2n^2} + \cdots\right)^n \to e^{pt}.$$

The last step consists of using Theorem 10.5.13 with

$$c_n = pt + \frac{pt^2}{2n} + \cdots \to pt.$$

We showed that the moment generating function of S_n/n converges to the moment generating function of random variable equal identically to p for every t (hence also in some neighborhood of $t = 0$). Theorem 10.5.11 allows us to infer that $S_n/n \xrightarrow{d} p$, and by Theorem 10.5.7, we have $S_n/n \xrightarrow{P} p$.

Weak Laws of Large Numbers

We will now prove some of weak laws of large numbers using the Chebyshev inequality (Theorem 8.8.1).

Theorem 10.5.14 *Let $X_1, X_2, \ldots$ be a sequence of iid random variables. Assume that $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 > 0$. Then for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right|\right\} = 0. \tag{10.50}$$

Proof. Based on the Theorem 10.1.1

$$E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu \quad \text{and} \quad \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}.$$

Applying the Chebyshev inequality to variable $(X_1 + \cdots + X_n)/n$, we obtain

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}, \tag{10.51}$$

which converges to 0 as $n \to \infty$. □

This theorem confirms, in some sense, our belief in the "law of averages": formula (10.51) tells us that if we take the averages $(X_1 + \cdots + X_n)/n$ of larger and larger numbers of observations (of some phenomenon, described by random variable X, whose "copies" $X_1, X_2, \ldots$ are being observed), then it becomes less and less likely that the average $(X_1 + \cdots + X_n)/n$ deviates by more than ϵ from the "true average" $\mu = E(X)$.

This assertion of Theorem 10.5.14 constitutes, to a large extent, the basis of the common understanding why "statistics works," that is, why increasing the number of observations pays off in the form of being able to make a better inference about some quantities.

The generality of the assertion of Theorem 10.5.14 is in stark contrast with narrowness of the assumptions specifying that $X_1, X_2, \ldots$ are iid random variables. The importance—practical and cognitive—of assertion (10.50) makes it worthwhile to analyze to which extent that assumption of Theorem 10.5.14 can be relaxed.

The full answer lies beyond the scope of this book. We will, however, analyze the question of to which extent the assumptions can be relaxed under the present proof, based on the Chebyshev inequality.

First, let us observe that we did not fully utilize the assumption that X_i's have the same distribution. In fact we used only a special consequence of this assumption, namely the fact that $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ are the same for all i. Thus we can drop the requirement that X_i's be identically distributed, as long as we retain the stationarity of the mean and variance. To use the example of measurement, we could take measurements of the same quantity μ with different measuring instruments or methods, provided that $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ (such measurements are called unbiased and having the same precision).

But even the requirement that $E(X_i) = \mu$ is not necessary: when $E(X_i) = \mu_i$, relation (10.50) can be replaced by

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \frac{\mu_1 + \cdots + \mu_n}{n}\right| \geq \epsilon\right\} \to 0, \tag{10.52}$$

which is the same as

$$P\left\{\left|\frac{U_1 + \cdots + U_n}{n}\right| \geq \epsilon\right\} \to 0,$$

where $U_i = X_i - \mu_i$ is the deviation of the ith random variable from its own mean.

Next, an inspection of the proof of Theorem 10.5.14 shows that it is not necessary that variances all be equal. What is required is that the variance of the average $(X_1 + \cdots + X_n)/n$ decreases to 0 as n increases. We can therefore formulate the following version of the law of large numbers:

Theorem 10.5.15 *Let $X_1, X_2, \ldots$ be iid random variables, with $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. If*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = 0, \tag{10.53}$$

then relation (10.52) *holds for every $\epsilon > 0$.*

Let us now see to what extent it is possible to relax the assumption of independence. Again, an inspection of the proof of Theorem 10.5.14 shows that the property which was used is the additivity of variance, specifically the fact that $\text{Var}(X_1 + \cdots + X_n) = \sigma_1^2 + \cdots + \sigma_n^2$. But this property holds under the assumption that the random variables are uncorrelated.

Finally, let us observe that even the latter condition can be relaxed: what we really need to make the proof valid is that $(1/n^2)\text{Var}(X_1 + \cdots + X_n) \to 0$. This in turn is implied by the assumption (10.53) in the case where all covariances are zero or negative. We therefore have

Theorem 10.5.16 *Let $X_1, X_2, \ldots$ be random variables with $E(X_i) = \mu_i$ and $\sigma_j^2 = \text{Var}(X_j)$, and such that $\text{Cov}(X_i, X_j) \leq 0$ for $i \neq j$, satisfying (10.53) as $n \to \infty$. Then for every $\epsilon > 0$,*

$$P\left\{\left|\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu_j)\right| \geq \epsilon\right\} \to 0$$

as $n \to \infty$.

This is as far as we will go using the techniques of proof based on the Chebyshev inequality. It is of some interest that in the case of identical distributions one can prove a weak law of large numbers without the assumption of existence of variance.

Theorem 10.5.17 *Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $E(X_i) = \mu$, and such that their common mgf exists in some neighborhood of $t = 0$. Then the relation* (10.50) *holds.*

Proof. From Theorem 10.5.7 it follows that it is enough to show that $(X_1 + \cdots + X_n)/n \xrightarrow{d} \mu$, and therefore (Theorem 10.5.11) that $m_{S_n/n}(t) \to e^{\mu t}$ in some neighborhood of $t = 0$.

Letting $m(t) = Ee^{tX}$, we have $m_{S_n/n} = [m(t/n)]^n$, since X_i's are iid. The existence of $\mu = E(X)$ implies the existence of the derivative m', and the relation $m'(0) = \mu$. Then the Taylor expansion of $m(t)$ is $m(t) = 1 + \mu t + o(|t|)$ (see Theorem 10.5.12) so that

$$m_{S_n/n}(t) = \left(1 + \frac{\mu t}{n} + o\left(\frac{|t|}{n}\right)\right)^n.$$

To complete the proof, it suffices to use Theorem 10.5.13 for $c_n = \mu t + no(|t|/n) \to \mu t$. □

Comment. Observe that the assumption of existence of an mgf is not necessary for the validity of the theorem. The proof of this strengthened version requires nothing more than replacing moment generating functions with characteristic functions. All the steps of the proof remain unchanged.

Strong Laws of Large Numbers

The strong laws of large numbers (SLLNs) are theorems that assert almost sure convergence of sequences of random variables obtained by averaging some underlying sequences of random variables. We will prove here two such theorems, both due to Kolmogorov.

Theorem 10.5.18 *Let $X_1, X_2, \ldots$ be independent, with $EX_i = \mu_i, X_i = \sigma_i^2$. If*

$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty,$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \to 0 \quad a.s. \tag{10.54}$$

Proof. To simplify notation, let $S_n = \sum_{j=1}^{n} X_i$ and $m_n = \sum_{i=1}^{n} \mu_i$. For a fixed $\epsilon > 0$ let

$$C_k = \left\{ \max_{2^{k-1} < n \leq 2^k} \frac{1}{n}|S_n - m_n| \geq \epsilon \right\}.$$

To prove the theorem, it suffices to show that $\sum P(C_k) < \infty$, since then, by the Borel-Cantelli lemma (Theorem 2.6.2), only finitely many events C_k will occur a.s., which implies (10.54).

If C_k occurs, then at least one of the inequalities

$$|S_n - E(S_n)| \geq \epsilon n \geq \epsilon 2^{k-1}, \qquad n = 2^{k-1} + 1, \ldots, 2^k,$$

occurs. By the Kolmogorov inequality (Theorem 8.8.5), we obtain

$$P(C_k) \leq \frac{\mathrm{Var}(S_{2^k})}{\epsilon^2(2^{2k-2})} = \frac{4}{\epsilon^2} \times \frac{\sum_{j=1}^{2^k} \sigma_j^2}{2^{2k}}.$$

We can therefore write

$$\begin{aligned} \sum_{k=1}^{\infty} P(C_k) &\leq \frac{4}{\epsilon^2} \sum_{k=1}^{\infty} \frac{1}{2^{2k}} \sum_{j=1}^{2^k} \sigma_j^2 = \frac{4}{\epsilon^2} \sum_{j=1}^{\infty} \sigma_j^2 \sum_{2^k \geq j} \frac{1}{2^{2k}} \\ &= \frac{16}{3\epsilon^2} \sum_{j=1}^{\infty} \frac{\sigma_j^2}{j^2} < \infty, \end{aligned}$$

since

$$\sum_{2^k \geq j} \frac{1}{2^{2k}} = \sum_{k \geq \log_2 j} 2^{-2k} = \frac{2^{-2\log_2 j}}{1 - \frac{1}{4}}.$$

□

We will now prove another strong law of large numbers, also due to Kolmogorov, that covers the iid case.

Theorem 10.5.19 *Let $X_1, X_2, \ldots$ be iid random variables, and let $S_n = X_1 + \cdots + X_n$. If $\mu = E(X_i)$ exists, then $S_n/n \to \mu$ a.s.*

Proof. Let us truncate the random variables $\{X_n\}$, by letting

$$Y_k = \begin{cases} X_k & \text{if } |X_k| \leq k \\ 0 & \text{if } |X_k| > k. \end{cases}$$

We have

$$\sum_{k=1}^{\infty} P\{Y_k \neq X_k\} = \sum_{k=1}^{\infty} P\{|X_k| > k\} = \sum_{k=1}^{\infty} P\{|X_1| > k\} < \infty$$

by the assumption that $E(X_1) < \infty$ and since X_k has the same distribution as X_1. Thus the inequality $X_k \neq Y_k$ will occur a.s. only a finite number of times, and it suffices to prove that $(1/n)\sum_{k=1}^{n} Y_k \to \mu$ a.s. To this end, we will use Theorem

10.5.19, proved above. We let $\sigma_k^2 = \text{Var}(Y_k) \le E(Y_k^2)$. Therefore we can write

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{\sigma_k^2}{k^2} &\le \sum_{k=1}^{\infty}\frac{E(Y_k^2)}{k^2} = \sum_{k=1}^{\infty}\frac{1}{k^2}\int_{-k}^{k}x^2dF(x)\\ &= \sum_{k=1}^{\infty}\frac{1}{k^2}\sum_{j=1}^{k}\int_{j-1<|x|\le j}x^2dF(x)\\ &= \sum_{j=1}^{\infty}\int_{j-1<|x|\le j}x^2dF(x)\sum_{k=j}^{\infty}\frac{1}{k^2} \le \sum_{j=1}^{\infty}\int_{j-1<|x|\le j}x^2dF(x)\times\frac{C}{j}\\ &\le C\sum_{j=1}^{\infty}\int_{j-1<|x|\le j}|x|dF(x) = C\int_{-\infty}^{\infty}|x|dF(x) < \infty.\end{aligned}$$

We use the estimate $1/j^2+1/(j+1)^2+\cdots < C/j$ valid for some C. It follows that $(1/n)\sum_{j=1}^{n}[Y_j - E(Y_j)] \to 0$ a.s. But $E(Y_j) \to E(X_1) = \mu$ as $j \to \infty$; hence also $(1/n)[E(Y_1)+\cdots+E(Y_n)] \to \mu$. Then we must have

$$\frac{1}{n}\sum_{j=1}^{n}Y_j \to \mu \quad \text{a.s.},$$

which completes the proof. □

PROBLEMS

10.5.1 Let $X_1, \ldots, X_n$ be a random sample from a POI(λ) distribution. Show that

$$e^{-\overline{X}_n} \xrightarrow{P} P(X_1 = 0).$$

10.5.2 Let $Z_1, \ldots, Z_n$ be a random sample from N(0, 1) distribution. Find the limiting distribution of $Y_n = \sum_{i=1}^{n}(Z_i + 1/n)/\sqrt{n}$.

10.5.3 Let $X_1, \ldots, X_n$ be a random sample from distribution with cdf $F(x) = 1 - x^{-2}$ for $1 \le x < \infty$, and 0 otherwise. Find the limiting distribution of: **(i)** $X_{1:n}$. **(ii)** $X_{1:n}^n$. **(iii)** $X_{n:n}/\sqrt{n}$.

10.5.4 Let $X_1, \ldots, X_n$ be a random sample from continuous distribution with a cdf F. Find the limiting distribution of: **(i)** $U_n = nF(X_{1:n})$. **(ii)** $W_n = n[1-F(X_{n:n})]$. **(iii)** $V_n = nF(X_{3:n})$.

10.5.5 Let $X_1, \ldots, X_n$ be a random sample from a logistic distribution with a cdf $F(x) = 1/(1+e^{-x})$, and let $V_n = X_{n:n}$. Then $V_n \xrightarrow{P} \infty$, but $V_n - \log n$ converge to a limiting distribution. Find $\lim_{n\to\infty} P\{V_n - \log n \le 0\}$ and $\lim_{n\to\infty} P\{|V_n - \log n| \le 1\}$.

10.5.6 A random variable has a Pareto distribution with parameters a, b ($a > 0, b > 0$) if its density is

$$f(x; a, b) = \frac{a}{b(1+x/b)^{a+1}}, x > 0.$$

Let $X_1, \ldots, X_n$ be a random sample from the Pareto distribution with density $f(x; 1, 1)$. **(i)** Find the limiting distribution of random variable $U_n = nX_{1:n}$. [*Hint:* Find cdf $F(x)$ of X_i, and then determine cdf of U_n in terms of $F(x)$.] **(ii)** Show that $V_n = X_{n:n}$ does not have a proper limiting distribution; specifically, $\lim_{n\to\infty} P\{V_n \leq t\} = 0$ for every t. **(iii)** Find the limiting distribution of V_n/n.

10.6 CENTRAL LIMIT THEOREM

The term *central limit theorem* (CLT) is a generic name used to designate any theorem that asserts that the sums of large numbers of random variables, after standardization (i.e., subtraction of the mean and division by standard deviation), have approximately a standard normal distribution.

As suggested by the adjective *central*, the search for conditions under which sums of a large number of components have an approximate normal distribution has been (and to a large extent still is) one of the leading research topics in probability theory for the last 200 years or so. We begin with the simplest case of iid sequences.

Theorem 10.6.1 (Lindeberg and Lèvy) *Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, where $0 < \sigma^2 < \infty$. Then, letting $S_n = X_1 + \cdots + X_n$, for every x*

$$\lim_{n\to\infty} P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{10.55}$$

Proof. Observe that $E(S_n) = n\mu$ and $\mathrm{Var}(S_n) = n\sigma^2$, so the left-hand side of (10.55) is simply the limit of the cdf's of the standardized sums

$$S_n^* = \frac{S_n - E(S_n)}{\sqrt{\mathrm{Var}(S_n)}}.$$

The right-hand side is the cdf of a standard normal random variable, denoted $\Phi(x)$. Thus the theorem asserts that the cdf's of S_n^* converge to $\Phi(x)$ for every x, hence at every point of continuity of $\Phi(x)$. Letting Z denote the standard normal random variable, Theorem 10.6.1 asserts that $S_n^* \xrightarrow{d} Z$.

When an the assertion is phrased in this way, it should be clear that a possible strategy of the proof is to use a moment generating function and Theorem 10.5.11. Let $m_X(t)$ be the common mgf of random variables X_i (assumed to exist). The existence of the first two moments suggests applying a Taylor expansion. We have

$$S_n^* = \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma\sqrt{n}}$$

so that

$$m_{S_n^*}(t) = \left[m_{(X-\mu)/\sigma\sqrt{n}}(t)\right]^n = \left[m_{(X-\mu)/\sigma}\left(\frac{t}{\sqrt{n}}\right)\right]^n. \tag{10.56}$$

Now

$$E\left(\frac{X-\mu}{\sigma}\right) = 0, \qquad E\left(\frac{X-\mu}{\sigma}\right)^2 = 1 \tag{10.57}$$

so that

$$m_{(X-\mu)/\sigma}(t) = 1 + \frac{t^2}{2} + o(t^2).$$

Consequently, using (10.56), we obtain

$$m_{S_n^*} = \left[1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n = \left[1 + \frac{t^2/2 + no(t^2/n)}{n}\right]^n.$$

By Theorem 10.5.13, for $c_n = \frac{t^2}{2} + no(\frac{t^2}{n}) \to \frac{t^2}{2}$ we obtain

$$m_{S_n^*}(t) \to e^{t^2/2},$$

which completes the proof in the special case where the underlying random variables X_i have mgf's. For the general case, see the comment following the proof of Theorem 10.5.17. □

■ **EXAMPLE 10.12**

Suppose that you buy a supply of 20 household batteries to be used for some specific purpose, one after another. Assume that the lifetime of each such battery is a random variable with a mean of 2 weeks and a standard deviation of 3 days. The batteries are replaced as needed, and the batteries that are unused do not keep on aging. What is the (approximate) probability that the supply of batteries will last more than 9 but less than 10 months? (i.e., more than 270 and fewer than 300 days?).

SOLUTION. The question here is about the probability that $S_{20} = X_1 + \cdots + X_{20}$ satisfies the inequality $270 < S_{20} < 300$. We have $n\mu = 20 \times 14 = 280$ and $\sigma\sqrt{n} = 3 \times \sqrt{20} = 13.4$. Thus $270 < S_{20} < 300$ occurs if and only if $(270 - 280)/13.4 < S_{20}^* < (300 - 280)/13.4$, (i.e., if $-0.75 < S_{20}^* < 1.49$).

For $Z \sim \mathrm{N}(0, 1)$ we have

$$\begin{aligned} P\{270 < S_{20} < 300\} &\approx P\{-0.75 < Z < 1.49\} = \Phi(1.49) - \Phi(-0.75) \\ &= 0.9319 - 0.2266 = 0.7053. \end{aligned}$$

Theorem 10.6.1 is a generalization of the oldest central limit theorem, due to Laplace, covering the binomial distribution.

Theorem 10.6.2 (Laplace) *If S_n has a binomial distribution* BIN(n, p), *then for any $z_1 < z_2$,*

$$\lim_{n\to\infty} \sum_{A_n \le j \le B_n} P\{S_n = j\} = \lim_{n\to\infty} P\{z_1 \le S_n^* \le z_2\} \tag{10.58}$$

$$= \Phi(z_2) - \Phi(z_1), \tag{10.59}$$

where $A_n = np + z_1\sqrt{npq}$, $B_n = np + z_2\sqrt{npq}$.

Proof. If S_n is the number of successes in n independent trials, each with probability of success p, then $S_n = X_1 + \cdots + X_n$, where $P\{X_i = 1\} = 1 - P\{X_i = 0\} = p$, $i = 1, 2, \ldots, n$. Consequently, $E(S_n) = np$ and $\text{Var}(S_n) = npq$ so that

$$S_n^* = \frac{S_n - np}{\sqrt{npq}}. \qquad \square$$

Suppose now that we want to find $P\{a \leq S_n \leq b\}$. According to Theorem 10.6.2, for large n we have

$$\begin{aligned} P\{a \leq S_n \leq b\} &= P\left\{\frac{a - np}{\sqrt{npq}} \leq S_n^* \leq \frac{b - np}{\sqrt{npq}}\right\} \\ &\approx \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np}{\sqrt{npq}}\right). \end{aligned} \tag{10.60}$$

However, approximation (10.60) can be improved somewhat if we observe that in the present case S_n is an integer-valued random variable, and for *integers* a and b the exact expression is

$$P\{a \leq S_n \leq b\} = \sum_{j=a}^{b} P\{S_n = j\}. \tag{10.61}$$

Each term on the right-hand side of (10.61) can be approximated by an area under the normal curve between $(j - 0.5 - np)/\sqrt{npq}$ and $(j + 0.5 - np)/\sqrt{npq}$.

Adding such approximations, the terms for neighboring j cancel, and we obtain the following formula:

$$P\{a \leq S_n \leq b\} \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{npq}}\right) \tag{10.62}$$

for any integers $a \leq b$.

■ **EXAMPLE 10.13**

A fair coin is tossed $n = 15$ times. Find the approximate probability that the number S_{15} of heads will satisfy the inequality $8 \leq S_{15} < 10$.

SOLUTION. Observe first that the inequality $8 \leq S_{15} < 10$ is the same as $8 \leq S_{15} \leq 9$. Since $np = 15 \times 0.5 = 7.5$, $\sqrt{npq} = \sqrt{15 \times 0.5 \times 0.5} = 1.94$, the approximation (10.58) gives

$$\begin{aligned} P\{8 \leq S_{15} \leq 9\} &\approx \Phi\left(\frac{9 + 0.5 - 7.5}{1.94}\right) - \Phi\left(\frac{8 - 0.5 - 7.5}{1.94}\right) \\ &= \Phi(1.03) - \Phi(0) = 0.8485 - 0.5 = 0.3485. \end{aligned}$$

The exact value is

$$\begin{aligned} \{S_{15} = 8\} + P\{S_{15} = 9\} &= \binom{15}{8}\left(\frac{1}{2}\right)^{15} + \binom{15}{9}\left(\frac{1}{2}\right)^{15} \\ &= 0.1964 + 0.1527 = 0.3491. \end{aligned}$$

Generally, the quality of approximation improves as n increases, and—for the same n—decreases as p moves away from $1/2$ in either direction. Also, observe that use of the *continuity correction*—addition and subtraction of $1/2$ from the limits a and b in formula (10.58)—makes sense only if $0.5/\sqrt{npq}$ exceeds the difference between the consecutive arguments in the tables of normal distribution that are actually being used.

■ EXAMPLE 10.14 Decision Problem

Suppose that we design a theater with 1000 seats. The theater has two entrances, A and B, situated with respect to a parking lot, public transportation, and so on, so that the patrons have equal chances of choosing any of the entrances.

Suppose also that our theater is to be located in a city where the climate calls for patrons to wear overcoats, which they can leave in a coatroom. There are to be two coatrooms, each located near one of the entrances, and while it is not impossible to enter through one entrance and leave the coat in a coatroom near to the other entrance, it is inconvenient to do so. How many coat hangers should each coatroom have?

SOLUTION. Clearly, the problem is not precise enough as stated: We have to specify the criterion which we want to attain. One of the objectives is minimization of the cost of equipping the coatroom in hangers, racks, and so on. We do not want to staff a coatroom that will remain empty. On the other hand, we do not want to inconvenience patrons by making them go to a distant coatroom. The two extremes, each satisfying one of the foregoing objectives, is to equip each coatroom with 1000 coat hangers, and to equip each with exactly 500 coat hangers.

A possible objective may be: We want to equip each coatroom with $500 + x$ hangers, where x is the smallest number such that (say) on 95% of nights when the theater is sold out, everyone will be able to leave his or her coat at the coatroom nearest to the entrance used.

To solve the problem, we have to make some assumptions about independence of choice of entrances A and B by the patrons. As the first approximation, assume that the patrons arrive one at a time and each chooses the entrance independent of other patrons. Let S_{1000} be the number of patrons (among $n = 1000$) who choose entrance A. We want to have $S_{1000} \leq 500+x$ (then everyone who enters through A is not inconvenienced), and also $1000 - S_{1000} \leq 500 + x$ (which is the analogous condition for those who choose entrance B). Thus we would like the event $500 - x \leq S_{1000} \leq 500 + x$ to occur with probability 0.95 or more. Since $p = 0.5$, $np = 1000 \times 0.5 = 500$, and $\sqrt{npq} = \sqrt{1000 \times 0.5 \times 0.5} = 15.8$, we have

$$\begin{aligned} P\{500 - x \leq S_{1000} \leq 500 + x\} &= P\left\{\frac{-x - 0.5}{15.8} \leq S^*_{1000} \leq \frac{x + 0.5}{15.8}\right\} \\ &\approx \Phi\left(\frac{x + 0.5}{15.8}\right) - \Phi\left(-\frac{x + 0.5}{15.8}\right) \geq 0.95. \end{aligned}$$

An inspection of Table A2. shows that $\Phi(1.96) - \Phi(-1.96) = 0.95$. We must therefore take x as the smallest integer for which $(x+0.5)/15.8 \geq 1.96$, which gives $x = 31$. Thus, to achieve our objective, we should install 532 coat hangers in each coatroom.

■ **EXAMPLE 10.15**

Continuing Example 10.14, a more realistic assumption is that people attend the theater in pairs, and both members of a pair come through the same entrance. We have now $n = 500$ pairs, and letting S_{500} denote the number of pairs who choose entrance A, we must have $2S_{500} \leq 500 + x$ and $1000 - 2S_{500} \leq 500 + x$. Now $E(S_{500}) = 500 \times 0.5 = 250$ and $\text{Var}(S_{500}) = 500 \times 0.5 \times 0.5 = 125$. The objective therefore becomes $250 - x/2 \leq S_{500} \leq 250 + x/2$. Using formula (10.58), we get

$$P\left\{250 - \frac{x}{2} \leq S_{500} \leq 250 + \frac{x}{2}\right\} = P\left\{\frac{-\frac{x}{2} - 0.5}{\sqrt{125}} \leq S^*_{500} \leq \frac{\frac{x}{2} + 0.5}{\sqrt{125}}\right\}$$
$$= \Phi\left(\frac{x+1}{22.3}\right) - \Phi\left(-\frac{x+1}{22.3}\right).$$

Again, x is the smallest integer for which $(x+1)/22.3 \geq 1.96$, so $x = 43$. We see that grouping (in this case into pairs, but the effect is the same for other groupings) a set of persons, with groups choosing the entrance independently, increases the variability: we now need to supply 543 hangers in each coatroom to meet the requirement.

The central limit theorem proved thus far concerns the rather narrow case of independent and identically distributed components. In this case the sum has asymptotically normal distribution, provided only that variance is finite. This theorem is often utilized to explain the frequent appearance of a normal distribution in nature. The argument typically goes along the lines of attribute, such as the height of a person, or an error of measurement of some quantity, (such as the speed of light). Whatever the case, the observed value is influenced by a large number of factors, some having a negative and some having a positive effect. Some such factors are known, but their effect cannot be predicted exactly, whereas other factors may not even be named. What matters is that all these factors operate largely independently of one another and each in isolation is small as compared with the total effect of all factors (i.e., factors that are known to have large effects are treated differently and are not included in these considerations). The central limit theorem therefore asserts that the total effect of such "small" factors is random and has approximately a normal distribution. We already know this to be true in case of factors that are iid. However, independence and identical distribution can hardly be justified in every real situation, and can at best be regarded as approximations to reality. Consequently, one can expect central limit theorems to be valid in wider classes of situations where the iid assumption does not hold. We state below a number of theorems that provide conditions for asymptotic normality in case of independent random variables that are not identically distributed.

Theorem 10.6.3 (Liapunov) *Let $X_1, X_2, \ldots$ be a sequence of random variables such that $E(X_i) = \mu_i$, $\mathrm{Var}(X_i) = \sigma_i^2$, and $\gamma_i = E|X_i - \mu_i|^3$. Moreover, put*

$$m_n = \sum_{j=1}^{n} \mu_j, \quad s_n^2 = \sum_{j=1}^{n} \sigma_j^2, \quad \Gamma_n = \sum_{j=1}^{n} \gamma_j,$$

and let $S_n = X_1 + X_2 + \cdots + X_n$ be the corresponding sequence of partial sums. If additionally $X_1, X_2, \ldots$ have finite third moments, then the condition

$$\lim_{n \to \infty} \frac{\Gamma_n}{s_n^3} = 0 \tag{10.63}$$

is sufficient for convergence

$$\frac{S_n - m_n}{s_n} \xrightarrow{d} N(0,1).$$

The proof of this theorem lies beyond the scope of this book.

As an illustration of the application of the Liapunov theorem, consider a sequence of independent trials, such that in nth trial the probability of success is p_n. We let S_n denote the number of successes in the n first trials so that $S_n = X_1 + \cdots + X_n$, where $X_i = 1$ or 0 depending on whether or not the ith trial leads to a success. We then have $E(X_i) = p_i$, $\mathrm{Var}(X_i) = p_i q_i$, while the third absolute central moment γ_i is

$$\gamma_i = E|X_i - p_i|^3 = |1 - p_i|^3 P(X_i = 1) + |0 - p_i|^3 P(X_i = 0) = q_i^3 p_i + p_i^3 q_i.$$

Thus $\gamma_i = p_i q_i (p_i^2 + q_i^2) \leq p_i q_i$, and

$$\frac{\Gamma_n}{s_n^3} \leq \frac{\sum_{i=1}^{n} p_i q_i}{(\sum_{i=1}^{n} p_i q_i)^{3/2}} = \frac{1}{(\sum_{i=1}^{n} p_i q_i)^{1/2}}.$$

Consequently, the Liapunov condition (10.63) holds if $\sum_{i=1}^{n} p_i q_i = \infty$, and we have

Theorem 10.6.4 *Consider a sequence of independent trials, with the probability of success in the ith trial being p_i. If S_n is the number of successes in the n first trials, then*

$$\frac{S_n - \sum_{i=1}^{n} p_i}{\sqrt{\sum_{i=1}^{n} p_i q_i}} \xrightarrow{d} N(0,1)$$

if $\sum_{i=1}^{\infty} p_i q_i = \infty$.

We close this chapter by stating a theorem that completed the long search for conditions implying limiting normality in case of independent random variables (the cases of dependent random variables are still the object of intense research).

Theorem 10.6.5 (Lindeberg and Feller) *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with finite second moments. Assume that $s_n^2 \to \infty$ and $\max_{1 \leq j \leq n} \sigma_j^2 / s_n^2 \to 0$ as $n \to \infty$. Then*

$$\frac{S_n - m_n}{s_n} \xrightarrow{d} N(0,1)$$

if and only if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|x-\mu_j|\geq \epsilon s_n} (x-\mu_j)^2 dF_j(x) = 0, \tag{10.64}$$

where F_j *is the cdf of* X_j.

The "if" part was proved by Lindeberg, so (10.64) is called the Lindeberg condition. The "only if" part is due to Feller.

We will now introduce a generalization of a central limit theorem that is very useful when only the first two moments, and not the entire probability distribution, are known. The theorem provides an approximate distribution of the transformed random variables obtained in random samples.

Theorem 10.6.6 (Delta Method) *Let* $X_1, X_2, \ldots$ *be a sequence of random variables such that*

$$\sqrt{n}(X_n - \theta) \to N(0, \sigma^2), \tag{10.65}$$

and let g *be a function with a nonzero derivative* $g'(\theta)$. *Then*

$$\sqrt{n}[g(X_n) - g(\theta)] \to N(0, \sigma^2[g'(\theta)]^2). \tag{10.66}$$

Proof. The Taylor expansion of $g(X_n)$ around $X_n = 0$ gives

$$g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + o_P(|X_n - \theta|), \tag{10.67}$$

where o_P was defined in Theorem 10.5.12. Rearranging (10.67) and multiplying by $\sqrt{n}$, we obtain

$$\sqrt{n}(g(X_n) - g(\theta)) = g'(\theta)\sqrt{n}(X_n - \theta) + o_P(\sqrt{n}|X_n - \theta|).$$

Based on (10.65), convergence of X_n to θ in probability, and Theorem 10.5.5, we obtain (10.66). □

Let us now illustrate the Delta method by an example.

■ **EXAMPLE 10.16**

Let X_n be a relative frequency of a success in n Bernoulli trials. Find the asymptotic distribution of $g(X_n) = 1/X_n$.

SOLUTION. Let p be the probability of success in a single trial. Then $g'(p) = -1/p^2$. Consequently, based on Theorems 10.6.5, 10.5.6, and the fact that

$$\sqrt{n}(X_n - p) \to \mathrm{N}(0, p(1-p)),$$

we immediately obtain

$$\sqrt{n}\left(\frac{1}{X_n} - \frac{1}{p}\right) \to \mathrm{N}\left(0, \frac{(1-p)}{p^3}\right).$$

Last we would like to mention two possible extensions of the Delta method. If $g'(\theta) = 0$, then one can take one more term in the Taylor expansion (10.67). Such approach will yield a second-order Delta method (for details see e.g., Casella and Berger, 2002). Also, the Delta method can be easily extended to a multivariate setting (see Oehlert, 1992).

PROBLEMS

10.6.1 Let $X_1, \ldots, X_n$ be a random sample from the distribution with density $f(x) = xe^{-x}, x > 0$. Find c if it is known that $P\{\overline{X}_n > c\} = 0.75$ for $n = 250$.

10.6.2 Assume that 500 students at a certain college will graduate on a given day. Because of space limitations the college offers each student two tickets for the commencement ceremony. From past experience it is known that 50% of the students will invite two guests to attend the ceremony, 20% students will invite one guest, and the remaining 30% will not attend at all, so they will invite no guests. How many chairs should the college order to have at least 95% chance that all attending guests will have seats?

10.6.3 Passengers on an international flight have a luggage weight limit B. The actual weight W of the passenger's luggage is such that W/B has a BETA(a, b) distribution where $a/(a+b) = 0.9$. Assume that the weights of luggage of different passengers are independent and that the plane has 220 seats. Find a and b if it is known that when the plane is fully booked, then there is a 5% chance that the total weight of baggage will exceed $200B$.

10.6.4 Let $X_1, \ldots, X_n$ be a random sample from the BETA(2, 3) distribution. Let $S_n = X_1 + \cdots + X_n$. Find the smallest n for which $P\{S_n \geq 0.75n\} \leq 0.01$.

10.6.5 A regular dodecahedron (12-sided Platonian solid) has six red and six white faces, with the faces of each color labeled $1, \ldots, 6$. If you toss a face with label k you pay or win $\$k$, depending on whether the color is red or white. Find the probability that after 50 tosses you are ahead by more than \$10.

10.6.6 Let $X_1, \ldots, X_{360}$ represent the outcomes of 360 tosses of a fair die. Let S_{360} be the total score $X_1 + \cdots + X_{360}$, and for $j = 1, \ldots, 6$, let Y_j be the total number of tosses that give outcome j. Use a normal approximation to obtain: **(i)** $P(55 < Y_3 < 62)$. **(ii)** $P(1200 < S_{360} < 1300)$. **(iii)** $P(1200 < S_{360} < 1300 | Y_1 = 55)$. **(iv)** $P(1200 < S_{360} < 1300 | Y_4 = 55)$. **(v)** $P(1200 < S_{360} < 1300 | X_1 = X_2 = \cdots = X_{55} = 4)$.

10.6.7 A fair coin is tossed $2n$ times. How large must n be if it is known that the probability of the equal number of heads and tails is less than 0.1?

10.6.8 Referring to Example 8.40, assume that a man's shoe has an average length of 1 foot and $\sigma = 0.1$ foot. Find the (approximate) probability that the mean of 16 lengths of men's shoes exceed 1 foot by more than 1 inch.

10.6.9 A die is unbalanced in such a way that the probability of tossing k ($k = 1, \ldots, 6$) is proportional to k. You pay \$4 for a toss, and win \$$k$ if you toss k. Find the approximate probability that you are ahead after 100 tosses.

10.6.10 Let $\overline{X_n}$ be a sample mean in a random sample of size n from POI(λ). Use the Delta method to find the limiting distribution of $g(\overline{X_n}) = \sqrt{n}(\overline{X_n} - \lambda)$.

10.6.11 Let X_n be a relative frequency of success in n Bernoulli trials. Use the Delta method to find the limiting distribution of $g(X_n) = X_n(1 - X_n)$.

CHAPTER 11

INTRODUCTION TO STATISTICAL INFERENCE

11.1 OVERVIEW

The role of this chapter is to provide an introduction to statistical inference covered in the remaining chapters. In the first part of the book, we introduced and developed techniques that lead to predicting the form of future observations (data). Given certain general information (e.g., about independence of some events), we could deduce the distributions of observed random variables. In practice, we answered questions such as: How many future observations will fall into a certain set? What will be the average of those observations?

As opposed to that, in statistics (more precisely, in inferential statistics), which will be now our main object of study, the question is: Given the data, what can we say about specific aspects of the stochastic mechanisms that govern the occurrence of those data? The actual data are regarded as a result of a random process, in the sense that if the data collection were to be repeated, the outcome would most likely be different, possibly even leading to a different conclusion. Consequently, whatever inference we make from the actual data, it is subject to error. This error—the central concept of statistics—is not meant to be a "mistake" of any kind (i.e., something that can be avoided).

At first one might think that this randomness will be eliminated if we increase the precision of measurement. Actually, the opposite is true. For instance, if we

measure the length of a table in integer number of feet, the result will be the same under repetition. When we increase the precision, to an inch, then to half of an inch, and so on, the variability of the results under repetition will become more and more pronounced.

The question then arises: What inference about the underlying phenomenon can be drawn from premises (data) which may differ from occasion to occasion? Viewed in this way, statistics is a part of theory of inductive inference. But this does not mean that the *theory* of statistics is itself inductive. As a theory, inferential statistics is a fragment of mathematics, in the same way as probability theory. Thus mathematical statistics has its own structure of specific concepts (motivated mostly by applications), and its own theorems.

Each theorem in mathematics asserts that some conclusions hold, provided that certain assumptions are satisfied, and theorems in statistics are no exception in this respect. Some of the assumptions refer to the process of data collection, or the properties of underlying random variables. The applicability of statistical methods (i.e., the empirical validity of the consequences of the appropriate theorems) depends crucially on the degree to which the assumptions are met in real situations. Sometimes this degree of validity is under the control of the experimenter (see Example 11.1); in some other cases, possibly after performing appropriate tests, we simply may have no conclusive evidence that the assumptions are violated; in particular situations (see Example 11.2) we may feel justified in disregarding the fact that the assumptions are false.

■ **EXAMPLE 11.1**

Suppose that we want to apply a method of statistical inference, for which we need an assumption that a certain random variable X is binomial. Often X represents the number of elements of some kind in the sample (number of defective items, number of patients who recovered after specific treatment, etc.). The assumption that the process of collecting observations is really a sequence of Bernoulli trials depends on various factors, some of which the experimenter can control. Of these, the principal factor is the independence in selecting the sample. In case of defective items, it may require sampling with replacement; in case of patients, it may require checking that the sample does not have identical twins or other persons whose reaction to the treatment in question may be similar because of some genetic reasons.

■ **EXAMPLE 11.2**

Imagine that we observe freshmen scores X on an aptitude test. Assuming that X has a normal distribution leads to sufficiently good approximations of the relevant probabilities. The fact remains, however, that test scores cannot be normally distributed, since X must be an integer, and it also cannot assume negative values.

In cases such as above, we typically feel justified in using the consequences of theorems that rely on the assumption that X has a normal distribution, even if we

know that this assumption is not satisfied. But there may be some "more serious" violations of the assumptions. Checking the validity of an assumption may involve using some elaborate statistical techniques. Often, however, it is sufficient to just have a glance at some preliminary graphical presentation of the data, or at the values of some crude statistics. In either case we deal with a summary or reduction of the data.

The methods of such initial reduction of the data belong properly to what is called *descriptive statistics*. Although the main object of the book is inferential (rather than descriptive) statistics, we present some basic ideas of descriptive statistics in the next section.

11.2 DESCRIPTIVE STATISTICS

A simple summary presentation of the data can lead to the discovery of surprising and important consequences. Descriptive statistics, available in most statistical packages, help one to better understand the data structure and therefore to avoid making incorrect assumptions, consequently choosing the appropriate method of data analysis. In our overview of descriptive statistics we do not attempt to be complete, since the field grows as more and more methods are being developed.

Let us begin with two examples, both concerning World War II.

■ EXAMPLE 11.3

The main route for supplying the Allied armies fighting Nazi Germany in Europe during World War II was through the Atlantic Ocean. The convoys were attacked regularly, mostly by German U-boats. As the war progressed, more and more data accumulated. It turned out that the average number of ships lost in an attack was relatively constant; in particular, it did not depend on the size of the convoy. This observation led to a simple conclusion: *to decrease losses, make convoys as big as possible*. Indeed, two separate convoys might expect to be detected and attacked about twice as many times as a convoy obtained by combining them. Since the average losses per attack are independent of the convoy size, such joining of two convoys cuts losses by half. This simple idea contributed substantially to winning the war.

■ EXAMPLE 11.4

Bombers were sent on missions over Germany. On route to and from, as well as over the target, they were subject to antiaircraft fire. The direct hits were not very frequent, but the AA shells were set to explode at specific altitudes, spraying the planes with shrapnel. The planes that returned from the mission were examined for locations of shrapnel hits, and all these locations were recorded on a silhouette of the plane. As more and more data became available, the silhouette was more densely covered with recorded locations of hits. There were, however, some areas that were hit less often than others.

A surprising order was then given: strengthen (by putting armor plates, etc.) those areas that were hit *seldom*. Here the argument is that the locations of

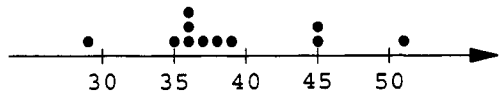


Figure 11.1 Dot diagram

shrapnel hits on bombers must have a uniform distribution over the silhouette of the plane. Any "white areas" in the data therefore indicate the locations of hits for the planes that did not return from missions.

We present these two examples not only because of their combination of simplicity of premises and unexpectedness of conclusion but also because they are both based on elementary ways of representing the data: plotting average losses per attack against convoy size, or making a scatter diagram of shrapnel hits. Certainly it is not often that one gets a chance of contributing to victory in war by a visual inspection of some descriptive statistics. Nevertheless, it is worth knowing some simple "tricks of the trade" in presenting the data so as to exhibit certain aspects of interest, or in making certain patterns visible. The ones mentioned in this section concern univariate data.

Dot Diagram

First, if the number of data points is small, one can often get good insight into the structure of the data by drawing a *dot diagram*. This is accomplished by marking the data values as dots on the horizontal axis, with repeated data represented by dots piled up on one another. For instance, the data points

$$28, 36, 37, 52, 36, 45, 39, 45, 38, 35, 36$$

would be represented as in the diagram of Figure 11.1. If the data are to be grouped into classes, with class boundaries and class counts replacing individual data values, then the dot diagram is a good device to help to choose class boundaries.

Stem-and-Leaf Display

This is another method of quick presentation of the data. Again, it is best explained by an example. Suppose that the data are

9.5	10.8	8.8	11.2	10.2	10.3	10.2	11.3	10.0	8.8
10.7	9.9	11.4	9.8	10.5	9.8	9.9	10.9	8.1	10.5
10.6	8.2	8.6	9.2	9.9	10.0	11.0	9.2	10.7	10.9.

We then choose "stems" and data values are presented as follows:

Stem	Leaves
8	8 8 1 2 6
9	5 9 8 8 9 2
10	8 2 3 2 0 7 5 9 5 6 0 7 9
11	2 3 4 0

Both the dot diagram and stem-and-leaf methods involve no reduction of information contained in the data. Most other methods involve partial reduction of the information; this concerns grouping and representing the data by various kinds of indices.

Box Plot

This is a very simple but informative representation of data sets (see Figure 11.2). The box plot is based on five values, and it provides a graphical presentation that

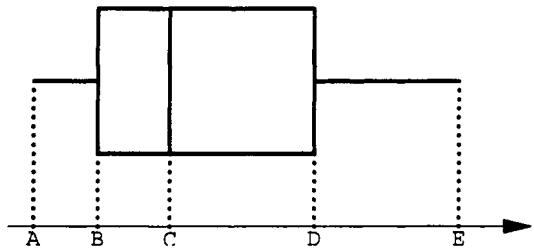


Figure 11.2 Box plot

allows us not only to visualize them, but also to gather information about skewness and outliers.

Thus, A and E are the smallest and largest data, respectively, so the difference $E - A$ is the range of the data. Next B and D are the lower and upper quartiles of the data, so the box (whose height is irrelevant) contains the central 50% of data points. Finally, C is the median of the data.

The median, as well as upper and lower quartiles, are the sample counterparts of the corresponding quantiles of the distribution. Let $y_1, \ldots, y_n$ be the data points, and let $y_{1:n} \leq y_{2:n} \leq \cdots \leq y_{n:n}$ be the ordered sample. There are several definitions of the sample pth quantile. Some texts define it as $y_{[np]+1:n}$, where $[x]$ stands for the largest integer not exceeding x. Another definition of the pth sample quantile is

$$y_{[(n+1)p]:n} + \{(n+1)p - [(n+1)p]\}(y_{[(n+1)p]+1:n} - y_{[(n+1)p]:n}). \qquad (11.1)$$

The method based on formula (11.1) uses the most obvious linear approximation. Rather than contemplating this formula, consider a simple example with $n = 4$ data points, which (arranged in increasing order) are

$$y_{1:4} = 5, \quad y_{2:4} = 8, \quad y_{3:4} = 15, \quad y_{4:4} = 20. \qquad (11.2)$$

Suppose that we want to compute the 37% sample quantile for the data. The four data points partition the range into five classes, determined by quantiles for 20%,

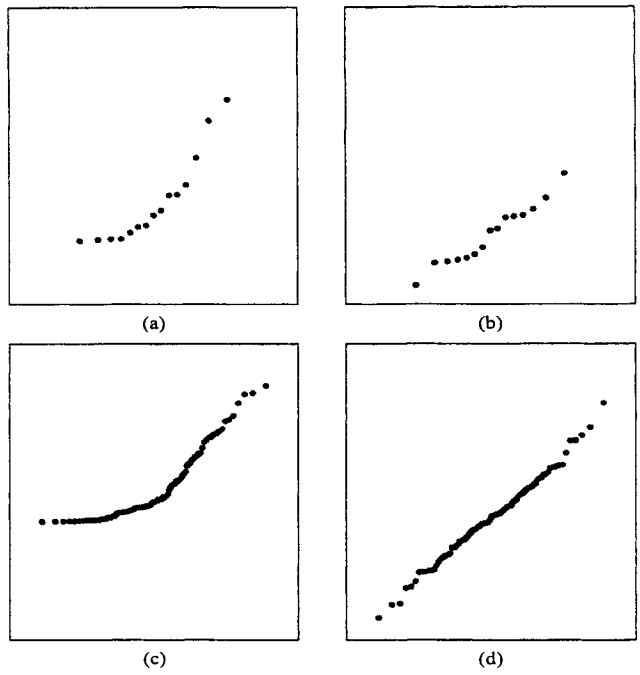


Figure 11.3 Normal quantile plots

40%, 60%, and 80%. The 37% quantile lies between the 20% and 40% quantile, hence between $y_{1:4} = 5$ and $y_{2:4} = 8$, at a point whose distance from 5 is $(37 - 20)/(40 - 20) = 0.85$ of the distance between 5 and 8. Thus the 37% quantile equals $5 + 0.85(8 - 5) = 7.55$.

Observe that formula (11.1) does not apply if p is close to either 0 or 1. The reason is that if $p < 1/(n+1)$, then $[(n+1)p] = 0$ and $y_{0:n}$ is not defined, whereas if $p > n/(n+1)$, then $y_{[(n+1)p]+1:n}$ does not exist.

For $p = 1/2$, formula (11.1) gives the sample median as

$$y_{[(n+1)/2]:n} + \left\{ \frac{n+1}{2} - \left[\frac{n+1}{2} \right] \right\} \times \left(y_{[(n+1)/2]+1:n} - y_{[(n+1)/2]:n} \right).$$

The formula for the median when n is odd simplifies to $y_{(n+1)/2:n}$, since in that case $[(n+1)/2] = (n+1)/2$. If n is even, we have $[(n+1)/2] = n/2$ and consequently the median is equal to $(y_{(n/2):n} + y_{(n/2)+1:n})/2$.

Sample quantiles are also used in a *normal quantile plot*, also known as *Q-Q plot*—a graphical method for checking the assumption that the data come from a normal population. The graph is obtained by plotting quantiles of standard normal distributions against corresponding (with the same p) sample quantiles obtained by formula (11.1). If the sample comes from a normal population, the Q-Q plot should form a straight line. Figure 11.3(a) and (b) shows Q-Q plots for data from exponential and normal distributions for small samples. As can easily be seen, random

fluctuations obscure the picture, so it is hard to decide which plot, if any, forms a straight line. Figure 11.3 (c) and (d) was obtained for large sample sizes ($n = 300$). Now the situation is obvious. A large sample represents the population adequately, and looking at the graphs one can readily conclude which data come from the normal distribution and which do not. The final conclusion about normality of the data is based on a subjective judgment resulting from a visual inspection. In Chapters 13, 15, and 16 we will show other methods of testing normality.

Grouping

Large sets of data can be unmanageable in their original form, so typically the data points are grouped into classes. One then gives the class boundaries (or equivalently, class midpoints and class widths) and the class counts (also called class frequencies). The only formal requirement here is that the classes be disjoint and cover all (actual, and also potentially possible) data values, so each data point falls into exactly one class.

Histograms

Graphical representation of the grouped data usually has the form of a histogram. These are formed as follows: The horizontal axis is divided into intervals centered at class midpoints and extending in either direction by one-half of the class width. Then rectangles are built on these intervals in such a way that the *areas of rectangles (not heights) are proportional to the class frequencies.* This distinction is vital in the case of unequal class widths.

What then is the "proper" number of classes? Should all classes be of the same width? How does grouping affect the values of summary indices, such as the mean?

Class Width and Number of Classes

There is no unique answer as to the proper class width (or number of classes). When the data are divided into too many classes (assume equal class widths), the histogram looks jagged. When there are too few classes, one loses too many details in the data. The appropriate number of classes depends not only on sample size but also on the shape of the distribution. One can formulate the criteria in different ways. For instance, Scott (1979) formulates the problem of the choice of class width as a problem of optimization of certain criteria and shows that the optimal width is of the order of $3.49\hat{\sigma}/\sqrt[3]{N}$, where N is the number of data points and $\hat{\sigma}$ is the estimate of a standard deviation. Thus for $N = 100$ data points with the shape of the histogram close to normal, the range[18] R is about $4\hat{\sigma}$, and the best class width is about $3.49(R/4)/\sqrt[3]{100} \approx R/5$, so that the histogram should have 5 classes. For $N = 1,000$, the range R is about $6\hat{\sigma}$, and the best class width is about $R/16$ (so 16 classes).

[18]The distribution of the range $R = X_{N:N} - X_{1:N}$ is given by (10.30). The exact formula for $E(R)$ is not available in the case of sampling from normal distribution. Crude approximations are $E(R) \approx 4\sigma$ for $N = 100$ and $E(R) \approx 6\sigma$ for $N = 1,000$. For exact values one needs to inspect the appropriate statistical tables.

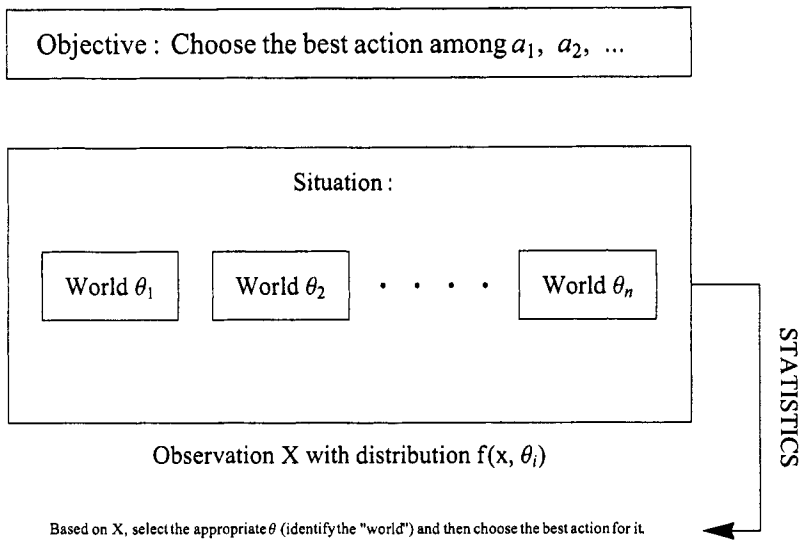


Figure 11.4 Decision scheme

Effect on Indices

Indices such as the mean are calculated for grouped data by replacing each data value within a class with the class midpoint. Consequently, the values of the same index, calculated from the raw data and from the grouped data, differ. Under the assumption of equal widths of classes, it is typically a relatively simple task to estimate the upper bound for an error of a given index due to grouping.

For instance, in case of the mean, the bound is as follows: Each of the N data points is off from its class midpoint by at most $L/2$, where L is the class width. If all these deviations are the largest possible and in the same direction, the difference between the exact and approximate mean is $N(L/2)/N = L/2$. Thus the error due to grouping is at most $L/2$.

11.3 BASIC MODEL

A typical user of statistical methods is someone who must choose an action in a situation of partial uncertainty with regard to the factors that affect the consequences of this action. The uncertainty may be alleviated to some extent by the fact that our decision maker can "spy on nature" by taking some observations that may help identify the relevant factors. The formal representation of this rather general description is actually the cornerstone of most of the theory of mathematical statistics. Figure 11.4 presents a general scheme of making decisions in a situation of uncertainty. Statistical inference constitutes only a part of this scheme.

First, with the exception of some special cases, the nature of the action to be selected is of no concern for statistics and will not be included in the formal structure

of the theory. Now let us imagine ourselves as decision makers. Using the terminology accepted in general decision theory, we say that we are in one of the situations labeled as "world θ_1," ..., "world θ_n," but we do not know which one. It is usually assumed that there is one action appropriate for each "world" (although this assumption is not really necessary). Thus, if we knew the "world" we are in (i.e., knew the value θ_j), we would know which action is the best to take.

The only way to identify the actual "world θ_j" is to perform some experiments. Let $\mathbf{X}$ be a generic symbol for the results of such experiments. It may be a single observation of some variable, a sequence of observations, a result of some complicated physical experiment, a score on a battery of tests, and so on, depending on the context. Naturally, to make the entire setup meaningful, there must be some relation between $\mathbf{X}$ and θ_j. This is expressed by the assumption that the observation $\mathbf{X}$, while being random, comes from a distribution that depends on θ_j. The randomness here is the crucial part of the entire setup. It may be due to sampling variability or be inherent in the phenomenon, or both.

It may happen, of course, that observation $\mathbf{X}$ will allow us to identify the value θ_j without ambiguity. This occurs when the sets of possible values of $\mathbf{X}$ are disjoint for different θ_j's. Such cases, however, are rare in real life. Most often we face a challenging case when the same outcome $\mathbf{X}$ is possible under several (or even all) θ_j but occurs with different probabilities, depending on θ_j.

■ **EXAMPLE 11.5**

The examples of situations falling under this scheme abound in science, engineering, management, and in everyday life. A doctor faces a patient who has a headache and high fever. Particular "worlds" are possible illnesses of the patient. The doctor orders some tests, and on the basis of their results $\mathbf{X}$ makes a diagnosis (identifies the illness, perhaps incorrectly) and chooses the best action in view of the illness diagnosed. As another example, take a pharmaceutical company that has developed a drug against a specific disease, hopefully superior to drugs used thus far. The "worlds" may be numbers θ describing the relative advantage of the new drug versus the best drug available so far, with $\theta > 1$ indicating that the new drug is better, and $\theta \leq 1$ indicating that it is no better (or even inferior) to the drug used thus far. Here the observation $\mathbf{X}$ is a series of studies, specified in great detail by the FDA standards.

Example 11.5 concerns medicine, but it is obvious that any new method is subject to the same scheme of tests before it is established whether or not (and to what extent) it is superior to some other method.

As seen in Figure 11.4, statistical theory is concerned with the problem of inference about θ on the basis of $\mathbf{X}$. In the case where there is a choice of the variable that we may observe, the statistical theory is also concerned with choosing the "best" variable (these are called problems of design of experiment).

11.4 BAYESIAN STATISTICS

An inspection of Figure 11.4 reveals that one level is missing, representing the possibility that the user may have some knowledge, previous experience, or perhaps other reasons (including prejudice or superstition) that make some "worlds" appear more likely to him than others, before any observations $\mathbf{X}$ are taken.

There is a general agreement among statisticians that in cases where the "world" is chosen randomly from the set of possible worlds, with probabilities having frequential interpretation, those probabilities should be taken into account in the process of deciding about θ_j on the basis of observation $\mathbf{x}$ (one can then simply compute the conditional probability of θ_j given $\mathbf{X} = \mathbf{x}$). For example, in the case of a physician seeing a patient with a headache and high fever, the possible "worlds" (in this case illnesses causing this particular set of symptoms) include initial stages of flu as well as the initial stages of the plague (Bubonic plaque). The incidence of the latter disease is so rare, however, that the doctor may feel perfectly justified in not ordering any test to check the possibility of the disease being the plaque. Here the doctor relies simply on his own and his colleagues' experience about the incidence of various diseases that might start with fever and headache at a given time and geographic location.

Statisticians differ in their opinion of whether the probabilities of various "worlds" reflecting the researcher's experience, intuition, "hunches," and so on, should be used in statistical methodologies. Those who allow such probabilities to be used are called *Bayesians*, and the resulting statistical methodology is called *Bayesian statistics*.

A rather strong argument for the Bayesian approach is provided by the theory of outliers (see Section 10.3). In a non-Bayesian setup, where $\theta = (\alpha, \lambda)$ is the parameter in the family of all gamma distributions, no configuration of data values can be rejected as containing an outlier, even if (say) $n - 1$ observations fall into interval $(0, 1)$ and the nth observation exceeds 10^6. This is because there exists a $\theta = (\alpha, \lambda)$, for which such a configuration of data points is very likely to occur. This counterintuitive example suggests that a statistician should eliminate some domains of the parameter space as "unlikely," using whatever information or experience he or she has. We will provide Bayesian solutions to various problems under consideration. This means that we will show how the problem is, or may be, solved if the prior probabilities of various θ_j's are available. A systematic presentation of the theory of Bayesian statistics lies, however, beyond the scope of this book.

11.5 SAMPLING

In the theory of statistics it is typically assumed that the data are observations of some random variable. The applicability of statistical methods depends therefore on how well the assumption of randomness and the assumption about the distribution are satisfied. In this section we will show some of the possible "traps" one may encounter in implementing statistical methods in practice.

Let us start with an example.

EXAMPLE 11.6

Suppose that the objective is to estimate an unknown parameter θ that is the average of some attribute in a population. It may be, for instance, the average yearly income of a family in a given region, the average age of a patient receiving certain treatment, and so on. In such cases one typically takes a sample of elements from the population and measures the values of the attribute being studied. Thus $\mathbf{X} = (X_1, \ldots, X_n)$, where $E(X_i) = \theta$ for $i = 1, \ldots, n$.

The following story illustrates some of the potential problems that one may encounter.

Three social science students, Jim, Joe, and Susan, were each assigned a task of estimating the average size of a class (number of students) in a given school district. Jim decided to make a card for each class in each school, shuffle the cards, sample one or more of them and then find the number of children in each class sampled.

Joe found a somewhat simpler scheme: He decided to prepare cards with names of schools and first sample a school (or several schools). Then for each school chosen, he decided to make cards with labels of classes and take a sample of those cards, at the end determining the numbers of children in each class sampled.

Susan applied a still simpler scheme: She decided to take a sample of children from the school district and ask each child about the size of the class that he or she attends. The question is: Which of the three students, if any, measured the parameter "average size of the class in a given school district"?

Since an increase of sample size affects only the precision of the estimator, not the parameter that is being estimated, we will consider only the cases where Jim, Joe, and Susan each take a single observation.

Suppose that there are k schools in the district in question, with the ith school having n_i classes, of sizes $C_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i$. Then the total number of classes is $N = \sum_{i=1}^{k} n_i$, and the average class size is

$$\theta = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} C_{ij}.$$

The objective is to estimate θ. If X, Y, and Z denote the random variables observed, respectively, by Jim, Joe, and Susan, then it is clear that $X = C_{ij}$ with probability $1/N$ (sampling is from the set of all classes). Thus $E(X) = \theta$.

Regarding random variable Y, we have $Y = C_{ij}$ if Joe selects ith school (probability $1/k$) and jth class in the ith school (probability $1/n_i$). Consequently

$$E(Y) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{1}{k n_i} C_{ij} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} C_{ij}.$$

We have $E(Y) \neq \theta$ except in a special case when all n_i's are equal, meaning each school has the same number of classes.

Finally, for the random variable Z, observed by Susan, the situation is as follows: Let $C = \sum_{i=1}^{k} \sum_{j=1}^{n_i} C_{ij}$ be the total number of children in all

classes. With probability C_{ij}/C a child from the jth class of the ith school will be selected, and then the value of Z will be C_{ij}. We have therefore

$$E(Z) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} C_{ij}P\{Z = C_{ij}\} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} C_{ij}\frac{C_{ij}}{C} = \frac{1}{C}\sum_{i=1}^{k}\sum_{j=1}^{n_i} C_{ij}^2.$$

Again, $E(Z) \neq \theta$ unless all classes are of the same size.

Thus it is only Jim whose method provides an estimate of the parameter θ. For random variables suggested by Joe and Susan, we generally have $E(Y) \neq \theta$ and $E(Z) \neq \theta$. The method suggested by Joe is known as *stratified sampling*: The population is divided into *strata*, and one first samples strata, and then takes a sample from each stratum (in this case the role of strata is played by schools).

It is worth mentioning here that if Joe decided to take a sample of schools, and then collect the data about *all* class sizes in selected schools, he would use what is known as *cluster sampling*.

With some prior information available (e.g., about relative sizes of the strata), one can easily adjust Joe's estimator (by taking appropriate weighted averages) to build random variable Y with $E(Y) = \theta$ (called an *unbiased* estimator of θ).

The situation with the Susan's method is not so straightforward, and cannot be easily remedied. It is related to *importance sampling*, where the probability of choosing an element with a larger value of the attribute is higher than the probability of the element with a smaller value.

The bias due to the phenomenon of importance sampling occurs quite often and evades notice. The following example (suggested by R. F. Green, personal communication) provides some surprising insight into the issue.

■ EXAMPLE 11.7 Siblings

Suppose that in a certain society, the distribution of the number of children in a family is Poisson with mean 4. What is the average number of siblings of a child in this society?

SOLUTION. An almost automatic response of most persons (including some statisticians) is 3. Actually, the answer is 4 in the special case of the Poisson distribution. In general, it is *more* than 3, except when all families have exactly 4 children. The situation is very much the same as with Susan's sampling Example 11.6.

Before proceeding to the solution, we should note that the distribution given in the problem concerns the population of *families*, but the question concerns the average in the population of *children*. Let $p_0, p_1, \ldots$ be the distribution of the number of children in the family, so that in the special case under analysis we have $p_k = (\lambda^k/k!)e^{-\lambda}$ for $\lambda = 4$. Then $\mu = p_1 + 2p_2 + \cdots$ and $\sigma^2 = \sum k^2 p_k - \mu^2$ are the mean and variance of the distribution $\{p_k\}$. Suppose that the population consists of a large number N of families. Clearly, Np_0 families

have no children, Np_1 families have one child, and so on. The total size of the population of children is therefore

$$M = Np_1 + 2Np_2 + 3Np_3 + \cdots = N\mu.$$

The probability of choosing a child from a family with k children is

$$\frac{kNp_k}{N\mu} = \frac{kp_k}{\mu}$$

for $k = 1, 2, \ldots$ and such a child has $k - 1$ siblings. Then the average number of siblings is

$$\begin{aligned} Q &= \sum_{k=1}^{\infty}(k-1)\frac{kp_k}{\mu} = \frac{1}{\mu}\left\{\sum_{k=1}^{\infty}k^2p_k - \sum_{k=1}^{\infty}kp_k\right\} \\ &= \frac{1}{\mu}\{\sigma^2 + \mu^2 - \mu\} = \frac{\sigma^2}{\mu} + \mu - 1. \end{aligned}$$

As can be seen, we have $Q \geq \mu - 1$ with $Q = \mu - 1$ in the case where $\sigma^2 = 0$ (i.e., when all families have the same number μ of children). For the Poisson distribution $\sigma^2 = \mu = \lambda$, so we have $Q = \mu = 4$.

A sampling bias closely related to the bias from importance sampling is connected with the following phenomenon, which caused some controversy before it became properly understood. To explain it, we will again use an anecdotal example.

■ **EXAMPLE 11.8 Renewal Paradox**

A statistical objective is to estimate the average lifetime of electric bulbs of a particular type, all produced by the same company. We assume that the distribution of the lifetime T is exponential with mean $E(T) = 1/\lambda$.

The usual procedure is to take a random sample of bulbs and observe their lifetimes $T_1, T_2, \ldots, T_N$. Such data can be used to estimate $E(T)$. If testing is run in parallel, it takes time $T_{N:N} = \max\{T_1, \ldots, T_N\}$ to collect the data. One could speed up the procedure by observing only the k shortest lifetimes $T_{1:N} < T_{2:N} < \cdots < T_{k:N}$ and then interrupt the data collection, recording only that $T_{k+1:N} > T^*$, where T^* is some threshold (so that $N - k$ lifetimes are not fully observed, but they are all known to exceed T^*). This is called *censoring of the data*.

In some cases another way can be used to collect data, that does not involve waiting. Suppose that there is a large building (e.g., an office skyscraper downtown) where the maintenance personnel use only bulbs of the type of interest for us. Whenever a bulb fails, it is immediately replaced by a new bulb, and this change is recorded. As a consequence we have access to the records, reaching into the past, of the time of replacement of every bulb. We might then select some time t^* (preferably in the past) and use all lifetimes of the bulbs that were operating at time t^* as the sample. Clearly, if t^* is sufficiently far back in the past, each of the bulbs that operated at t^* has already been

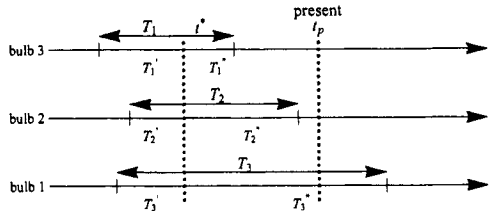


Figure 11.5 Sampling interreplacement times

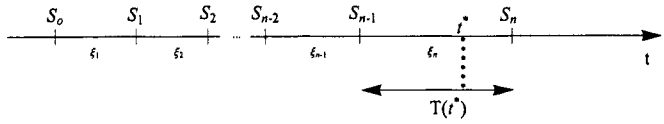


Figure 11.6 Renewal process

replaced, and its lifetime $T(t^*)$ is known. Otherwise, we would have to wait for some time to observe the value of $T(t^*)$ for that bulb. The situation is best explained in Figure 11.5.

In general, T_i' is the *spent* lifetime at t^* for the ith bulb, and similarly T_i'' is the *residual* lifetime at t^*. It is clear that if $t_p - t^*$ is large enough, all residual times will be observable. For bulbs 1 and 2, the lifetimes are $T_1(t^*) = T_1' + T_1''$ and $T_2(t^*) = T_2' + T_2''$. The value T_3 is not observable at present but is at least $T_3' + (t_p - t^*)$. The sample $T_1, T_2, \ldots, T_N$ of all N bulbs that were operating at t^* can be observed and the average $(1/N)\sum_{i=1}^{N} T_i(t^*)$ can serve as an estimator $E(T)$.

The obvious question is: Is such a method of collecting data correct? It should be stressed that we are *not* concerned here with the practical implementation of the scheme. We assume that the time t^* is chosen without the knowledge of the replacement records (a condition preventing conscious or unconscious bias) and disregard the fact that in reality light bulbs are used only part of the time, and that at some locations lights are used more often or are subject to different conditions (outdoor and indoor lights, etc.). Since the problem concerns theory, we deliberately idealize the situation and assume that the data concern only bulbs that operate constantly and under the same conditions. This way the interreplacement times along each time axis in Figure 11.5 are sampled independently from the same exponential distribution.

To find the answer, consider a single process of changing of bulbs in one place (see Figure 11.6). The process starts at time $t = 0$; the consecutive lifetimes are $\xi_1, \xi_2, \ldots$, and the times of replacements are $S_0 = 0, S_n = S_{n-1} + \xi_n$ for $n \geq 1$. The lifetime recorded in the sample, $T(t^*)$, is the value

$\xi_n = S_n - S_{n-1}$ such that $S_{n-1} \le t^* < S_n$. We assume that $\xi_1, \xi_2, \ldots$ are iid with exponential distribution.

The distribution of the time S_n of the nth replacement was derived in Section 9.3. It is a gamma distribution with parameters n and λ so that the density of S_n is

$$g_n(t) = \frac{\lambda^n}{\Gamma(n)} t^{n-1} e^{-\lambda t} = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}. \tag{11.3}$$

For further use, observe that we have

$$\sum_{n=1}^{\infty} g_n(t) = \lambda \sum_{n=1}^{\infty} \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t} = \lambda. \tag{11.4}$$

To find $E[T(t^*)]$, we must first find $F_{T(t^*)}(x) = P\{T(t^*) \le x\}$, the cdf of $T(t^*)$. Assume that $x < t^*$. Then a replacement would have occurred before time t^*, since otherwise the original bulb would still be working at t^*, and we would have $T(t^*) = \xi_1 > t^* > x$.

Thus, for some $n = 2, 3, \ldots$ we must have $S_{n-1} = z \le t^* < S_n = S_{n-1} + \xi_n$. In this case $\xi_n > t^* - z$ so that $t^* - z < \xi_n \le x$. The condition $\xi_n \le x$ implies $t^* - z \le x$, which means that the time z of the $(n-1)$st replacement satisfies $t^* - x \le z \le t^*$. Partitioning with respect to $n = 2, 3, \ldots$, conditioning on time z (time of the $(n-1)$st replacement), and using (11.4), we obtain

$$\begin{aligned}
F_{T(t^*)}(x) &= P\{T(t^*) \le x\} = \sum_{n=2}^{\infty} \int_{t^*-x}^{t^*} P\{t^* - z < \xi_n \le x\} g_{n-1}(z)dz \\
&= \sum_{n=2}^{\infty} \int_{t^*-x}^{t^*} [e^{-\lambda(t^*-z)} - e^{-\lambda x}] g_{n-1}(z)dz \\
&= \int_{t^*-x}^{t^*} [e^{-\lambda(t^*-z)} - e^{-\lambda x}] \sum_{n=2}^{\infty} g_{n-1}(z)dz \\
&= \lambda \int_{t^*-x}^{t^*} [e^{-\lambda(t^*-z)} - e^{-\lambda x}]dz = 1 - e^{-\lambda x} - \lambda x e^{-\lambda x}.
\end{aligned}$$

If $x > t^*$, the derivation above has to be modified in two ways. First, we have to add the probability that the first bulb is still working at t^*, and that its lifetime ξ_1 satisfies $t^* < \xi_1 \le x$ (see Figure 11.7). Second, if a replacement occurred before t^*, the time $S_{n-1} = z$ of last replacement before t^* satisfies the inequality $0 < z \le t^*$. Consequently we have

$$\begin{aligned}
P\{T(t^*) \le x\} &= e^{-\lambda t^*} - e^{-\lambda x} + \sum_{n=2}^{\infty} \int_0^{t^*} [e^{-\lambda(t^*-z)} - e^{-\lambda x}] g_{n-1}(z)dz \\
&= 1 - e^{-\lambda x} - \lambda t^* e^{-\lambda x}.
\end{aligned}$$

Differentiating, we obtain the density of $T(t^*)$:

$$f_{T(t^*)}(x) = \begin{cases} \lambda^2 x e^{-\lambda x} & \text{for } x < t^* \\ \lambda(1 + \lambda t^*) e^{-\lambda x} & \text{for } x \ge t^*. \end{cases} \tag{11.5}$$

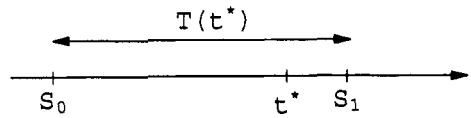


Figure 11.7

An easy integration gives the formula for $E[T(t^*)]$, and we can show that

$$\lim_{t^* \to \infty} E[T(t^*)] = \frac{2}{\lambda}.$$

Therefore, on average, the lifetime of the bulb that is in operation at time t^* is about *twice as long* as the average lifetime of other bulbs. So the described method of sampling is biased.

The conclusion above about sampling procedure being biased remains valid as long as the lifetimes $\xi_1, \xi_2, \ldots$ are iid. The specific assumption of exponentiality of distribution implies the form (11.5) of the density, but the fact that $\lim_{t^* \to \infty} E[T(t^*)] > E(\xi)$ is valid for any nondegenerate distribution of ξ_i's (e.g., see Karlin and Taylor, 1975).

Since $T(t^*) = T' + T''$, meaning $T(t^*)$ is the spent lifetime at t^* plus the residual lifetime at t^*, we can try to derive our result as follows: The residual time T'' is exponential, in view of the memoryless property of exponential distribution, so $E(T'') = 1/\lambda$. For the spent lifetime T', it cannot exceed t^*, and it exceeds x (where $x < t^*$) if there are no replacements between $t^* - x$ and t^*. Since replacements form a Poisson process, the latter probability is $e^{-\lambda(t^* - x)}$. Consequently

$$P\{T' > x\} = \begin{cases} e^{-\lambda(t^* - x)} & \text{for } x < t^* \\ 0 & \text{for } x \geq t^*. \end{cases}$$

Thus, by Theorem, 8.2.2

$$E(T') = \int_0^\infty P\{T' > x\} dx = \int_0^{t^*} e^{-\lambda(t^* - x)} dx = \frac{1}{\lambda}(1 - e^{-\lambda t^*}),$$

and therefore $E[T(t^*)] = E(T') + E(T'') = 2/\lambda - (1/\lambda)e^{-\lambda t^*}$. This calculation, however, relies on the assumption that the memoryless property of the exponential distribution is valid for residual waiting time counted from a *randomly selected* moment (and not only from a fixed moment).

For example, suppose that we observe the waiting time for nearest replacement from the moment *preceding* a given replacement by some fixed constant, say 5 hours. Such a moment is random, and yet the waiting time is at most 5 hours by definition, and therefore does not have exponential distribution. This

simple example shows that the notion of a "random moment" has to be qualified: We are not allowed to know the future at the time we select our moment to start the observations.

11.6 MEASUREMENT SCALES

To explain the motivation for the topics of the last three chapters, it is necessary to introduce some concepts related to the level of measurements. Consider a set of objects of some kind. Typically, for the purpose of analysis, statistical description, and so on, each of these objects can be represented by a number $b(x)$ assigned to object x. Often that number represents the result of measurement in some units, but sometimes it identifies only the class to which the object belongs.

If $b(x) > b(y)$, then x has "more ..." than y (is heavier, longer, warmer, older, etc.). The question we want to address is which type of statements expressed through the values of b are meaningful and which are not. To take an example, if b represents length, and $b(x) = 10, b(y) = 5$, we say that x is *twice as long* as y, since $b(x)/b(y) = 2$. This statement will remain valid whether we express the length in inches, centimeters, or miles: the values $b(x)$ and $b(y)$ will change, but their ratio will remain 2. However, if b represents temperature and $b(x) = 10, b(y) = 5$, it makes no sense to say that x is "twice as warm" as y. Indeed, it is enough to express temperature on a different scale (e.g., change from Fahrenheit to Celsius): The ratio of temperatures will change as the scale changes. The question therefore is: Why are the ratios of scale values meaningful for length, duration, or weight, but not meaningful[19] for temperature?

The full impact of such questions became apparent only when physical, mathematical, and statistical methods started to be used in the social sciences. Here the attributes that one considers are typically "soft," and the use of certain methods can lead to conclusions that are illegitimate, but have the deceptive appearance of being very precise (the statement that a new brand of instant coffee "tastes 11.3% better" may be effective in commercials, mainly because of a deceptive use of a decimal point, but its meaning, if any, is obscure).

It is clear that the decision which statements expressed through the values $b(x)$ are allowed and which are not must lie in the analysis of the nature of a measured attribute. Such an analysis has the following general form: The starting point is an *empirical relational system*, consisting of a set, A, of objects under consideration and a number of relations on A, say $R_1, R_2, \ldots, R_k$. These relations represent some empirically observable relationships between the objects in A: pairs, triplets, and so on.

The measurement is an assignment of numbers to objects in A (hence it is a function on A). This function must satisfy conditions that "mimic" the empirical relations $R_1, R_2, \ldots$. If such a function exists, we say that measurement exists. The

[19]Observe that the precision of measurements has nothing to do with the answer. Ratios of lengths make sense even if the lengths are determined imprecisely; the ratios then are simply subject to bigger errors. But the ratios of the temperatures make no sense even if the temperature is measured with the most precise devices available.

level of freedom of choice of such a function (if it exists) determines the type of the measurement scale.

We will now illustrate the description above by examples.

■ **EXAMPLE 11.9 Scale of Hardness**

Let the objects under consideration be minerals, and let the empirical relation $\succ$ be defined as $x \succ y$ if mineral x scratches mineral y. We will assign numbers $h(x)$ to objects x in A (minerals) to represent their hardness, in the sense that $h(x) > h(y)$ whenever $x \succ y$. One such assignment gives the value 10 to diamonds and lower values to other minerals. This choice of assignment of numbers to minerals is arbitrary, except that the relation $\succ$ between values of function h must mimic the relation $\succ$ between arguments of h (i.e., minerals). It is precisely this degree of arbitrariness of h that makes it meaningless to claim that "x is twice as hard as y" if $h(x)/h(y) = 2$.

But even in such a simple case as above, the possibility of assigning numerical scale values $h(x)$ to minerals x in A results from the fact that the relation $\succ$ (of scratching) is transitive: if $x \succ y$ and $y \succ z$, then $x \succ z$. Still the existence of a function h that mimics the relation $\succ$ is not obvious. Imagine that infinitely many new minerals, each of different hardness, are suddenly discovered. Could one still find enough distinct numbers to label those minerals? The answer is positive, but we will not provide the proof here.

■ **EXAMPLE 11.10**

Relations that order pairs do not have to be transitive. Consider the set A consisting of versions of the state budget, with the relation of preference $\succ$ among them defined as $x \succ y$ if majority (of some voting body) prefers x to y. We can then have a disturbing possibility that $x \succ y, y \succ z$ but $z \succ x$. To see this, take the simplest case, where there are three voters a, b, and c, such that a prefers x to y to z, b prefers z to x to y, while c prefers y to z to x. Then a and b (hence a majority) prefer x to y; a and c prefer y to z, while b and c prefer z to x. We have therefore $x \succ y \succ z \succ x$ and there is no numerical assignment of scale values to x, y, and z that mimics the relation $\succ$.

■ **EXAMPLE 11.11**

Consider now an attribute such as length and a possible measurement system for it. We have here the relation of comparisons with respect to length, accomplished empirically by putting the objects parallel to each other, with one end lined up (like pencils). We then obtain a relation, $\succeq$, interpreted as "not shorter than." The ultimate goal is to show the existence of a function b (length) defined on A, such that $x \succeq y$ if, and only if, $b(x) \geq b(y)$. Clearly, we must require that $\succeq$ satisfies some conditions (axioms), such as transitivity: *if* $x \succeq y$ *and* $y \succeq z$, *then* $x \succeq z$, and so on.

However, one relation $\succeq$ is not sufficient to define length as we know it, since there may be many functions b satisfying the requirement that $\succeq$ agrees with the order induced by the values of b. Clearly, something more is needed to force all functions b to differ one from another only by the unit of measurement (i.e., such that if b and b^* are two assignments of lengths, then $b^* = \alpha b$ for some $\alpha > 0$).

In this case we need a ternary relation describing the operation of concatenation, that is, putting the objects end to end. Letting $\circ$ denote such an operation, we obtain new objects, such as $x \circ y$, corresponding to x and y aligned one after another. This operation can be identified with a relation that holds between x, y, and z, if $x \circ y \sim z$, where $\sim$ is defined by $a \sim b$ if $a \succeq b$ and $b \succeq a$. Clearly, we want $b(x \circ y) = b(x) + b(y)$, and to achieve that, the relations $\succeq$ and $\circ$ jointly must satisfy a number of conditions, such as the most obvious ones: *if* $x \succeq y$, *then* $x \circ z \succeq y \circ z$ *for every* z, and $(x \circ y) \circ z \sim (z \circ x) \circ y$, for all x, y, z. Less obvious is the requirement of the Archimedean property:

For every x and y there exists n such that $x^{\circ n} \succeq y$,

where $x^{\circ n} = x \circ \cdots \circ x$ (n times).

A measurement theory of length is then a relational system consisting of set A, relations $\succeq$ and $\circ$, a set of conditions for $\succeq$ and $\circ$ (referred to as *axioms*), and a theorem asserting that if these axioms are satisfied, then:

1. *There exists a real-valued strictly positive function b on A such that*

$$b(x \circ y) = b(x) + b(y) \text{ and } x \succeq y \text{ if and only if } b(x) \geq b(y).$$

2. *If b^* is any other function satisfying condition 1, then $b^* = \alpha b$ for some $\alpha > 0$.*

The first part of the assertion provides the existence of measurement b, and the second part provides information about its uniqueness. In the case of length, the measurement is unique up to the choice of unit of measurement, so that the ratios of lengths are invariant. This means that for any objects x, y the ratio $b(x)/b(y)$ does not depend on the choice of function b.

To "purists" every attribute that one wants to measure (e.g., represent numerically) ought to be analyzed in this way (for a discussion of the topic, see Valleman and Wilkinson, 1993). The systems vary in two respects: (1) the nature of relations in the relational systems and the axioms that they must satisfy depend on what is being observed and how, and (2) the uniqueness part of the conclusion about measurement scale varies depending on what is being measured.

Each scale has its own level of "uniqueness," described by a condition corresponding to condition (2) in Example 11.11. Although theoretically there may be infinitely many types of scales, in practical situations one encounters only four major types of scales, as specified in Definition 11.6.1.

Consider a relational system, where b and b^* are any two assignments of numbers to objects in the set A—measurement scales specified by the axioms of the relational system.

Definition 11.6.1

(i) If there exists $\alpha > 0$ such that $b^*(x) = \alpha b(x)$, then the measurement is on the *ratio* scale.

(ii) If there exists $\alpha > 0$ and β such that $b^*(x) = \alpha b(x) + \beta$, then the measurement is on the *interval* scale.

(iii) If there exists a monotone increasing function u such that $b^*(x) = u[b(x)]$, then the measurement is on the *ordinal* scale

(iv) If there exists a one-to-one function v such that $b^*(x) = v[b(x)]$, then the measurement is on the *nominal* scale.

In other words, for ratio scales (see Example 11.11), the measurement is unique up to a choice of unit (examples are length, duration, etc.). For interval scales, one can choose not only the unit but also the zero of the scale (e.g., temperature). For ordinal scales only the order matters. For instance, if $b(x) = 10$ and $b(y) = 2$, then all that we can say is that x is "more ..." than y, on an attribute designated by .... We cannot meaningfully say that the "difference" between x and y is 8, or that x is "five times ..." as y. The reason is that $b(x) = 10$ and $b(y) = 2$ can be replaced by $b^*(x) = 100, b^*(y) = -3$, or $b^*(x) = 1, b^*(y) = 0.99$, or any other two values, as long as the first exceeds the second.

In physical measurement, the best such scale known is that of hardness (see Example 11.9) when the main empirical relation is $x \succeq y$ if "x scratches y." In the social sciences, the situation is not so clear. The relations between stimuli elicited by asking subjects to evaluate them "on the scale of 1 to 10" or by marking responses such as "strongly agree," "agree," and so on, are of the ordinal nature but often are treated as if they are expressed on an interval or ratio scale.

The same concerns education, where grades are averaged and compared as if they were measured on an interval scale. Thus a student who takes six classes and gets one A and five B's as the semester grades has a GPA of 3.16. Another student, who also takes six courses, gets four A's, and two C's, has a GPA that is higher (3.33). If the scoring (function b defined on grades $A, B, C, etc.$) were changed to $b(A) = 4, b(B) = 3, b(C) = 1$, that student's GPA would be 3.00 and his grades would be judged as worse than these of the first student.

There is nothing unique or objective in assigning the values 4, 3, and 2 to grades A, B, and C, considering that the process of averaging grades usually includes different subjects, grades by different teachers, and with criteria often formulated rather vaguely. Nevertheless, tradition and the practical need to assess students' performance force one to fix the scoring system for grades, and regard them as a measurement on an interval scale.

Finally, the nominal scale—the weakest of the four—can hardly be regarded as measurement, since the numbers serve only for the purpose of identification (e.g., numbers on jerseys of football players).

The four scales mentioned above form an order: any transformation $b^*(x) = \alpha b(x), \alpha > 0$ allowed for the ratio scale is a particular case of a transformation $b^*(x) = \alpha b(x) + \beta, \alpha > 0$ that defines the interval scale. This transformation is

monotone, hence allowed for the ordinal scale, and in turn, each monotone transformation is one to one, allowed for the nominal scale. One can always lower the level of measurement, while to use a higher measurement level is incorrect.

The theory of measurement scales, as outlined above, was introduced by Stevens (1946). It rapidly gained popularity in psychology and the social sciences and was developed into a highly sophisticated theory (e.g., see Krantz et al., 1971, or Roberts, 1979). The four types of scales, as defined above, are most commonly encountered and popularly known. Unfortunately, this knowledge also contributed to the popular belief that *all* possible scale types (as defined by classes of allowed transformations) are ordered from "stronger" to "weaker." In fact the order is only partial, with some scales being noncomparable. The noncomparable scales correspond to classes of transformations that are not contained one in another.

The types of statistics that one can sensibly use depend on the type of the scale. While the mean μ and standard deviation σ are defined in both cases, one should not use the coefficient of variation σ/μ in the case of data values measured on the interval (but not ratio) scale. This is because under transformation $y = \alpha x + \beta$ ($\alpha > 0$) the standard deviation becomes multiplied by α, while the mean becomes multiplied by α and shifted by β. Therefore the value σ/μ depends on the choice of zero of the scale.

On the other hand, for data measured on the ordinal scale only, the mean and standard deviation are not invariant, and only statistics expressed through ranks (e.g., median and other quantiles) should be used. Finally, for data on the nominal scale, one is allowed to use only class frequencies.

In most of what follows (Chapters 12–14), it is assumed that the data represent measurements on the interval scale. This allows us to use such characteristics as the mean and standard deviation, and to assume that the data follow a normal distribution.[20] Chapter 15 concerns methods of handling the data measured on ordinal scales only (more precisely, only when ordinal relations are taken into account). Finally, in Chapter 16 we discuss the case of *categorical* data, expressed on the nominal and/or ordinal scale with data grouped into classes.

[20] Strictly speaking, if X is measured on a ratio scale, it cannot have normal distribution (since negative values are ruled out for ratio scales). Such level of adherence to the rules, imposed by scale types, would drastically impoverish the scope of statistical applications by absurdly disallowing to treat attributes such as height, weight, size, duration, etc., in certain populations as normally distributed, merely because these attributes cannot be represented by a negative number.

This Page Intentionally Left Blank

CHAPTER 12

ESTIMATION

12.1 INTRODUCTION

Estimation is the process of extracting information about the value of a certain population parameter θ from the data. Consecutive observations (data points) $x_1, x_2, \ldots$ are selected at random from the population being studied and are considered values of some random variables $X_1, X_2, \ldots$. So if the same experiment of taking n observations were to be repeated, the new data points would most likely be different. The statistical laws that govern the variability of data under repetition (actual or hypothetical) can be used to build the theory of statistical inference. In estimation theory it is usually assumed that the distribution of each observation X_i is known, except for the value of some parameter θ. An estimator of θ is then a rule that allows us to calculate an approximation of θ, based on sample $X_1, \ldots, X_n$.

■ **EXAMPLE 12.1**

Let $X_1, X_2, \ldots$ represent the weights of successive trout caught at a certain location. We are interested in a parameter such as $\theta = P\{X_i > w_0\}$, where w_0 is some fixed weight, so that θ is the fraction of trout whose weight exceeds a given threshold w_0.

Given the observed weights $X_1, \ldots, X_n$, we want to estimate θ as

$$T_n = T(X_1, \ldots, X_n) = \frac{\text{number of } i \text{ such that } X_i \geq w_0}{n}.$$

The formula above gives an obvious, and rather unsophisticated estimator of θ as a relative frequency.

There is another method of estimating the same parameter. Assume that the weights of trout follow a normal distribution $\mathrm{N}(\mu, \sigma^2)$. Given the sample $X_1, \ldots, X_n$, we then calculate the sample mean $\overline{X} = (1/n)(X_1 + \cdots + X_n)$ and sample variance $S^2 = [1/(n-1)] \sum (X_i - \overline{X})^2$. Both $\overline{X}$ and S^2 can be regarded as approximations of μ and σ^2, respectively. Consequently, we estimate θ as

$$U_n = U(X_1, \ldots, X_n) = 1 - \Phi\left(\frac{w_0 - \overline{X}}{S}\right), \tag{12.1}$$

where Φ is the cdf of a standard normal variable. Here the rationale is that

$$\theta = P\{X \geq w_0\} = P\left\{Z \geq \frac{w_0 - \mu}{\sigma}\right\} = 1 - \Phi\left(\frac{w_0 - \mu}{\sigma}\right),$$

which is approximated by (12.1).

The example above shows that there can be several estimators of the same parameter (i.e., several distinct rules of calculating an approximation of θ, given the sample). Since every estimator, such as T_n and U_n in Example 12.1, is a random variable, the obvious questions are:

1. How to assess the performance of estimators and to choose the best one?

2. Are there methods for obtaining estimators other than "ad hoc" methods, used to obtain estimators T_n and U_n in Example 12.1?

A systematic attempt to answer these two questions resulted in a theory with a clear conceptual structure, supported by powerful theorems. The empirical situation is such that we observe values of independent random variables $X_1, X_2, \ldots$, sampled from the same distribution $f(x, \theta)$, where f is a density or a probability function, depending on whether the X_i's are continuous or discrete. This distribution depends on a parameter that assumes some value θ (unknown to the observer) from a parameter space Θ. The examples abound: the X_i's can be Bernoulli observations for an unknown probability of success θ, or normally distributed observations with an unknown mean θ and a known standard deviation, or with a known mean and an unknown standard deviation θ, and so on.

To answer question 1: How to judge estimators? Let $T_n = T_n(X_1, \ldots, X_n)$ be the estimator for a random sample of size n, namely some function of observations $X_1, \ldots, X_n$, selected to approximate θ. The first requirement for a "good" estimator is its consistency, defined by the requirement that $T_n \to \theta$. Since T_n's are random variables, we must specify the type of convergence (e.g., in probability or almost surely). The need of consistency is obvious: estimators that are not consistent do

not guarantee that one gets closer to the true value of the parameter by increasing the sample size.

If $E(T_n) = \theta$ for every n, then estimator T_n is called unbiased. This means that if the estimation were to be repeated several times for different samples (but with the use of the same estimator), the results would be, on average, "on target." The "quality" of an unbiased estimator may then be defined as its variance. This corresponds to taking $(t_n - \theta)^2$ as the "penalty" or "loss" due to the error of accepting the value t_n of T_n as the approximation of θ. While the squared error is not the only loss function possible, it is realistic in many problems.

To sketch this development: First, one has a powerful Rao-Cramèr inequality, which states that there exists a lower bound for variances of all unbiased estimators of θ (for any fixed sample size n). This bound shows therefore the best that can be achieved in estimating a given parameter, in a sense of providing a yardstick by which one can tell how close a given estimator is to the "ideal" (i.e., to the estimator with the smallest possible variance).

To answer question 2: How to construct estimators? There are methods (maximum likelihood, moments, etc.) of finding an estimator, and there are also methods of modifying an estimator in order to improve it (e.g., to make its variance closer to the possible minimum value). One such improvement method is based on the concept of a sufficient statistic that retains information in the data that is relevant for estimating parameter θ. The Rao-Blackwell theorem says that if T is an unbiased estimator of θ, and S is a sufficient statistic for θ, then $T^* = E(T|S)$ is an estimator of θ that is better (or not worse) than T. Therefore one gets a powerful tool of improving estimators: start with any unbiased estimator T, and find its conditional expectation T^* with respect to a sufficient statistic. If this new estimator T^* is not the best, then continue the process, conditioning with respect to another sufficient statistic, and so on. Instead of such conditioning "one step at a time," one can condition T with respect to the so-called minimal sufficient statistic (i.e., maximal reduction of the data that preserves information about θ). The estimator T^* obtained by such conditioning cannot be improved any further. This raises the obvious questions: Is T^* the best possible estimator? Does the estimator T^* depend on the initial starting estimator T? These two questions turn out to be closely related. If the family of distributions $\{f(x, \theta), \theta \in \Theta\}$ is complete, then T^* attains the lower bound given by the Rao-Cramèr theorem, and also, T^* can be obtained by conditioning any unbiased estimator of θ with respect to the minimal sufficient statistic.

As already mentioned, the sketch above gives a "success story with a happy ending" in statistical methodology. Presented in this manner it may appear easy and effortless. But one should remember that bringing the theory to its present form required analysis of countless examples, proving or disproving various conjectures, formulating weaker and weaker sets of assumptions under which such or other assertion is true, and this took about half a century of effort on the part of many statisticians.

In this chapter we will also present some other topics and results in estimation: asymptotic properties of maximum likelihood estimators, Bayesian estimation, the bootstrap approach, and so on.

Let us begin with a simple estimation problem and a number of possible solutions. This will allow us to formulate various questions, as well as to suggest some natural generalizations to be discussed later in this chapter.

In many situations where statistical methods apply, one observes a simple random sample $X_1, \ldots, X_n$ from a distribution whose functional form is known, except for the value of a certain parameter θ. This means that the actual observations $x_1, \ldots, x_n$ are the values of iid random variables $X_1, \ldots, X_n$, each with a distribution that will be denoted $f(x, \theta)$, where f stands either for the density or for the probability function depending on whether X_i's are continuous or discrete.

We also assume that θ is an element of some *parameter space* Θ. In a simple scenario θ is a single number, so Θ is a subset of the real line, but in general, Θ may be a multidimensional space, or even a space with a more complicated structure.

In what follows, we often assume that Θ is an interval of the real line, so that we can employ standard optimization techniques, e.g., involving the differentiation of various quantities with respect to θ.

■ EXAMPLE 12.2

A politician needs to estimate the proportion of voters who favor a certain issue. In a public opinion poll, n persons are sampled from the population and their responses $X_1, \ldots, X_n$ are noted, where $X_i = 1$ or 0, depending whether or not the ith person polled favors the issue in question. Letting θ denote the proportion of voters in favor of the issue, we have

$$f(x, \theta) = P\{X_i = x|\theta\} = \begin{cases} \theta & \text{for } x = 1 \\ 1 - \theta & \text{for } x = 0. \end{cases}$$

Here $0 \leq \theta \leq 1$ (i.e., $\Theta = [0, 1]$), and $f(x, \theta)$ can be written as:

$$f(x, \theta) = \theta^x (1 - \theta)^{1-x}, \qquad x = 0, 1.$$

■ EXAMPLE 12.3

Measurement with error is often represented as an estimation problem. We are to measure the value of an attribute of an object (its weight, dimension, temperature, content of some substance, etc.). The true value of this attribute is θ, but the measurements are subject to error. For instance, it is often assumed that the ith measurement is $X_i = \theta + \epsilon_i$, where $\epsilon_1, \epsilon_2, \ldots$ are iid random variables. If we assume that $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, then $X_i \sim N(\theta, \sigma^2)$, and

$$f(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{(x-\theta)^2/2\sigma^2},$$

where σ is the standard deviation of measurements (assumed known). In this case the parameter space Θ is the subset of a real line representing the possible values of the measured attribute.

If the standard deviation of measurement, σ, is unknown, then θ is a two-dimensional parameter $\theta = (\mu, \sigma)$, and we may be interested in estimating

one or both components of θ. In this case the parameter space Θ is a subset of the half-plane $\{(\mu, \sigma) : \sigma > 0\}$.

Before proceeding further, a few comments about notation and terminology are in order. First, we will typically use a symbol such as T, to denote a statistic, even if it refers to various possible sample sizes. For instance, we may consider a sample mean $\overline{X} = (1/n)(X_1 + \cdots + X_n)$. In fact, however, we are considering here a *sequence* of statistics, each being a function of a different number of arguments. To formulate this properly, it will sometimes be convenient to use the notation involving the sample size. We will write T_n for the estimator based on a sample of size n so that $T_n = T_n(X_1, \ldots, X_n)$.

Second, we will call a statistic an *estimator* (of θ), when this statistic is *used to estimate* θ. Thus, formally, when sample size n is not specified, an estimator is a sequence of statistics, the nth one depending on observations $X_1, \ldots, X_n$. The value of an estimator, obtained for a particular sample, will be called an *estimate* of θ.

Finally, in presenting the general theory, we will use the symbol θ for the unknown parameter. However, we will also be using traditional notation, such as p for probability of success and σ for standard deviation.

The following example will show several estimators of the same parameter:

■ **EXAMPLE 12.4**

Suppose that we take a random sample $X_1, \ldots, X_n$ from the U$[0, \theta]$ distribution so that

$$f(x, \theta) = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

The objective is to estimate the range θ.

SOLUTION. We will suggest several estimators of θ. First, we get some information about θ from the largest element of the sample. We feel (and will justify it later) that as the sample size increases, the largest value should get closer and closer to θ. This suggests taking

$$T_1 = X_{n:n} \tag{12.2}$$

as an estimator of θ. A disadvantage of T_1 is that it always underestimates θ. We can remedy this in a number of ways. For instance, we can argue as follows: In the sample of size n, observed values $X_1, \ldots, X_n$ partition the interval $[0, \theta]$ into $n + 1$ intervals. Since $X_1, \ldots, X_n$ tend to be "evenly dispersed" over [0, θ], each of these $n + 1$ intervals will have, on average, the same length $\theta/(n + 1)$. Thus we should "push" T_1 to the right by $(1/n)T_1$, which suggests using the estimator

$$T_2 = \left(1 + \frac{1}{n}\right) T_1 = \frac{n+1}{n} X_{n:n}. \tag{12.3}$$

Another way of "adjusting" T_1 may be based on the fact that by symmetry, the maximal observation is, on average, at the same distance from the upper

bound θ as the minimal observation is from the lower bound 0. Thus we can use the estimator

$$T_3 = X_{1:n} + X_{n:n}. \tag{12.4}$$

We will also argue that the minimum of the sample allows us to estimate θ. The same argument as above, with a partition of the range $[0, \theta]$ into $n+1$ parts of about equal length, suggests the estimator

$$T_4 = (n+1)X_{1:n}. \tag{12.5}$$

This estimator does not seem very reliable. In particular, it may happen that $T_4 < X_{n:n}$, in which case using T_4 would make no sense. Moreover, maximum and minimum of the sample have—by symmetry—the same variance. The estimator T_4 magnifies this variance by the factor $(n+1)^2$. We will show later how this fact affects the precision of the estimators.

Finally, we take

$$T_5 = 2\overline{X}, \tag{12.6}$$

arguing that the average $\overline{X}$ should be close to the midpoint $\theta/2$.

Example 12.4 shows that one can have several estimators for the same parameter. This poses a natural question of establishing criteria for a choice. The next few sections give criteria to evaluate the performance of estimators.

Another question concerns the methods of finding estimators, especially "good" ones: Rather than rely on intuition and common sense, it is desirable to have a scheme that can produce estimators in an "automatic" fashion. We will also present some modern generalizations and extensions of estimation procedures, especially computer-intensive techniques called *resampling methods*.

12.2 CONSISTENCY

One of the basic properties of a good estimator is that it provides more precise information about θ with the increase of the sample size n. We introduce the following definition:

Definition 12.2.1 The estimator $T = \{T_n, n = 1, 2, \ldots\}$ of parameter θ is called *consistent*, if T_n converges to θ in probability, that is,

$$\lim_{n\to\infty} P\{|T_n - \theta| \le \epsilon\} = 1 \tag{12.7}$$

for every $\epsilon > 0$. The estimator T_n will be called *strongly consistent* if T_n converges to θ almost surely,

$$P\{\lim_{n\to\infty} T_n = \theta\} = 1. \tag{12.8}$$

When both kinds of consistency are considered at the same time, estimators satisfying (12.7) are called *weakly* consistent. □

We will now analyze the consistency of the five estimators T_1–T_5 introduced in Example 12.4.

■ **EXAMPLE 12.5**

The distribution of $T_1 = X_{n:n}$ is easy to obtain. Indeed, since X_i's are uniform on $[0, \theta]$, we have

$$P\{X_i \le t\} = F(t) = \begin{cases} 0 & \text{if } t < 0 \\ t/\theta & \text{if } 0 \le t \le \theta \\ 1 & \text{if } t > \theta. \end{cases}$$

We are interested only in probabilities for t between 0 and θ. Since $X_{n:n} \le t$ if and only if $X_i \le t$ for all i, by independence, we have

$$P\{T_1 \le t\} = P\{X_1 \le t, \ldots, X_n \le t\} = \left(\frac{t}{\theta}\right)^n. \tag{12.9}$$

For $0 < \epsilon < \theta$,

$$P\{|T_1 - \theta| \le \epsilon\} = P\{T_1 \ge \theta - \epsilon\} = 1 - \left(\frac{\theta - \epsilon}{\theta}\right)^n.$$

Since $[(\theta - \epsilon)/\theta]^n \to 0$ as $n \to \infty$, estimator T_1 is consistent. Using notation from Chapter 10, we may write

$$T_1 \xrightarrow{P} \theta.$$

■ **EXAMPLE 12.6**

Next, $T_2 = [(n+1)/n]\, T_1$, and since $(n+1)/n \to 1$, we have $T_2 \xrightarrow{P} \theta$.

■ **EXAMPLE 12.7**

Regarding the consistency of estimator T_3, observe first that for $0 < \epsilon < \theta$,

$$P\{X_{1:n} > \epsilon\} = \prod_{i=1}^{n} P\{X_i > \epsilon\} = \prod_{i=1}^{n} [1 - F(\epsilon)] = \left(1 - \frac{\epsilon}{\theta}\right)^n \to 0, \tag{12.10}$$

which means that $X_{1:n} \xrightarrow{P} 0$. Consequently, since

$$T_3 = T_1 + X_{1:n} \xrightarrow{P} \theta,$$

T_3 is also consistent.

■ **EXAMPLE 12.8**

Estimator T_4 seems inferior to the others, as remarked in Example 12.4. We will show that T_4 is not consistent. For $0 < \epsilon < \theta$, we have

$$\begin{aligned} P\{|T_4 - \theta| \le \epsilon\} &= P\{|(n+1)X_{1:n} - \theta| \le \epsilon\} = P\left\{\frac{\theta - \epsilon}{n+1} \le X_{1:n} \le \frac{\theta + \epsilon}{n+1}\right\} \\ &= P\left(X_{1:n} \le \frac{\theta + \epsilon}{n+1}\right) - P\left(X_{1:n} \le \frac{\theta - \epsilon}{n+1}\right). \end{aligned}$$

The cdf of $X_{1:n}$ can be obtained from formula (12.10). Using the fact that $(1 + x/n)^n \to e^x$, we then have

$$\begin{aligned} P\{|T_4 - \theta| \le \epsilon\} &= 1 - \left(1 - \frac{\theta + \epsilon}{\theta(n+1)}\right)^n - \left[1 - \left(1 - \frac{\theta - \epsilon}{\theta(n+1)}\right)^n\right] \\ &= \left(1 - \frac{\theta - \epsilon}{\theta(n+1)}\right)^n - \left(1 - \frac{\theta + \epsilon}{\theta(n+1)}\right)^n \\ &\to e^{-(\theta-\epsilon)/\theta} - e^{-(\theta+\epsilon)/\theta} < 1. \end{aligned}$$

Thus $\lim_{n\to\infty} P\{|T_4 - \theta| > \epsilon\} > 0$, which shows that T_4 is not a consistent estimator of θ.

■ **EXAMPLE 12.9**

By the law of large numbers, we have

$$\overline{X} \xrightarrow{P} E(X_i) = \frac{\theta}{2}, \tag{12.11}$$

so

$$T_5 = 2\overline{X} \xrightarrow{P} 2 \times \frac{\theta}{2} = \theta, \tag{12.12}$$

which shows that the estimator T_5 is also consistent.

From the strong law of large numbers in the iid case (Theorem 10.5.19) it immediately follows that the estimator T_5 is strongly consistent. Showing strong consistency of estimators T_1, T_2 and T_3 will be left as exercises.

PROBLEMS

12.2.1 Show that the estimators T_1, T_2, and T_3 in Example 12.4 are strongly consistent.

12.2.2 Let $X_1, \ldots, X_n$ be a random sample from the distribution with density $f(x;\theta) = e^{-(x-\theta)}$ for $x \ge \theta$, and $f(x,\theta) = 0$ otherwise. Check if $T = X_{1:n}$ is a consistent estimator of θ.

12.2.3 The density of a Pareto distribution is $f(x,\alpha,\theta) = \alpha\theta^\alpha x^{-(\alpha+1)}$ for $x \ge \theta$ and equals 0 otherwise. Show that $T = X_{1:n}$ is a consistent estimator of θ.

12.2.4 Assume that the observations are taken from the U$[0,\theta]$ distribution, and let U_n be the number of observations (out of first n) that are less than 5. Show that if $\theta > 5$, then $T_n = 5n/U_n$ is a consistent estimator of θ.

12.2.5 Observations are randomly sampled from the U$[0,\theta]$ distribution. After the sample size reaches n, the experimenter starts recording the minimum observations $X_{1:n}, X_{1:n+1}, \ldots$. He will continue until he gets X_{n+N} such that

$$X_{1:n} = X_{1:n+1} = \cdots = X_{1:n+N-1} > X_{1:n+N}.$$

Suggest an estimator of θ based on observing $X_{1:n}, \ldots, X_{1:n+N}$. [*Hint:* Find the probability distribution of N and express $E(N)$ as a function of θ.]

12.3 LOSS, RISK, AND ADMISSIBILITY

Assume again that n independent observations $X_1, \ldots, X_n$ of a random variable with distribution[21] $f(x, \theta)$ are taken in order to estimate θ. The closer θ^* and θ will be, the more successfully will θ^* subsequently serve some specific purpose.

In the simplest case this degree of success may be a decreasing function of the error $|\theta^* - \theta|$ so that the smaller the error, the better. One can easily imagine, however, situations where the error of overestimating θ (i.e., $\theta^* > \theta$) is less serious than the error of underestimating it ($\theta^* < \theta$). To build a general theory, we assume that such a situation can be adequately represented by specifying the loss function $\mathcal{L}(\theta^*, \theta)$, which describes the negative consequences of proceeding as if the value of the parameter were θ^* while in reality it is θ (we consider "loss" and negative consequences, but by changing the sign, we can convert the considerations to "rewards" and positive consequences).

Suppose now that the experimenter decides to use the estimator $T = T(X_1, \ldots, X_n)$ with the loss equal to $\mathcal{L}(T, \theta)$. The performance of estimator T can be evaluated as the average loss

$$R_T(\theta) = E_\theta\{\mathcal{L}(T(X_1, \ldots, X_n), \theta)\}, \tag{12.13}$$

where E_θ stands for expected value with respect to the distribution $f_n(x, \theta)$. The function R given by formula (12.13) above is called a *risk function*.

We now need to introduce two important concepts:

Definition 12.3.1 The estimator T_1 is *R-dominating* estimator T_2, or is *R-better* than T_2, if for all $\theta \in \Theta$ we have

$$R_{T_1}(\theta) \leq R_{T_2}(\theta),$$

and the inequality is strict for at least one value of θ.

Moreover, an estimator T is called *R-inadmissible* if there exists an estimator T' that is R-better that T. Otherwise, T will be called *R-admissible.* □

The basis for evaluation of an estimator is the risk function R, which depends on the unknown parameter θ. Clearly, if we have two estimators, and their risk functions are such that one of them is below the other (or equal to it) *regardless of the value of* θ, we can decide that the corresponding estimator is better. Quite often, however, the two risk functions cross each other (i.e., one is below the other for some θ, and above it for some other θ). In such a case the estimators are *not comparable*: Because we do not know θ, and we do not know which risk function is smaller at the actual (true) value of θ, we cannot decide which estimator is better. Thus we obtain only a *partial order* of estimators; estimators that are not dominated by any other are admissible. Within the class of admissible estimators, by definition, none dominates the others, and therefore we still need other criteria for choice. But at least the problem becomes reduced in the sense that all inadmissible estimators are ruled out. In practice, the search for the class of admissible estimators for a specific loss function can be difficult.

[21] We remind the reader that $f(x, \theta)$ is either the density or the probability function.

To build a general theory, it is necessary to choose a loss function that could is acceptable. One such a loss function is the *squared error*, that is,

$$\mathcal{L}(\theta^*, \theta) = (\theta^* - \theta)^2. \tag{12.14}$$

The use of this loss function[22] makes sense only for estimators with finite second moment (i.e., finite variance). We shall tacitly make this assumption. The risk of an estimator T for loss (12.14) is therefore $E_\theta[T(X_1, \ldots, X_n) - \theta]^2$. This risk appears so frequently that it has acquired its own name.

Definition 12.3.2 The risk of an estimator T computed for the loss function (12.14) is called the *mean squared error* of T, and is typically denoted as

$$\mathrm{MSE}_\theta(T) = E_\theta[T(X_1, \ldots, X_n) - \theta]^2. \tag{12.15}$$

Also, when no risk function is specified, admissibility will always mean admissibility with respect to the mean squared error. □

The following definition and theorem show the important role of the first two moments of T in the theory based on mean squared error.

Definition 12.3.3 An estimator T such that

$$E_\theta(T) = \theta \tag{12.16}$$

for every θ will be called *unbiased*, and the difference

$$B_\theta(T) = E_\theta(T) - \theta \tag{12.17}$$

will be called the *bias* of estimator T. □

Clearly, $B_\theta(T) = 0$ if and only if T is unbiased. The estimator T will be called positively (or negatively) biased, depending on whether $B_\theta(T) > 0$ or $B_\theta(T) < 0$. More generally, if

$$\lim_{n\to\infty} E_\theta[T(X_1, \ldots, X_n)] = \theta, \tag{12.18}$$

then T will be called *asymptotically unbiased.*

We have the following theorem:

Theorem 12.3.1 *The mean squared error* (MSE) *of an estimator is the sum of its variance and square of the bias.*

[22] Another appealing loss function is $\mathcal{L}(\theta^*, \theta) = |\theta^* - \theta|$. This function, suggested by Laplace at the beginning of the nineteenth century, leads to a theory that is much less tractable mathematically, and has begun to be developed only recently. One could argue, however, that an adequate loss function should reflect consequences of errors made *after* the parameter has been estimated (see Gafrikova and Niewiadomska-Bugaj, 1992). The starting point in the latter approach is that we use the distribution $f(x, \theta)$ for some purpose, such as prediction, calculation of some probabilities, and so on. Thus the loss $\mathcal{L}(\theta^*, \theta)$ should depend on how much the (estimated) distribution $f(x, \theta^*)$ differs from the true distribution $f(x, \theta)$. Such "difference" between distributions can be expressed through a measure of difficulty in discriminating between the two distributions.

Proof.

$$\begin{aligned}\mathrm{MSE}_\theta(T) &= E_\theta(T-\theta)^2 = E_\theta\{[T-E_\theta(T)]+[E_\theta(T)-\theta]\}^2 \\ &= E_\theta[T-E_\theta(T)]^2 + [E_\theta(T)-\theta]^2 \\ &\quad + 2E_\theta\{[T-E_\theta(T)]\times[E_\theta(T)-\theta]\} \\ &= \mathrm{Var}_\theta(T) + [B_\theta(T)]^2. \end{aligned}$$

□

Consequently the MSE of an unbiased estimator is equal to its variance.

We will now find bias and MSE for the estimators $T_1 - T_5$ introduced in Example 12.4.

Observe first that for n=1 we have $T_1 = X_1, T_2 = T_3 = T_4 = T_5 = 2X_1$. Next $E_\theta(T_1) = \theta/2$, whereas the remaining estimators are unbiased. Finally, $\mathrm{Var}(T_1) = \mathrm{Var}(X_1) = \theta^2/12$; hence the mean squared errors of all of those estimators are the same, $\theta^2/3$. They are either all admissible or all inadmissible if there exists a better estimator. In the examples below we assume that $n > 1$.

■ EXAMPLE 12.10

Let us begin with the estimator $T_1 = X_{n:n}$. Clearly, since $T_1 < \theta$, the estimator is biased. The cdf of T_1 is given by (12.9), so the density of T_1 is

$$f_{T_1}(t) = \frac{nt^{n-1}}{\theta^n}, \qquad 0 \le t \le \theta, \tag{12.19}$$

and $f_{T_1}(t) = 0$ otherwise. Thus

$$E_\theta(T_1) = \frac{n}{\theta^n}\int_0^\theta t^n dt = \frac{n}{n+1}\theta, \tag{12.20}$$

a result that we derived in Example 10.3 and used to justify the need to "adjust" T_1 to obtain unbiased estimators T_2 and T_3.

The bias of T_1 is negative:

$$B_\theta(T_1) = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}. \tag{12.21}$$

Using the formula for $\mathrm{Var}(T_1)$ obtained in Example 10.3, we obtain

$$\mathrm{MSE}_\theta(T_1) = \mathrm{Var}_\theta(T_1) + [B_\theta(T_1)]^2 = \frac{2\theta^2}{(n+1)(n+2)}. \tag{12.22}$$

■ EXAMPLE 12.11

Estimator $T_2 = [(n+1)/n]T_1$ is unbiased. Therefore

$$\mathrm{MSE}_\theta(T_2) = \mathrm{Var}_\theta(T_2) = \frac{(n+1)^2}{n^2}\mathrm{Var}_\theta(T_1) = \frac{\theta^2}{n(n+2)}. \tag{12.23}$$

Comparing with $\mathrm{MSE}_\theta(T_1)$, we have

$$\frac{\mathrm{MSE}_\theta(T_2)}{\mathrm{MSE}_\theta(T_1)} = \frac{n+1}{2n} < 1,$$

which means that the risk of T_2 is, for large n, about half of the risk of T_1. This result shows that the estimator T_1 is not admissible: There exists an estimator that is better for all θ, namely T_2 (this does *not* imply that T_2 is admissible!).

■ **EXAMPLE 12.12**

Let us now consider the estimator $T_3 = X_{1:n} + X_{n:n}$. This modification of T_1 is again intended to remove the bias so that $B_\theta(T_3) = 0$. Indeed, $E(X_{n:n}) = [n/(n+1)]\theta$, and by symmetry, $E(X_{1:n}) = [1/(n+1)]\theta$. It was shown in Section 10.3 that

$$\mathrm{Var}_\theta(X_{1:n}) = \mathrm{Var}_\theta(X_{n:n}) = \mathrm{Var}_\theta(T_1).$$

Based on formula (10.35), we obtain

$$\begin{aligned} \mathrm{MSE}_\theta(T_3) &= \mathrm{Var}_\theta(T_3) \\ &= \mathrm{Var}_\theta(X_{1:n}) + \mathrm{Var}_\theta(X_{n:n}) + 2\mathrm{Cov}_\theta(X_{1:n}, X_{n:n}) \\ &= 2\mathrm{Var}_\theta(T_1) + 2\mathrm{Cov}_\theta(X_{1:n}, X_{n:n}) \\ &= \frac{2n\theta^2}{(n+1)^2(n+2)} + \frac{2\theta^2}{(n+1)^2(n+2)} \\ &= \frac{2\theta^2}{(n+1)(n+2)}. \end{aligned} \tag{12.24}$$

Thus $\mathrm{MSE}_\theta(T_3) > \mathrm{MSE}_\theta(T_2)$; again, T_3 is not admissible.

■ **EXAMPLE 12.13**

Estimator $T_4 = (n+1)X_{1:n}$ is unbiased so that

$$\mathrm{MSE}_\theta(T_4) = \mathrm{Var}_\theta(T_4) = (n+1)^2\mathrm{Var}_\theta(X_{1:n}) = \frac{n\theta^2}{n+2}.$$

Compared with T_2, we see that T_4 is inadmissible—observe that the variance of T_4 does not even tend to 0.

■ **EXAMPLE 12.14**

For the estimator T_5, since $\mathrm{Var}(X_i) = \theta^2/12$, we have

$$\mathrm{Var}(T_5) = \mathrm{Var}\left[\frac{2}{n}(X_1 + \cdots + X_n)\right] = \frac{4}{n^2} \times n\mathrm{Var}(X_i) = \frac{\theta^2}{3n}.$$

This time, $\mathrm{MSE}_\theta(T_2)/\mathrm{MSE}_\theta(T_5) = 3/(n+2)$, which for $n > 1$ is less than 1, and therefore T_5 is not admissible.

It follows from the examples above that for $n > 1$, the best out of the five suggested estimators is T_2: In each case the risk is a parabola $k\theta^2$, with the coefficient k smallest for estimator T_2. As we concluded, T_1, T_3, T_4, and T_5 are inadmissible.

This, however, does not imply automatically that T_2 is admissible. There may exist an estimator with a smaller risk than the risk of T_2. To determine the admissibility of T_2, let us simply find the coefficient α such that $T = \alpha X_{n:n} = \alpha T_1$ has the MSE smaller than T_2. We have

$$E_\theta(T) = \alpha E(T_1) = \frac{\alpha n}{n+1}\theta,$$

and therefore

$$B_\theta(T) = \frac{\alpha n}{n+1}\theta - \theta = \frac{\alpha n - (n+1)}{n+1}\theta.$$

Since

$$\mathrm{Var}_\theta(T) = \alpha^2 \mathrm{Var}_\theta(T_1) = \frac{\alpha^2 n}{(n+1)^2(n+2)}\theta^2,$$

we have

$$\begin{aligned} \mathrm{MSE}_\theta(T) &= \left[\frac{\alpha^2 n}{(n+1)^2(n+2)} + \frac{\alpha^2 n^2 - 2\alpha n(n+1) + (n+1)^2}{(n+1)^2}\right]\theta^2 \\ &= \left[\frac{\alpha^2 n + (n+2)(\alpha^2 n^2 - 2\alpha n(n+1) + (n+1)^2)}{(n+1)^2(n+2)}\right]\theta^2. \end{aligned}$$

Differentiating the numerator with respect to α, we obtain $\alpha = (n+2)/(n+1)$ as the value that minimizes $\mathrm{MSE}_\theta(T)$. For such α, the estimator

$$T = \frac{n+2}{n+1} X_{n:n} \tag{12.25}$$

has the risk

$$\mathrm{MSE}_\theta(T) = \frac{\theta^2}{(n+1)^2} < \frac{\theta^2}{n(n+2)} = \mathrm{MSE}_\theta(T_2).$$

Thus T_2 is also inadmissible, and the best estimator (of those considered) is the estimator T given by (12.25).

PROBLEMS

12.3.1 Let T_1 and T_2 be two unbiased estimators of θ with variances σ_1^2, σ_2^2, respectively. Find values a and b such that: **(i)** Estimator $aT_1 + bT_2$ is unbiased. **(ii)** Unbiased estimator $aT_1 + bT_2$ has a minimum variance assuming that T_1 and T_2 are independent. **(iii)** Unbiased estimator $aT_1 + bT_2$ has a minimum variance if $\mathrm{Cov}(T_1, T_2) = C$.

12.3.2 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{N}(\theta, \sigma^2)$ distribution with σ^2 known. Show that the estimator T of θ, defined as $T(X_1, \ldots, X_n) \equiv 3$ ($T = 3$ regardless of the observations), is admissible.

12.3.3 Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{EXP}(1/\theta)$ distribution. Compare the mean squared errors of two estimators of θ: $T_1 = \overline{X}$ and $T_2 = [n/(n+1)]\overline{X}$.

12.3.4 Let $X_1, \ldots, X_n$ be n Bernoulli trials with probability of success θ, and let $S = \sum_{i=1}^{n} X_i$. Compare the mean squared errors of two estimators of θ: $T_1 = S/n$ and $T_2 = (S+1)/(n+2)$.

12.3.5 Let $X - 1, \ldots, X_n$ be a random sample of size n from the discrete distribution with probability function $f(x, \theta) = \theta(1-\theta)^x, x = 0, 1, \ldots$. Compare the MSE's of two estimators of θ: $T_1 = \overline{X}$ and $T_2 = [n/(n+1)]\overline{X}$.

12.3.6 Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution (μ and σ^2 are unknown), and let

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

be two estimators of σ^2. **(i)** Compare the MSE's of S^2 and S_1^2. **(ii)** Consider

$$S_k^2 = k\sum_{i=1}^{n}(X_i - \overline{X})^2$$

as estimators of σ^2 and find k for which S_k^2 has smallest MSE. Explain why, in practice, the only values of k used are (suboptimal): $k = 1/n$ and $k = 1/(n-1)$.

12.3.7 Let $U = X_{1:n}$ and let $V = X_{n:n}$ in a random sample from $U[\theta - 1, \theta + 1]$ distribution. **(i)** Show that $\overline{X}$ and $(U+V)/2$ are both unbiased estimators of θ. **(ii)** Determine the MSE's of estimators in (i).

12.3.8 Let $X_1, \ldots, X_4$ be a random sample from $U[0, \theta]$ distribution. Compare the mean squared errors of four estimators of θ: $T_1 = 5X_{1:4}$, $T_2 = (5/2)X_{2:4}$, $T_3 = (5/3)X_{3:4}$, and $T_4 = (5/4)X_{4:4}$.

12.4 EFFICIENCY

We assume that the observations $X_1, X_2, \ldots$ are iid random variables, with $E(X_i) = \theta$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, the latter assumed known.

The obvious choice of estimator of the mean θ is $T = (X_1 + \cdots + X_n)/n$, traditionally denoted by $\overline{X}$ or $\overline{X}_n$, if we need to stress the dependence on the sample size. Since

$$E(\overline{X}_n) = \frac{1}{n}E(X_1 + \cdots + X_n) = \theta, \tag{12.26}$$

$\overline{X}_n$ is always an unbiased estimator of the mean θ. Furthermore, by the Chebyshev inequality and the fact that $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ for every $\epsilon > 0$,

$$P\{|\overline{X}_n - \theta| > \epsilon\} \leq \frac{\mathrm{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0, \tag{12.27}$$

and therefore $\overline{X}_n$ is a consistent estimator of θ if only $\mathrm{Var}(X_i) < \infty$. In view of unbiasedness, the risk of $\overline{X}_n$ is equal to its variance, that is,

$$\mathrm{MSE}_\theta(\overline{X}_n) = \frac{\sigma^2}{n}. \tag{12.28}$$

Let us now compare the situation with that of estimating the range θ of uniform distribution, by using estimators such as T_1 or T_2 given by (12.2) and (12.3). There are two main points of difference here. The first is that in the case of estimating the mean by $\overline{X}_n$, the risk for a fixed n is constant—it does not depend on the estimated parameter θ. In the case of estimators T_1 and T_2, the risk given by (12.22) and (12.23) is a quadratic function of θ: the larger the value estimated is, the larger is the variance of the results.

The second difference—and much more important—is the rate at which the risk changes with the sample size n. In the case of estimating the mean by the sample average $\overline{X}_n$, the risk changes inversely with n, whereas in estimating the range of uniform distribution by T_1, or T_2, it changes (approximately) inversely with n^2. Naturally the latter case is more desirable practically: increasing the sample size by the factor of 10 reduces the mean squared error by factor of 100—that is, to about 1% of the mean squared error for the original sample size. As opposed to that, in estimating the mean by $\overline{X}_n$, an increase of sample size by the factor of 10 decreases the mean squared error by the factor of 10 only—that is, to about 10% of the mean squared error for the original sample size.

So why do statisticians use the sample average as an estimator, if there are other estimators that are so much better?

The answer is that a situation such as the one with the sample mean—when the mean squared error decreases inversely with n—is much more common than the situation where the mean squared error decreases with n^2. In most cases the best estimators that one can obtain have a mean squared error that decreases in proportion to $1/n$.

The concept that we will attempt to formalize is that of the average "amount of information about θ contained in a single observation of X." Let $X_1, X_2, \ldots$ be a random sample, selected from the distribution $f(x, \theta)$, where θ takes values in some open domain, and the set of points x at which $f(x, \theta) > 0$ does not depend on θ. (The last assumption rules out the case of observations from the U$[0, \theta]$ distribution, where the set of points at which $f(x, \theta)$ is positive depends on θ.) The reason for distinguishing such "regular" case is that if this assumption were violated, a single observation may eliminate some values of θ with certainty (e.g., in the case of uniform distribution on $(0, \theta]$, observation $X = 3$ rules out all $\theta < 3$). Such elimination is typically unattainable in practical situations; hence we concentrate the theory on "regular" cases.

Let us start from some intuitions. Suppose one gets a message that some event A occurred. The amount of information in this message depends, in an obvious way, on what event A is and who receives the message (think here of a message that Delta flight 503 will arrive 2 hours late). Apart from such semantic and personal information, there is also some information contained in the message (that A occurred) depending only on how likely is the event A. If $P(A)$ is close to 1, the amount of information is close to 0, while for $P(A)$ close to 0, the amount of information is high.

This interpretation agrees with common intuition, as well as with practice (e.g., in newspaper publishing). The fact that Mr. Smith found a dead roach in the can of beer he drank is of interest and worth reporting precisely because such an event

is rare; hence its occurrence carries lots of information. If, on the other hand, dead roaches were commonly found in cans of beer, no one would care to report it.

An interesting and coherent theory is obtained if the amount of information in the occurrence of an event with probability p is $\log(1/p) = -\log p$.

The situation in the case of estimation is somewhat more complicated, since we want to define the amount of information *about* θ in the event $X = x$, where X is a random variable with distribution (density or pdf) $f(x, \theta)$. Here $f(x, \theta)$ plays the role of probability p of the event $X = x$. Since we are interested in information about θ, it appears natural to consider the rate of change of $\log(1/[f(x, \theta)]) = -\log f(x, \theta)$ under varying θ at the point x, that is, the derivative

$$\frac{\partial}{\partial\theta}[-\log f(x,\theta)] = -\frac{\frac{\partial}{\partial\theta}f(x,\theta)}{f(x,\theta)}. \tag{12.29}$$

We eliminate the effect of the sign by considering the square of quantity (12.29), and take the expectation to avoid the restriction to specific value x.

These considerations lead to the following definition:

Definition 12.4.1 Let X be a random variable with distribution $f(x, \theta)$, such that the set of points x at which $f(x, \theta) > 0$ is the same for all θ. We assume that the function $f(x, \theta)$ is twice differentiable with respect to θ for every x. Then the *Fisher information* about θ in a single observation X is defined as

$$I(\theta) = E\{J(X,\theta)\}^2, \tag{12.30}$$

where

$$J(X,\theta) = \frac{\partial}{\partial\theta}[\log f(X,\theta)], \tag{12.31}$$

provided that the expectation (12.30) exists. □

Thus, in the case of a continuous random variable X, the quantity (12.30) is

$$I(\theta) = \int_{-\infty}^{+\infty} [J(x,\theta)]^2 f(x,\theta)dx = \int_{-\infty}^{+\infty} \left[\frac{\partial f(x,\theta)}{\partial\theta}\right]^2 \frac{1}{f(x,\theta)}dx,$$

while in the discrete case integration is replaced by summation.

■ **EXAMPLE 12.15**

Let X have normal distribution $\mathrm{N}(\theta, \sigma^2)$. Then

$$\log f(x,\theta) = -\log(\sigma\sqrt{2\pi}) - \frac{(x-\theta)^2}{2\sigma^2}$$

and

$$J(X,\theta) = \frac{\partial}{\partial\theta}[\log f(X,\theta)] = \frac{X-\theta}{\sigma^2}. \tag{12.32}$$

Thus

$$I(\theta) = E\left(\frac{X-\theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^4}E(X-\theta)^2 = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

■ **EXAMPLE 12.16**

Consider now a single Bernoulli trial with probability θ. We have $P\{X = 1|\theta\} = \theta$ and $P\{X = 0|\theta\} = 1 - \theta$ so that

$$f(x,\theta) = \theta^x(1-\theta)^{1-x}, \quad x = 0, 1. \tag{12.33}$$

Thus $\log f(x,\theta) = x\log\theta + (1-x)\log(1-\theta)$ and

$$\left[\frac{\partial}{\partial\theta}\log f(x,\theta)\right]^2 = \left(\frac{x}{\theta} - \frac{1-x}{1-\theta}\right)^2 = \begin{cases} \frac{1}{(1-\theta)^2} & \text{if } x = 0 \\ \frac{1}{\theta^2} & \text{if } x = 1. \end{cases} \tag{12.34}$$

Taking the expectation, we obtain

$$I(\theta) = \frac{1}{(1-\theta)^2} \times (1-\theta) + \frac{1}{\theta^2} \times \theta = \frac{1}{\theta(1-\theta)}. \tag{12.35}$$

■ **EXAMPLE 12.17**

Let us slightly modify Example 12.16: Suppose now that $P\{X = 1|\theta\} = \theta^2$ and $P\{X = 0|\theta\} = 1 - \theta^2$; thus we have a single Bernoulli trial, but now the probability of success is θ^2. However, we are still interested in the amount of information about θ (not θ^2). A practical example here may be found in genetics, where θ is a frequency of a recessive gene (e.g., a gene causing a person to have blue eyes) so that $X = 1$ corresponds to finding a person with some features caused by a recessive gene (requiring both parents to transmit this gene). We now have

$$f(x,\theta) = (\theta^2)^x(1-\theta^2)^{1-x};$$

hence

$$\begin{aligned} \left[\frac{\partial}{\partial\theta}\log f(x,\theta)\right]^2 &= \left\{\frac{\partial}{\partial\theta}\left[2x\log\theta + (1-x)\log(1-\theta^2)\right]\right\}^2 \\ &= \left\{\frac{2x}{\theta} - 2\theta\frac{(1-x)}{1-\theta^2}\right\}^2 \\ &= \begin{cases} \frac{4\theta^2}{(1-\theta^2)^2} & \text{if } x = 0 \\ \frac{4}{\theta^2} & \text{if } x = 1. \end{cases} \end{aligned}$$

Consequently

$$I(\theta) = \frac{4\theta^2}{(1-\theta^2)^2}(1-\theta^2) + \frac{4}{\theta^2}\theta^2 = \frac{4\theta^2}{1-\theta^2} + 4 = \frac{4}{1-\theta^2}. \tag{12.36}$$

Comparison of this result with the result in Example 12.16 is quite instructive. In a single Bernoulli trial with probability of success θ, the average amount of information in a trial is a function that assumes its minimal value 4 at $\theta = 1/2$. As θ moves away from $1/2$ towards either $\theta \approx 0$ (success very rare) or $\theta = 1$

(failure very rare), the average amount of information about θ increases to infinity.

On the other hand, if we can only observe whether or not an event with probability θ^2 has occurred, then the average amount of information about θ is close to 4 if $\theta \approx 0$, and increases to infinity when $\theta \to 1$. The symmetry of the preceding example is lost. Indeed, only when θ is close to 1 is θ^2 also close to 1, so successes that are sufficiently frequent provide a good estimator of θ^2, and hence also of θ. For small θ, we have a poor estimate of θ^2, and hence also a poor estimate of θ.

Alternative formulas for $I(\theta)$ can also be derived. Observe first that

$$\begin{aligned} E_\theta[J(X,\theta)] &= \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial\theta}\log f(x,\theta)\right] f(x,\theta)dx \\ &= \int_{-\infty}^{\infty} \left[\frac{\frac{\partial}{\partial\theta}f(x,\theta)}{f(x,\theta)}\right] f(x,\theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)dx \\ &= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \frac{\partial}{\partial\theta}(1) = 0. \end{aligned} \tag{12.37}$$

This is valid under the conditions that allow to interchange differentiation with respect to θ and integration with respect to x. Thus, in view of (12.37),

$$\mathrm{Var}_\theta[J(X,\theta)] = E_\theta[J(X,\theta)]^2 - \{E[J(X,\theta)]\}^2 = I(\theta). \tag{12.38}$$

Still another formula for $I(\theta)$ can be obtained by noting that

$$\begin{aligned} \frac{\partial}{\partial\theta}J(x,\theta) = \frac{\partial}{\partial\theta}\left[\frac{\frac{\partial}{\partial\theta}f(x,\theta)}{f(x,\theta)}\right] &= \frac{\left(\frac{\partial^2}{\partial\theta^2}f(x,\theta)\right) - \left(\frac{\partial}{\partial\theta}f(x,\theta)\right)^2}{[f(x,\theta)]^2} \\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)} - [J(x,\theta])^2. \end{aligned}$$

Taking expectations of both sides, we get

$$E_\theta\left(\frac{\partial}{\partial\theta}J(x,\theta)\right) = \int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2}f(x,\theta)dx - E_\theta\left[J(X,\theta)\right]^2 = -I(\theta),$$

since, assuming one can interchange integration and differentiation,

$$\int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2}f(x,\theta)dx = \frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{\infty} f(x,\theta)dx = \frac{\partial^2}{\partial\theta^2}(1) = 0.$$

We have therefore the following theorem:

Theorem 12.4.1 *If the density $f(x,\theta)$ is twice differentiable in θ and the equality $\int f(x,\theta)dx = 1$ can be differentiated twice under the integral sign, we have*

$$I(\theta) = E_\theta[J(X,\theta)]^2 = \mathrm{Var}_\theta[J(X,\theta)] = -E_\theta\left(\frac{\partial}{\partial\theta}J(x,\theta)\right). \tag{12.39}$$

□

The following theorem determines the average information contained in a random sample of size n:

Theorem 12.4.2 *The information $I_n(\theta)$, in a random sample of n observations, is n times the information of a single observation:*

$$I_n(\theta) = nI(\theta). \tag{12.40}$$

Proof. The density of $\mathbf{X}$ is $f(\mathbf{X}, \theta) = f(X_1, \theta) \cdots f(X_n, \theta)$ so that

$$J(\mathbf{X}, \theta) = \frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f(X_i, \theta) = \sum_{i=1}^{n} J(X_i, \theta).$$

Differentiating both sides, we have

$$\frac{\partial}{\partial \theta} J(\mathbf{X}, \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} J(X_i, \theta). \tag{12.41}$$

□

The next theorem connects the amount of information about θ contained in a single observation, with the variance of an estimator of θ.

Theorem 12.4.3 (Rao-Cramér Inequality) *For any estimator T_n of θ, such that $E_\theta(T_n) = m(\theta)$, we have*

$$\mathrm{Var}_\theta(T_n) \geq \frac{[m'(\theta)]^2}{nI(\theta)}. \tag{12.42}$$

In particular, if T is unbiased, then $m(\theta) = \theta, m'(\theta) = 1$, and

$$\mathrm{Var}_\theta(T_n) \geq \frac{1}{nI(\theta)}. \tag{12.43}$$

Proof. Let $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ be an arbitrary estimator of θ, assumed to have a finite variance. Letting $\mathbf{x} = (x_1, \ldots, x_n)$, we have

$$m(\theta) = E_\theta(T) = \int \cdots \int T(\mathbf{x}) f(\mathbf{x}, \theta) dx_1 \cdots dx_n.$$

Consequently, assuming again that we can differentiate under the integral sign, we obtain

$$\begin{aligned} m'(\theta) &= \int \cdots \int T(\mathbf{x}) \left[\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta) \right] dx_1 \cdots dx_n \\ &= \int \cdots \int T(\mathbf{x}) \left[\frac{\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta)}{f(\mathbf{x}, \theta)} \right] f(\mathbf{x}, \theta) dx_1 \cdots dx_n \\ &= \int \cdots \int T(\mathbf{x}) J(\mathbf{x}, \theta) f(\mathbf{x}, \theta) dx_1 \cdots dx_n \\ &= E_\theta\{T(\mathbf{X}) J(\mathbf{X}, \theta)\} = \mathrm{Cov}_\theta\{T(\mathbf{X}), J(\mathbf{X}, \theta)\}. \end{aligned} \tag{12.44}$$

The last equality follows from the fact that $E[J(\mathbf{X}, \theta)] = 0$, which can be established in the same way as for a single random variable X, in view of the relation

$$J(\mathbf{X}, \theta) = \sum_{i=1}^{n} J(X_i, \theta).$$

Now, by the Schwarz inequality (8.6.7) and (12.44), we may write

$$[m'(\theta)]^2 = [\mathrm{Cov}_\theta(T(\mathbf{X}), J(\mathbf{X}, \theta)]^2 \leq \mathrm{Var}_\theta(T) \times \mathrm{Var}_\theta[J(\mathbf{X}, \theta)]. \tag{12.45}$$

□

Observe that we have equality in (12.42) if and only if we have equality in (12.45), which means that the correlation coefficient between $T(X)$ and $J(X, \theta)$ satisfies the condition $|\rho_{T(X),J(X,\theta)}| = 1$. By Theorem 8.5.6, this is equivalent to linear relationship between $T(X)$ and $J(X, \theta)$, that is,

$$T(X) = \gamma_1(\theta) J(X, \theta) + \gamma_2(\theta) \tag{12.46}$$

for some functions $\gamma_1(\theta)$ and $\gamma_2(\theta)$.

An important point here is that T is an estimator of θ, so the left-hand side does not depend on θ. Consequently, the right-hand side does not depend on θ either; that is, functions γ_1 and γ_2 cancel the dependence of $J(X, \theta)$ on the parameter θ. The set of conditions that allow all differentiations under the integral sign, combined with the condition of independence of θ of the set of x for which $f(x, \theta) > 0$, will be called the regularity conditions.

We have the following definition:

Definition 12.4.2 Any unbiased estimator T that satisfies the regularity conditions and whose variance attains the minimum equal to the right-hand side of (12.43) is called *efficient*. The ratio $nI(\theta)/\mathrm{Var}_\theta(T)$ is called the *efficiency* of T.

More generally, given two unbiased estimators T_1 and T_2 of θ, the ratio of their variances $\mathrm{Var}_\theta(T_1)/\mathrm{Var}_\theta(T_2)$ is called the *relative efficiency* of T_2 with respect to T_1. □

■ **EXAMPLE 12.18**

Consider the problem of estimating the mean θ of normal distribution $\mathrm{N}(\theta, \sigma^2)$. From Example 12.15 we have $I(\theta) = 1/\sigma^2$. The most obvious estimator for θ, $\overline{X}_n$, is unbiased, and $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$. The lower bound of variances of all unbiased estimators of θ, based on samples of size n, is $1/[nI(\theta)] = \sigma^2/n$, which shows that $\overline{X}_n$ is efficient.

■ **EXAMPLE 12.19**

Continuing Example 12.16, suppose that we want to estimate probability of success θ on the basis of n Bernoulli trials. Since $I(\theta) = 1/[\theta(1-\theta)]$, the information in the sample of size n is $n/[\theta(1-\theta]$. For $S_n = X_1 + \cdots +$

X_n being the total number of successes, $T = S_n/n$ will be an estimator of θ. S has binomial distribution; $E(T) = (1/n)E(S_n) = \theta$, and $\text{Var}(T) = (1/n^2)\text{Var}(S_n) = \theta(1-\theta)/n$, which shows that T is efficient.

■ **EXAMPLE 12.20**

In Section 12.2 we studied five estimators of θ in U[0, θ] distribution. We found that variances of estimators T_1, T_2, and T_3 decrease to zero like $1/n^2$ (and not at the rate $1/n$, as for efficient estimators). Thus those estimators are superefficient. This is possible since $f(x, \theta)$ does not satisfy the regularity conditions. The set of points x at which the density $f(x, \theta)$ is positive is [0, θ]; hence it depends on θ.

Under the regularity conditions, the right-hand side of the Rao-Cramér inequality (12.43) gives the lower bound for variances of unbiased estimators, and therefore, if an efficient estimator of θ is found, it is known that a better unbiased estimator does not exist. It is necessary to point out, however, that the Rao-Cramér bound is not always attainable. It may happen that the best possible unbiased estimator has a variance larger than the Rao-Cramér bound. For a thorough discussion of these topics see, for example, Mood et al. (1974).

PROBLEMS

12.4.1 Let $\overline{X}_{2k}$ be the sample mean of $2k$ independent observations from a normal distribution with mean θ and known variance σ^2. Find the efficiency of $\overline{X}_k$ (i.e., of estimator that uses only half of the sample).

12.4.2 Let X have EXP(λ) distribution. Find the Fisher information $I(\lambda)$.

12.4.3 Let $X_1, \ldots, X_n$ be a random sample from EXP(λ) distribution. Propose an efficient estimator of $1/\lambda$ and determine its variance.

12.4.4 Find Fisher information $I(\theta)$ in a random sample of size n from the Cauchy distribution with density $f(x, \theta) = \{\pi[1 + (x-\theta)^2]\}^{-1}$.

12.4.5 Let X_1, X_2 be a random sample of size 2 from N(μ, σ^2) distribution. Determine the amount of information about μ and about σ^2 contained in: **(i)** $X_1 + X_2$. **(ii)** $X_1 - X_2$.

12.4.6 Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with an unknown p. Show that the variance of any unbiased estimator of $(1-p)^2$ must be at least $4p(1-p)^3/n$.

12.4.7 Show that the estimator $T = \overline{X}$ satisfies relation (12.46) and determine functions γ_1 and γ_2 if a random sample $X_1, X_2, \ldots, X_n$ is selected from: **(i)** N(θ, σ^2) distribution with σ known. **(ii)** BIN(1, θ) distribution.

12.5 METHODS OF OBTAINING ESTIMATORS

Before introducing further criteria for evaluating performance of estimators (besides consistency, unbiasedness and efficiency), we shall now present methods of constructing estimators.

Method of Moments Estimators

In this approach estimators are obtained by equating sample quantities to the corresponding population quantities as functions of the parameter to be estimated. The estimator is then found as a solution of the resulting equation with respect to the parameter, and the general approach is also known as a *plug-in principle*

Since the most commonly taken quantity is a moment of the random variable, this method of constructing estimators is generally known as *method of moments*.

■ **EXAMPLE 12.21**

Let $X_1, \ldots, X_n$ be a random sample from the EXP(θ) distribution. Then $E(X_i) = 1/\theta$. The sample counterpart of the first moment is the sample mean $\overline{X}_n = (1/n)(X_1 + \cdots + X_n)$. Equating the empirical and theoretical mean, we obtain the equation $\overline{X}_n = 1/\hat{\theta}$, which gives the estimator $T_1 = 1/\overline{X}_n$ of the parameter θ.

We could, however, use the second moment as well. Then we have $E(X^2) = \mathrm{Var}(X) + [E(X)]^2 = 2/\theta^2$. The empirical counterpart of the second moment is $(1/n)\sum_{i=1}^n X_i^2$. We therefore obtain the equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{2}{\hat{\theta}^2},$$

which, solved for $\hat{\theta}$, gives an estimator

$$T_2 = \sqrt{\frac{2n}{\sum_{i=1}^n X_i^2}}. \tag{12.47}$$

Still another possibility would be to use the median m, the point at which cdf equals $1/2$, so $F(m) = 1 - e^{-\theta m} = 1/2$, implying $m = (\log 2)/\theta$. The empirical counterpart of the median m is the sample median. We will assume for simplicity, that the sample size is odd ($n = 2k + 1$). Then the sample median is the $(k+1)$st order statistic $X_{k+1:n}$. We therefore have still another estimator of θ,

$$T_3 = \frac{\log 2}{X_{k+1:n}}. \tag{12.48}$$

Finally, suppose that we want to estimate the probability

$$p = P\{X \geq 3\} = e^{-3\theta}.$$

The method of moments suggests using the estimator e^{-3T}, where T is any estimator of θ. Estimators T_1, T_2, and T_3 now give three estimators of p,

namely

$$\hat{p}_1 = \exp\left\{-\frac{3}{\overline{X}_n}\right\}, \quad \hat{p}_2 = \exp\left\{-3\sqrt{\frac{2n}{\sum X_i^2}}\right\}, \quad \hat{p}_3 = \exp\left\{\frac{-3\log 2}{X_{k+1:n}}\right\}.$$

■ **EXAMPLE 12.22**

Let us observe that the estimator $T_5 = 2\overline{X}_n$ from Example 12.4 is also an example of estimator obtained by the method of moments. Since $E(X) = \theta/2$, doubling the sample mean gives a moment estimator of θ.

The method of moments can also be applied in the case of estimating several parameters at once, as illustrated by the following example:

■ **EXAMPLE 12.23**

Suppose that we want to estimate both μ and σ^2 based on a random sample $X_1, \ldots, X_n$ from some distribution with mean μ and variance σ^2.

SOLUTION. We could choose the following two expressions:

$$E(X) = \mu, \qquad E(X^2) = \sigma^2 + \mu^2. \tag{12.49}$$

This suggests that we compare the first and second empirical moment with μ and $\sigma^2 + \mu^2$ by obtaining the equations

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}, \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\sigma}^2 + \hat{\mu}^2.$$

Solving for $\hat{\mu}$ and $\hat{\sigma}^2$, we obtain

$$\hat{\mu} = \overline{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{12.50}$$

Thus the sample mean and the sample variance (with divisor n) are method-of-moment estimators of the population mean and variance in the general case.

■ **EXAMPLE 12.24**

As the last example of the method-of-moment estimator, consider the problem of the randomized response, discussed in Section 1.2. Generally, the empirical problem concerns some attribute, call it Q, that a person might be reluctant to admit having. The objective is to estimate the frequency θ of persons in a given population whose true reply to a question "Are you a Q-person?" is "yes." To collect the data on such questions one can use a questionnaire with

a randomized response (see Example 1.9). The respondent activates two random mechanisms, one generating the event A or its complement A^c, the other generating the event B or its complement B^c. These two mechanisms operate independently. The probabilities $P(A) = \alpha$ and $P(B) = \beta$ are known; however, only the respondent knows which of the events A and B occurred in a given instance. He is instructed to respond to the question "Are you a Q-person?" if A occurs, and otherwise respond to the question "Did B occur?" The answer "yes" or "not" is recorded by the experimenter who does not know which question was actually answered.

Conditioning on the occurrence or nonoccurrence of event A, we have

$$\begin{aligned} P(\text{"yes"}) &= P(\text{"yes"}|A)P(A) + P(\text{"yes"}|A^c)P(A^c) \\ &= \theta\alpha + \beta(1-\alpha). \end{aligned}$$

If now X respondents out of N tested replied "yes," then X/N is an estimator of $P(\text{"yes"})$, and we have an approximate equality

$$\frac{X}{N} \approx \theta\alpha + \beta(1-\alpha),$$

which suggests using as an estimator of θ the random variable

$$T = \frac{X/N - \beta(1-\alpha)}{\alpha}.$$

A problem with the estimators obtained by method of moments is that they are not unique, since one can take moments of various orders, or even different quantities. However, as a rule, one should take the moments of lowest orders that depend on θ, as illustrated by the following example:

■ **EXAMPLE 12.25**

Suppose that the observations are known to have $\mathrm{U}[-\theta, \theta]$ distribution. Then $E(X) = 0$, and the first moment contains no information about θ. One may use here the fact that

$$E(X^2) = \int_{-\theta}^{\theta} x^2 \times \frac{1}{2\theta} dx = \frac{\theta^2}{3},$$

which gives the estimator

$$\hat{\theta} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}.$$

The moment estimators are consistent under some very mild conditions. Indeed, the strong law of large numbers for the iid case asserts that $(1/n)(X_1^k + X_2^k + \cdots +$

X_n^k) converges with probability 1 to $E_\theta(X^k)$ if only $E_\theta(|X|^k) < \infty$. Consequently, the empirical moments converge also in probability to the corresponding theoretical moments, and if only the parameter θ is a continuous function of moments (as is usually the case), the consistency of method-of-moment estimators follows.

The situation is not so straightforward for estimators built on quantities other than the moments. Consistency has to be studied separately in each such case, by establishing whether or not the sample analogue of a given quantity converges in probability (or almost surely) to the corresponding theoretical quantity. However, the main issue with method-of-moment estimators is that they either coincide with estimators obtained by the maximum likelihood method (discussed below), or they are inferior to them.

Maximum Likelihood Estimators

Let $X_1, \ldots, X_n$ be a random sample from distribution $f(x, \theta)$, where $f(x, \theta)$ may stand for the density or for the probability function. If the actual observations are $X_1 = x_1, \ldots, X_n = x_n$, then the probability of this sample (or the joint density) is

$$f(x_1, \theta) \cdots f(x_n, \theta). \tag{12.51}$$

The product (12.51), regarded as a function of parameter θ, is called the *likelihood function* of the sample, or simply the likelihood function. We will use the symbol

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i, \theta) = f_n(\mathbf{x}, \theta), \tag{12.52}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$.

Definition 12.5.1 Given the sample $\mathbf{x} = (x_1, \ldots, x_n)$, the value $\hat{\theta} = \hat{\theta}(\mathbf{x})$ that maximizes the likelihood function (12.52), is called the *maximum likelihood estimate* (MLE) of θ. □

Let us begin with some examples:

■ **EXAMPLE 12.26**

We will find MLE of the probability of success θ, if in five independent Bernoulli trials with probability of success θ, three successes and two failures were observed.

We have here $f(x, \theta) = \theta^x(1-\theta)^{1-x}$ where $x = 0$ or 1 represent failure and success, respectively. Thus the likelihood is

$$L(\theta) = \prod_{i=1}^{5} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^3(1-\theta)^2, \tag{12.53}$$

where $0 \leq \theta \leq 1$. The information as to which trials led to successes and which to failures does not affect the likelihood and is irrelevant for the estimation of θ.

To find the maximum of function (12.53) we differentiate, obtaining the equation

$$\begin{aligned} L'(\theta) &= 3\theta^2(1-\theta)^2 - 2\theta^3(1-\theta) = \theta^2(1-\theta)[3(1-\theta) - 2\theta] \\ &= \theta^2(1-\theta)(3-5\theta) = 0. \end{aligned}$$

The solutions are: $\theta = 0, \theta = 1$, and $\theta = 3/5$. An inspection of $L(\theta)$ shows that this function attains its maximum at the last solution, while $\theta = 0$ and $\theta = 1$ give the minima. Thus the maximum likelihood estimate of θ is $\hat{\theta} = 3/5$.

■ **EXAMPLE 12.27**

Suppose that we take n=3 observations from POI(θ) distribution, obtaining values $x_1 = 2, x_2 = 0, x_3 = 5$. The likelihood is

$$L(\theta) = \left(\frac{\theta^2}{2!}e^{-\theta}\right)\left(\frac{\theta^0}{0!}e^{-\theta}\right)\left(\frac{\theta^5}{5!}e^{-\theta}\right) = \frac{\theta^7}{2!5!}e^{-3\theta}.$$

The derivative now is

$$L'(\theta) = \frac{1}{2!5!}[7\theta^6 e^{-3\theta} - 3\theta^7 e^{-3\theta}],$$

and $L'(\theta) = 0$ for $\theta = 0$ and for $\theta = 7/3$. An inspection of L shows that the maximum occurs at the second solution, so the MLE of θ is now $7/3$.

■ **EXAMPLE 12.28**

Suppose that we observe values $x_1 = 3$ and $x_2 = -2$ from a N$(0, \theta^2)$ distribution. The likelihood now is

$$L(\theta) = \frac{1}{\theta\sqrt{2\pi}}e^{-9/2\theta^2} \times \frac{1}{\theta\sqrt{2\pi}}e^{-4/2\theta^2} = \frac{1}{2\pi\theta^2}e^{-13/2\theta^2}.$$

Since $L(\theta)$ is maximized at the same point at which its logarithm is maximized, we will take the logarithm first and then differentiate. Taking natural logarithms, we have

$$\log L(\theta) = -2\log\theta - \log(2\pi) - \frac{13}{2\theta^2}.$$

Hence

$$\frac{d}{d\theta}\log L(\theta) = -\frac{2}{\theta} + \frac{13}{\theta^3},$$

and we obtain the equation

$$-\frac{1}{\theta}\left(2 - \frac{13}{\theta^2}\right) = 0$$

with solution $\theta = \sqrt{13/2}$. After checking that the likelihood is maximized at $\sqrt{13/2}$, we obtain $\hat{\theta} = \sqrt{13/2}$ as the MLE of θ.

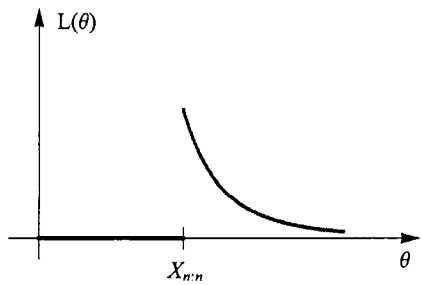


Figure 12.1 Likelihood function for the range R in uniform distribution

Some obvious questions arise here. First, the *statistical* questions concerning the properties of the suggested procedure. The point is that the maximum likelihood estimate depends on the sample, so it is a random quantity. Our procedure therefore defines an estimator. What are its properties, such, as consistency, bias and efficiency?

Second, the *mathematical* questions: Does MLE always exist? If so, is the maximum of the likelihood function unique? Can it always be obtained by differentiation of likelihood or of its logarithm and solving the resulting equation?

There are cases when, formally speaking, MLE does not exist, but these cases can often be modified in a natural way so as to remedy the situation.

■ **EXAMPLE 12.29**

Consider the problem of estimating θ based on observations from the U(0, θ) distribution, with density

$$f(x,\theta) = \begin{cases} 1/\theta & \text{for } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Given the sample $x_1, \ldots, x_n$, the likelihood is

$$L(\theta) = f(x_1,\theta)\cdots f(x_n,\theta) = \begin{cases} 1/\theta^n & \text{if } 0 < x_i < \theta \text{ for all } i \\ 0 & \text{otherwise.} \end{cases}$$

Thus the likelihood function is discontinuous: it is the function depicted in Figure 12.1, with discontinuity at the point $t = \max(x_1, \ldots, x_n)$, and decreasing for $\theta > t$. However, for $\theta = t$ we have $L(t) = 0$, so there is no point at which this function attains its maximum, and MLE does not exist.

The cause of the trouble here is the choice of the definition of the density. If we define $f(x,\theta)$ to be $1/\theta$ in the *closed* interval $0 \le x \le \theta$, the likelihood would actually reach its maximum at $\hat{\theta} = \max(x_1, \ldots, x_n)$.

The example above shows a type of situation where MLE does not exist because of the choice of the density function. Since the density function can be modified at

a single point (or at any finite or countable set of points), such a modification can affect the likelihood function. Consequently, we may (and will) always assume that the densities are defined in such a way that the maximum of the likelihood exists.

■ **EXAMPLE 12.30**

Assume that we are estimating θ based on a random sample $X_1, \ldots, X_n$, with X_i's being distributed uniformly on the interval $[\theta - 1/2, \theta + 1/2]$. We have

$$f(x, \theta) = \begin{cases} 1 & \text{for } \theta - \frac{1}{2} \leq x \leq \theta + \frac{1}{2} \\ 0 & \text{otherwise .} \end{cases}$$

Consequently, $L(\theta) = 1$ if $X_{n:n} \leq \theta + 1/2$ and $X_{1:n} \geq \theta - 1/2$, and is equal 0 otherwise. All values of $L(\theta)$ in a certain interval are equal 1, and the maximum is not unique. All values between $X_{n:n} - 1/2$ and $X_{1:n} + 1/2$ are MLE's of θ.

Despite the above-mentioned shortcomings, maximum likelihood estimates are reasonable in most cases appearing in practice. We will continue with the example of the MLE in the case of a fixed number of Bernoulli trials.

■ **EXAMPLE 12.31**

Let us consider, as in Example 12.26, five Bernoulli trials with probability of success θ. The information on which trials led to success and which led to failures is not essential: what matters is the total number S of successes. The likelihood function for $S = s$ is

$$L(\theta, s) = \theta^s (1 - \theta)^{5-s}.$$

For $0 \leq \theta \leq 1$, the function $L(\theta, s)$ is maximized at $s/5$ if $s = 1, 2, 3$, or 4, as can be seen by taking derivatives and solving the equation $L'(\theta) = 0$. For $s = 0$, we have $L(\theta, 0) = (1 - \theta)^5$, and the maximum occurs at $\theta = 0$. Similarly, for $s = 5$, we have $L(\theta, 5) = \theta^5$, and the maximum occurs at $\theta = 1$. Thus the MLE of θ, for given s, is $s/5$ (although for $s = 0$ and $s = 5$ the maximum occurs at the boundary and therefore cannot be established by taking derivatives).

Graphically the six likelihood functions corresponding to various numbers s of successes are presented in Figure 12.2. If we now regard s as a value of random variable S, then the likelihood function $L(\theta, S)$ becomes random. The corresponding point at which $L(\theta, 5)$ attains its maximum equal to $S/5$, is also random. Thus the MLE becomes a random variable, dependent on the sample $(X_1, \ldots, X_n)$ through the statistic $S = X_1 + \cdots + X_5$.

We will regard the likelihood function as a random function of θ, the randomness being induced by the sample $X_1, \ldots, X_n$. The value $\hat{\theta}$ that maximizes the likelihood will be called a *maximum likelihood estimator* of θ. Following tradition, we will use the same MLE symbol for both the maximum likelihood estimator (random variable) and for the maximum likelihood estimate (its value for a particular sample).

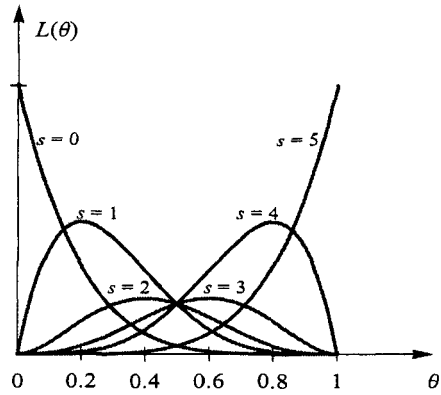


Figure 12.2 Likelihood function for five Bernoulli trials

Let us begin by stating some properties of maximum likelihood estimators. The property that makes these estimators quite convenient is known as *invariance*.

Invariance Principle *If* $\hat{\theta}$ *is the* MLE *of parameter* θ, *then* $h(\hat{\theta})$ *is the* MLE *of parameter* $h(\theta)$.

Let us begin with the case where the mapping h is one-to-one. Then there exists the inverse mapping g of Θ' onto Θ such that $\theta' = h(\theta)$ whenever $\theta = g(\theta')$. If the likelihood $L(\theta, \mathbf{x})$ is maximized at the point $\hat{\theta} = \hat{\theta}(\mathbf{x})$, then the function $L(g(\theta'), \mathbf{x})$ is maximized when $g(\theta') = \hat{\theta}(\mathbf{x})$, and hence when $\theta' = h[\hat{\theta}(\mathbf{x})]$. Perhaps the most common application of the invariance principle is the fact that if $\hat{\theta}$ is the MLE of the variance σ^2, then $\sqrt{\hat{\theta}}$ is the MLE of the standard deviation σ.

The invariance principle is valid in the case of multidimensional parameters, and one-dimensional functions of such parameters. Consider the following example:

■ **EXAMPLE 12.32**

Suppose that the sample is taken from a distribution $f(x; \mu, \sigma)$, where μ and σ are the mean and standard deviation (we may consider f as normal, but it is not necessary). Thus $\theta = (\mu, \sigma)$ is a two-dimensional parameter. Assume that the parameter space is $\Theta = \{\mu > 0, \sigma > 0\}$. Suppose that we want to estimate the coefficient of variation $\nu = \sigma/\mu$.

The invariance principle asserts that if $\hat{\mu}$ and $\hat{\sigma}$ are MLE's of μ and σ, then $\hat{\nu} = \hat{\sigma}/\hat{\mu}$ is the MLE of the coefficient of variation. It should be recognized, however, that this conclusion does not follow from previous reasoning, since the function that maps $\theta = (\mu, \sigma)$ into $h(\theta) = \sigma/\mu$ is not one-to-one.

The argument showing that the invariance principle for the MLE's is also valid in the multidimensional case is as follows: Assume that $\theta = (\eta_1, \eta_2, \ldots, \eta_m)$ is an m-dimensional parameter, and let $h(\theta) = h(\eta_1, \eta_2, \ldots, \eta_m)$ be a function of θ to be estimated. Then find $m - 1$ functions $h_2(\theta) = h_2(\eta_1, \eta_2, \ldots, \eta_m), \ldots, h_m(\theta) = h_m(\eta_1, \eta_2, \ldots, \eta_m)$ such that the vector

$$H(\theta) = (h(\theta), h_2(\theta), \ldots, h_m(\theta))$$

is a one-to-one mapping of m-dimensional parameter space Θ into a subset Θ' of the m-dimensional space. By the preceding argument, if $\hat{\theta} = (\hat{\eta}_1, \ldots, \hat{\eta}_m)$ is the MLE of θ, then

$$H(\hat{\theta}) = (h(\hat{\eta}_1, \ldots, \hat{\eta}_m), h_2(\hat{\eta}_1, \ldots, \hat{\eta}_m), \ldots, h_m(\hat{\eta}_1, \ldots, \hat{\eta}_m))$$

is the MLE of $H(\theta)$. It follows therefore that $h(\hat{\eta}_1, \ldots, \hat{\eta}_m) = h(\hat{\theta})$ is the MLE of $h(\theta)$.

The second important property of maximum likelihood estimators is that they do not depend on the design of the experiment. To explain the issues involved here, we start by formulating the likelihood principle.

Likelihood Principle *Consider two sets of data,* $\mathbf{x}$ *and* $\mathbf{y}$*, obtained from the same population, although possibly according to different sampling plans. If the ratio of their likelihoods,* $L_1(\theta, \mathbf{x})/L_2(\theta, \mathbf{y})$*, does not depend on* θ*, then both data sets provide the same information about the parameter* θ *and consequently should lead to the same conclusion about* θ*.*

Consequently, what matters are the ratios of likelihoods rather than the likelihoods themselves. We will explain by some examples how this principle makes the MLE's independent on the design of experiment.

■ **EXAMPLE 12.33**

Assume that we want to estimate θ, the probability of success in Bernoulli trials. One experimental design consists of fixing the number n of observations and recording the number of successes. If we observe x successes, the likelihood is $L_1(\theta, x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$.

Suppose now that one decides to fix x and take observations until x successes are recorded. Now the probability that the observations will end on the nth trial is given by negative binomial distribution, so the likelihood is

$$L_2(\theta, n) = \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}.$$

Observe that in the first case x was random and n fixed; in the second it is the other way around. Nevertheless, the ratio of these two likelihoods is $\binom{n}{x}/\binom{n-1}{x-1}$, which does not depend on θ. The MLE of θ is $\hat{\theta} = x/n$ in either case. Therefore the additional information that in the second case the last experiment led to success does not affect our estimate of θ.

■ **EXAMPLE 12.34**

Consider tennis players A and B who from time to time play a match against each other. Let us assume that the probability of A winning a set against B is p, and the results of sets are independent (This is a somewhat oversimplified assumption; probability p may change over time, and results of sets within a match may be dependent. We will, however, take this assumption as a starting point for analysis.) We want to estimate the probability p, which reflects the relative strengths of players A and B.

SOLUTION. Assume that A and B are men.[23] Therefore the probability that A will win in 3 sets is $\alpha_3 = p^3$. The probability that he will win in 4 sets is $\alpha_4 = 3p^3(1-p)$, since he must win the last set (probability p) and two of the first three sets [probability $\binom{3}{2}p^2(1-p)$]. Similarly the probability of winning in 5 sets is $\alpha_5 = 6p^3(1-p)^2$. The analogous probabilities of winning by player B in 3, 4, and 5 sets are β_3, β_4, and β_5, obtained by interchanging the roles of p and $1-p$.

Now assume that the sport section of a newspaper provides data for the last year, in the form of six numbers $(a_3, a_4, a_5, b_3, b_4, b_5)$, where a_i is the number of matches won by A (against B) in i sets, and similarly for b_i. Letting $q = 1-p$, the likelihood of the data is

$$\begin{aligned} L(p; a_3, \ldots, b_5) &= \alpha_3^{a_3}\alpha_4^{a_4}\alpha_5^{a_5}\beta_3^{b_3}\beta_4^{b_4}\beta_5^{b_5} \\ &= (p^3)^{a_3}(3p^3q)^{a_4}(6p^3q^2)^{a_5}(q^3)^{b_3}(3q^3p)^{b_4}(6q^3p^2)^{b_5} \\ &= Cp^{3a_3+3a_4+3a_5+b_4+2b_5}q^{3b_3+3b_4+3b_5+a_4+2a_5} \\ &= Cp^aq^b, \end{aligned}$$

where C is a constant independent of p, while a and b are the total numbers of sets won only by A and by B, respectively. We obtain

$$\log L = \log C + a\log p + b\log(1-p),$$

and the MLE of p can be easily found as

$$\hat{p} = \frac{a}{a+b} = \frac{\text{number of sets won by } A}{\text{number of sets played}}.$$

Observe that the number of matches, equal to $a_3 + a_4 + a_5 + b_3 + b_4 + b_5$, does not appear in the estimator of p. The fact that the last set in each match must be won by the winner of the match is not relevant here: the estimator is the same whether or not we fix in advance the number of sets (not matches!) to be played, and record the number of sets won by A.

Still another important property of maximum likelihood estimators is that they can be obtained from samples in which some data are only partially observable. To illustrate this property, we will find the MLE from a censored sample.

[23] This assumption is important since men play "best out of five sets," whereas women play "best out of three sets."

■ **EXAMPLE 12.35**

Suppose that we want to estimate parameter λ in the EXP(λ) distribution. For convenience of terminology, let us interpret the observations as lifetimes of some pieces of equipment. A typical experiment consists here of putting n pieces of the equipment to test and observing the lifetimes $X_1, X_2, \ldots$. Suppose that the experiment is interrupted after some time T, and some tested items are still working. The data then have the form of a certain number of values $X_1, \ldots, X_m$ of observed lifetimes, while about the remaining $n - m$ items we know only that $X_{m+1} > T, \ldots, X_n > T$.

The likelihood of such a sample is obtained by multiplying the values of density function at points $X_1, \ldots, X_m$ and the probabilities $P\{X_i > T\} = e^{-\lambda T}$ for $i = m + 1, \ldots, n$. Therefore the likelihood is

$$L(\lambda) = \lambda e^{-\lambda X_1} \times \cdots \times \lambda e^{-\lambda X_m} \left(e^{-\lambda T}\right)^{n-m} = \lambda^m e^{-\lambda[X_1 + \cdots + X_m + (n-m)T]}.$$

Since

$$\log L(\lambda) = m \log \lambda - \lambda[X_1 + \cdots + X_m + (n - m)T],$$

the MLE of λ is easily found to be

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{m} X_i + (n - m)T}. \tag{12.54}$$

Finally, we will discuss consistency, asymptotic unbiasedness, and asymptotic efficiency of MLE's. We will not prove, or even formulate, the theorem, as it lies beyond the scope of this book. It is not difficult, however, to provide an informal explanation why MLE's, under some regularity assumptions, have the properties asserted.

The main idea is as follows: In Section 12.4 we introduced the function $J(X, \theta) = \frac{\partial}{\partial \theta}[\log f(X, \theta)]$, used to define the information $I(\theta)$ and the concept of efficient estimators. If $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ is a random vector from distribution $f(x, \theta)$, then $J(\mathbf{X}^{(n)}, \theta) = \sum_{i=1}^{n} J(X_i, \theta)$, with components $J(X_i, \theta)$ being iid random variables. We know from Section 12.4 that $E[J(X_i, \theta)] = 0$ and $\mathrm{Var} J(X_i, \theta) = I(\theta)$. Consequently $E[J(\mathbf{X}^{(n)}, \theta)] = 0$ and $\mathrm{Var}[J(\mathbf{X}^{(n)}, \theta)] = nI(\theta)$. By the central limit theorem of Lindeberg-Lévy (Theorem 10.6.1), we have

$$\frac{J(\mathbf{X}^{(n)}, \theta)}{\sqrt{nI(\theta)}} \xrightarrow{d} Z, \tag{12.55}$$

where Z is the standard normal random variable.

Let $T = T(\mathbf{X}^{(n)})$ be an unbiased efficient estimator of θ so that $E_\theta(T) = \theta$ and $\mathrm{Var}_\theta(T) = 1/nI(\theta)$. We know also by (12.46) that

$$T = \gamma_1(\theta) J(\mathbf{X}^{(n)}, \theta) + \gamma_2(\theta),$$

so

$$E_\theta(T) = \gamma_1(\theta) E_\theta[J(\mathbf{X}^{(n)}, \theta)] + \gamma_2(\theta) = \gamma_2(\theta)$$

and

$$\mathrm{Var}_\theta(T) = \gamma_1^2(\theta)E[J(\mathbf{X}^{(n)},\theta)]^2 = \gamma_1^2(\theta)\mathrm{Var}[J(\mathbf{X}^{(n)},\theta)] = \gamma_1^2(\theta)nI(\theta).$$

Consequently, we must have

$$\gamma_2(\theta) = \theta, \qquad \gamma_1^2(\theta) = \frac{1}{n^2I^2(\theta)},$$

which implies that $\gamma_1(\theta) = \pm 1/nI(\theta)$. It follows now that

$$T = \pm\frac{1}{nI(\theta)}J(\mathbf{X}^{(n)},\theta) + \theta,$$

which means that

$$\frac{J(\mathbf{X}^{(n)},\theta)}{\sqrt{nI(\theta)}} = \pm\sqrt{nI(\theta)}(T-\theta). \tag{12.56}$$

The left-hand side of (12.56) has an asymptotic standard normal distribution by (12.55), and the same must be true for the right-hand side (under any choice of sign).

Consequently, we proved the following theorem:

Theorem 12.5.1 *If $T = T(X_1,\ldots,X_n)$ is an efficient estimator of θ based on a random sample from the distribution $f(x,\theta)$, then the random variable*

$$\sqrt{nI(\theta)}(T-\theta)$$

asymptotically has the standard normal distribution.

Actually, the same theorem holds if efficient estimators T are replaced by maximum likelihood estimators $\hat{\theta}_n$ of θ, provided that these estimators are obtained by solving the equation

$$\frac{\partial L(\theta,\mathbf{X}^{(n)})}{\partial\theta} = J(\mathbf{X}^{(n)},\theta) = 0$$

so that

$$J(\mathbf{X}^{(n)},\hat{\theta}_n) = 0. \tag{12.57}$$

The proof requires imposing a number of technical assumptions, and is quite complicated, but the main idea is simple. We first expand the function $J(\mathbf{X}^{(n)},\theta)$ about the point $\hat{\theta}_n$ so that

$$\begin{aligned} J(\mathbf{X}^{(n)},\theta) &= J(\mathbf{X}^{(n)},\hat{\theta}_n) + (\theta-\hat{\theta}_n)\left.\frac{\partial J}{\partial\theta}\right|_{\theta=\hat{\theta}_n} \\ &\quad + \frac{1}{2}(\theta-\hat{\theta}_n)^2\left.\frac{\partial^2 J}{\partial\theta^2}\right|_{\theta=\hat{\theta}_n} + \cdots \end{aligned}$$

We want to show that the second order term converges to 0 in probability so that it can be neglected in the limit, and the same is true for the sum of all higher order terms. Using (12.57), we then write

$$J(\mathbf{X}^{(n)},\theta) \approx (\theta-\hat{\theta}_n)\left.\frac{\partial J}{\partial\theta}\right|_{\theta=\hat{\theta}_n}.$$

The left-hand side is now a sum of iid components, as in the proof of Theorem 12.5.1, and it obeys the central limit theorem. Finally, studying the behavior of the derivatives

$$\left.\frac{\partial J(\mathbf{X}^{(n)}, \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_n},$$

we establish that Theorem 12.5.1 holds also with T replaced by estimators $\hat{\theta}_n$ (with the same norming constants). This shows that $\hat{\theta}_n$ are asymptotically unbiased, asymptotically efficient, and have an asymptotic normal distribution with mean θ and variance $1/nI(\hat{\theta}_n)$. These properties make maximum likelihood estimators a preferred choice.

Bayesian Estimators

In both methods of constructing estimators presented thus far, there was no provision for using any information about θ that could be available prior to taking the sample. We will now describe a possible setup that is a special case of the setup considered thus far.

As before, the observations $X_1, \ldots, X_n$ are assumed to be a random sample from the distribution $f(x, \theta)$, and the objective is to estimate the value of θ. This time, however, assume that θ is an element of some parameter space Θ, sampled according to probability distribution $\pi(\theta)$. For instance, if Θ is the real line, and Ξ is a continuous random variable with density $\pi(\theta)$, then $P\{\Xi \in A\} = \int_A \pi(\theta)d\theta$.

These assumptions describe the situation where the considered statistical problem of estimation is just one of a series of analogous estimation problems that differ by the value of the estimated parameter θ, and the variability of the parameter over similar estimation problems is described by density π, referred to as the *prior* density of θ.

■ **EXAMPLE 12.36**

A grocery store receives shipments of some merchandise, (e.g., oranges). The quality of each shipment can be described by a certain parameter θ.

The examples of such parameters may be the percentage of spoiled fruits, the average diameter of a fruit in shipment, the average sugar content in the juice, and so on. Typically it is not practical to determine the value of θ for a given shipment exactly, because of cost, or because sampling is destructive. In such cases one takes a sample from a given shipment and uses it to estimate θ for this shipment.

If the store has been buying the shipments from the same company for some time, it usually accumulated some data about variability of values of θ from different shipments. It is natural, therefore, that such historical data, in addition to the actual sample, should be used in assessment of the parameter θ for a given shipment.

■ **EXAMPLE 12.37**

To some, the example of purchasing oranges may appear not serious enough to justify introducing a new concept (in fact, a new branch of statistics!). Con-

sider therefore the case of a physician who has to diagnose a patient. The data are various symptoms (or their lack), results of tests, and so on. The values of the parameter θ are different possible diagnoses, such as θ_1= common cold, θ_2 = tuberculosis, and θ_3 = AIDS. The physician knows the distributions $f(\mathbf{x}, \theta)$ of the results $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of n specific tests for every θ (here x_1 may be temperature, x_2 and x_3 may be systolic and diastolic blood pressure, x_4 may be glucose level, etc.). At some time the physician reaches a diagnosis (i.e., reaches the decision, perhaps tentative, about θ and begins treatment). He may also order further tests. Estimation problems considered thus far vary only by the nature of the parameter space and the character of data $x_1, x_2, \ldots, x_n$, which may be qualitative. Observe at this point how the discrete and qualitative nature of the parameter space affects the problem in the sense of definition of error: while in estimation the error can be quantified as the difference $T(\mathbf{x}) - \theta$, in medical diagnosis such quantification is not possible (e.g., consider the "error" of treating a patient with TB as having a common cold, and vice versa, treating common cold as TB).

It is clear, however, that the experience and intuition of the diagnosing physician play a crucial role here. In particular, the physician may have some idea as to the prior probabilities $\pi(\theta)$ of various diseases.

The information from the sample, $\mathbf{x} = (x_1, \ldots, x_n)$, and prior information about the distribution of θ can be combined into the conditional density of θ given $\mathbf{x}$, referred to as the *posterior* density. We have here

$$P\{\Xi \in A | \mathbf{X} = \mathbf{x}\} = \int_A \pi(\theta|\mathbf{x}) d\theta,$$

where

$$\pi(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}; \theta)\, \pi(\theta)}{\int_\Theta f_n(\mathbf{x}; u)\, \pi(u)\, du}. \tag{12.58}$$

Actually, the integral in the denominator of (12.58) is the reciprocal of the normalizing constant, and we often do not need to determine its value. What really matters is the fact that the posterior density of the parameter given the data is proportional to the likelihood of the data multiplied by the prior density:

$$\pi(\theta|\mathbf{x}) = C f_n(\mathbf{x}, \theta)\, \pi(\theta) = C L(\theta, \mathbf{x})\, \pi(\theta).$$

■ **EXAMPLE 12.38**

Suppose that we take observations from independent Bernoulli trials with the same probability of success θ, while θ follows a beta distribution with parameters α and β, that is,

$$\pi(\theta) = C\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 \le \theta \le 1, \tag{12.59}$$

where C is the normalizing constant, given in (9.67). Letting $s = \sum x_i$, we have

$$f_n(\mathbf{x}, \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^s (1-\theta)^{n-s}.$$

Consequently, the posterior density of θ, given $\mathbf{x}$, is

$$\pi(\theta|\mathbf{x}) \simeq K\theta^s(1-\theta)^{n-s}\theta^{\alpha-1}(1-\theta)^{\beta-1} = K\theta^{\alpha+s-1}(1-\theta)^{n+\beta-s-1},$$

which we recognize as a BETA($\alpha+s, \beta+n-s$) distribution.

■ **EXAMPLE 12.39**

Assume that the observations $X_1, X_2, \ldots, X_n$ are independent, each with POI(λ) distribution, while θ has a GAM(α, λ) distribution so that

$$\pi(\theta) = C\theta^{\alpha-1}e^{-\lambda\theta}, \quad \theta > 0, \tag{12.60}$$

where C is the normalizing constant given in (9.49). Letting $s = \sum x_i$, the posterior density of θ is

$$\pi(\theta|\mathbf{x}) = \left[\prod_{i=1}^{n}\left(\frac{\theta^{x_i}}{x_i!}e^{-\theta}\right)\right] C\theta^{\alpha-1}e^{-\lambda\theta} = K\theta^{s+\alpha-1}e^{-(n+\lambda)\theta}. \tag{12.61}$$

Here the constant K also incorporates the factorials $x_1! \cdots x_n!$ (in general, we need to keep track of only the terms involving θ). We recognize (12.61) as a gamma distribution with a shape parameter $\alpha+s$ and a scale parameter $\lambda+n$.

■ **EXAMPLE 12.40**

Suppose that $X_1, \ldots, X_n$ is a random sample from N(θ, σ^2) distribution, with σ^2 known. Regarding θ, it has a prior normal distribution N(μ, τ^2). We will find the posterior distribution of θ, given the sample.

SOLUTION. The likelihood function can be transformed as follows:

$$\begin{aligned} f_n(\mathbf{x}, \theta) &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\right\} \\ &\simeq \exp\left\{-\frac{1}{2\sigma^2}\left[\sum(x_i-\overline{x}_n)^2 + n(\theta-\overline{x}_n)^2\right]\right\} \\ &\simeq \exp\left\{-\frac{n}{2\sigma^2}(\theta-\overline{x}_n)^2\right\}. \end{aligned}$$

Here the sign $\simeq$ means proportionality up to constants that do not depend on any unknown parameters—such dropped constants are $\sigma^{-n}(2\pi)^{-n/2}$ in the first proportionality sign, and $\exp\{-(n/2\sigma^2)(x_i-\overline{x}_n)^2\}$ in the second proportionality sign. The posterior density $\pi(\theta|\mathbf{X})$ is therefore proportional to

$$\exp\left\{-\frac{n}{2\sigma^2}(\theta-\overline{x}_n)^2\right\} \times \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\}.$$

Now we need to separate the terms involving θ; all other terms will be absorbed in the proportionality constant. Expanding the squares and leaving only terms with θ, we obtain, after some algebra,

$$\pi(\theta|\mathbf{x}) \simeq \exp\left\{-\frac{1}{2q^2}(\theta-m)^2\right\},$$

where

$$m = \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu + \frac{n\tau^2}{\sigma^2 + n\tau^2}\overline{x}_n \tag{12.62}$$

and

$$q^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}. \tag{12.63}$$

This means that the posterior density of θ is again normal with mean m and variance q^2 given by (12.62) and (12.63), respectively.

■ **EXAMPLE 12.41**

Suppose finally that the observations $X_1, \ldots, X_n$ are independent normal $N(0, \theta^2)$, with θ again having a gamma distribution as in Example 12.39. We have

$$\pi(\theta|\mathbf{x}) = K\left(\prod_{i=1}^{n} \frac{1}{\theta\sqrt{2\pi}} e^{-\frac{x_i^2}{2\theta^2}}\right)\theta^{\alpha-1}e^{-\lambda\theta} = K^*\theta^{\alpha-n-1}e^{-\lambda\theta-\frac{\sum x_i^2}{2\theta^2}}. \tag{12.64}$$

This time the density (12.64) is not a member of any known family of distributions.

The following definition is related to the situation in Examples 12.38 through12.40 (as opposed to that in Example 12.41):

Definition 12.5.2 A family $\mathcal{F}$ of distributions of parameter θ is said to provide *conjugate* priors for the distribution $f(x, \theta)$ of observations if, whenever the prior distribution of θ is in $\mathcal{F}$, the posterior density $\pi(\theta|\mathbf{x})$ also belongs to $\mathcal{F}$ for any sample $\mathbf{x}$. □

The essence of Examples 12.38 through 12.40 is formulated in the following theorems:

Theorem 12.5.2 *Beta densities are conjugate priors for the binomial distribution.*

Theorem 12.5.3 *Gamma densities are conjugate priors for the Poisson distribution.*

Theorem 12.5.4 *Normal densities are conjugate priors for normal distribution.*

We are now in a position to define Bayes estimators. The choice of the estimator depends, naturally, on the loss function. We will present the general theory, and then focus on the special case of the squared loss function.

In general, as in Section 12.3, the penalty for accepting the value of the parameter as θ^*, while in fact it is θ, is expressed by the loss function $\mathcal{L}(\theta^*, \theta)$. In the present case we know the distribution of θ, given the observed sample $\mathbf{x}$. Thus we should choose the value θ^* so as to minimize the expected loss

$$E\{\mathcal{L}(\theta^*, \Xi)|\mathbf{x}\} = \int_{\Theta} \mathcal{L}(\theta^*, \theta)\,\pi(\theta|\mathbf{x})\,d\theta. \tag{12.65}$$

Note that the left-hand side of (12.65) *does not depend on* θ. It is a function of θ^* and $\mathbf{x}$. For each $\mathbf{x}$ we can therefore try to minimize it, that is, find a value $\theta^*(\mathbf{x})$ such that

$$E\{\mathcal{L}(\theta^*(\mathbf{x}), \Xi)|\mathbf{x}\} \leq E\{\mathcal{L}(\theta^{**}, \Xi)|\mathbf{x}\}$$

for every θ^{**} in Θ. Such a value $\theta^*(\mathbf{x})$ is the best choice of estimate given the sample $\mathbf{x}$. When $\mathbf{x}$ varies according to the distribution $f(\mathbf{x}, \theta)$, we obtain a statistic $\theta^*(\mathbf{X})$. We will now introduce the following definition:

Definition 12.5.3 The statistic $\theta^*(X)$ minimizing the left-hand side of (12.65) is called the *Bayes estimator of* θ *for the loss function* $\mathcal{L}(\theta^*, \theta)$. □

Minimizing (12.65) in the general case may not be easy. However, when $\mathcal{L}(\theta^*, \theta) = (\theta^* - \theta)^2$, the minimizing value θ^* is well known. It is the mean of the posterior distribution of θ given $\mathbf{x}$. As in the case of other estimators, we will tacitly take the squared error as the loss function (unless we explicitly specify some other loss function). This is the customary choice. It is motivated primarily by the fact that the quadratic loss function allows further development of the theory. Therefore, *if the loss function is not explicitly specified as other than quadratic, then the Bayes estimator of parameter* θ *is understood as the mean of posterior distribution:*

$$T(\mathbf{X}) = E\{\Xi|\mathbf{X}\} = \int \theta\pi(\theta|\mathbf{X})d\theta. \tag{12.66}$$

■ **EXAMPLE 12.42**

If the observations $\mathbf{X} = (X_1, \ldots, X_n)$ form a random sample from a Bernoulli distribution (so that $S = \sum_{i=1}^{n} X_i$ is BIN(n, θ)), and θ varies according to BETA(α, β), then the Bayes estimator of θ is

$$T(X_1, \ldots, X_n) = \frac{\alpha + S}{\alpha + \beta + n}. \tag{12.67}$$

This follows from the fact, established in Example 12.38, that the posterior distribution of θ is BETA$(\alpha+S, \beta+n-S)$ with the mean $(\alpha+S)/(\alpha+\beta+n)$.

■ **EXAMPLE 12.43**

Recall from Example 9.20 that Laplace estimated the probability that the sun will rise tomorrow if it is known to have done so on the past n consecutive days. The observations form a binomial random variable S, with all trials resulting in success, so $S = n$. The Laplace estimate $(n + 1)/(n + 2)$ of θ is the Bayes estimate for the uniform prior distribution, which is BETA$(1, 1)$.

■ **EXAMPLE 12.44**

In a similar way we can use Theorem 12.4.1, which asserts that gamma densities are conjugate priors for Poisson distribution. If $X_1, \ldots, X_n$ is a sample from POI(θ) distribution, and θ has prior density gamma with parameters α

and λ, then the Bayes estimator of θ is

$$T(X_1, \ldots, X_n) = \frac{\alpha + S}{\lambda + n}. \tag{12.68}$$

This follows from the fact that the mean of a gamma distribution is the ratio of the shape parameter to the scale parameter.

■ **EXAMPLE 12.45**

From (12.62) in Example 12.40 it follows that the Bayes estimator in case of the normal distribution $N(\theta, \sigma^2)$, with known σ^2 and normal prior $N(\mu, \tau^2)$, is

$$T = \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu + \frac{n\tau^2}{\sigma^2 + n\tau^2}\overline{X}_n. \tag{12.69}$$

Thus T is a weighted average of prior mean μ and sample mean $\overline{X}_n$. With the increase of the sample size n, the estimator puts more and more weight to the sample size $\overline{X}_n$, and in the limit we have $T = \overline{X}_n$. In other words, ultimately the empirical evidence always prevails over prior convictions. Finally, the same limiting conclusion is obtained if $\tau^2 \to \infty$, that is, when the prior information is more vague (the prior variance τ^2 is interpretable as a measure of uncertainty of the prior information). Going toward the opposite extreme, where $\sigma^2 \to \infty$ or $\tau^2 \to 0$ (observations are subject to large errors, or prior knowledge has high certainty), the Bayes estimator of θ attaches more and more weight to the prior information that the mean of θ is μ.

The last three examples illustrate the convenience of using conjugate prior distributions. This convenience lies basically in the fact that we have a simple formula for the Bayes estimator, and therefore we can quickly adjust our estimates when new observations become available. To illustrate this point, we consider the following example:

■ **EXAMPLE 12.46**

Continuing Example 12.44, assume that we first take m observations, with the total number of successes $S = X_1 + \cdots + X_m$, and then n observations, with the total number of successes $S' = X_1' + \cdots + X_n'$. The prior density of θ, before taking the first set of observations, is BETA(α, β).

The situation at the end of second series of observations can be regarded in two ways:

1. We have the total of $S + S'$ successes in $m + n$ trials, so the Bayes estimator is

$$\frac{\alpha + (S + S')}{\alpha + \beta + (m + n)}. \tag{12.70}$$

2. We have the total of S' successes in the last n trials, where the new prior distribution becomes a posterior distribution after we take the first series of observations. That is, the posterior density is beta with parameters $\alpha_1 = \alpha + S$ and

$\beta_1 = \beta + m - S$. The Bayes estimator is now

$$\frac{\alpha_1 + S'}{\alpha_1 + \beta_1 + n},$$

the same as (12.70).

Essentially what we observed here is an instance of a general theorem on updating evidence (Theorem 4.4.2), which says that if we have two independent (sets of) observations $\mathbf{x}$ and $\mathbf{x}'$, we can use them either "at once," to determine the posterior distribution of θ given $(\mathbf{x}, \mathbf{x}')$, or we can do it in steps. That is, we find the posterior density of θ, given one data set (e.g., $\mathbf{x}$), and use it as a new prior to find the posterior density of θ given $\mathbf{x}'$. The results will still be the same. Moreover, the order of choice between $\mathbf{x}$ and $\mathbf{x}'$ is irrelevant, and does not need to coincide with the order in which the data $\mathbf{x}$ and $\mathbf{x}'$ were collected.

All the examples above concern estimators for the squared error loss function. If the loss function is $\mathcal{L}(\theta^*, \theta) = |\theta^* - \theta|$, then the Bayes estimator is the *median* of the posterior distribution (see Theorem 8.6.3). For the normal case (Example 12.45), the mean and median coincide; hence (12.69) is also the Bayes estimator for the absolute error loss. For Examples 12.42, 12.43, and 12.44, the posterior distributions are beta or gamma, and their medians are not expressible by simple formulas in term of the parameters.

It is clear that the class of *all* distributions on the parameter space Θ is always a set of conjugate priors for any distributions. Such a statement, however, is totally pointless. In fact, a class of conjugate priors is useful only if it leads to a simple formula for the posterior density, allowing explicit formulas for the means, hence for Bayesian estimators against mean squared loss. In this perspective the fact that a given class of distributions is a class of conjugate priors is just a mathematical curiosity without much significance. Indeed, referring again to Example 12.46, we consider there a situation where observations $S = \sum X_i$ are binomial with parameter θ, while θ has a beta distribution. The first assumption is defensible: we often can *make* S have a binomial distribution by using an appropriate sampling scheme. But the law that governs the variability of θ from case to case is beyond our control, and the class of situations described above, where the parameter θ is a value of a random variable Ξ with distribution π, is rather restricted. Most often, the situation that a statistician faces is "one of a kind," characterized by an unknown value of θ and it does not make sense to think of a prior distribution π of θ as telling us "how often" we had analogous statistical problems in which the value of the parameter satisfied the inequality of the form $a \leq \theta \leq b$.

There is a view, accepted by some statisticians, that even in such "one of a kind" situations it makes sense to consider and use the prior distribution of the parameter. Actually, the issue of whether one allows using prior distributions—even if they do not represent frequencies of occurrence of situations characterized by some values of θ—is the issue dividing statisticians into Bayesians and non-Bayesians.

The philosophical points of this division are beyond the scope of this book. One could, however, consider the following two competing principles.

1. Statistical conclusions should depend on data only. When two statisticians analyze the same data using the same method, they should reach the same conclusion.

2. Statistical conclusions may depend on the experience, intuition, and insight of the statistician who analyzes a given set of data.

Very roughly, statisticians who adhere strictly to principle 1 are non-Bayesians, and those who favor 2 are Bayesians. The latter use the prior distribution π as a means to express their knowledge, intuition, and so forth. In this respect, having a class of conjugate priors is usually of great help, primarily to express one's own prior experience or convictions, or to elicit information about the analyzed problem from the practitioners whom they advise. However, the first of the examples below shows something more fundamental, namely that the concept of prior distribution, reflecting one's personal judgments about a "one-of-a-kind" case, is sometimes unavoidable. Consider the following situation, which without this concept appears paradoxical:

■ **EXAMPLE 12.47**

Imagine yourself playing the following game: There are two envelopes, each containing a check. The amount on one check is twice as big as the amount on the other. You choose an envelope and inspect the check. At this moment you are offered an option to choose the other envelope. What should you do?

SOLUTION. The standard reasoning is as follows: Let a be the amount on the check that was in the first envelope you selected. The other envelope then contains a check for either $2a$ or $a/2$, each with the same probability. Thus, if you change your decision, the expected outcome is

$$\frac{1}{2}(2a) + \frac{1}{2}\left(\frac{a}{2}\right) = \frac{5}{4}a > a, \tag{12.71}$$

which means that *you should always change* the envelope.

This may seem paradoxical, since money appears to be created out of nowhere, just by changing the decision. The explanation lies in the fact that calculation (12.71) of the expected value uses probabilities 0.5 that the other envelope contains checks for $2a$ or $a/2$. In fact, one should use here the conditional probabilities, given the observed value a, and these calculations involve prior probabilities.

Indeed, suppose that $\pi(x)$ is the prior probability that the envelopes contain checks for the amounts x and $2x$. Assuming that one has no clairvoyant abilities, and therefore always has the chance 0.5 of selecting an envelope with the lesser amount on a check, the unconditional probability of observing the amount a is $0.5\pi(a/2) + 0.5\pi(a)$. Given the observed amount a, the probability that the check in the other envelope is for the amount $a/2$ equals

$$\frac{0.5\pi\,(a/2)}{0.5\pi\,(a/2) + 0.5\pi(a)} = \frac{\pi\,(a/2)}{\pi\,(a/2) + \pi(a)}.$$

The analogous probability for the amount $2a$ is

$$\frac{0.5\pi(a)}{0.5\pi(a/2)+0.5\pi(a)} = \frac{\pi(a)}{\pi(a/2)+\pi(a)}.$$

But the condition

$$\frac{\pi(a/2)}{\pi(a/2)+\pi(a)} = \frac{\pi(a)}{\pi(a/2)+\pi(a)}$$

implies $\pi(a/2) = \pi(a)$, which cannot be satisfied for all a (regardless of whether π represents a density or a discrete probability distribution).

This argument refers to some prior distribution π on the possible amount on the lesser check. Whether this distribution has any frequential interpretation (referring to analogous games played before), or not (if such a game is played only once) is irrelevant here. The only way to escape the paradox is to realize that everyone has some idea as to the probable range of values that may appear on the checks in the game. If the check in hand shows a very "small" value, then the other is probably for a higher value. If the check in hand shows a very "high" value, the other is probably smaller. The concepts "very small," "very large," and "probably" are subjective here and refer to the player's idea about the distribution π.

This example provides a rather powerful argument for the need of the Bayesian approach to statistical problems.

■ **EXAMPLE 12.48**

A piece of rock (e.g., taken from the moon) is sent to a laboratory to determine its radioactivity level. Assume that the measurement is simply Geiger count N_t, recorded for certain time t. The role of the number of observations n is now played by observation time t. (If this feature should be confusing to the reader, assume simply that the experiment is run in such a way that Geiger counts $X_1, X_2, \ldots$ are recorded, where X_i is the total count in ith hour of observation. Then $N_t = X_1 + \cdots + X_t$ if t is an integer number of hours.)

We know that N_t has a Poisson distribution with mean θt, where θ is the radiation intensity expressed in average number of emissions per hour. Suppose that the initial estimate of θ is needed urgently, so that observations can be carried out only for a limited time T. In other words, we have at our disposal a single observation of N_t. The likelihood function here is

$$L(\theta) = \frac{(\theta T)^{N_T}}{(N_T)!} e^{-\theta T},$$

so

$$\log L(\theta) = C + N_T \log\theta - \theta T,$$

and the MLE is easily seen to be

$$\hat{\theta} = \frac{N_T}{T}$$

(note that this result concerns the MLE in the case of observations running continuously in time; we no longer have sample of size n). To fix the ideas, assume that a total of 50 counts was recorded in $T = 100$ hours of observations so that the MLE of θ is 0.5.

Suppose now that there are two physicists in the laboratory, and each has his or her own ideas about what the radioactivity level θ of the specimen tested might be. Dr. Brown favors a certain theory of how the moon was formed, and how and when its rocks became initially radioactive. She thinks that moon rocks of the kind she analyzes should have their level of radioactivity θ about 1, but she is willing to incorporate a fair amount of uncertainty in her judgment, allowing the standard deviation of θ to be as much as 50% of the mean.

On the contrary, Dr. Smith favors a theory which predicts that the moon should have uniformly low radioactivity, say $\theta = 0.4$ on average, with standard deviation not exceeding 5% of the mean.

Let us see how these prior convictions will affect the estimates of θ for the specimen in question. We assume that the prior densities belong to the gamma family. If the parameters of a gamma distribution are α (shape) and λ (scale), then the mean is α/λ, and the variance is α/λ^2; see (9.50) and (9.51). Consequently, the ratio of the standard deviation to the mean (the so-called coefficient of variation) is

$$CV = \frac{\sqrt{\alpha/\lambda^2}}{\alpha/\lambda} = \frac{1}{\sqrt{\alpha}}.$$

Thus, for Dr. Brown, we have

$$CV = \frac{1}{\sqrt{\alpha}} = 0.5, \qquad \frac{\alpha}{\lambda} = 1,$$

which gives $\alpha = \lambda = 4$.

For Dr. Smith, we have

$$CV = \frac{1}{\sqrt{\alpha}} = 0.05, \qquad \frac{\alpha}{\lambda} = 0.4,$$

so $\alpha = 400, \lambda = 1000$.

The Bayes estimator is (see Example 12.44) $(\alpha + N_T)/(\lambda + T)$, so that

$$\hat{\theta}_B = \frac{4 + 50}{4 + 100} = 0.519 \quad \text{and} \quad \hat{\theta}_S = \frac{400 + 50}{1000 + 100} = 0.409$$

for Dr. Brown and Dr. Smith, respectively.

We can see the effects of two factors. One is that the MLE (in this case equal 0.5) is being "pulled" toward the mean of the prior distribution. The amount of pull depends on the variance of the prior distribution, reflecting the strength of conviction in the prior distribution: Dr. Smith, whose prior has much smaller variance, ends up with an estimate much closer to his prior mean than does Dr. Brown.

■ EXAMPLE 12.49 Are Birds Bayesians?

The example below concerns the behavior of certain species of birds in their search for food. The complete theory, including optimization aspects, has been developed by Green (1980). We present here only a fragment concerning assessment (estimation) by a bird.

Assume that the species in question can find food only in "patches," each consisting of a certain number of places where prey can be found. To fix the ideas, we will think of birds that prey on worms living in pine cones. We assume that a pine cone has n "holes," each of them containing a prey with probability θ, independently of other holes. Thus, given θ, the number of prey in a cone has BIN(n, θ) distribution. The bird has a fixed search pattern, so it does not search the same hole twice, and we assume that the prey cannot hide or escape to another hole during the search. Finally, we assume that θ, the rate of infestation of cones, varies between cones in such a way that θ has beta distribution with parameters α and β.

Let us consider now what could be the best strategy for a bird. First, it is reasonable to assume that the bird is trying to optimize[24] the rate of food intake per unit of time. Specifically, a strategy will tell the bird when to leave a cone and start searching the next one. The bird will optimize the rate of food intake, taking into account the average catch at a cone, average time spent on it, and the average time of flying to another cone.

Intuitively, if α and β are large, the variability of θ is small: the variance of the beta distribution is $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. In such cases, all cones are about the same: the variability between the cones is due mainly to variability in the binomial distribution with parameters n and $\theta = \alpha(\alpha+\beta)$. In such cases there is very little incentive to fly to the next cone before the current one is searched to the end.

However, if α and β are small, in particular, if $\alpha < 1$ and $\beta < 1$, the variability of θ is large. Actually, in the latter case, the distribution of θ has density unbounded at $\theta = 0$ and at $\theta = 1$. This means that most cones will be of two categories only: very rich in prey, when θ is close to 1, and very poor in prey, when θ is close to 0. The optimal strategy is then to assess—as quickly as possible—whether θ is close to 1 or to 0, and behave accordingly, leaving the cone in the second case.

Now, if after searching k holes the bird found x worms, its assessment of θ is $(\alpha+x)/(\alpha+\beta+k)$ (i.e., equals the Bayes estimate of θ given x successes in k trials). Without going into detail, the optimal strategy specifies, for each k, the threshold for x, below which the bird ought to leave the cone.

There is an empirical problem to determine whether or not the birds follow the optimal strategy. The experiments involve the use of artificial cones and observation of birds' behavior depending on the findings in the holes searched previously. The preliminary results indicate that birds follow some kind of

[24] We are using here a convenient terminology based on an analogy with humans. In reality, birds cannot be expected to solve optimization problems, which require computers for humans (see Green, 1980). What we mean here is that in the process of evolution, any mutation toward a better search strategy is likely to become established, and it is possible that birds use a strategy that is close to optimal.

strategy of breaking or continuing the search of a cone depending on the outcome of the search so far. Whether or not this is an optimal strategy is unclear. But the truly fascinating problem here is that if the birds use a Bayesian strategy, whether they are capable of changing the prior distribution. In other words, are birds born with knowledge of a search strategy that is optimal against some fixed α and β characterizing infestation of cones prevalent in the last hundred years (say), or can an individual bird change its search strategy in years of higher infestation of its habitat?

Let us investigate briefly the problem of consistency of Bayes estimators. First, let us observe that in the cases of estimators analyzed (of θ in the binomial case with a beta prior, of θ in the Poisson case with a gamma prior, and of μ in the normal distribution with a normal prior), as the sample size increases, the effect of prior distribution decreases to zero. Indeed, if S_n denotes the binomial random variable with probability of success θ, the Bayes estimator of θ for beta prior satisfies

$$\hat{\theta} = \frac{\alpha + S_n}{\alpha + \beta + n} = \frac{\alpha/n + S_n/n}{(\alpha + \beta)/n + 1} \xrightarrow{P} \theta,$$

since S_n/n converges to θ in probability (and also almost surely).

Similarly, if $X_1, \ldots, X_n$ is a random sample from the Poisson distribution with mean θ, and θ has a gamma distribution with parameters α and λ, then the Bayes estimator of θ satisfies

$$\hat{\theta} = \frac{\alpha + \sum_{j=1}^{n} X_j}{\lambda + n} = \frac{\alpha/n + \overline{X}_n}{\lambda/n + 1} \xrightarrow{P} \theta,$$

by the law of large numbers, which asserts that the sample average $\overline{X}_n$ converges to θ in probability (and also almost surely).

Finally, an analogous conclusion for the normal case has already been obtained. We see therefore that Bayes estimators are consistent; in fact they become increasingly closer to MLE's of the same parameters, regardless of the prior distribution. This property is true for Bayes estimators under some very general conditions, which we will not state here.

One of the problems that a statistician faces quite often is the determination of the sample size: "How big should n be in order that ...." Various conditions may appear in place of dots; in the case of estimation, these conditions typically state the precision of the estimate, in the sense of the probability of errors of a given size. In case of Bayes estimators the situation is relatively simple. One of the criteria for determining the sample size may be expressed through the posterior distribution. In the most typical case, one may wish to have a sample size that ensures that the posterior variance is below a certain minimum (of course, such criteria make sense only if the estimator used is unbiased or has a small bias).

Least Squares Estimators

This method of estimation is dating back to Lagrange and Gauss—for an interesting account of its discovery see Stigler (1986). The basic setup is now different—the data are independent but they are not coming from the same distribution. For a

given value u of some *independent* variable (random or not) U, observations (one or perhaps more) of some random variable Y_u are taken. It is assumed that

$$Y_u = Q(u) + \epsilon,$$

where $Q(u)$ is some function of u, and ϵ is a random variable (usually called *error*), such that $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2 < \infty$. It follows that

$$E(Y_u) = Q(u) \qquad \text{and} \qquad \text{Var}(Y_u) = \sigma^2.$$

The function $Q(u)$ is usually called the *regression* of Y on u.

■ **EXAMPLE 12.50**

One of the common situations falling under the scheme above arises when we analyze some system that changes in time. Thus, u is the time of taking the observation of some attribute of the system, and $Q(u)$ is the expected value of the observed random variable Y_u, interpreted also as the "true" state of the system.

In some cases, we may take only one observation at any given time u; in other cases, we may have a number of observations made at the same time.

■ **EXAMPLE 12.51**

In some cases variable U can be controlled by the experimenter. For example, a chemist may study the rate of a certain reaction in different temperatures. He then chooses the temperature level u_1 and observes the reaction rate one or more times. Then he changes the level to u_2, repeats the observations, and so on. The numbers of observations need not be the same for different values u. In general, the choice of distinct values $u_1, \ldots, u_k$ (as well as the choice of k) and the choice of numbers of observations $n_1, n_2, \ldots, n_k$ to be made at selected points $u_1, \ldots, u_k$, belongs to the *design* of the experiment.

Assume now that we have the experimental data for the design that can be described by the set of pairs

$$(u_i, n_i), \qquad i = 1, \ldots, k, \tag{12.72}$$

where u_i's are distinct values of variable U and n_i's are positive integers.

The data have form of the array $\{y_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i\}$, where y_{ij} is the value of random variable Y_{ij}, representing the jth observation for the value u_i of variable U. We assume that all Y_{ij}'s are independent, with

$$E(Y_{ij}) = Q(u_i) \qquad \text{and} \qquad \text{Var}(Y_{ij}) = \sigma^2. \tag{12.73}$$

In most typical cases, the functional form of Q is postulated and assumed to depend on some parameters. For instance, in the case of a linear regression model, we assume that

$$Q(u) = \beta u + \alpha, \tag{12.74}$$

where β and α are the slope and intercept of the regression line. In more complicated setups, we may postulate a quadratic regression

$$Q(u) = \beta_2 u^2 + \beta_1 u + \alpha,$$

or some other functional form of Q. Generally,

$$Q(u) = \varphi(u; \theta_1, \theta_2, \ldots, \theta_r),$$

where φ is a known function and $\theta_1, \ldots, \theta_r$ are some parameters to be estimated. The method of *least squares* is based on the quadratic form

$$S(\theta_1, \ldots, \theta_r) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [y_{ij} - \varphi(u_i; \theta_1, \ldots, \theta_r)]^2. \tag{12.75}$$

The values $\hat{\theta}_1, \ldots, \hat{\theta}_r$ that minimize S given by (12.75) are called least squares (LS) estimates of $\theta_1, \ldots, \theta_r$. As usual, those estimators, regarded as functions of the random variables $\{Y_{ij}\}$ are called *LS-estimators*.

The usual way of finding $\hat{\theta}_1, \ldots, \hat{\theta}_r$ is by solving the set of simultaneous equations

$$\frac{\partial S}{\partial \theta_l} = 0, \qquad l = 1, \ldots, r, \tag{12.76}$$

which in the present case take the form

$$\sum_{j=1}^{n_i} [y_{ij} - \varphi(u_i; \theta_1, \ldots, \theta_r)] \sum_{i=1}^{k} \frac{\partial \varphi(u_i; \theta_1, \ldots, \theta_r)}{\partial \theta_l} = 0. \tag{12.77}$$

■ **EXAMPLE 12.52 Linear Regression**

Suppose that $Q(u) = \beta u + \alpha$. In this case the algebra will simplify somewhat if we order all Y_{ij}'s into a single sequence $y_1, \ldots, y_n$, where $n = \sum_{i=1}^{k} n_i$, and relabel the corresponding values as $u_1, \ldots, u_n$ (n values u_i of variable U, not necessarily all distinct). The quantity to be minimized is

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_j - \beta u_i - \alpha)^2. \tag{12.78}$$

Differentiating with respect to α and β and setting the derivatives equal to 0, we obtain

$$\beta \sum_{i=1}^{n} u_i^2 + \alpha \sum_{i=1}^{n} u_i = \sum_{i=1}^{n} u_i y_i \quad \text{and} \quad \beta \sum_{i=1}^{n} u_i + n\alpha = \sum_{i=1}^{n} y_i.$$

The solution can be written as

$$\hat{\beta} = \frac{\sum (u_i y_i - \overline{u} \times \overline{y})}{\sum (u_i - \overline{u})^2}, \tag{12.79}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{u}, \tag{12.80}$$

provided that $\sum(u_i - \overline{u})^2 > 0$. The latter condition is ensured if there are at least two distinct values of u used in the experiment, that is, if not all observations are made for the same value of S.

An easy check shows that (12.79) does minimize function U.

Robust Estimators

An important issue in estimation problems is to obtain estimators, generally termed *robust*, that would be relatively unaffected by deviations from the assumed model. We will now briefly sketch two such approaches.

To grasp the main issues involved here, consider the problem of estimating the mean μ of a distribution. Then the sample mean $\overline{X}$ is an estimator of μ, and we know that $\overline{X}$ has a number of desirable properties, such as consistency, unbiasedness, and efficiency. These properties, however, are valid under specific assumptions about the underlying distribution. If the actual distribution differs in some way from the assumed one, $\overline{X}$ may no longer have the same properties. We say that $\overline{X}$ is sensitive to the deviation from the model, or *contamination*. This concept is intended to describe the occurrence of *outliers*, that is, observations that follow a distribution different from the one assumed in the model. Most of the outlier distribution is usually concentrated around much larger (or smaller) values than those typically encountered in the model. Formally, the model assumes the distribution $f(x, \theta)$, where θ is (say) the mean, while the actual sample is taken from the distribution

$$\varphi(x, \theta) = (1 - \epsilon) f(x, \theta) + \epsilon g(x), \tag{12.81}$$

where most of the mass of $g(x)$ is far away from the range of "typical" values under $f(x, \theta)$.

The two important approaches to robust estimation are L-estimators and M-estimators.

L-Estimators

L-estimators, linear functions of order statistics, provide good estimators of location and scale parameters.

Formally, θ will be called a *location parameter* if

$$f(x, \theta) = h(x - \theta)$$

for some probability density (or probability mass function) h.

Similarly θ is called a *scale parameter* if

$$f(x, \theta) = \frac{1}{\theta} h\left(\frac{x}{\theta}\right)$$

for some probability density (or probability mass function) h.

The mean of a distribution may or may not be its location parameter. For instance, it is so in the case of normal distribution. However, in the case of exponential distribution, the mean is a scale parameter, while in the case of a gamma distribution, the mean is neither a scale or a location parameter.

Let $X_1, \ldots, X_n$ be a random sample from the distribution $f(x, \theta)$, and let

$$X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$$

denote the order statistics of the sample. The L-estimator will be any statistic T of the form

$$T = \sum_{k=1}^{n} \gamma_{n,k} Y_{k:n},$$

where $\gamma_{n,k}$ for $k = 1, \ldots, n$ is a double array of coefficients. The class of L-estimators contains many well-known estimators: Choosing $\gamma_{n,k} = 1/n$ for $k = 1, \ldots, n$ gives $T = \overline{X}$. The choice $\gamma_{n,1} = 1, \gamma_{n,k} = 0$ for $k \geq 2$, or $\gamma_{n,n} = 1, \gamma_{n,k} = 0$ for $k < n$, gives two extreme order statistics $Y_{1:n}$ and $Y_{n:n}$. In a similar way one can obtain any sample quantile. Choosing $\gamma_{n,[3n/4]+1} = 1, \gamma_{n,[n/4]+1} = -1, \gamma_{n,k} = 0$ for remaining k, one obtains a sample interquartile range, and so on.

Perhaps the most important L-estimators are the trimmed and Winsorized means, defined as follows:

Definition 12.5.4 Let $0 < \alpha < 1/2$. Then the *α-trimmed mean* is

$$U_n = \frac{1}{n - 2[n\alpha]} \sum_{k=[n\alpha]+1}^{n-[n\alpha]} Y_{k:n}, \tag{12.82}$$

while the *α-Winsorized* mean is

$$V_n = \frac{1}{n} \left([n\alpha] Y_{[n\alpha]+1:n} + \sum_{k=[n\alpha]+1}^{n-[n\alpha]} Y_{k:n} + [n\alpha] Y_{n-[n\alpha]:n} \right). \tag{12.83}$$

□

On the one hand, α-trimming consists of removing the $100\alpha\%$ of lowest and highest observations from the sample, and taking an average of the remaining ones (the middle $100(1 - 2\alpha)\%$ of observations). On the other hand, α-Winsorizing consists of replacing each observation in the lower $100\alpha\%$ and in the upper $100\alpha\%$ of the sample by the sample quantile of order α and $1 - \alpha$, respectively, so that the sample size will stay unchanged. The Winsorized mean is then calculated as the mean of the Winsorized sample.

It is clear that the purpose of trimming (or Winsorizing) is to eliminate (or decrease) the effect of outliers in the sample. An important issue is to define the notion of optimality and then to determine the optimal level α at which the mean should be trimmed or Winsorized.

Among the main advantages of L-estimators is their asymptotic normality when the weights $\gamma_{n,k}$ are reasonably smooth or if they are nonzero for a certain number of central order statistics. For more details, see Serfling (1980) and Arnold et al. (1993).

M-Estimators

Another class of estimators is obtained as follows: Let $h(x, u)$ be a function of two arguments. Given a sample $x_1, x_2, \ldots, x_n$ from the distribution $f(x, \theta)$, take as an estimator of θ the solution of the equation

$$\sum_{k=1}^{n} h(x_k, u) = 0. \tag{12.84}$$

Such estimators are most often obtained by solving an approximate minimization problem. Suppose that we have a "distance" of some sort (not necessarily satisfying any conditions for metric), say $H(x, u)$. As an estimator of θ we choose a point u^* that minimizes the sum

$$\sum_{k=1}^{n} H(x_k, u), \tag{12.85}$$

interpreted as the sum of distances from u to all sample points. In a sense, u^* is the point closest to the sample, with closeness being expressed by the function H. Differentiating (12.85) with respect to u and setting the derivative equal to 0, we obtain equation (12.84) with $h(x, u) = \frac{\partial}{\partial u} H(x, u)$.

This formulation comprises two important classes of estimators, namely MLE's and LS-estimators. Indeed, if we define the function $H(x, u)$ as $-\log f(x; u)$, then $h(x, u) = -\frac{\partial}{\partial u} \log f(x, u)$ and the M-estimator corresponding to this choice is the maximum likelihood estimator (the minus sign is implied by the fact that now the problem is formulated as a minimization problem).

By taking appropriate functions H and h, we can obtain different variants of least square estimators. Similarly, trimmed or Winsorized means can be obtained by appropriate choices of the functions H and h. For example, we may take $H(x, u) = H(x - u)$ for some function H of one argument. The M-estimator then minimizes the sum $\sum_{i=1}^{n} H(X_i - u)$. For $H(x) = x^2$ we have the simplest least square estimator. If $H(x) = x^2$ for $|x| \leq k$ and $H(x) = k^2$ for $|x| > k$, we obtain a form of the Winsorized mean.

As with L-estimators, the main direction of research is to study the asymptotic properties (e.g., normality) of M-estimators under some general assumptions on functions H or h, and distributions of X_i.

PROBLEMS

12.5.1 Let $X_1, \ldots, X_n$ be a random sample from GAM(α, λ) distribution. Find: **(i)** The MME of $\theta = (\alpha, \lambda)$, using the first two moments. **(ii)** The MME of α when λ is known, and the MME of λ when α is known.

12.5.2 Find the MME of parameter θ in the distribution with density $f(x, \theta) = (\theta + 1)x^{-(\theta+2)}$, for $x > 1, \theta > 0$.

12.5.3 Let $X_1, \ldots, X_n$ be a random sample from the distribution uniform on the union of the two intervals: $[-2, -1]$ and $[0, \theta]$. Find: **(i)** The MME of θ. **(ii)** The MLE of θ. **(iii)** The MLE of θ if positive X_i's are recorded exactly, and negative X_i's can only be counted. **(iv)** The MLE of θ if X_i's cannot be observed, and one can only count the numbers of positive and negative ones.

12.5.4 Let $X_1, \ldots, X_n$ be a random sample from Poisson distribution with mean λ. Find the MLE of $P(X = 0)$.

12.5.5 (Bragging Tennis Player) As in Example 12.34, consider tennis players A and B who from time to time play matches against each other. The probability that A wins a set against B is p.

Suppose now that we do not have complete data on all matches between A and B; we learn only of A's victories, so we know numbers a_3, a_4, and a_5 of matches won in 3, 4, and 5 sets (we do not even know whether he lost any matches with B).

Show that one can find the MLE of p, and find the MLE of the total number of matches and of the number of sets that A lost against B (do not attempt the algebraic solution).

12.5.6 Some phenomena (e.g., headway in traffic) are modeled to be a distribution of a sum of a constant and an exponential random variable. Then the density of X has the form

$$f(x; a, b) = \begin{cases} 0 & \text{for } x < b \\ ae^{-a(x-b)} & \text{for } x \geq b, \end{cases}$$

where $a > 0$ and $b > 0$ are two parameters. Find: **(i)** The MME of $\theta = (a, b)$. **(ii)** The MLE of θ.

12.5.7 A single observation of a random variable X with a geometric distribution results in $X = k$. Find the MLE of the probability of success θ if: **(i)** X is the number of failures preceding the first success. **(ii)** X is the number of trials up to and including the first success.

12.5.8 Find the distribution of the MLE of the probability of success θ based on two Bernoulli trials.

12.5.9 Show that the family of gamma distributions provides conjugate priors for the exponential distribution. Determine the the posterior distribution.

12.5.10 Let $X_1, \ldots, X_n$ be a random sample from the EXP(θ) distribution, and let the prior distribution of θ be EXP(β). Find Bayes estimator of θ and $\mu = 1/\theta$ using: **(i)** Squared error loss. **(ii)** Absolute error loss.

12.5.11 Suppose that there were 15 successes in 24 trials. Find the MLE of the probability of success θ if it is known that $\theta \leq 1/2$.

12.5.12 Two independent Bernoulli trials resulted in one failure and one success. What is the MLE of the probability of success θ if it is known that: **(i)** θ is at most $1/4$. **(ii)** θ exceeds $1/4$.

12.5.13 Let $X_1, \ldots, X_n$ be a random sample from POI(λ) distribution. Find the MLE of λ assuming that: **(i)** $X_1 + \cdots + X_n > 0$. **(ii)** $X_1 + \cdots + X_n = 0$.

12.5.14 Find the MME and MLE of the standard deviation of a Poisson distribution.

12.5.15 Let $X_1, \ldots, X_n$ be a random sample from N(μ, σ^2) distribution, μ is known. Find the MLE of σ: **(i)** Directly. **(ii)** First finding the MLE of variance σ^2 and then using the invariance property.

12.5.16 Let X_1, X_2 be a random sample from a N(μ, σ^2) distribution with μ and σ^2 unknown. Find the MLE of σ^2 if the only available information is that the difference between observations equals 3.

12.5.17 Find the MLE of the mean of a U$[\theta_1, \theta_2]$ distribution based on a random sample of size n.

12.5.18 For n observations taken from the U$[0, \theta]$ distribution, let U_n be the number of the ones that are less than 3. Find the MLE of θ.

12.5.19 Suppose that the median of 20 observations, taken from a normal distribution with an unknown mean and variance, is 5 and that only one observation differs from the median by more than 3. Suggest an estimate of the probability that the next two observations will both be between 4 and 5.

12.5.20 Let $X_1, \ldots, X_n$ be a random sample from a log-normal distribution with a parameters μ and σ^2 [this means that $\log X_i \sim \text{N}(\mu, \sigma^2)$]. Find the MLE of μ and σ^2.

12.5.21 Let $X_1, \ldots, X_m$ be a random sample from the N(μ_1, σ_1^2) distribution and let $Y_1, \ldots, Y_n$ be a random sample from the N(μ_2, σ_2^2) distribution, with X_i's being independent from Y_j's. Find the MLE of: **(i)** μ_1, μ_2, σ^2 if $\sigma_1 = \sigma_2 = \sigma$. **(ii)** $\mu, \sigma_1^2, \sigma_2^2$ where $\mu_1 = \mu_2 = \mu$.

12.5.22 $Y_1, \ldots, Y_n$ are independent variables. Assuming that $x_1, \ldots, x_n$ are such that $\sum(x_i - \overline{x})^2 > 0$, compare the MSE's of the MLE and LS-estimators of parameter θ, if: **(i)** $Y_i \sim \text{EXP}(\theta x_i)$. **(ii)** $Y_i \sim \text{POI}(\theta x_i)$.

12.5.23 For independent variables $Y_1, \ldots, Y_n$ with distribution N$(\alpha + \beta x_i, \sigma^2)$, where $x_1, \ldots, x_n$ are fixed, show that the LS-estimator and ML-estimator of $\theta = (\alpha, \beta)$ coincide.

12.5.24 Let $X_1, \ldots, X_n$ be a random sample from U$[\theta, \theta+1]$ distribution. **(i)** Show that $T = c(X_{n:n} - 1) + (1-c)X_{1:n}$, $0 < c < 1$, is the MLE of θ, and find the value of c that minimizes its MSE. **(ii)** Determine the asymptotic distribution of T.

12.6 SUFFICIENCY

The considerations of this section are motivated by the following observation: each of the estimators analyzed in this chapter is dependent on the random sample $X_1, \ldots, X_n$ through some statistic (e.g., $\overline{X}$ or $X_{n:n}$), that reduces the data to a single number. From a purely formal viewpoint, any such reduction involves some loss of information.

The concept of a sufficient statistic is intended to cover situations where the information lost in reducing data to the value of a statistic is not relevant for the purpose of estimating parameter θ. Thus the definition of sufficiency of a statistic is relative to a given parameter θ.

As before, we consider a random sample $X_1, \ldots, X_n$ from the distribution $f(x, \theta)$, where θ belongs to some parameter space Θ. Again, f can be either a density or a probability function of a discrete distribution.

The definition below introduces a statistic that conveys the same information about θ as the whole sample $(X_1, \ldots, X_n)$.

Definition 12.6.1 The statistic T is said to be *sufficient* for θ if the conditional distribution of $X_1, \ldots, X_n$, given $T = t$, does not depend on θ for any value t. □

Before exploring the consequences of the definition let us analyze some examples.

■ **EXAMPLE 12.53**

Consider two independent Bernoulli trials X_1, X_2, with probability of success p, and let $T = X_1 + X_2$. The conditional distribution of (X_1, X_2) given T is

$$\begin{aligned} P\{X_1 = 0, X_2 = 0 | T = 0\} &= 1, \\ P\{X_1 = 1, X_2 = 1 | T = 2\} &= 1. \end{aligned}$$

Since $P(T = 1) = P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1) = 2p(1-p)$, for $T = 1$ we have

$$P\{X_1 = 1, X_2 = 0 | T = 1\} = \frac{P\{X_1 = 1, X_2 = 0\}}{P(T = 1)} = \frac{p(1-p)}{2p(1-p)} = \frac{1}{2}.$$

Thus, whatever the value of T (0, 1, or 2), the conditional distribution of (X_1, X_2) does not depend on p.

■ **EXAMPLE 12.54**

Let X_t and X_s be the numbers of events in a Poisson process with intensity λ, observed between 0 and t, and between t and $t + s$. We will show that $U = X_t + X_s$ is a sufficient statistic for λ.

Indeed, since X_t and X_s are independent, we have for $k = 0, 1, \ldots, n$,

$$\begin{aligned} P\{X_t = k, X_s = n - k | U = n\} &= \frac{P\{X_t = k\} P\{X_s = n - k\}}{P\{U = n\}} \\ &= \frac{\frac{(\lambda t)^k}{k!} e^{-\lambda t} \times \frac{(\lambda s)^{n-k}}{(n-k)!} e^{-\lambda s}}{\frac{[\lambda(t+s)]^n}{n!} e^{-\lambda(t+s)}} \\ &= \binom{n}{k} \left(\frac{t}{t+s}\right)^k \left(\frac{s}{t+s}\right)^{n-k}. \end{aligned}$$

Thus X_t, given $U = n$, is binomial with parameters n and $t/(t+s)$, and does not depend on λ.

If the conditional distribution of observations $X_1, \ldots, X_n$, given the statistic $T = t$, does not depend on the parameter θ that we want to estimate, then (once we know that $T = t$) the additional knowledge of a particular configuration of $X_1, \ldots, X_n$ observed in the data is irrelevant in estimating θ. For instance, in Example 12.53, $T = 1$ means that we had one success and one failure. The fact that the first trial was a failure and the second was a success is of no additional help in estimating p.

Since the process of determining the conditional distribution of observations given the value of statistic T is sometimes cumbersome, it is desirable to have another method for verifying that a statistic is sufficient. Such a method is given by the following theorem, due to Neyman:

Theorem 12.6.1 (Factorization Criterion) *Let $X_1, \ldots, X_n$ be a random sample from the distribution $f(x, \theta)$, with $\theta \in \Theta$. A statistic $T = T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if for all $\mathbf{x} = (x_1, \ldots, x_n)$ and all $\theta \in \Theta$ the joint distribution*

$$f_n(\mathbf{x}, \theta) = f(x_1, \theta) \cdots f(x_n, \theta)$$

can be written as

$$f_n(\mathbf{x}, \theta) = u[T(\mathbf{x}), \theta] v(\mathbf{x}), \tag{12.86}$$

where u is a nonnegative function that depends on both θ and $\mathbf{x} = (x_1, \ldots, x_n)$, but dependence on $\mathbf{x}$ is only through the function T, and v is a function of $\mathbf{x} = (x_1, \ldots, x_n)$ that does not depend on θ.

Proof. We will give the proof in the discrete case; the proof for the continuous case requires careful consideration because densities are defined only up to sets of probability zero. Thus we have now $f(x, \theta) = P_\theta\{X_i = x\}, i = 1, \ldots, n$.

Let $Q(t)$ be the set of all $\mathbf{x} = (x_1, \ldots, x_n)$ such that $T(\mathbf{x}) = t$. Then

$$P_\theta(T = t) = \sum_{\mathbf{x} \in Q(t)} f_n(\mathbf{x}, \theta).$$

Suppose first that T is a sufficient statistic. Let $T = t$, and consider any point $\mathbf{x} \in Q(t)$. The conditional probability $P_\theta\{\mathbf{X} = \mathbf{x} | T = t\}$ does not depend on θ; let us call it $v(\mathbf{x})$. Letting $u(t, \theta) = P_\theta\{T = t\}$, we obtain, for any fixed $\mathbf{x} \in Q(t)$,

$$f_n(\mathbf{x}, \theta) = P_\theta\{\mathbf{X} = \mathbf{x}\} = P_\theta\{\mathbf{X} = \mathbf{x} | T = t\} P_\theta\{T = t\} = v(\mathbf{x}) u(t, \theta),$$

which is the factorization (12.86).

Conversely, suppose that $f_n(\mathbf{x}, \theta)$ satisfies formula (12.86) for some functions u and v. Let us fix t and $\theta \in \Theta$ and compute the conditional probability of a point $\mathbf{x}$ given $T = t$. Clearly, for $\mathbf{x} \notin Q(t)$, we have $P_\theta\{\mathbf{X} = \mathbf{x} | T = t\} = 0$. For $\mathbf{x} \in Q(t)$, we have

$$\begin{aligned} P_\theta\{\mathbf{X} = \mathbf{x} | T = t\} &= \frac{P_\theta\{\mathbf{X} = \mathbf{x}\}}{P_\theta\{T = t\}} = \frac{f_n(\mathbf{x}, \theta)}{\sum_{\mathbf{y} \in Q(t)} f_n(\mathbf{y}, \theta)} \\ &= \frac{u(t, \theta) v(\mathbf{x})}{\sum_{\mathbf{y} \in Q(t)} u(t, \theta) v(\mathbf{y})} = \frac{v(\mathbf{x})}{\sum_{\mathbf{y} \in Q(t)} v(\mathbf{y})}. \end{aligned}$$

□

We will now apply Theorem 12.6.1 to find sufficient statistics in various families of distributions.

■ **EXAMPLE 12.55**

Consider again the Bernoulli trials with probability p of success. We have here, letting $t = \sum x_i$,

$$\begin{aligned} f_n(x_1, \ldots, x_n, p) &= \prod_{i=1}^{n} p^{x_i} (1 - p)^{1 - x_i} = p^t (1 - p)^{n - t} \\ &= \left(\frac{p}{1 - p} \right)^t (1 - p)^n. \end{aligned}$$

In this case a sufficient statistic is the total number of successes $T = \sum X_i$, while the function $v(x_1, \ldots, x_n) \equiv 1$.

■ **EXAMPLE 12.56**

In a Poisson distribution, we have, letting $t = \sum x_i$,

$$f_n(x_1, \ldots, x_n, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \lambda^t e^{-n\lambda} \frac{1}{x_1! \cdots x_n!},$$

which again shows that $T = \sum X_i$ is a sufficient statistic. Here $u(t, \lambda) = \lambda^t e^{-n\lambda}$, while $v(x_1, \ldots, x_n) = (x_1! \cdots x_n!)^{-1}$.

■ **EXAMPLE 12.57**

Now let $(x_1, \ldots, x_n)$ be observations from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ where σ^2 is known. Then the joint density of the sample can be written as

$$\begin{aligned} f_n(x_1, \ldots, x_n, \mu) &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right\} \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum x_i^2 - 2\mu \sum x_i + n\mu^2 \right) \right\} \\ &= \exp\left\{ \frac{\mu}{\sigma^2} \sum x_i - \frac{n\mu^2}{2\sigma^2} \right\} v(x_1, \ldots, x_n). \end{aligned}$$

Thus $\sum X_i$ is a sufficient statistic for μ, with

$$v(x_1, \ldots, x_n) = \frac{\exp\{-1/2\sigma^2 \sum x_i^2\}}{\sigma^n (2\pi)^{n/2}}.$$

If now μ is known, but σ^2 is the unknown parameter, we take $v(x_1, \ldots, x_n) \equiv 1$, and the sufficient statistic is $\sum (X_i - \mu)^2$.

■ **EXAMPLE 12.58**

Consider finally the case of sampling from a $\mathrm{U}[0, \theta]$ distribution. The joint density of the sample is

$$f(x_1, \ldots, x_n, \theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } X_{n:n} \le \theta \\ 0 & \text{otherwise,} \end{cases}$$

or equivalently

$$f(x_1, \ldots, x_n, \theta) = \frac{1}{\theta^n} I(\theta, \max(x_1, \ldots, x_n)),$$

where $I(\theta, c) = 1$ if $\theta \ge c$ and 0 if $\theta < c$. This shows the function $u(\theta, T)$ (with $v(x_1, \ldots, x_n) \equiv 1$) and that $X_{n:n}$ is a sufficient statistic.

Sufficient statistics are not unique. Every one-to-one transformation of a sufficient statistic is again a sufficient statistic. For instance, in Example 12.57, an alternative sufficient statistic to $\sum X_i$ is $\overline{X} = (1/n)\sum X_i$.

The concept of sufficient statistic can be generalized in an obvious way to a set of jointly sufficient statistics to cover a multidimensional case. The intuitions motivating the concept are the same as in a one-dimensional case.

We will now formulate the definition using the factorization theorem, that is, use the assertion of the theorem as a basis for the definition.

Definition 12.6.2 The statistics $T_1, \ldots, T_k$ are *jointly sufficient* for θ if the distribution $f_n(\mathbf{x}, \theta)$ satisfies the following condition: for every $\mathbf{x} = (x_1, \ldots, x_n)$ and $\theta \in \Theta$ we have

$$f_n(\mathbf{x}, \theta) = u[T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}), \theta]v(\mathbf{x}),$$

where the nonnegative function $v(\mathbf{x})$ does not depend on θ while the function u depends on θ, and depends on $\mathbf{x}$ only through the values of statistics $T_1, \ldots, T_k$. □

■ **EXAMPLE 12.59**

In a random sample $X_1, \ldots, X_k$ from a $N(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown so that $\theta = (\mu, \sigma^2)$, we have

$$\begin{aligned} f_n(\mathbf{x}, \theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2\right\} \\ &= \frac{\sigma^{-n}}{(2\pi)^{n/2}} \exp\left\{-\frac{\sum x_i^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2}\right\}; \end{aligned}$$

statistics $T_1 = \sum_{i=1}^n X_i, T_2 = \sum_{i=1}^n X_i^2$ are jointly sufficient for θ.

If $T_1, \ldots, T_k$ are jointly sufficient for θ, and $T_1^*, \ldots, T_k^*$ are obtained from $T_1, \ldots, T_k$ by a one-to-one transformation, then $T_1^*, \ldots, T_k^*$ are also jointly sufficient for θ. Thus another pair of jointly sufficient statistics in Example 12.59 is $\overline{X}$ and $S^2 = (1/n)\sum_{i=1}^n (X_i - \overline{X})^2$. This can also be seen from the representation

$$\begin{aligned} f_n(\mathbf{x}, \theta) &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \mu)^2\right\} \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \overline{X})^2 - \frac{n(\overline{X} - \mu)^2}{2\sigma^2}\right\} \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}[S^2 - (\overline{X} - \mu)^2]\right\}. \end{aligned}$$

Obviously the sample $(X_1, \ldots, X_n)$ is always jointly sufficient. It suffices to write formally $T_i(X_1, \ldots, X_n) = X_i$ for $i = 0, 1, \ldots, n$. This is true but useless. It is also true that the order statistics $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ are jointly sufficient, since

$$f_n(\mathbf{x}, \theta) = \prod_{j=1}^n f(x_j, \theta) = \prod_{i=1}^n f(x_{i:n}, \theta),$$

where the factors of the last product are a permutation of the factors of the middle product.

Thus a set of jointly sufficient statistics always exists, and it is natural to look for the maximal reduction of data that retains sufficiency. Here reduction is meant both in the sense of reducing the dimensionality, and in the sense of using transformations that are not one-to-one, such as squaring.

The definition of maximal reduction, hence minimality of sufficient statistic, is as follows:

Definition 12.6.3 A sufficient statistic T is called *minimal* if T is a function of any other sufficient statistic. □

The definition covers the idea of maximal reduction of the data, still leaving sufficiency intact. Indeed, suppose that T is minimal in the sense of Definition 12.6.3 and still can be reduced. This means that there exists a function h which is not one-to-one such that $U = h(T)$ is a sufficient statistic. Then T is not a function of U, which gives a contradiction.

Definition 12.6.3 extends to the case of a minimal sufficient set of statistics.

Definition 12.6.4 A set $\{T_1, \ldots, T_k\}$ of sufficient statistics is called *minimal* if $T_1, \ldots, T_k$ are jointly sufficient and they are functions of any other set of jointly sufficient statistics. □

The intuition here is as follows: if $\{T_1, \ldots, T_m\}$ and $\{T_1', \ldots, T_k'\}$ are both jointly sufficient, and the second set is a function of the first, that is, if $T_i' = \varphi_i(T_1, \ldots, T_m)$ for $i = 1, \ldots, k$ and for some functions $\varphi_1, \ldots, \varphi_k$, then

$$(T_1, \ldots, T_m) \succeq (T_1', \ldots, T_k'),$$

where $\succeq$ is an ordering among jointly sufficient sets of statistics. The minimal jointly sufficient statistics are at the "base" of this ordering, while the fact that $(X_{1:n}, \ldots, X_{n:n})$ is always a set of jointly sufficient statistics means that the set of statistics being ordered by relation $\succeq$ is not empty. Example 12.59 shows that the minimal sufficient set of statistics is not unique; however, all these minimal sufficient sets are obtained as one-to-one functions of other minimal sufficient sets.

■ **EXAMPLE 12.60**

Let $X_1, \ldots, X_n$ be random sample from the distribution uniform on $[-\theta, \theta]$ so that

$$f(x; \theta) = \begin{cases} \frac{1}{2\theta} & \text{if } -\theta \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Proceeding as in Example 12.58, we can easily show that the pair of extreme order statistics $(X_{1:n}, X_{n:n})$ is jointly sufficient for θ. However, this is not a minimal set: the statistic $T(X_{1:n}, X_{n:n}) = \max(|X_{1:n}|, |X_{n:n}|)$ is a single sufficient statistic for θ.

Typically the dimensionality of the minimal set of jointly sufficient statistics equals the number of dimensions of the parameter θ. Thus, in the case of a normal

distribution with the parameter being the pair (μ, σ^2), the minimal sets of jointly sufficient statistics consist of two statistics (e.g., $\sum X_i$ and $\sum X_i^2$). However, there exist distributions that depend on a single parameter, and yet the only jointly sufficient statistics (hence the minimal sets of jointly sufficient statistics) are the order statistics $(X_{1:n}, \ldots, X_{n:n})$. This is true, for instance, for the Cauchy distribution with density

$$f(x;\theta) = \frac{1}{\pi[1+(x-\theta)^2]}.$$

Consider now the role that sufficient statistics play in estimation. Observe first that the likelihood function is

$$L(\theta, \mathbf{x}) = v(\mathbf{x})u[T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}), \theta], \tag{12.87}$$

where $\{T_1, \ldots, T_k\}$ is a set of jointly sufficient statistics (minimal or not). Since the factor $v(\mathbf{x})$ plays no role in the determination of θ at which the likelihood attains its maximum, we can expect that the MLE $\hat{\theta}$ of θ will be a function of $T_1, \ldots, T_k$. Since the same kind of representation of the likelihood is valid if $T_1, \ldots, T_k$ form a minimal sufficient set, we can expect that $\hat{\theta}$ will be a function of a minimal sufficient set of statistics.

This argument is valid provided that the MLE is unique (which is usually the case). When the MLE is not unique (for such situations, see Example 12.30), one cannot claim that they are all functions of sufficient statistics, but one of them will be.

We can therefore state the following theorem:

Theorem 12.6.2 *If the MLE of θ exists and is unique, then it is a function of minimal sufficient set of statistics. If the MLE is not unique, then there exists an MLE that is a function of a minimal sufficient set of statistics.*

A very similar situation exists in the case of Bayes estimators. Indeed, if $T_1, \ldots, T_k$ are jointly sufficient for θ, then the posterior density of θ given $\mathbf{x}$ is, for prior density $\pi(\theta)$,

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{f_n(\mathbf{x},\theta)\pi(\theta)}{\int f_n(\mathbf{x},\eta)\pi(\eta)d\eta} = \frac{v(\mathbf{x})u(T_1,\ldots,T_k,\theta)\pi(\theta)}{\int v(\mathbf{x})u(T_1,\ldots,T_k,\eta)\pi(\eta)d\eta} \\ &= \frac{u(T_1,\ldots,T_k,\theta)\pi(\theta)}{H(T_1,\ldots,T_k)},\end{aligned}$$

since the factor $v(\mathbf{x})$ cancels; here H is some function of statistics $T_1, \ldots, T_k$. Consequently, the Bayes estimator, equal to the mean of posterior distribution, depends on $\mathbf{x}$ only through statistics $T_1, \ldots, T_k$. This proves the following theorem:

Theorem 12.6.3 *The Bayes estimator of θ is a function of the minimal sufficient set of statistics for θ.*

The main importance of sufficient statistics is, to a large extent, related to the next theorem, according to which any estimator that is not based on sufficient statistic can be improved, and hence is not admissible (see Definition 12.3.1). The theorem is formulated in the case of a single sufficient statistic T. However, it can be easily generalized to the case of the minimal sufficient set of statistics.

Theorem 12.6.4 (Rao-Blackwell) *Assume that in a family of distributions* $\{f(x,\theta), \theta \in \Theta\}$ *a sufficient statistic* T *exists, and let* W *be an estimator of* $g(\theta)$. *Then*

$$W^*(T) = E_\theta(W|T) \tag{12.88}$$

is also an estimator of $g(\theta)$, *such that* $E(W) = E(W^*)$, *and its risk* $R(W^*,\theta) = E\{[W^* - g(\theta)]^2\}$ *satisfies the condition*

$$R(W^*,\theta) \leq R(W,\theta) \tag{12.89}$$

for all $\theta \in \Theta$. *Moreover, the inequality in* (12.89) *is strict for some* θ *unless* W *is a function of* T *and is an unbiased estimator of* $g(\theta)$.

Proof. We need to show first that formula (12.88) indeed defines an estimator of $g(\theta)$, that is, that the left-hand side does not depend on θ but only on the sample $X_1,\ldots,X_n$. We will also show that this dependence is through the value of T only so that the obtained estimator is a function of the sufficient statistic T. The second part of the proof will consist of showing that inequality (12.89) holds for every θ.

To show that formula (12.88) defines an estimator, we will carry the argument in the discrete case, with $f(x,\theta) = P\{X = x|\theta\}$. Let t be a possible value of T, and let $Q(t) = \{\mathbf{x} : T(\mathbf{x}) = t\}$.

For any fixed t and $\mathbf{x} \in Q(t)$, the conditional distribution of $\mathbf{X}$ given $T = t$ is, by the factorization theorem,

$$\begin{aligned} P\{\mathbf{X} = x|T = t,\theta\} &= \frac{f_n(\mathbf{x},\theta)}{\sum_{\mathbf{y}\in Q(t)} f_n(\mathbf{y},\theta)} = \frac{v(\mathbf{x})u(t,\theta)}{\sum_{\mathbf{y}\in Q(t)} v(\mathbf{y})u(t,\theta)} \\ &= \frac{v(\mathbf{x})}{\sum_{\mathbf{y}\in Q(t)} v(\mathbf{y})}, \end{aligned}$$

which is independent of θ. On the other hand, $P\{\mathbf{X} = \mathbf{x}|T = t,\theta\} = 0$ if $\mathbf{x} \notin Q(t)$, which is also independent of θ. Consequently we have

$$E_\theta(V|T = t) = \sum_{\mathbf{x}\in Q(t)} W(\mathbf{x})P\{\mathbf{X} = \mathbf{x}|T = t,\theta\} = \frac{\sum_{\mathbf{x}\in Q(t)} W(\mathbf{x})v(\mathbf{x})}{\sum_{\mathbf{y}\in Q(t)} v(\mathbf{y})},$$

which is independent of θ, as asserted.

The estimator $E_\theta(W|T)$ depends on the actual sample observed in the following sense: after observing the sample $\mathbf{X} = \mathbf{x}^*$, compute $T(\mathbf{x}^*) = t$, and use (12.90) to obtain $E_\theta(W|T)$.

Clearly, if $T(\mathbf{x}^*) = T(\mathbf{x}^{**}) = t$, then the value of the estimator $E_\theta(W|T)$ is the same for samples $\mathbf{x}^*$ and $\mathbf{x}^{**}$. This means that this value depends on the sample only through the value of the sufficient statistic T.

It remains now to prove the inequality (12.89). Based on Theorem 8.7.2,

$$\mathrm{Var}(X) = E_Y\{\mathrm{Var}(X|Y)\} + \mathrm{Var}_Y\{E(X|Y)\}. \tag{12.90}$$

Letting $E_\theta(W) = k(\theta)$, the risk of estimator W is

$$\begin{aligned} R(W,\theta) &= E\{[W - g(\theta)]^2\} = E\{([W - k(\theta)] + [k(\theta) - g(\theta)])^2\} \\ &= E\{[W - k(\theta)]^2\} + [k(\theta) - g(\theta)]^2 \\ &= \mathrm{Var}(W) + [k(\theta) - g(\theta)]^2, \end{aligned} \tag{12.91}$$

since the cross product is easily seen to vanish. The second term is simply the square of the bias [recall that W is used to estimate $g(\theta)$, rather than θ].

We will now use the formula (12.90), taking $X = W$ and $Y = T$. Thus for every θ we have:

$$\begin{aligned} \mathrm{Var}(W) &= E_T\{\mathrm{Var}(W|T)\} + \mathrm{Var}_T\{E(W|T)\} \\ &\geq \mathrm{Var}_T\{E(W|T)\} = \mathrm{Var}(W^*). \end{aligned} \tag{12.92}$$

Now, by the fact that

$$k(\theta) = E(W) = E_T\{E(W|T)\} = E(W^*), \tag{12.93}$$

we can write, using (12.91), (12.92), and (12.93),

$$\begin{aligned} R(W,\theta) &\geq \mathrm{Var}(W^*) + [k(\theta) - g(\theta)]^2 = \mathrm{Var}(W^*) + [E(W^*) - g(\theta)]^2 \\ &= E[W^* - g(\theta)]^2 = R(W^*, \theta). \end{aligned}$$

It remains to determine the conditions for equality in (12.89). It is clear that the equality occurs if and only if $k(\theta) \equiv g(\theta)$ and $E_T\{\mathrm{Var}(W|T)\} = 0$. The first condition means that W is an unbiased estimator of $g(\theta)$. The second condition means that $\mathrm{Var}(W|T) \equiv 0$. That is, for every value t of T, the random variable W is constant; hence W is a function of T. □

What the Rao-Blackwell theorem says is that in order to decrease the risk of an estimator V, one should decrease its bias. Then, if the estimator is not a function of sufficient statistic (or statistics), replace it by the estimator defined as the conditional expectation of V given sufficient statistics.

We will now illustrate the second part of this principle. Thus, in the examples below we will start from some unbiased estimators and improve them by conditioning with respect to sufficient statistics.

■ **EXAMPLE 12.61**

Let $X_1, \ldots, X_n$ be a random sample from POI(λ) distribution. Suppose that we want to estimate $P\{X = 0\} = e^{-\lambda}$.

We know (see Example 12.56) that the sufficient statistic for λ is $\overline{X}$, or equivalently, $T = \sum_{i=1}^n X_i$, which has POI($n\lambda$) distribution. Let us estimate $P\{X_i = 0\} = e^{-\lambda} = g(\lambda)$ by $W = U/n$—a relative frequency of X_i's in the sample that are equal to 0. To compare the risks of W and $W^* = E(W|T)$, it is enough to compare their variances, since it will be shown that W and W^* are both unbiased estimators for $g(\lambda)$.

Clearly, U is binomial with parameters n and $g(\lambda)$, so $E(W) = (1/n)E(U) = g(\lambda)$. Thus W is unbiased for $g(\lambda)$, and its risk equals

$$R(W,\lambda) = \mathrm{Var}(U/n) = \frac{e^{-\lambda}(1 - e^{-\lambda})}{n}. \tag{12.94}$$

Now let W^* be defined as

$$W^* = E(W|T) = \frac{1}{n}E(U|T). \tag{12.95}$$

Clearly, the number of X_i's which are equal to 0 cannot exceed n. If $T = 1$, then $U = n - 1$, so it remains to consider cases where $T = t > 1$.

Let us start by determining the conditional distribution of one of the observations, say X_1, given $T = t$. We have, for $j \leq t$,

$$\begin{aligned} P\{X_1 = j | \sum_{i=1}^{n} X_i = t\} &= \frac{P\{X_1 = j, \sum_{i=2}^{n} X_i = t - j\}}{P\{\sum_{i=1}^{n} X_i = t\}} \\ &= \frac{\frac{\lambda^j e^{-\lambda}}{j!} \times \frac{[(n-1)\lambda]^{t-j}}{(t-j)!} e^{-(n-1)\lambda}}{\frac{(n\lambda)^t}{t!} e^{-n\lambda}} \\ &= \binom{t}{j} \left(\frac{1}{n}\right)^j \left(1 - \frac{1}{n}\right)^{t-j}, \end{aligned}$$

which shows that given $\sum X_i = t$, the variable X_1 is binomial with parameters t and $1/n$.

Now let

$$Y_i = \begin{cases} 1 & \text{if } X_i = 0 \\ 0 & \text{if } X_i > 0, \end{cases}$$

so that for each Y_j we have

$$E(Y_j|T = t) = P(X_j = 0|T = t) = P(X_1 = 0|T = t) = \left(1 - \frac{1}{n}\right)^t.$$

Since $U = Y_1 + \cdots + Y_n$, we obtain

$$\begin{aligned} W^* &= \frac{1}{n} E(U|T) = \frac{1}{n} E(Y_1 + \cdots + Y_n|T) \\ &= \frac{1}{n} \sum_{j=1}^{n} E(Y_j|T) = \left(1 - \frac{1}{n}\right)^T, \end{aligned} \tag{12.96}$$

a result that is somewhat unexpected.

Finding the risk of W^* (in order to compare it with the risk of W) requires some work. We know that W^* must be unbiased because $W = U/n$ was unbiased, and therefore we need to compare only variances of W^* and W. Recalling that $T = \sum_{i=1}^{n} X_i$ has POI($n\lambda$) distribution, we have

$$\begin{aligned} Ez^T &= \sum_{t=0}^{\infty} z^t \times P\{T = t\} = \sum_{t=0}^{\infty} z^t \frac{(n\lambda)^t}{t!} e^{-n\lambda} \\ &= e^{-n\lambda} e^{zn\lambda} \sum_{t=0}^{\infty} \frac{(zn\lambda)^t}{t!} e^{-zn\lambda} = e^{n\lambda(z-1)}. \end{aligned}$$

On the other hand, $E[z^T]^2 = E[(z^2)]^T = e^{n\lambda(z^2-1)}$. Consequently

$$\mathrm{Var}(z^T) = E[z^T]^2 - [E(z^T)]^2 = e^{n\lambda(z^2-1)} - e^{2n\lambda(z-1)}.$$

The estimator W^* is obtained by putting $z = 1 - 1/n$, which gives

$$\begin{aligned} \mathrm{Var}(W^*) &= e^{n\lambda[((n-1)/n)^2-1]} - e^{2n\lambda[(n-1)/n-1]} \\ &= e^{-\lambda(2-1/n)} - e^{-2\lambda} = e^{-2\lambda}(e^{\lambda/n} - 1). \end{aligned}$$

To compare the risks (variances) of the original estimator $W = U/n$ and the improved estimator W^*, observe that the risk of W, given by (12.94), can be written as

$$\frac{e^{-\lambda}(1-e^{-\lambda})}{n} = \frac{\lambda e^{-2\lambda}}{n} + \frac{e^{-2\lambda}}{n}(e^{\lambda} - \lambda - 1).$$

On the other hand, for the risk of W^* we have

$$\begin{aligned} e^{-2\lambda}(e^{\lambda/n} - 1) &= e^{-2\lambda}\left(\frac{\lambda}{n} + \frac{\lambda^2}{2n^2} + \frac{\lambda^3}{6n^3} + \cdots\right) \\ &= e^{-2\lambda}\frac{\lambda}{n} + \frac{e^{-2\lambda}}{n^2}\left(\frac{\lambda^2}{2} + \frac{\lambda^3}{6n} + \cdots\right), \end{aligned}$$

which is smaller if n is sufficiently large.

■ **EXAMPLE 12.62**

Consider now estimators T_1–T_5 of θ in the U$[0, \theta]$ distribution introduced in Example 12.5. Three of them, T_3, T_4, and T_5, are not functions of a statistic $T_1 = X_{n:n}$ sufficient for θ (see Example 12.58). Theorem 12.6.4 will be applied to improve estimators T_3, T_4, and T_5.

SOLUTION. The improved estimator T_3 is

$$T_3^* = E_\theta(T_3|T_1) = E_\theta\{T_1 + X_{1:n})|T_1\} = T_1 + E_\theta\{X_{1:n}\}.$$

Given T_1, other $n-1$ observations are iid, each with a U$[0, T_1]$ distribution. Conditional expectation $E(X_{1:n}|T_1)$ is the same as the expectation of a minimum of $X_1', \ldots, X_{n-1}'$, each X_i' being uniform on $[0, T_1]$. This conditional expectation equals $(1/n)T_1$. Consequently,

$$T_3^* = T_1 + \frac{1}{n}T_1 = \frac{n+1}{n}T_1,$$

which is equal to the estimator T_2 considered in Example 12.5.

Next

$$T_4^* = E_\theta\{(n+1)X_{1:n}|T_1\} = (n+1)\frac{1}{n}T_1,$$

where we used the results obtained in determining T_3^*. Finally, for T_5 we have

$$\begin{aligned} T_5^* = E_\theta\{T_5|T_1\} &= E_\theta\left\{\frac{2}{n}(X_1 + \cdots + X_n)|T_1\right\} \\ &= \frac{2}{n}E(X_1 + \cdots + X_n|T_1). \end{aligned}$$

In this case, given T_1, the sum $X_1 + \cdots + X_n$ must have one term equal to T_1, while the others have a uniform distribution on $[0, T_1]$. So again, letting X_i' be uniform on $(0, T_1)$, we have

$$\begin{aligned} E(X_1 + \cdots + X_n|T_1) &= T_1 + E(X_1' + \cdots + X_{n-1}') \\ &= T_1 + (n-1)\frac{T_1}{2} = \frac{n+1}{2}T_1, \end{aligned}$$

and therefore

$$T_5^* = \frac{2}{n} \times \frac{n+1}{2} T_1 = \frac{n+1}{n} T_1.$$

Observe that improving three different estimators (T_3, T_4, and T_5) by conditioning with respect to sufficient statistic T_1 led in each case to the same unbiased estimator $T_2 = [(n+1)/n]T_1$. The natural question arises: Was it a coincidence, or a rule? If such a result occurs as a rule, we could always select an unbiased estimator for which the conditioning would be especially simple. Thus we would find the unique unbiased estimator that could not be improved any further, hence with minimal risk. We will now introduce one more concept that connects all the pieces together.

The situation at present may be summarized as follows: On the one hand, we have the Rao-Cramér inequality, which tells us how much reduction of the variance (hence of the risk for unbiased estimators) is possible. This bound may be effectively calculated in textbook cases, but in real-life situations it can be hard or impossible to determine.

On the other hand, we have the Rao-Blackwell theorem, which tells us that unbiased estimators that are not based on sufficient statistics can be improved; to be more specific, their variance can be reduced by conditioning on sufficient statistics.

Again, the technical difficulties of such conditioning can be formidable. But here the situation can often be simplified. Rather than try to improve an estimator by finding its conditional expectation on a sufficient statistic, we can simply abandon the effort and refrain from using estimators that are not functions of sufficient statistics. Instead, we can try to find an unbiased estimator that is built only on sufficient statistics. This may still be technically difficult but often easier than determining conditional expectations.

Once we find an unbiased estimator that is a function of a sufficient statistic (or a minimal sufficient set of statistics), the Rao-Blackwell theorem tells us that it cannot be improved any further (at least by conditioning). We therefore are tempted to conclude that we found the minimum variance unbiased estimator. Such a conclusion is justified if the estimator in question has variance equal to the right-hand side of the Rao-Cramér inequality. Otherwise, the reasoning on which it is based still has a flaw and has to be amended. Indeed, this reasoning works if there is only one unbiased estimator based on a minimal sufficient statistic (or minimal sufficient set of statistics). But what if there are several such estimators? We know that none of the estimators can be improved by further conditioning, but this does not guarantee that their risks are all minimal.

This observation makes it necessary to single out the cases where such an "anomaly" cannot occur, that is, situations where an unbiased estimator based on a sufficient statistic is unique.

Accordingly we introduce the following definition:

Definition 12.6.5 A family of distributions $\{f(t;\theta), \theta \in \Theta\}$ of variable T is called *complete* if for all $\theta \in \Theta$ the condition $E_\theta[u(T)] = 0$ implies that $u(T) = 0$ with probability 1. A sufficient statistic with a distribution in a complete family is said to be a *complete sufficient* statistic. □

We have the following theorem:

Theorem 12.6.5 *If the family of distributions* $\{f(t;\theta), \theta \in \Theta\}$ *is complete, then any unbiased estimator* $u(T)$ *of parameter* $g(\theta)$ *is unique.*

Proof. If $u_1(T)$ and $u_2(T)$ are unbiased estimators of $g(\theta)$, then

$$E_\theta[u_1(T)] = E_\theta[u_2(T)] = g(\theta),$$

and $E_\theta\{u_1(T) - u_2(T)\} = 0$ for all $\theta \in \Theta$. Hence $P_\theta\{u_1(T) - u_2(T) = 0\} = 1$ for all θ, which means that the functions u_1 and u_2 coincide. □

Theorem 12.6.6 (Lehmann and Scheffé) *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution* $\{f(x;\theta), \theta \in \Theta\}$, *and let* $\{T_1, \ldots, T_k\}$ *be jointly complete sufficient statistics for* θ. *If the statistic* $U^* = U^*(T_1, \ldots, T_k)$ *is an unbiased estimator of* $g(\theta)$, *then* U^* *is the minimum variance unbiased estimator (MVUE) of* $g(\theta)$.

Proof. By completeness, any statistic $U = U(T_1, \ldots, T_k)$ that is unbiased for $g(\theta)$ must satisfy the condition $U = U^*$ with probability 1 (for every θ). On the other hand, for any statistic $W = W(X_1, \ldots, X_n)$ that is unbiased for $g(\theta)$ and is not a function of $T_1, \ldots, T_k$, $U = E\{W|T_1, \ldots, T_k\}$ will also be unbiased and will be a function of $T_1, \ldots, T_k$ such that $\mathrm{Var}_\theta(U) \leq \mathrm{Var}_\theta(W)$, which shows that variance of U^* is minimal for each θ. □

The concept of completeness pertains to a family of distributions. It is usually stated in a phrase appealing to intuition and easy to remember: A family is complete if "there are no nontrivial unbiased estimators of zero" (a trivial unbiased estimator of zero is a random variable identically equal to zero). We will now give some examples of complete families.

■ **EXAMPLE 12.63**

Consider random sampling from a POI(λ) distribution. A sufficient statistic for λ, $T = \sum_{i=1}^n X_i$, has POI($n\lambda$) distribution. Let $u(T)$ be a statistic, and suppose that $E_\theta[u(T)] = 0$ for all θ. This means that

$$E_\theta[u(T)] = \sum_{n=0}^{\infty} u(n)\frac{\theta^n}{n!}e^{-\theta} = e^{-\theta}\sum_{n=0}^{\infty} u(n)\frac{\theta^n}{n!} = 0.$$

Since $e^{-\theta} \neq 0$, we obtain $g(\theta) = \sum u(n)\theta^n/n! = 0$ for all θ. However, the Taylor expansion of $g(\theta)$ is $\sum[g^{(n)}(0)/n!]\theta^n$, we have $u(n) = g^{(n)}(0)$. But the latter quantity is zero for all n.

Observe here that we used the fact that the family contains *all* Poisson distributions as indexed by θ (or at least sufficiently many of them). If we consider the family of some selected Poisson distributions, say for $\theta_1, \theta_2, \ldots, \theta_r$, we can choose some negative and some positive values of $u(t)$ so as to obtain appropriate cancelations. The simplest example here is to take a family consisting of just one Poisson distribution, with mean $\theta = 5$ (say). Then the function $u(T) = T - 5$ is not identically zero, yet its expectation is zero for all members of the family (since there is only one distribution in the family, and the expectation of this distribution is 5).

■ **EXAMPLE 12.64**

In a U[0, θ] distribution, the sufficient statistic for θ is $T_1 = X_{n:n}$, and its density $f(t, \theta)$ is

$$f(t;\theta) = \begin{cases} \frac{nt^{n-1}}{\theta^n} & 0 \le t \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we have a function u that satisfies the condition

$$E_\theta[u(T)] = \int_0^\theta \frac{u(t)nt^{n-1}}{\theta^n}\, dt = 0.$$

This means that

$$\int_0^\theta t^{n-1}u(t)dt = 0,$$

and by differentiation with respect to θ we get $\theta^{n-1}u(\theta) = 0$ for all $\theta > 0$. Consequently, we must have $u(t) \equiv 0$, and therefore the class of all distributions uniform on $[0, \theta]$, $\theta > 0$, is complete.

This explains why in Example 12.62, by conditioning the unbiased estimators T_3, T_4, and T_5 on sufficient statistic T_1, we obtain the (unique, in view of completeness) unbiased estimator $[(n+1)/n]T_1$ of θ.

Proving completeness of a family starting directly from the definition may not be easy. We will therefore present a rather general and easily verifiable sufficient condition for completeness:

Definition 12.6.6 The distribution $f(x, \theta)$ is in an *exponential class*, if

$$f(x;\theta) = \begin{cases} c(\theta)b(x)\exp[\sum_{i=1}^m q_i(\theta)k_i(x)] & \text{if } x \in A \\ 0 & \text{otherwise,} \end{cases} \tag{12.97}$$

where $\theta = (\theta_1, \ldots, \theta_m)$ and $c(\theta), b(x)$ are strictly positive functions. It is assumed here that the parameter space is a generalized rectangle (possibly infinite in some dimensions) of the form $\Theta = \{\theta : a_i \le \theta_i \le b_i, i = 1, \ldots, m\}$.

Moreover:

1. The set A where $f(x;\theta)$ is positive does not depend on θ.
2. The functions $q_1(\theta), \ldots, q_m(\theta)$ are continuous and functionally independent (i.e., none is a function of others).
3. The functions $k_1(x), \ldots, k_m(x)$ are linearly independent (i.e., no linear combination of them is identically zero unless all coefficients are zero), and when $f(x;\theta)$ is a density of a continuous random variable, they are differentiable.

□

We have the following theorem:

Theorem 12.6.7 *Let $X_1, \ldots, X_n$ be a random sample from a $\{f(x,\theta), \theta \in \Theta\}$ distribution that belongs to the exponential class* (12.97). *Then the set of statistics $\{T_1, \ldots, T_m\}$ such that*

$$T_j(X_1, \ldots, X_n) = \sum_{i=1}^{n} k_j(X_i), \text{ for } j = 1, \ldots, m$$

is a minimal set of complete sufficient statistics for $\theta = (\theta_1, \ldots, \theta_m)$.

The joint sufficiency of statistics $T_1, \ldots, T_m$ follows immediately from (12.97). For the proof of completeness see Lehmann and Romano (2005).

Distributions from most common families (e.g., binomial, Poisson, normal, or gamma) belong to the exponential class. We will show it for the binomial and normal distributions, leaving the proofs for the remaining families as exercises.

■ **EXAMPLE 12.65**

For the Bernoulli distribution we have for $x = 0, 1$

$$\begin{aligned} f(x;p) &= p^x(1-p)^{1-x} = (1-p)\left(\frac{p}{1-p}\right)^x \\ &= (1-p)\exp\left\{x\log\left(\frac{p}{1-p}\right)\right\}. \end{aligned}$$

Thus $A = \{0,1\}$, $m = 1, q_1(p) = \log[p/(1-p)]$, and $k_1(x) = x$. The sufficient statistic is $T = \sum k_1(X_i) = \sum X_i$.

■ **EXAMPLE 12.66**

For a normal distribution we have, after letting $\theta = (\mu, \sigma)$,

$$\begin{aligned} f(x,\theta) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x^2-2\mu x+\mu^2)}{2\sigma^2}\right\} \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-\mu^2}{2\sigma^2}\right\}\exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right\}, \end{aligned}$$

so we have $c(\theta) = (1/\sigma\sqrt{2\pi})e^{-\mu^2/2\sigma^2}$, $b(x) \equiv 1$, $q_1(\theta) = -1/2\sigma^2$, $q_2(\theta) = \mu/\sigma^2$, $k_1(x) = x^2$, $k_2(x) = x$. A minimal sufficient set of statistics is therefore $T_1 = \sum_{i=1}^n X_i^2$ and $T_2 = \sum_{i=1}^n X_i$.

We end this section with one more concept whose notion is complementary to the notion of a sufficient statistic.

Definition 12.6.7 Let $X_1, \ldots, X_n$ be a random sample from a distribution $f(x,\theta)$. Statistic $T = T(X_1, \ldots, X_n)$ is *ancillary* for parameter θ, if its distribution does not depend on θ. □

■ **EXAMPLE 12.67**

In a random sample of size n obtained from the N$(\theta, 1)$ distribution, statistic $T = \sum_{i=1}^{n}(X_i - \overline{X})^2$ has a chi-square distribution with $n - 1$ degrees of freedom, so it is ancillary for θ.

The following theorem, due to Basu, establishes the independence of complete sufficient and ancillary statistics:

Theorem 12.6.8 (Basu) *Let $X_1, \ldots, X_n$ be a random sample from the distribution $\{f(x, \theta), \theta \in \Theta\}$. If $\{T_1, \ldots, T_m\}$ are jointly complete sufficient statistics for θ, then $\{T_1, \ldots, T_m\}$ and any ancillary statistic $V = V(X_1, \ldots, X_n)$ are independent.*

PROBLEMS

12.6.1 Find sufficient statistic(s) for parameter θ in the following distributions: **(i)** $f(x, \theta) = (x/\theta^2)e^{-x^2/2\theta^2}$ for $x > 0$ (Rayleigh). **(ii)** $f(x, \theta) = (1/2\theta)e^{-|x|\theta}$ (double exponential). **(iii)** BETA$(\theta, 2\theta)$. **(iv)** U$[\theta, 2\theta]$.

12.6.2 Generalizing Example 12.53, let $X_1, \ldots, X_n$ be n independent Bernoulli trials. Show that $T = \sum_{i=1}^{n} X_i$ is sufficient for probability of success p by finding the joint distribution of $(X_1, \ldots, X_n)$ given $T = t$. Find the marginal distribution $P\{X_1 = 1 | T = t\}$.

12.6.3 Find a sufficient statistic for θ if observations are uniformly distributed on the set of integers $0, 1, \ldots, \theta$.

12.6.4 Show that the N$(0, \theta)$ family is not complete.

12.6.5 Let $X_1, \ldots, X_n$ be a random sample from the distribution with a density $f(x; \lambda, \theta) = \lambda e^{-\lambda(x-\theta)}$ for $x \geq \theta$ and 0 otherwise. Determine a pair of jointly sufficient statistics for parameters λ and θ.

12.6.6 Show that the following families of distributions are in the exponential class: **(i)** POI(λ). **(ii)** EXP(λ). **(iii)** NBIN(r, θ), r known. **(iv)** BETA(θ_1, θ_2). **(v)** WEI(θ, λ), λ known.

12.6.7 Show that the family of GAM(α, λ) distributions is in an exponential class, and find the minimal jointly sufficient statistics.

12.6.8 Suppose that a random sample is taken from a distribution with density $f(x; \theta) = 2x/\theta^2$ for $0 \leq x \leq \theta$ and $f(x; \theta) = 0$ otherwise. Find the MLE of the median of this distribution, and show that this estimator is a minimal sufficient statistic.

12.6.9 Let $X_1, \ldots, X_n$ be a random sample from the EXP(λ) distributions. Suppose that only first k order statistics $X_{1:n}, \ldots, X_{k:n}$ are observed. Find a minimal sufficient statistic for λ.

12.6.10 Let X_1, X_2 be a random sample of size $n = 2$ from the EXP(λ) distribution. Show that statistic $T = X_1/X_2$ is ancillary for λ.

12.6.11 Let $X_1, \ldots, X_n$ be a random sample from the U$[\theta, \theta+1]$ distribution. Show that statistic $T = X_{n:n} - X_{1:n}$ is ancillary for θ.

12.7 INTERVAL ESTIMATION

Thus far we have dealt with point estimation. An estimator was used to produce a single number, hopefully close to an unknown parameter. The natural question arises: What can be said about the distance between the value produced by the estimator and the true value of the parameter?

Here again the approaches are different, depending on whether or not one can regard the value of θ as a realization of some random variable.

Bayesian Intervals

In the Bayesian scheme the true value θ in a given situation is regarded as the observed value of a random variable Ξ with prior distribution (e.g., density) π. Given the observation $\mathbf{x}$, the posterior density of Ξ is $\pi(\theta|\mathbf{x})$, and we can assess the probability that Ξ lies between two values a and b as

$$P\{a \leq \Xi \leq b \,|\mathbf{x}\} = \int_a^b \pi(\theta|\mathbf{x})d\theta. \tag{12.98}$$

■ **EXAMPLE 12.68**

Suppose that we observe $n = 3$ Bernoulli trials with unknown probability of success θ, where θ has the prior distribution BETA(2, 2). Assume that we record $x = 2$ successes. Then the posterior distribution of the parameter is again beta, with parameters $\alpha + x = 4$ and $n + \beta - x = 3$. So the posterior density is

$$\frac{\Gamma(7)}{\Gamma(4)\Gamma(3)}\theta^3(1-\theta)^2 = 60\theta^3 - 120\theta^4 + 60\theta^5, \qquad 0 < \theta < 1.$$

The expected value of the posterior distribution is $4/7 = 0.57$, which is the Bayes estimate of θ. The probability that the true value of θ lies below 0.2 equals

$$\int_0^{0.2} (60\theta^3 - 120\theta^4 + 60\theta^5)d\theta = 0.017.$$

The probability that the true value of θ lies above 0.8 equals

$$\int_{0.8}^1 (60\theta^3 - 120\theta^4 + 60\theta^5)d\theta = 0.099.$$

Consequently, with probability $1 - 0.017 - 0.099 = 0.884$ the true value of θ lies between 0.2 and 0.8.

Confidence Intervals

In many cases the value of θ cannot be regarded as a value of a random variable with a prior distribution that can be assigned a frequential interpretation. In such cases non-Bayesian statisticians do not consider it meaningful to speak of "the probability that the value of parameter θ belongs to some set." According to this point of

view, all randomness is in the process of selection of the sample. The relevant statements on interval estimation are now expressed through the concept of *confidence intervals*.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from the distribution $f(x, \theta)$. Let α be a fixed number with $0 < \alpha < 1$ (typically α is a small number such as $\alpha = 0.05$ or $\alpha = 0.01$).

Definition 12.7.1 A pair of statistics $L = L(\mathbf{X})$ and $U = U(\mathbf{X})$ is an $(1 - \alpha)$-*level confidence interval (CI)* for θ if for all $\theta \in \Theta$,

$$P_\theta\{L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\} = 1 - \alpha. \tag{12.99}$$

Similarly, a statistic $L(\mathbf{X})$ is a $(1 - \alpha)$-*level lower confidence bound* if

$$P\{L(\mathbf{X}) \leq \theta\} = 1 - \alpha. \tag{12.100}$$

□

The upper confidence bound $U(\mathbf{X})$ is defined in a similar way. It is important to notice that the meaning of formula (12.99) is quite different from the meaning of the left-hand side of formula (12.98).

To point out the difference in the most concise way, observe that the left-hand side of (12.98) refers to posterior probability given $\mathbf{x}$. This means that the sample has already been taken, and the probability statement refers to the mechanisms of sampling the particular value of θ. However, according to non-Bayesian philosophy this sampling does not take place, and therefore probability statements after $\mathbf{x}$ has been observed make no sense.

On the other hand, formula (12.99) refers to randomness in the sample (for a given fixed θ). The interval $[L(\mathbf{X}), U(\mathbf{X})]$ varies from sample to sample and is therefore random. By definition, the probability that it covers the unknown (fixed) value of θ is $1 - \alpha$. This probability, however, refers to the process of taking the sample only. After a specific sample $\mathbf{x}$ has been observed, the observed interval $[L(\mathbf{x}), U(\mathbf{x})]$ either covers or does not cover θ. All randomness is gone, and it makes no sense to speak of "the probability that θ belongs to the interval $[L(\mathbf{x}), U(\mathbf{x})]$" because this probability is now either 0 or 1. Since we do not know which is the case, the term *confidence* is used. We say that the $(1-\alpha)$-level confidence interval for θ is $[L(\mathbf{x}), U(\mathbf{x})]$. In frequential terms, the last statements means "the confidence interval $L[(\mathbf{x}), U(\mathbf{x})]$ has been obtained by a method that produces randomly generated intervals, of which, on average, $100(1 - \alpha)\%$ cover the unknown value θ."

It is perhaps regrettable that the terminology here, as it developed historically, is a little ambiguous. The term *confidence interval* refers to interval $L[(\mathbf{X}), U(\mathbf{X})]$ with random endpoints as well as to the interval $L[(\mathbf{x}), U(\mathbf{x})]$ with endpoints obtained from the sample observed. Sometimes the first interval is called the *probability* interval, and the second is called the *sample* confidence interval. We will follow a more common tradition, referring to confidence intervals and sample confidence intervals.

In constructing confidence intervals, we use quantiles of some known distributions, such as standard normal, Student's t, and chi-square. For typographical reasons, as well as to follow the tradition, we use different notation for distributions symmetric about 0 (N(0, 1) and Student's t).

Thus, if Z is a standard normal random variable, then z_α will denote the $(1-\alpha)$-quantile of Z (or, upper α-percentile), that is,

$$P\{Z \le z_\alpha\} = 1 - \alpha. \tag{12.101}$$

Then $z_{1-\alpha} = -z_\alpha$, and $P\{|Z| \le z_{\alpha/2}\} = 1 - \alpha$. Similarly the $(1-\alpha)$-quantile of Student's t distribution with ν degrees of freedom will be denoted by $t_{\alpha,\nu}$.

For asymmetric distributions, such as chi-square, the subscript α will be used for upper quantiles of order α so that $\chi^2_{\alpha,\nu}$ denotes the quantile of order $1-\alpha$ of a chi-square distribution with ν degrees of freedom.

■ **EXAMPLE 12.69**

Consider the random sample $X_1, \ldots, X_n$ from the distribution $\mathrm{N}(\theta, \sigma^2)$, where θ is the unknown parameter while σ^2 is known. Then the interval

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is a $(1-\alpha)$-level confidence interval for θ. Indeed, the inequality

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

is equivalent to

$$\theta - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < \theta + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

so both inequalities have the same probabilities. Since $\mathrm{Var}(\overline{X}) = \sigma^2/n$, we have

$$\begin{aligned} &P\left\{\overline{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}} < \theta < \overline{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\right\} \\ &= P\left\{-z_{\alpha/2} < \tfrac{\overline{X}-\theta}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\} = 1 - \alpha, \end{aligned}$$

which was to be shown.

■ **EXAMPLE 12.70**

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density $f(x, \theta)$ depending on a one-dimensional parameter θ. Furthermore, let $F(x, \theta)$ be the cdf of X, and let $G = 1 - F$. Then each of the random variables $-2\log F(X, \theta)$ and $-2\log G(X, \theta)$ has an exponential distribution with parameter $1/2$. Indeed

$$P\{-2\log F(X,\theta) > t\} = P\{F(X,\theta) < e^{-t/2}\} = e^{-t/2},$$

since $F(X, \theta)$ is uniform on $[0, 1]$. The argument for $G(X, \theta)$ is analogous. Consequently, the sums

$$U = -2\sum_{i=1}^{n} \log F(X_i, \theta) \quad \text{and} \quad W = -2\sum_{i=1}^{n} \log G(X_i, \theta) \tag{12.102}$$

have GAM$(n, 1/2)$ distribution, which is the same as chi-square distribution with $2n$ degrees of freedom. We can therefore write

$$P\left\{\chi^2_{1-\alpha/2,2n} < -2\sum_{i=1}^{n} \log F(X_i, \theta) < \chi^2_{\alpha/2,2n}\right\} = 1 - \alpha \tag{12.103}$$

and

$$P\left\{\chi^2_{1-\alpha/2,2n} < -2\sum_{i=1}^{n} \log G(X_i, \theta) < \chi^2_{\alpha/2,2n}\right\} = 1 - \alpha. \tag{12.104}$$

If the inequality that defines the event in either (12.103) or (12.104) can be solved to the form $L(X_1, \ldots, X_n) < \theta < U(X_1, \ldots, X_n)$, we obtain, by definition, a confidence interval for θ at level $1 - \alpha$.

■ **EXAMPLE 12.71**

Continuing Example 12.70, if $X_1, \ldots, X_n$ have an exponential distribution with density $\theta e^{-\theta x}$, then $F(x, \theta) = 1 - e^{-\theta x}$ and $G(x, \theta) = e^{-\theta x}$. Formula (12.104) gives

$$P\left\{\chi^2_{1-\alpha/2,2n} < 2\theta\sum_{i=1}^{n} X_i < \chi^2_{\alpha/2,2n}\right\} = 1 - \alpha,$$

and we obtain the $(1 - \alpha)$-level confidence interval

$$\frac{\chi^2_{1-\alpha/2,2n}}{2n\overline{X}} < \theta < \frac{\chi^2_{\alpha/2,2n}}{2n\overline{X}}. \tag{12.105}$$

Note that if we used F instead of G, we would obtain the inequality

$$\chi^2_{1-\alpha/2,2n} < -2\sum_{i=1}^{n} \log(1 - e^{-\theta X_i}) < \chi^2_{\alpha/2,2n},$$

which cannot easily be solved for θ. Thus, even when we have an explicit formula for F (and hence for G), use of one of these functions often leads to simpler results than use of the other.

In general, how are confidence intervals obtained? For large n, the answer lies in the limit theorem 12.5.1 about the asymptotic distributions of MLE's obtained from setting the derivative of the likelihood equal to 0 and solving the resulting equation. Indeed, since $\sqrt{nI(\theta)}(\hat{\theta}_n - \theta)$ with $I(\theta)$ defined by (12.27) is asymptotically standard normal (as in Example 12.69), we have

$$-z_{\alpha/2} < \sqrt{nI(\theta)}(\hat{\theta}_n - \theta) < z_{\alpha/2}$$

with probability close to $1 - \alpha$. This inequality is equivalent to the inequality

$$\hat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{nI(\theta)}} < \theta < \hat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{nI(\theta)}}. \tag{12.106}$$

The latter is not a confidence interval, since the bounds depend on the unknown value of θ. However, for large n we can replace θ by $\hat{\theta}_n$ in $I(\theta)$. Consequently, we can state the following theorem:

Theorem 12.7.1 *Let $\hat{\theta}_n$ be the* MLE *estimator of θ in a problem for which* MLE*'s satisfy assumption of Theorem 12.5.1. Then for large n,*

$$\left(\hat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_n)}},\ \hat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_n)}}\right)$$

is an approximate $(1-\alpha)$-level confidence interval for parameter θ.

■ **EXAMPLE 12.72**

As an illustration we consider the problem of estimating probability p in Bernoulli distribution. The observations $X_1, X_2, \ldots, X_n$ are iid random variables indicating whether consecutive trials lead to success or failure (i.e., $X_i = 1$ or 0 with probabilities p and $1-p$). The sum $S_n = X_1 + \cdots + X_n$ is the total number of successes in n trials. The MLE of p is $\hat{p} = S_n/n$, and (12.35) gives $I(p) = 1/p(1-p)$ so that

$$\text{Var}(\hat{p}) = \frac{1}{n^2}\text{Var}(S_n) = \frac{p(1-p)}{n} = \frac{1}{nI(p)}.$$

By the central limit theorem for the binomial distribution, the random variable

$$\frac{S_n - np}{\sqrt{np(1-p)}} = \sqrt{nI(p)}(\hat{p} - p)$$

is asymptotically standard normal. Consequently, we have the approximate relation

$$P\{-z_{\alpha/2} < \sqrt{nI(p)}(\hat{p} - p) < z_{\alpha/2}\} = 1 - \alpha. \tag{12.107}$$

The inequality on the left-hand side of (12.107) is equivalent to

$$|\hat{p} - p| < \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{p(1-p)},$$

or

$$\left(1 + \frac{z_{\alpha/2}^2}{n}\right)p^2 - \left(2\hat{p} + \frac{z_{\alpha/2}^2}{n}\right)p + \hat{p}^2 < 0. \tag{12.108}$$

The solution of the quadratic inequality (12.108),

$$\frac{\hat{p} + z_{\alpha/2}^2/2n - z_{\alpha/2}\sqrt{\frac{4\hat{p}(1-\hat{p}) + z_{\alpha/2}^2/n}{4n}}}{1 + z_{\alpha/2}^2/n}$$

$$< p < \frac{\hat{p} + z_{\alpha/2}^2/2n + z_{\alpha/2}\sqrt{\frac{4\hat{p}(1-\hat{p}) + z_{\alpha/2}^2/n}{4n}}}{1 + z_{\alpha/2}^2/n}, \tag{12.109}$$

gives an approximate $(1-\alpha)$-level confidence interval for probability p.

If we disregard the terms of the order higher than $1/\sqrt{n}$, we obtain a more commonly used further approximation, namely

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{12.110}$$

which coincides with the confidence interval given in Theorem 12.7.1

The method of obtaining confidence intervals for any n is based on the pivotal variables defined as follows.

Definition 12.7.2 A random variable W is called *pivotal* for θ if it depends only on the sample $\mathbf{X}$ and on unknown parameter θ, while its distribution does not depend on any unknown parameters (including θ). □

This means that in principle one should be able to calculate the numerical values of probabilities for W. It is *not enough* to know that W has a normal distribution (say) with one or more parameters unspecified.

If $W = W(\mathbf{X}, \theta)$ is a pivotal random variable, then the $(1-\alpha)$-level confidence interval can be obtained as follows: First, for given α, one determines the values q_α^* and q_α^{**} such that

$$P\{q_\alpha^* \le W(\mathbf{X}, \theta) \le q_\alpha^{**}\} = 1 - \alpha, \tag{12.111}$$

which is possible since W is pivotal.

It now remains to convert the inequality $q_\alpha^* \le W(\mathbf{X}, \theta) \le q_\alpha^{**}$ into the form

$$L(\mathbf{X}, q_\alpha^*, q_\alpha^{**}) \le \theta \le U(\mathbf{X}, q_\alpha^*, q_\alpha^{**}) \tag{12.112}$$

or, more generally, into the form

$$\theta \in S(\mathbf{X}^{(n)}, q_\alpha^*, q_\alpha^{**}), \tag{12.113}$$

where S is some set (not necessarily an interval). But as is often the case, if the pivotal quantity W is monotone in θ for every value of $\mathbf{X}$, a solution of the form (12.112) is attainable, and then the statistics L and U provide the lower and upper endpoints of the $(1-\alpha)$-level confidence interval for θ.

■ **EXAMPLE 12.73**

The random variables (12.102) from Example 12.70 are pivotal, which shows that in a continuous case with a one–dimensional parameter there always exist at least two pivotal variables.

The next example shows that pivotal variables can also exist in the case of multi-dimensional parameters.

■ **EXAMPLE 12.74**

As in Example 12.69, assume that $X_1, \ldots, X_n$ is a random sample from the distribution $N(\theta, \sigma^2)$, but σ^2 is now unknown. We are still interested in estimating the mean θ. In such cases the statisticians use the term *nuisance* parameter for σ^2.

In Example 12.69 the random variable $(\overline{X} - \theta)/(\sigma/\sqrt{n})$ was *pivotal*, since σ was known. However, if σ is unknown, we have to proceed differently. The idea is to cancel σ in the denominator of $(\overline{X} - \theta)/(\sigma/\sqrt{n})$.

Recall that when $X_i \sim N(\theta, \sigma^2)$ and $S^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2$, the random variable nS^2/σ^2 has a chi-square distribution with $n - 1$ degrees of freedom (Theorem 10.2.1). Moreover, S^2 is independent of $\overline{X}$. Consequently, the ratio

$$\begin{aligned} T &= \frac{(\overline{X} - \theta)/(\sigma/\sqrt{n})}{\sqrt{nS^2/\sigma^2(n-1)}} = \frac{\overline{X} - \theta}{S}\sqrt{n-1} \\ &= \frac{\overline{X} - \theta}{\sqrt{(1/n)\sum(X_i - \overline{X})^2}}\sqrt{n-1} \end{aligned} \tag{12.114}$$

has the Student's t distribution with $n-1$ degrees of freedom. Thus (12.114) is a pivotal random variable (observe that σ is canceled). If now $t_{\alpha/2,n-1}$ is the $(1 - \alpha/2)$-quantile of Student's t distribution with $n - 1$ degrees of freedom (see Table A3.), then we have

$$P\{-t_{\alpha/2,n-1} \leq \frac{\overline{X} - \theta}{S}\sqrt{n-1} \leq t_{\alpha/2,n-1}\} = 1 - \alpha,$$

which gives

$$\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n-1}},\ \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n-1}}\right) \tag{12.115}$$

as a $(1 - \alpha)$-level confidence interval for θ.

■ **EXAMPLE 12.75 Confidence Intervals for Variance**

Consider again a random sample with normal distribution $N(\mu, \sigma^2)$, this time assuming that the parameter to be estimated is σ^2. We distinguish two cases: when μ is known, and when μ is unknown. To provide some motivation for considering these cases, consider the problem of assessing the quality of some measuring device. A typical procedure here consists of repeating the measurements of the same object. Assuming that the measurement errors are normal, the two cases above correspond to measurements of an object whose true value μ of the measured attribute is otherwise known, and measurements of an object whose true value μ of the measured attribute is unknown.

The pivotal random variables are respectively,

$$U = \sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 \quad \text{and} \quad V = \sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2.$$

Since U is a sum of n independent squares of standard normal random variables, it has a chi-square distribution with n degrees of freedom. We know also from Theorem 10.2.1 that V has a chi-square distribution with $n-1$ degrees of freedom. Thus

$$P\left\{\chi^2_{1-\alpha/2,n} < \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} < \chi^2_{\alpha/2,n}\right\} = 1-\alpha$$

and

$$P\left\{\chi^2_{1-\alpha/2,n-1} < \frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha.$$

We obtain therefore the $(1-\alpha)$-level confidence intervals

$$\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\alpha/2,n}} < \sigma^2 < \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{1-\alpha/2,n}} \tag{12.116}$$

and

$$\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\chi^2_{1-\alpha/2,n-1}}. \tag{12.117}$$

Two obvious questions arise that concern the intervals (12.116) and (12.117). First, when μ is unknown, one *must* use the interval (12.117). But when μ is known, one can either utilize this knowledge and use interval (12.116) or not utilize it and use interval (12.117). Which is the proper procedure? The answer is that a shorter interval is obtained, on average, if one uses the available information about the expectation μ.

Second, to get the probability $1-\alpha$ of covering the unknown value of σ^2, it is not necessary to choose values $\chi^2_{1-\alpha/2,n}$ and $\chi^2_{\alpha/2,n}$, which cut equal probabilities at both tails of the chi-square distribution. One could choose two other points a and b such that the probability of a chi-square random variable assuming a value between a and b is $1-\alpha$. There are infinitely many such pairs (a,b), the pair $(\chi^2_{1-\alpha/2,n}, \chi^2_{\alpha/2,n})$ being just one of them. An obvious criterion would be to choose a pair a, b that minimizes the length of the confidence interval. It is easy to see that this length is proportional to $1/a - 1/b$. We have a similar discussion when μ is unknown.

To solve this optimization problem, let g_m be the density of the chi-square distribution with m degrees of freedom (in our case, we use either $m=n$ or $m=n-1$). Then the problem can be formulated as

$$\text{minimize} \quad \left(\frac{1}{a}-\frac{1}{b}\right) \quad \text{subject to} \quad \int_a^b g_m(x)dx \geq 1-\alpha.$$

This problem can be solved by using Lagrange multipliers: Differentiating the function

$$\frac{1}{a}-\frac{1}{b}+\lambda\left[\int_a^b g_m(x)dx-(1-\alpha)\right]$$

with respect to a, b, and λ, we obtain three equations:

$$-\frac{1}{a^2} - \lambda g_m(a) = 0, \qquad \frac{1}{b^2} + \lambda g_m(b) = 0, \tag{12.118}$$

and

$$\int_a^b g_m(x)dx = 1 - \alpha. \tag{12.119}$$

As a solution choose a, b satisfying constraint (12.119) and such that $a^2 g_m(a) = b^2 g_m(b)$. These can be solved numerically for each m (see Table A5.). The solution, for large m, is close to the "symmetric" solution, where $a = \chi^2_{1-\alpha/2,m}$, $b = \chi^2_{\alpha/2,m}$. For small sample sizes it is better to use an exact solution of the optimization problem above, as it gives the shortest possible confidence interval.

Table A5. gives the optimal left cutoff probability α, and the corresponding upper quantiles $a = \chi^2_{\alpha_1,m}$ and $b = \chi^2_{\alpha_2,m}$, where $\alpha_2 = (1-\alpha) - \alpha_1$, for $\alpha = 0.1$ and $\alpha = 0.05$ and various numbers of degrees of freedom. The table also gives the relative gain resulting from using the shortest confidence intervals (with length proportional to $1/a - 1/b$), as compared to the confidence interval with equal cutoff probabilities, hence with length proportional to $1/\chi^2_{1-\alpha/2,m} - 1/\chi^2_{\alpha/2,m}$.

■ **EXAMPLE 12.76**

Suppose that five observations from a normal distribution with unknown μ and σ^2 are such that $\sum(X_i - \overline{X})^2 = c$. To obtain a 95%CI for σ^2, the usual procedure based on cutting off 2.5% at both ends of the chi-square distribution with 4 degrees of freedom uses thresholds $\chi^2_{0.975,4} = 0.484$ and $\chi^2_{0.025,4} = 11.143$ (see Table A4.). This leads to the confidence interval $c/11.143 < \sigma^2 < c/0.484$, or $(0.090c, 2.066c)$, of length $1.976c$.

The shortest 95% confidence interval involves the upper quantiles $a = \chi^2_{0.0497,4} = 0.708$ and $b = \chi^2_{0.9003,4} = 21.047$. The confidence interval is now $0.0475c = c/21.047 < \sigma^2 < c/0.708 = 1.412c$. The length is $1.3649c$, which is 69% of the length of the interval based on quantiles corresponding to 2.5% probability on each end.

Two natural questions arise: Do pivotal random variables always exist, and if not, can one construct confidence intervals in some other way? Regarding the first question, the criteria for existence of pivotal random variables were given by Antle and Bain (1969). In analogy with examples concerning the normal distribution, we have the following theorem:

Theorem 12.7.2

(a) *If θ is a location parameter and $\hat{\theta}$ is MLE of θ, then $\hat{\theta} - \theta$ is pivotal.*

(b) *If θ is a scale parameter, and $\hat{\theta}$ is MLE of θ, then $\hat{\theta}/\theta$ is pivotal.*

(c) *Let θ_1 be a location parameter and θ_2 be a scale parameter so that $f(x, \theta_1, \theta_2) = (1/\theta_2)h[(x - \theta_1)/\theta_2]$ for some density h. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are MLE's of θ_1 and θ_2, respectively, then $(\hat{\theta}_1 - \theta_1)/\hat{\theta}_2$ is pivotal.*

The answer to the second question is that even if pivotal random variables do not exist, we can construct confidence intervals. We will give here examples of confidence intervals for quantiles.

■ **EXAMPLE 12.77**

Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F(t)$, and let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be the order statistics of the sample. To simplify, assume that F is continuous and strictly increasing, and let $\xi_{1/2}$ be the median of the distribution F. Thus $\xi_{1/2}$ is the unique solution of the equation $F(t) = 1/2$. We will construct confidence intervals for $\xi_{1/2}$. More precisely, we will consider (random) intervals of the form $[X_{a:n}, X_{b:n}]$ for integers a, b satisfying the condition $1 \leq a < b \leq n$ and assess, for each of such intervals, the probability that it covers the parameter $\xi_{1/2}$.

The event $X_{a:n} \leq \xi_{1/2}$ occurs if and only if at least a elements in the sample are below the median. Similarly the event $\xi_{1/2} \leq X_{b:n}$ occurs if and only if at least $n - b$ elements in the sample are above the median. If we call an observation below the median a success and let S denote the number of successes in the sample, then the event $X_{n,a} \leq \xi_{1/2} \leq X_{n,b}$ occurs if and only if we have at least a successes and at least $n - b$ failures so that $a \leq S \leq b$. Since S has distribution BIN$(n, 1/2)$, we obtain

$$P\{X_{a:n} \leq \xi_{1/2} \leq X_{b:n}\} = \sum_{k=a}^{b} \binom{n}{k} \frac{1}{2^n}.$$

For instance, if $n = 10$, then the interval $[X_{3:10}, X_{7:10}]$ between the third and seventh order statistic is a confidence interval for the median with the confidence level

$$\sum_{k=3}^{7} \binom{10}{k} \frac{1}{2^{10}} = 1 - 2\left[\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right] \times \frac{1}{2^{10}} = 0.891.$$

On the other hand, the interval $[X_{2:10}, X_{6:10}]$ is a confidence interval for $\xi_{1/2}$ with confidence level $\sum_{k=2}^{6} \binom{10}{k}(1/2)^{10} = 0.709$.

One-sided confidence bounds for $\xi_{1/2}$ are obtained in the same way: for instance, the order statistic $Y_{a:n}$ is a lower bound for $\xi_{1/2}$ with the confidence level $\sum_{k=a}^{n} \binom{n}{k}(1/2)^n$.

■ **EXAMPLE 12.78**

Continuing Example 12.77, let us now fix p and let ξ_p be the pth quantile of the population distribution. Again letting S be the number of successes, where now "success" is an observation below ξ_p, S has a BIN(n, p) distribution. The

interval $[X_{a:n}, X_{b:n}]$ is a confidence interval for ξ_p with the confidence level

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

Again, taking $n = 10$, the same interval $[X_{3:10}, X_{7:10}]$ can serve as a confidence interval for the first quartile $\xi_{1/4}$, except that now the confidence level will be

$$\sum_{k=3}^{7} \binom{10}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{10-k} = 0.526.$$

However, in this case it is better to take confidence intervals based on the order statistics where the indices are not equidistant from the extremes.

We complete this chapter with the critique of the concept of confidence intervals. The point is that while a pair of statistics $L = L(\mathbf{X}), U = U(\mathbf{X})$ may satisfy the condition

$$P\{U < \theta < L\} = 1 - \alpha,$$

there are cases where the sample $\mathbf{X}$ may provide additional information, allowing one to claim higher probability of coverage than $1 - \alpha$. This is illustrated by the following example from De Groot (1986).

■ **EXAMPLE 12.79**

Suppose that two independent observations X_1, X_2 are taken from the distribution uniform on $[\theta - 1/2, \theta + 1/2]$. Then order statistics $X_{1:2}$ and $X_{2:2}$ provide a 50% confidence interval for θ. Indeed, the interval $[X_{1:2}, X_{2:2}]$ covers θ if and only if one of the observations is below θ and the other is above θ, an event with probability $1/2$. However, if $X_{2:2} - X_{1:2} > 1/2$, then the interval $[X_{1:2}, X_{2:2}]$ covers θ with probability 1. Still, strict adherence to the definition of confidence interval requires reporting such an interval as a 50% confidence interval rather than as a 100% confidence interval.

This example (as well as other examples of this type, e.g., Example 12.47) gives some weight to the Bayesian approach to statistical inference.

Bootstrap Intervals

The bootstrap approach is a robust alternative to inference based on assumptions that may be either not valid or impossible to verify (e.g., the sample size is too small), or where calculations of standard errors are very complex. Bootstrap intervals can be obtained for the mean, median, proportion, correlation coefficient, regression coefficient—practically any population parameter. The quality of the procedure is not always the same, and therefore the effectiveness of the bootstrap approach has been extensively studied by carefully designed simulations for different parameters, types of population distributions, and sample sizes.

The idea underlying the bootstrap approach is that the sampling distribution of an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ can be "imitated" by the sampling distribution of

the analogous estimator $\hat{\theta}^* = \hat{\theta}(X_1^{(B)}, \ldots, X_n^{(B)})$ applied to so–called *bootstrap samples*.

Let us assume that $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the original sample consisting of n values randomly selected from $f(x, \theta)$ distribution, and let $\hat{\theta} = \hat{\theta}(\mathbf{x})$, be an estimator selected to estimate unknown parameter θ.

The bootstrap random variable, $X^{(B)}$, is defined as

$$P\{X^{(B)} = x_i\} = \frac{1}{n}, \qquad i = 1, \ldots, n, \tag{12.120}$$

and each bootstrap sample, $\mathbf{x}_j^*, j = 1, \ldots, B$, will be a random sample from the distribution (12.120) (in practice, this means that we are sampling the elements of the original sample with replacement). Next for each of the B bootstrap samples we obtain the estimate

$$\hat{\theta}_j^* = \hat{\theta}(\mathbf{x}_j^*), \qquad j = 1, \ldots, B.$$

It has to be mentioned that the number B of bootstrap samples needs to be rather large, usually at least a thousand.

We introduce two kinds of bootstrap interval estimation.

Bootstrap t Confidence Interval

This interval is especially recommended when the distribution of $\hat{\theta}^*$ is close to normal, which can be verified by a Q-Q plot or any test for normality. The *bootstrap standard error* is estimated as

$$SE_{BOOT} = \sqrt{(1/B)\sum(\hat{\theta}_j^* - \hat{\theta}_{BOOT})^2},$$

where $\hat{\theta}_{BOOT} = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^*$, and consequently the $(1-\alpha)100\%$ bootstrap t confidence interval has the form

$$\hat{\theta} \pm t_{\alpha/2,n-1} SE_{BOOT}.$$

Bootstrap Percentile Interval

This interval does not require normality of bootstrap estimator $\hat{\theta}^*$. Its endpoints are sample quantiles of order $\alpha/2$ and $1-\alpha/2$ obtained from $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

The natural question now is: Is there any difference in these procedures, and if yes, then how to decide which one to use? We recommend that you obtain both the bootstrap t confidence interval and the bootstrap percentile interval, and compare them. If they are not close, then probably neither of them should be used. If the bootstrap estimate of bias, $\hat{\theta} - \hat{\theta}_{BOOT}^*$, is small, then a bootstrap percentile interval is recommended. Also, since almost all variation in the bootstrap sample is caused by the selection of the original sample (resampling adds very little variation), bootstrap intervals based on small samples can be unreliable. For statistics such as the median or quartiles they should be used with caution even for moderate samples. On the other hand, bootstrap t confidence intervals for the population mean θ, based on a trimmed mean as an estimator $\hat{\theta}$, can be a good choice.

In Example 12.80 we illustrate both types of bootstrap intervals.

■ **EXAMPLE 12.80**

We compare different confidence intervals for an unknown population mean θ in the EXP$(1/\theta)$ distribution. Initially we assume a certain value of θ and use it to generate a random sample of size n (we select $n = 20$, and $\theta = 1$). A generated random sample will later be used to generate bootstrap samples. As an estimator $\hat{\theta}$ of a population mean θ we take the sample mean $\hat{\theta} = \overline{X}$.

The generated random sample $x_1, \ldots, x_{20}$ yields a sample mean $\hat{\theta} = 1.1167$. Assuming that the information about the population distribution (exponential) is available, we apply formula (12.105) from Example 12.71 and obtain

$$\frac{\chi^2_{0.975,40}}{2 \times 20 \times 1.1167} < \frac{1}{\theta} < \frac{\chi^2_{0.025,40}}{2 \times 20 \times 1.1167},$$

or equivalently $0.7527 < \theta < 1.828$. After generating $B = 1000$ bootstrap samples, we obtain $\hat{\theta}^*_{BOOT} = 1.1083$ and $SE_{BOOT} = 0.2038$. The distribution of the bootstrap estimator $\hat{\theta}^*$ does not indicate lack of normality. Consequently we obtain 95% bootstrap t confidence interval as $1.1167 \pm 2.093 \times 0.238$, or $(0.6901, 1.5433)$, and the 95% bootstrap percentile interval as $(0.7478, 1.5551)$. All three confidence intervals cover the true value $\theta = 1$, and their widths are 1.0753, 0.8532, and 0.8073, respectively. Bootstrap percentile interval seems to work very well here, but this should not be surprising, since, as mentioned, the bootstrap percentile interval should be recommended when the estimated bias is small. Here $\hat{\theta} - \hat{\theta}^*_{BOOT} = 1.1167 - 1.1083 = 0.0084$—less than 1% of a true value of the parameter.

Several methods to make bootstrap intervals more accurate are proposed in the literature. We will not discuss them here, but interested readers can see Efron and Tibshirani (1993). In Chapter 13 we will use bootstrap techniques for testing hypotheses.

PROBLEMS

12.7.1 Six randomly selected adults are asked if they favor additional taxes to help fund more affordable health care. Four of them respond "yes." Assuming that the prior distribution is BETA$(1, 1)$, determine the posterior density and obtain the probability that in the population the proportion of people favoring a tax increase is between 60% and 70%.

12.7.2 Continuing Problem 12.5.10, obtain the 90% Bayesian interval for θ and for $\mu = 1/\theta$ if $n = 6, \overline{x} = 4$, and $\beta = 2$.

12.7.3 Let $X_1, \ldots, X_n$ be a random sample selected from the Pareto distribution with density $f(x; \theta) = \theta 2^\theta x^{-(\theta+1)}$. Find: **(i)** The sufficient statistic for θ. **(ii)** A 95% CI for θ.

12.7.4 Based on a random sample 1.23, 0.36, 2.13, 0.91, 0.16, 0.12 selected from the GAM$(2.5, \theta)$ distribution, find an exact 95% CI for parameter θ.

12.7.5 Seven measurements of the concentration of some chemical in cans of tomato juice are $1.12, 1.18, 1.08, 1.13, 1.14, 1.10, 1.07$. Assume that these numbers represent a random sample from the distribution $\mathrm{N}(\theta, \sigma^2)$. **(i)** Find the shortest 95% and 99% CI's for θ, if σ^2 is unknown. **(ii)** Answer part (i) if $\sigma^2 = 0.0004$. **(iii)** Use the data to obtain a 90% CI for σ^2 and for σ.

12.7.6 A large company wants to estimate the fraction p of its employees who participate in a certain health program. It has been decided that if p is below 25%, a special promotion campaign will be launched. In a random sample of 85 employees the number of those who participated in the program was 16. **(i)** Find a 95% CI for p using formulas (12.109) and (12.110). Should the campaign be launched? **(ii)** Answer the same question if the data are 340 and 64, respectively.

12.7.7 Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$ distribution with both parameters unknown. Let L_α be the length of the shortest confidence interval for μ on confidence level $1 - \alpha$. **(i)** Find $E(L_\alpha^2)$ as a function of n, σ^2 and α. **(ii)** Find the smallest n such that $E(L_\alpha^2) \leq \sigma^2/2$ for a given α.

12.7.8 Obtain a $(1 - \alpha)100\%$ CI for θ if $X_1, \ldots, X_n$ is a random sample from a: **(i)** $\mathrm{N}(\mu, \theta)$ distribution with μ and θ unknown. **(ii)** $\mathrm{BETA}(1, \theta)$ distribution.

12.7.9 Find the probability that the length of a 95% confidence interval for the mean of normal distribution with unknown σ is less than σ, $n = 25$.

12.7.10 Suppose that the largest observation recorded in a sample of size $n = 35$ from a distribution uniform on $[0, \theta]$ is 5.17. Find a 90% CI for θ.

12.7.11 **(i)** Use the large sample distribution of MLE of mean λ in Poisson distribution to construct an approximate $(1 - \alpha)$-level CI for λ. **(ii)** Assuming that the numbers of new cars of a given make sold per week in 15 consecutive weeks—5, 5, 6, 3, 5, 8, 1, 4, 7, 7, 5, 4, 3, 0, 9—form a random sample from a $\mathrm{POI}(\lambda)$ distribution, find a 90% CI for λ.

12.7.12 Based on a random sample of size n selected from the $\mathrm{WEI}(\theta, 4)$ distribution, derive a 95% confidence interval for θ based on $X_{1:n}$.

12.7.13 Suppose that the lifetime T of a certain kind of device (e.g., a fuel pump in a car) has an $\mathrm{EXP}(\lambda)$ distribution. The observed lifetimes of a sample of the devices are $350, 727, 615, 155, 962$ (in days). Find a 95% CI's for: **(i)** λ. **(ii)** $E(T)$. **(iii)** The standard deviation of the lifetime of the device. **(iv)** The probability that the two copies of the device that you bought will each last more than two years.

12.7.14 Based on a random sample $X_1, \ldots, X_n$ from the $\mathrm{U}[0, \theta]$ distribution find: **(i)** $P\{X_{n:n} < \theta < 2X_{n:n}\}$. **(ii)** Such k that the interval $(X_{n:n}, kX_{n:n})$ is a $(1 - \alpha)100\%$ CI for θ.

12.7.15 Suppose that the arrivals at a checkout counter in a supermarket (i.e., times of arriving at the counter or joining the queue, whichever is earlier) form a Poisson process with arrival rate λ. Counting from noon, the thirteenth customer arrived at

12:18 p.m. Find a 90% CI for: **(i)** λ. **(ii)** The variance of interarrival times between consecutive customers.

12.7.16 A sample of 200 trees in a forest has been inspected for a presence of some bugs, out of which 37 trees were found to be infested. **(i)** Assuming a binomial model, give a 90% confidence interval for the probability p of a tree being infested. Use the exact and the approximated formulas. **(ii)** Usually, if a tree is infested, one might expect some of the neighboring trees to be infested too. Thus whether the binomial model is adequate depends on the way the sample was selected. Describe a way of selecting 200 trees so that the binomial model is realistic.

CHAPTER 13

TESTING STATISTICAL HYPOTHESES

13.1 INTRODUCTION

In this chapter we present the basic concepts and results of the theory of testing statistical hypotheses.

Assume that a random sample $X_1, X_2, \ldots, X_n$ was selected from distribution $f(x, \theta)$, where θ is an element of a parameter space Θ. Unlike in the estimation problems, we are now not interested in approximating the true value of θ. What we need to know is only whether θ belongs to some specific subset $\Theta_0 \subset \Theta$. A procedure that allows us to decide whether or not θ is in $\Theta_0 \subset \Theta$ is called a *test of the hypothesis* that $\theta \in \Theta_0$.

The problem is trivial in the case where the sets of possible values of X are disjoint for $\theta \in \Theta_0$ and $\theta \notin \Theta_0$ ($\theta \in \Theta_1$). However, in the more realistic cases some (or even all) values of $\mathbf{X}$ can occur for both $\theta \in \Theta_0$ and for $\theta \notin \Theta_0$, and there is no procedure that will *always* allow us to reach the correct decision. Indeed, it may happen that while θ is in Θ_0, we will observe a value of $\mathbf{X}$ that will lead us to the conclusion that $\theta \notin \Theta_0$. On the other hand, it may happen that θ is not in Θ_0, and we will reach the conclusion that $\theta \in \Theta_0$. These two possibilities are called errors of type I and type II, and the objective is to choose the decision rule that will minimize their probabilities.

■ EXAMPLE 13.1

A consumer protection agency decides to investigate complaints that some boxes contain less of a product (e.g., cereal) than the amount printed on the box indicates. The boxes are filled and then sealed by a machine. Even with the most sophisticated equipment available, the weight of the cereal put into boxes will vary. Suppose that it is established that these amounts follow a normal distribution with some mean μ (which can be set in the packing plant) and a standard deviation σ (which is the attribute of the machine). For example, let $\mu = 20$ oz and $\sigma = 1.5$ oz.

If the mean μ is set at 20 oz, then about 50% of the boxes will contain less than the nominal amount, and at least for buyers of these boxes, the fact that the other 50% of boxes contain more than the nominal amount may be of no relevance. Consequently, the packing company must set the average μ at a value *above* the nominal weight 20 oz. Since no value μ will guarantee that the content of a box will *always* be above 20 oz, a reasonable requirement may be that 99% of boxes must contain at least 20 oz of cereal. If X denotes the weight of the cereal in a randomly chosen box, then

$$0.99 = P(X \geq 20) = P\left(Z \geq \frac{20-\mu}{1.5}\right) = 1 - \Phi\left(\frac{20-\mu}{1.5}\right),$$

and $\mu = 23.5$ oz. Thus, in order to satisfy the requirement, the company should set the average weight of the content of the box at least at a value 23.5 oz.

The agency that investigates the customers' complaints has to determine whether or not the average weight content is at least 23.5 oz. Since opening all boxes and weighting their content is not feasible, the agency must decide how to do it in a more practical way. For example, one may agree (and such a decision should be made in advance, prior to any sampling) that $n = 16$ boxes of the cereal in question will be bought; the stores will be sampled, and in each store one box will be chosen at random from the shelf. Then the content of all 16 boxes will be weighed and the average $\overline{x}$ will be calculated. Next some threshold has to be established (e.g., $\overline{x} = 23$) so that if $\overline{x} \leq 23$ oz, the hypothesis $\mu \geq 23.5$ will be rejected; otherwise, it will be accepted.

The probabilities of two types of errors can be calculated as follows: It may happen that $\mu \geq 23.5$, but the sample average $\overline{x}$ falls below 23 oz. We then declare the company to be at fault (of *not* putting, on average, at least 23.5 oz of cereal in boxes), although in fact the company is not at fault.

It may also happen that we observe $\overline{x} > 23$ (hence we declare the company as not being at fault), whereas in fact $\mu < 23.5$. To compute the chance of these two kinds of errors, we will find the probability that the company is declared at fault while the true mean is μ,

$$P\{\overline{X} \leq 23|\mu\} = \Phi\left(\frac{23-\mu}{1.5/\sqrt{16}}\right),$$

which is a decreasing function of μ.

Declaring the company at fault is an error if in fact $\mu \geq 23.5$ (otherwise, it is a correct decision). The probability of such an error attains the smallest value for $\mu = 23.5$:

$$\Phi\left(\frac{23 - 23.5}{1.5/\sqrt{16}}\right) = 0.09.$$

The chances of the other type of error—declaring the company as not at fault, when in fact $\mu < 23.5$—for any specific μ are equal to

$$P\{\overline{X} > 23|\mu\} = 1 - \Phi\left(\frac{23 - \mu}{1.5/\sqrt{16}}\right).$$

For instance, if $\mu = 22$, then the probability above is less than 0.004.

To make a decision "$\theta \in \Theta_0$" or "$\theta \notin \Theta_0$," we specify a set C of points in the space of values of $\mathbf{X}$ with the instruction: if the actual observation $\mathbf{x}$ is in C, then decide that $\theta \notin \Theta_0$; otherwise, decide that $\theta \in \Theta_0$. Such a set C is called a *critical* (or *rejection*) *region* for the hypothesis that $\theta \in \Theta_0$.

In Example 13.1 we have $\mathbf{X} = (X_1, \ldots, X_{16})$, and the critical region C consisting of all points $(x_1, \ldots, x_{16})$ is such that $(1/16)(x_1 + \cdots + x_{16}) \leq 23$.

Ideally the best critical region C would minimize the probabilities of errors of type I and type II, but this principle is too general. One has to realize the following two points:

1. If $\theta \in \Theta_0$, then deciding that $\theta \notin \Theta_0$ is the error of type I. This error occurs whenever the observation $\mathbf{X}$ falls into the critical region C. So the probability of an error of type I is

$$\alpha(\theta) = P_\theta\{\mathbf{X} \in C\}. \tag{13.1}$$

On the other hand, if $\theta \notin \Theta_0$, then deciding that $\theta \in \Theta_0$ is the error of type II; this occurs whenever $\mathbf{X} \notin C$. So the probability of an error of type II is

$$\beta(\theta) = P_\theta\{\mathbf{X} \notin C\} = 1 - P_\theta\{\mathbf{X} \in C\}. \tag{13.2}$$

Probabilities of errors are not single numbers but functions of the parameter θ, both expressible in terms of $\pi(\theta) = P_\theta\{\mathbf{X} \in C\}$, called the *power function* of the test C. The minimization principle requires that power function be "as small as possible" for $\theta \in \Theta_0$ and "as large as possible" for $\theta \notin \Theta_0$. Obviously the best test is such that $P_\theta\{\mathbf{X} \in C\} = 0$ if $\theta \in \Theta_0$ and $P_\theta\{\mathbf{X} \in C\} = 1$ if $\theta \notin \Theta_0$. However, in most cases of practical importance and interest, $P_\theta\{\mathbf{X} \in C\}$ is a continuous function of θ, and such a "best" test does not exist.

2. Looking at formulas (13.1) and (13.2), we see that any modification of C has opposite effects on the probabilities of errors of type I and type II: if one of them decreases, then the other increases.

Example 13.2, which is a continuation of Example 13.1, provides more explanation.

■ **EXAMPLE 13.2**

The properties of the decision procedure as to whether the cereal-producing company should be declared at fault depend on the choice of the threshold (23 oz), and the sample size n. Changing the threshold up or down changes the probabilities of the two types of errors, with the changes always going in opposite directions. The only way to decrease probabilities of both types of errors at the same time is to increase the sample size n. The "best" test must result from a compromise between the cost of sampling, and the consequences of the two types of errors. In the case of boxes of cereal, the cost of increasing n from 16 to 100 is of a little concern. Of course, increasing the sample size in other experiments might be more expensive, time-consuming, or even impossible.

The consequences of declaring the company at fault, when it is not, can again be expressed by the cost (the company may have to pay a fine, unnecessarily reset the mean to a higher value, etc.). The consequences of declaring company not at fault, when in fact it is, are hard to express in terms of cost; they involve many very small losses suffered by individual buyers.

The fact that the probabilities of errors of types I and II are negatively related, and that each is a function of θ (defined on sets Θ_0 and Θ_1, respectively), makes formulation of the criterion to be optimized a difficult and challenging task. The conceptual structure of the theory is as follows. First, the problem of optimization is solved for the simplest case, where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. The solution is given by the Neyman-Pearson lemma, which determines the test (i.e., critical region) with the preassigned probability of error of type I, and a minimal probability of error of type II (or equivalently, a maximum power at value θ_1). Such test is called *most powerful*. In some sense the Neyman-Pearson lemma plays a role analogous to that of the Rao-Cramèr inequality in estimation theory: both set the standard by showing how much can potentially be achieved.

In some classes of testing problems, in which Θ_1 (and often also Θ_0) consist of more than one element, there exists a test C that is the most powerful against any $\theta \in \Theta_1$; such tests are called *uniformly most powerful*, (UMP).

There are, however, classes of testing problems in which the UMP test does not exist. In such cases the idea is to eliminate tests with some undesirable properties. The UMP test will often exist in the reduced class.

We discuss only one such reduction, reduction to the class of unbiased tests. The UMP unbiased tests may not achieve maximum possible power specified by the Neyman-Pearson lemma, but they are uniformly most powerful among the unbiased tests (UMPU).

A following analogy can help one understand why such a reduction can lead to the existence of a UMP test. Suppose that we want to find the strongest athlete, in the sense of the strength of his right arm and strength of his left arm (considered separately). It may happen that there will be no strongest athlete in the group. The one with strongest right arm will have weak left arm, and vice versa, and no winner will be found. But one may restrict the competition to only those athletes that have equal strength in both arms. Among those, the winner (or winners) can always be found. Such a winner may be called the most powerful "unbiased" athlete.

This analogy immediately suggests another solution: Why not compare the athletes with respect to combined strengths of right and left arms? Or, according to the sum of appropriate scores, possibly different for left and right arms? (similar to deciding the winner in events such as the decathlon, which is based on sum of scores for different events).

While in our example the "most powerful unbiased athlete(s)" could always be found, it is not necessarily so with UMP tests. To understand why, suppose that the competition involves comparisons of strengths of right and left arms, and also right and left legs (separately). If no "absolute" winner (strongest on each limb) exists, one could reduce the competition to "unbiased" athletes, those whose strengths of right and left arms are the same, and whose strengths of right and left legs are the same. But now there is no guarantee that "most powerful unbiased athlete" exists, since an athlete with strongest arms need not have strongest legs, and vice versa.

In short, UMP unbiased tests exist in some classes of problems and do not exist in other classes, so some further reduction may be needed to single out the UMP tests in this reduced class. The best exposition of the theory can be found in the monograph of Lehmann and Romano (2005).

The following example shows how reality can force one to use tests that are suboptimal:

■ **EXAMPLE 13.3**

The testing procedure suggested in Example 13.1 declares the company at fault when $\overline{x} < 23$. Such a test may well be optimal from the point of view of statistical criteria involving probabilities of errors, but it has a disturbing feature: it may happen that we declare the company at fault when all boxes tested actually contain more than the nominal 20 oz amount. Such a decision may not be defensible in court, in case the company decides to appeal. One should then restrict the consideration to procedures, that have the threshold value set below the nominal value 20 oz, and such procedures may be suboptimal.

The next example shows the nature of technical difficulties that appear when the term "parameter" is interpreted in a somewhat nonstandard sense.

■ **EXAMPLE 13.4**

Suppose that a pharmaceutical company develops a drug to treat some disease, hoping that it is superior to the drug used thus far. After running the toxicity tests, studying the side effects, and so on (as required by the FDA), the only issue that remains is to compare the merits of the two drugs. In a typical setup one chooses two groups (samples) of patients. One group is then given the "old" drug, while the other receives the drug just developed. The data will consist of two sets of observations $\mathbf{X} = X_1, \ldots, X_m$ and $\mathbf{Y} = Y_1, \ldots, Y_n$—the results of some medical tests where m and n represent numbers of people in each group.

Setting the hypotheses in the form discussed above can be accomplished as follows: Let F and G be the cdf's of observations in the first and second

sample, respectively. The joint cdf of the data (X, Y) may be taken as

$$P\{X_1 \le x_i, \ldots, X_m \le x_m, Y_1 \le y_1, \ldots, Y_n \le y_n\} = \prod_{i=1}^{m} F(x_i) \times \prod_{j=1}^{n} G(y_j).$$

This joint distribution is characterized by the pair $\theta = (F, G)$ of two cdf's. Thus the parameter space is the class of all pairs of distributions, and the hypothesis tested is that θ belongs to the "diagonal" in this space, namely to the set of all pairs of the form (F, F), where F is some cdf.

13.2 INTUITIVE BACKGROUND

As already mentioned, we start from the same setup as in Chapter 12: We observe a random sample $X_1, X_2, \ldots, X_n$ from a certain distribution. We know that this distribution belongs to some family $f(x, \theta)$ and we know that θ is an element of a parameter space Θ. As before, we do not know what is the true value of θ that governs the probabilities of the sample $X_1, X_2, \ldots, X_n$.

The difference between the problem of estimation and the problem of testing is simply that now we are not interested in approximating the true value of θ. What we need to know is only whether θ belongs to some specific subset $\Theta_0 \subset \Theta$.

■ **EXAMPLE 13.5**

Consider a politician who will win or lose the election, depending whether the proportion of voters who will cast their votes for him exceeds 50% or not. If θ is the fraction of voters who favor the politician in question, then $0 \le \theta \le 1$, so that the parameter space is $\Theta = [0, 1]$, and the set of interest is $\Theta_0 = (\frac{1}{2}, 1]$. A survey may give a sample $X_1, \ldots, X_n$ of random variables with Bernoulli distribution $f(x, \theta) = \theta^x(1 - \theta)^{1-x}, x = 0, 1$, and the question is how to use the observation to determine whether or not $\theta \in \Theta_0$. A typical question may be: suppose that out of n=400 voters sampled, only 195 will vote for the politician in question, and 205 will vote against him. Should one reject the hypothesis that $\theta > 1/2$? The answer depends on many factors, the most crucial being: how likely is it to observe the data so much (195 vs 205) or even more in favor of rejecting the claim $\theta > 1/2$, if in fact the true value of θ is above $1/2$?

In the generally accepted terminology, the statement that θ belongs to Θ_0 and the statement that θ does not belong to Θ_0 are called *hypotheses*.

Let $\Theta_1 = \Theta \setminus \Theta_0$ so that out of the two hypotheses, $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, exactly one must be true. Logically speaking, denying the truth of one hypothesis is equivalent to accepting the truth of the other. The actually observed sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ provides some evidence, typically pointing in favor of one of the two hypotheses. A *decision rule* is a rule that tells us what to do in case of observing $\mathbf{x} = (x_1, \ldots, x_n)$. These actions may be "accept H_0," "reject H_0 (and therefore accept H_1)," "take another observation," or even "leave the matter

unresolved." A decision rule will be called a *test*[25] (more precisely, a *nonrandomized* test) if the only actions allowed are "accept H_0" and "accept H_1."

Since only two actions are allowed in a test, we must choose one of them regardless of the value of $\mathbf{x}$. Therefore any decision rule is equivalent to a partition of the sample space into two subsets: the set of those points $\mathbf{x}$ at which H_0 is to be accepted, and its complement, the set of those points $\mathbf{x}$ at which H_0 is to be rejected (i.e., H_1 accepted).

Accordingly we designate one of the hypotheses as the *null hypothesis*, denoted H_0, with the corresponding subset of parameter space denoted Θ_0. The complementary hypothesis is called the *alternative hypothesis*, and is usually denoted by H_1 or H_a. The corresponding subset of parameter space is $\Theta_1 = \Theta \setminus \Theta_0$.

The testing procedure is specified by the subset C of the sample space, called the *critical* region, and the rule

$$\text{reject } H_0 \text{ if } \mathbf{x} \in C.$$

Observe that in this formulation the two hypotheses, H_0 and H_1, are treated symmetrically, and the designation of one of them as the null hypothesis is arbitrary. Subsequently this symmetry will be lost, and the null hypothesis will play a different role than the alternative.

To formulate the meaningful criteria that will eventually allow to choose the best test, we introduce some auxiliary definitions.

Definition 13.2.1 A hypothesis $H\colon \theta \in \Theta_0$ is called *simple* if it completely specifies the distribution of the sample; otherwise, it is called *composite*. □

As mentioned before, two types of error can be made as a result of our decision: a true null hypothesis can be rejected, or a false null hypothesis can be not rejected.

It goes without saying that we would like to minimize the probabilities of both types of errors. The trouble here is twofold. First, the probability of errors depends typically on the value of θ, so the phrase "probability of rejecting null hypothesis when it is true" does not have unique meaning if the null hypothesis is composite. In fact, there are many values of θ for which null hypothesis is true, and the probability in question depends on the particular value of θ. The same remark applies to the alternative hypothesis if it is composite.

The second source of difficulty lies in the fact that under any reasonable definition of the probability of errors, the two probabilities are inversely related. If the probability of one type of error decreases, then the probability of error of the other type increases. To obtain a convenient tool for expressing the probabilities of errors associated with a test, we introduce the following definition, central for the whole theory of testing statistical hypotheses:

[25] It is important to realize that the phrases "accept H" and "reject H" are to be interpreted as decisions regarding future behavior rather than regarding truth or falsehood of H. Thus, to "accept H" means (in majority of practical situations) to proceed "*as if H were true*." In fact it would be unrealistic to expect more: our decision is based on observation of a random sample; hence we are always exposed to a risk of a wrong decision, namely rejecting H when in reality H is true, or accepting H when in reality it is false. The best one can do is to control the probabilities of these errors. Instead of stating that "hypothesis H_0 is accepted," the verdict is often phrased in a more cautious way such as "there is not enough evidence to reject H_0."

Definition 13.2.2 If C is the critical region of a test of the null hypothesis H_0: $\theta \in \Theta_0$, the function $\pi_C(\theta)$, defined on parameter space Θ as

$$\pi_C(\theta) = P_\theta\{\mathbf{X} \in C\}, \tag{13.3}$$

is called the *power* (or *power function*) of test C. □

Thus, power is the probability of rejecting the null hypothesis if the value of the parameter is θ. Qualitatively speaking, a good test should have high power when $\theta \notin \Theta_0$, since for such θ the null hypothesis is not true, and rejecting it is a correct decision. On the other hand, for $\theta \in \Theta_0$ the power of a good test should be low, since for such θ the null hypothesis is true, and rejecting it is a wrong decision. Let us now consider a few examples.

■ **EXAMPLE 13.6**

A supermarket buys oranges from a certain company, that claims that the fraction of unacceptable fruit (e.g., rotten) in each shipment does not exceed 3%. The supermarket is willing to accept such (but not higher) percentage of bad fruit. The procedure agreed upon specifies that a random sample of 30 oranges will be selected from each shipment (the method of sampling is typically also a part of the protocol). The shipment will be accepted if there is no more than one unacceptable fruit in this sample, and rejected (or bought at a discounted price, etc.) otherwise.

First, assume that the results of sampling constitute independent Bernoulli trials with a probability of "success" (finding a rotten orange) equal to θ.

The range of θ is (at least theoretically) the whole interval $\Theta = [0, 1]$. The random sample $X_1, \ldots, X_{30}$ is drawn from the distribution $f(x, \theta) = \theta^x(1-\theta)^{1-x}, x = 0, 1$. The supplier's claim is that $\theta \leq 0.03$. Let us take this claim as the null hypothesis so that $\Theta_0 = [0, 0.03]$, while the alternative is $\Theta_1 = (0.03, 1]$.

The agreed procedure specifies that the supplier's claim (null hypothesis) is rejected if there are two or more bad oranges in the sample, that is, if

$$S_{30} = \sum_{i=1}^{30} X_i \geq 2.$$

This means that the critical region is the set $C = \{2, 3, \ldots, 30\}$. The elements in C are values of S_{30} that lead to rejection of the supplier's claim that $\theta \in \Theta_0$.

The power function of this test is

$$\begin{aligned} \pi_C(\theta) &= P_\theta\{S_{30} \geq 2\} = 1 - P_\theta\{S_{30} = 0\} - P_\theta\{S_{30} = 1\} \\ &= 1 - (1-\theta)^{30} - 30\theta(1-\theta)^{29} = 1 - (1-\theta)^{29}(1 + 29\theta). \end{aligned}$$

Function $\pi_C(\theta)$ is strictly increasing (see Figure 13.1) from $\pi_C(0) = 0$ to $\pi_C(1) = 1$, and its highest value on the set Θ_0 (null hypothesis) is attained at $\theta = 0.03$, and equals

$$\pi_C(0.03) = \max_{\theta \in \Theta_0} \pi_C(\theta) = 0.227.$$

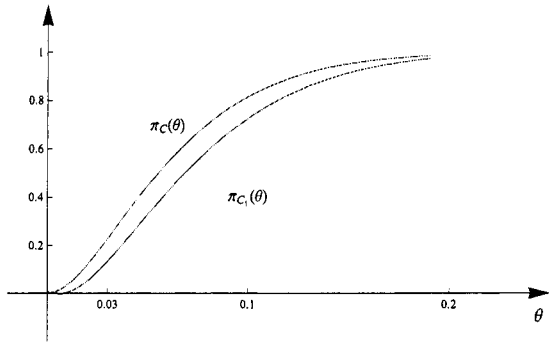


Figure 13.1 Power functions $\pi_C(\theta)$ and $\pi_{C_1}(\theta)$

The supermarket manager would probably be content with such a procedure: if the alternative is true, that is, if the fraction θ of unacceptable fruits exceeds 0.03, then he has at least a 22.7% chance of detecting it. In fact this chance increases rather fast with θ. For instance, for $\theta = 0.05$, we have $\pi_C(\theta) = 0.446$, and for $\theta = 0.1$, we have already $\pi_C(\theta) = 0.816$.

On the other hand, the supplier will not be too happy. The chances of having a shipment rejected while the standard of $\theta \leq 0.03$ is actually met can be as high as 22.7% (if θ is very close to 0.03).

■ **EXAMPLE 13.7**

Continuing Example 13.6, a possible solution to the problem of accepting orange shipments is to change the procedure and reject the shipment if three or more (rather than two or more) fruits are unacceptable. The critical region, C_1, for rejecting the claim $\theta \leq 0.03$ is now $\{S_{30} \geq 3\}$, so

$$\begin{aligned}\pi_{C_1}(\theta) &= 1-(1-\theta)^{30}-30\theta(1-\theta)^{29}-435\theta^2(1-\theta)^{28} \\ &= 1-(1-\theta)^{28}(1+28\theta+406\theta^2).\end{aligned}$$

Again, this is an increasing function, and (see Figure 13.1)

$$\pi_{C_1}(0.03) = \max_{\theta\in\Theta_0} \pi_{C_1}(\theta) = 0.060.$$

This time the supplier is happier, but the manager is not. He might feel that the chance of accepting a shipment with $\theta = 0.1$, equal to $1 - \pi_{C_1}(0.1) = 0.411$, is too high.

These two examples illustrate the fact that an attempt to decrease the probabilities of one type of error by changing the critical region leads to an increase of the probabilities of error of the other type. The only way to decrease both probabilities of error at the same time is to increase the sample size.

■ **EXAMPLE 13.8 Randomization**

In the case of orange shipments discussed in Examples 13.6 and 13.7, one could also suggest the following procedure, which does not involve any increase of the sample size and yet leads to a procedure that may be acceptable to both the manager and the supplier.

Since rejection of the shipment if $S_{30} \geq 2$ is too favorable for the manager, while rejection if $S_{30} \geq 3$ is too favorable for the supplier, one could suggest the following procedure: if S_{30} is 0, 1, or 2, the shipment is accepted by the store; if S_{30} is 4, 5, ..., 30, the shipment is rejected by the store. However, if $S_{30} = 3$, a coin is tossed, and the shipment is accepted or rejected depending on the outcome of the toss. Actually, it is not necessary to use a fair coin. One can also activate some other random mechanism such that the probability of the shipment being rejected if $S_{30} = 3$ is some fixed number γ.

The power function $\pi(\theta) = P_\theta\{H_0 \text{ is rejected}\}$ now equals

$$\pi(\theta) = \gamma P\{S_{30} = 3\} + P\{S_{30} \geq 4\},$$

where $0 < \gamma < 1$, and it lies between power functions of the two procedures considered in Examples 13.6 and 13.7. It is possible that the manager and the supplier can negotiate a value of γ that is acceptable for both of them.

The randomized procedure described in Example 13.8 is somewhat controversial. If testing statistical hypotheses is regarded as a process aimed at establishing the truth or falsehood of some statements about the experimental situation under analysis, then indeed, declaring the truth of one hypothesis on the basis of a flip of a coin may appear appalling and ridiculous. However, in testing statistical hypotheses according to the original intention of Neyman and Pearson (who built the foundations of this theory), the rejection and acceptance of the hypothesis are not statements about truth and falsehood. They are intended to be the guidelines for future actions. As Neyman (1950, pp. 259–260) wrote: "The terms "accepting" and "rejecting" a statistical hypothesis are very convenient and are well established. It is important, however, to keep their exact meaning in mind and to discard various additional implications which may be suggested by intuition. Thus, to accept a hypothesis H means only to take an action A rather than B. This does not mean that we necessarily believe that the hypothesis H is true. Also, if the application of the rule of inductive behavior "rejects" H, this means only that the rule prescribes action B and does not mean that we believe that H is false."

In light of this interpretation, the randomization of a decision is justifiable.

■ **EXAMPLE 13.9**

Let us consider a random sample from the U$[0, \theta]$ distribution. We have here $\Theta = (0, \infty)$. Suppose that we want to test the null hypothesis $H_0 : 5 \leq \theta \leq 6$. The alternative hypothesis is $H_1 : \theta < 5$ or $\theta > 6$. Thus $\Theta_0 = [5, 6]$ and $\Theta_1 = (0, 5) \cup (6, \infty)$.

We consider two tests and determine their power. First, we know (see Example 12.58) that $T_1 = X_{n:n}$ is the sufficient statistic for θ, and we might try to base the testing procedure on it. Clearly, if $T_1 \geq 6$, we should reject

the null hypothesis, since it simply cannot be true in this case. Similarly we should reject the null hypothesis if T_1 is too small, for example, $T_1 < 4.6$. Finally, we may argue that the values of T_1 slightly below 6 are also a good indication that H_0 may be false, since we always have $T_1 < 6$. The actual thresholds will depend on n. For the critical region

$$C = \{(x_1, \ldots, x_n) : \ T_1 < 4.6 \text{ or } T_1 > 5.9\}, \tag{13.4}$$

the power of the test is

$$\pi_C(\theta) = P_\theta\{T_1 < 4.6\} + P_\theta\{T_1 > 5.9\}, \tag{13.5}$$

where

$$P_\theta\{T_1 < 4.6\} = \begin{cases} 1 & \text{if } \theta \leq 4.6 \\ \left(\frac{4.6}{\theta}\right)^n & \text{if } \theta > 4.6. \end{cases}$$

Similarly

$$\begin{aligned} P_\theta\{T_1 > 5.9\} &= 1 - P_\theta\{T_1 \leq 5.9\} \\ &= \begin{cases} 0 & \text{if } \theta \leq 5.9 \\ 1 - \left(\frac{5.9}{\theta}\right)^n & \text{if } \theta > 5.9. \end{cases} \end{aligned}$$

Thus

$$\pi_C(\theta) = \begin{cases} 1 & \text{for } \theta \leq 4.6 \\ \left(\frac{4.6}{\theta}\right)^n & \text{for } 4.6 < \theta \leq 5.9 \\ \left(\frac{4.6}{\theta}\right)^n + 1 - \left(\frac{5.9}{\theta}\right)^n & \text{for } \theta > 5.9. \end{cases}$$

The graph of this power function for $n = 15$ is presented in Figure 13.2.

Let us also consider a test based on statistic $\overline{X}$. We know that $E(\overline{X}) = \theta/2$. To have a comparison with the test (13.4), consider now a test with the critical region

$$C_1 = \{(x_1, \ldots, x_n) : \ \overline{x} < 2.30 \ \text{ or } \ \overline{x} > 2.95\}. \tag{13.6}$$

To determine the exact distribution of $\overline{X}$, while possible in principle, is very cumbersome, so we will rely on the approximation provided by the central limit theorem, using the fact that $\mathrm{Var}(X_i) = \theta^2/12$. The distribution of $\overline{X}$ is therefore approximately normal $\mathrm{N}(\theta/2, \theta^2/12n)$, and the power of the test (13.6) is

$$\begin{aligned} \pi_{C_1}(\theta) &= P\{\overline{X} < 2.30\} + P\{\overline{X} > 2.95\} \\ &\approx 1 - P\left\{\frac{2.30 - \theta/2}{\theta/\sqrt{12n}} \leq Z \leq \frac{2.95 - \theta/2}{\theta/\sqrt{12n}}\right\} \\ &= 1 - \left[\Phi\left(\frac{2.95 - \theta/2}{\theta/\sqrt{12n}}\right) - \Phi\left(\frac{2.30 - \theta/2}{\theta/\sqrt{12n}}\right)\right]. \end{aligned}$$

The graph of this function, for $n = 15$, is presented in Figure 13.2.

As already mentioned, the expectation of a "good" test is that it should have the power as low as possible on the null hypothesis, and as high as possible on the alternative hypothesis. A glance at Figure 13.2 shows that the test C, based on the sufficient statistic, is much better than the test C_1 based on the sample mean $\overline{X}$.

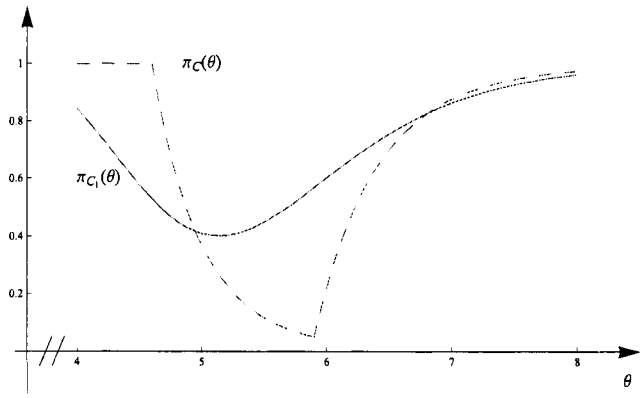


Figure 13.2 Power functions of tests C and C_1

As can be seen, for all θ in the null hypothesis H_0 we have $\pi_C(\theta) < \pi_{C_1}(\theta)$. Actually,

$$\sup_{\theta \in \Theta_0} \pi_C(\theta) = \max(\pi_C(5), \pi_C(6)) = \max(0.2863, 0.2414) = 0.2863.$$

On the other hand,

$$\sup_{\theta \in \Theta_0} \pi_{C_1}(\theta) = \max(\pi_{C_1}(5), \pi_{C_1}(6)) = \max(0.4094, 0.6033) = 0.6033.$$

For the alternative hypothesis, when we want the power to be as high as possible, we have $\pi_C(\theta) > \pi_{C_1}(\theta)$, except for two rather narrow ranges immediately to the left of $\theta = 5$ and to the right of $\theta = 6$. The test based on critical region C may be considered "better" than the test based on critical region C_1.

■ **EXAMPLE 13.10**

Finally, consider one of the classical examples, a test hypotheses about the mean in a normal distribution with known variance. We assume that $X_1, \ldots, X_n$ is a random sample from the $N(\mu, 1)$ distribution (assumption that $\sigma = 1$ is not a restriction of generality, as it simply amounts to choosing the appropriate unit for X_i's). Suppose that we want to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$. In this case the null hypothesis is simple while the alternative is composite.

The procedure will be based on the sample mean $\overline{X}$. If this average deviates "too much" from μ_0, then the null hypothesis should be rejected. This means taking the critical region of the form

$$C = \{(x_1, \ldots, x_n) : |\overline{X} - \mu_0| > k\}, \tag{13.7}$$

where k is some constant.

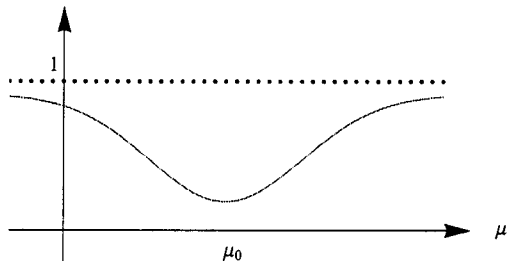


Figure 13.3 Power of the two-sided test C

Recalling that $\text{Var}(\overline{X}) = 1/2$, the power function of this test is

$$\begin{aligned}
\pi_C(\mu) &= P_\mu\{|\overline{X} - \mu_0| > k\} = 1 - P_\mu\{|\overline{X} - \mu_0| \le k\} \\
&= 1 - P_\mu\{\mu_0 - k \le \overline{X} \le \mu_0 + k\} \\
&= 1 - P\left\{\frac{-k + \mu_0 - \mu}{1/\sqrt{n}} \le \frac{\overline{X} - \mu}{1/\sqrt{n}} \le \frac{k + \mu_0 - \mu}{1/\sqrt{n}}\right\} \\
&= 1 - P\left\{\frac{-k + \mu_0 - \mu}{1/\sqrt{n}} \le Z \le \frac{k + \mu_0 - \mu}{1/\sqrt{n}}\right\}.
\end{aligned}$$

If H_0 is true, then $\mu = \mu_0$ and the power equals

$$\pi_C(\mu_0) = 1 - P\{|Z| \le k\sqrt{n}\},$$

a quantity that depends on the chosen threshold k. Also, for any fixed $k > 0$, we have $\pi_C(\mu_0) \to 0$ as $n \to \infty$. On the other hand, for fixed $k > 0$ and n, we have $\pi_C(\mu) \to 1$ when $\mu \to \pm\infty$.

This means that for a fixed sample size n we can choose the critical region so that the probability of the type I error is equal to any preassigned level. The probability $1 - \pi_C(\mu)$ of the type II error approaches 0 as μ moves away from the null hypothesis (see Figure 13.3). How quickly it happens (i.e., how fast the power function approaches 1 as $|\mu - \mu_0|$ increases) depends on n.

The main issues raised by the examples of this section may be summarized as follows:

1. A test, being a procedure that ultimately leads to a choice between two hypotheses, is equivalent to specifying a critical region C in the space of observations.

2. The performance of the test C is described by the power function defined as $\pi(\theta) = P\{\text{rejecting } H_0 \text{ if } \theta \text{ is the true parameter value}\}$. All criteria of choice of a test should ultimately be expressed in terms of the power function. Any such criterion should conform to the intuitive requirement that the power

of a test should be as high as possible for θ in the alternative hypothesis, and as low as possible for θ in the null hypothesis.

3. The class of all possible tests (i.e., class of all possible critical regions C) is very rich, and *any* reduction of this class will facilitate the search for the best set (in whichever way the optimality is ultimately defined).

PROBLEMS

13.2.1 Let X_1, X_2 be a random sample from the U$[\theta, \theta + 1]$ distribution. In the test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$, H_0 is rejected when $X_1 + X_2 > k$. Find the power function of the test that has probability of the type I error equal 0.05

13.2.2 Consider the following procedure for testing the hypothesis $H_0 : p \geq 0.5$ against the alternative $H_1 : p < 0.5$ in BIN$(10, p)$ distribution. We take observation X_1, and reject H_0 if $X_1 = 0$ or accept H_0 if $X_1 \geq 9$; otherwise, we take another observation X_2 (with the same distribution as X_1 and independent of it). Then we accept or reject H_0 depending on whether $X_1 + X_2 \geq 5$ or $X_1 + X_2 < 5$. Determine and graph the power function of this procedure.

13.2.3 Consider three tests C_1, C_2, and C_3 of the same hypothesis, performed independently (e.g., for each of these tests the decision is based on a different sample). Consider now the following three procedures:
A: Reject H_0 only if all three tests reject it; otherwise, accept H_0,
B: Reject H_0 only if at least two tests reject it; otherwise, accept H_0,
C: Reject H_0 only if at least one test rejects it; otherwise, accept H_0.
(i) Express the power functions of procedures A, B, and C through power functions of tests C_1–C_3. **(ii)** Assuming that power functions of tests C_1–C_3 are the same ($\pi_{C_1}(\theta) = \pi_{C_2}(\theta) = \pi_{C_3}(\theta) = \pi(\theta)$), graph the power functions of procedures A, B, and C and determine where each the procedures A, B, and C performs better than the "component" test C_i.

13.2.4 Let X_1, X_2 be a random sample of size 2 from the U$[0, \theta]$ distribution. We want to test $H_0 : \theta = 3$ against $H_1 : \theta = 2$ (observe that H_0 and H_1 do not exhaust all possibilities). **(i)** H_0 will be rejected if $\overline{X} < c$. Find c such that the probability of a type I error of the test is $\alpha = 0.05$ and determine its power function. **(ii)** Answer (i) if H_0 will be rejected when $X_{2:2} < c$. **(iii)** Compare the power functions of both tests.

13.2.5 An urn contains five balls, r red and $5 - r$ white. The null hypothesis states that all balls are of the same color (i.e., $H_0 : r = 0$ or $r = 5$). Suppose that we take a sample of size 2 and reject H_0 if the balls are of different colors. Find the power of this test for $r = 0, \ldots, 5$ if the sample is drawn: **(i)** Without replacement. **(ii)** With replacement. **(iii)** In each case find the probability of a type I error. **(iv)** Answer (i)–(iii) if the null hypothesis is now $H_0 : r = 1$ or $r = 2$, and it is rejected when all k balls selected are white.

13.2.6 Let $X_1, \ldots, X_9$ be a random sample from the N$(\mu, 1)$ distribution. To test the hypothesis $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$, one uses the test "reject H_0 if $3 \leq \overline{X} \leq 5$." Find the power function and show that this is a bad test.

13.3 MOST POWERFUL TESTS

If in the parametric situation considered in this chapter both hypotheses, null and alternative, are simple, it means that we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ where θ_0 and θ_1 are two parameter values.

To simplify the notation, let us write $f(x, \theta_0) = f_0(x)$ and $f(x, \theta_1) = f_1(x)$. The only assumption about f_0 and f_1 is that they represent different probability distributions. In the discrete case, this means that $P\{X = x|H_0\} = f_0(x) \neq f_1(x) = P\{X = x|H_1\}$ for some x. In the continuous case, f_0 and f_1 are densities, and it is not enough that f_0 and f_1 differ at some isolated point, or even on a countable set of points. Thus in the continuous case we assume that

$$\int_A f_0(x)dx \neq \int_A f_1(x)dx$$

for some set A.

We present the motivation for steps of the proof of the reduction principle theorem that will be introduced later. We consider the case of a single observation, but the extension to the case of n observations will be obvious.

Suppose that we determined some critical region C. Thus the test is

$$\text{reject } H_0 \text{ if } x \in C. \tag{13.8}$$

Since both hypotheses are simple, we can determine the probabilities of errors of both types. Thus the probability of rejecting H_0 if it is true (type I error) equals

$$\alpha = P\{x \in C|H_0\} = \begin{cases} \int_C f_0(x)dx \\ \sum_{x \in C} f_0(x), \end{cases} \tag{13.9}$$

depending on whether we deal with a continuous or a discrete case. Similarly the probability of not rejecting H_0 if H_1 is true (type II error) is

$$\beta = P\{x \notin C|H_1\} = \begin{cases} 1 - \int_C f_1(x)dx \\ 1 - \sum_{x \in C} f_1(x). \end{cases} \tag{13.10}$$

Since to decrease β we must increase the integral in (13.10), the problem is to choose C so as to maximize $\int_C f_1(x)dx$. It will be convenient to partition the set of values of x into four sets A_0, A_1, A_2, A_3, depending on which of the two densities is zero and which is positive (see Figure 13.4).

Let us first consider the set

$$A_1 = \{x : f_0(x) = 0, f_1(x) > 0\}. \tag{13.11}$$

If any part of this set lies outside C, then we can improve the test by including it into C, that is, take the test with critical region $C^* = C \cup A_1 = C \cup (C^c \cap A_1)$.

Indeed, for the test C^* (in the continuous case) we have

$$\alpha^* = \int_{C^*} f_0(x)dx = \int_C f_0(x)dx + \int_{C^c \cap A_1} f_0(x)dx = \alpha,$$

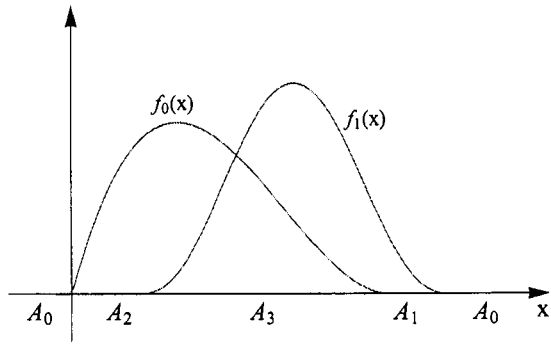


Figure 13.4 Partition into sets A_0, A_1, A_2, A_3

since $f_0(x) \equiv 0$ on $C^c \cap A_1$. Thus test C^* has the same probability of a type I error as test C. However (again for the continuous case), we have

$$\begin{aligned} \beta^* = 1 - \int_{C^*} f_1(x)dx &= 1 - \int_C f_1(x)dx - \int_{C^c \cap A_1} f_1(x)dx \\ &\leq 1 - \int_C f_1(x)dx = \beta, \end{aligned}$$

which means that the probability of a type II error has decreased. This argument shows that the set A_1 should be totally contained in the critical region C.

On the other hand, consider the set

$$A_2 = \{x : f_0(x) > 0, f_1(x) = 0\}. \tag{13.12}$$

A reasoning analogous to that carried for A_1 shows that the entire set A_2 should lie outside the critical region C. Next, the set

$$A_0 = \{x : f_0(x) = f_1(x) = 0\}$$

plays no role in our reasoning. The way in which this set is partitioned between C and its complement has no effect on α and β.

Now let

$$A_3 = \{x : f_0(x) > 0 \text{ and } f_1(x) > 0\}. \tag{13.13}$$

The problem reduces to finding the best way of partitioning A_3 between C and its complement, in order to improve C.

To simplify the argument, consider the discrete case. Suppose that we have already selected some set C as a candidate for a critical set, and that $C \supset A_1$ and $A_2 \subset C^c$.

Let $x^* \notin C$ (with $x^* \in A_3$), and consider the consequences of changing C by including x^* in it. The probability of a type I error will change to

$$\sum_{x \in C} f_0(x) + f_0(x^*) = \alpha + f_0(x^*) > \alpha.$$

On the other hand, the probability of a type II error will change to

$$1 - \left[\sum_{x \in C} f_1(x) + f_1(x^*)\right] = \beta - f_1(x^*) < \beta.$$

Thus α increased by $f_0(x^*)$ while β decreased by $f_1(x^*)$.

In a similar way, suppose that $x^{**} \in C$ and that we remove it from C. Then α will decrease by $f_0(x^{**})$ while β will increase by $f_1(x^{**})$.

Consequently, if $x^* \notin C$ and $x^{**} \in C$, switching the role of these points results in the following change:

$$\begin{aligned} \alpha \quad &\text{changes to} \quad \alpha + (f_0(x^*) - f_0(x^{**})), \\ \beta \quad &\text{changes to} \quad \beta + (f_1(x^{**}) - f_1(x^*)). \end{aligned}$$

Since we want to minimize both α and β, such a change should *always* be carried out if $f_0(x^*) - f_0(x^{**}) \leq 0$ and $f_1(x^{**}) - f_1(x^*) \leq 0$, with at least one inequality being strict.

The inequalities above mean (remembering that x^* and x^{**} are in A_3) that $f_1(x^{**}) \leq f_1(x^*)$ and $f_0(x^{**}) \geq f_0(x^*) > 0$. So since one inequality is strict

$$\frac{f_1(x^{**})}{f_0(x^{**})} < \frac{f_1(x^*)}{f_0(x^*)}. \tag{13.14}$$

We therefore obtained the following principle:

Reduction Principle In choosing the critical regions, one should restrict the considerations to sets based on the *likelihood ratio* $f_1(x)/f_0(x)$ of the form

$$C_k = \left\{x : \frac{f_1(x)}{f_0(x)} \geq k\right\}. \tag{13.15}$$

□

Indeed, if a critical set is not of the form (13.15), then there exist points x^* and x^{**}, with $x^* \notin C$ and $x^{**} \in C$ and such that (13.14) holds. Then a better critical region can be obtained by exchanging the roles of x^* and x^{**}.

It should be noticed that we did not solve the problem of finding the best test. In fact we did not even specify the criterion to be optimized. The reduction principle above tells us only which tests should not be used—in other words—it specifies the class of tests from which the choice should be made, provided only that the optimality criterion is compatible with the general motive to decrease *both* probabilities of errors.

Two obvious ways of defining the criterion to be optimized are as follows:

1. Impose an upper bound on one of the probabilities of errors, and minimize the probability of the other kind of error.

2. Minimize a linear combination of the two error probabilities.

These two approaches are closely related, the first being used in the original Neyman-Pearson lemma dating back to 1933. In formulating the theorem below, we will consider a general situation of a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from one of the

two distributions f_0 or f_1. The null hypothesis asserts that the distribution is f_0. For $i = 0, 1$ we will write $f_i(\mathbf{x}) = f_i(x_1) \times \cdots \times f_i(x_n)$. The critical regions under consideration are now subsets of the n-dimensional Euclidean space.

Theorem 13.3.1 (Neyman-Pearson Lemma) *Let C^* be a critical region that has the following properties: There exists a constant $k > 0$ such that:*

(a) *If $f_1(\mathbf{x})/f_0(\mathbf{x}) > k$, then $\mathbf{x} \in C^*$.*

(b) *If $f_1(\mathbf{x})/f_0(\mathbf{x}) < k$, then $\mathbf{x} \notin C^*$ with points at which $f_1(\mathbf{x})/f_0(\mathbf{x}) = k$ partitioned between C^* and its complement in some way.*

Let C be any critical region. Then $\alpha(C) \leq \alpha(C^*)$ implies $\beta(C) \geq \beta(C^*)$, and if $\alpha(C) < \alpha(C^*)$, then $\beta(C) > \beta(C^*)$.

Proof. Without danger of confusion we can use the symbols A_0, A_1, A_2, and A_3 for the four sets in $\mathbf{R}^n$, depending on which of the joint densities $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ is zero and which is strictly positive. If we assume that $c/0 = \infty$ for any $c > 0$, then the critical region C^* in the lemma is such that

$$A_1 \subset C^*, \quad A_2 \subset (C^*)^c. \tag{13.16}$$

Indeed, we have $f_1(\mathbf{x})/f_0(\mathbf{x}) = \infty > k$ on A_1 and $f_1(\mathbf{x})/f_0(\mathbf{x}) = 0 < k$ on A_2. The theorem follows from the argument for the one-dimensional case preceding this proof: if C does not meet conditions (13.16), then one or both of its error probabilities can be improved, as stated in the assertion of the Neyman-Pearson lemma.

Next, the set A_0 plays no role, and the question remains about points $\mathbf{x}$ in A_3. In the discrete case, the argument given for one-dimensional case remains valid: if C is not of the form specified in the lemma, then there exist $\mathbf{x}^* \notin C$ and $\mathbf{x}^{**} \in C$ such that (13.15) holds, and by switching these points we improve both error probabilities. It remains therefore to prove the theorem in the continuous case. Consider the sets

$$C^* \cap C^c = \left\{ \mathbf{x} : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} > k, \mathbf{x} \notin C \right\}$$

and

$$(C^*)^c \cap C = \left\{ \mathbf{x} : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} < k, \mathbf{x} \in C \right\}.$$

We have here

$$\begin{aligned}
\alpha(C^*) - \alpha(C) &= \int_{C^*} f_0(\mathbf{x})d\mathbf{x} - \int_C f_0(\mathbf{x})d\mathbf{x} \\
&= \int_{C^* \cap C^c} f_0(\mathbf{x})d\mathbf{x} - \int_{C \cap (C^*)^c} f_0(\mathbf{x})d\mathbf{x} \\
&\leq \int_{C^* \cap C^c} \frac{1}{k} f_1(\mathbf{x})d\mathbf{x} - \int_{C \cap (C^*)^c} \frac{1}{k} f_1(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{k} \int_{C^c} f_1(\mathbf{x})d\mathbf{x} - \frac{1}{k} \int_{(C^*)^c} f_1(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{k}[\beta(C) - \beta(C^*)].
\end{aligned}$$

The inequality between the extreme terms completes the proof: if $\alpha(C) \leq \alpha(C^*)$, then the left-hand side is nonnegative and so must be the right-hand side, which means that $\beta(C) \geq \beta(C^*)$. If the left-hand side is strictly positive, so must be the right-hand side. □

We will now formulate the analogue of the Neyman-Pearson lemma in the case where the criterion to be minimized is a linear combination of the two error probabilities. The proof, which follows closely the reasoning in the proof of Neyman-Pearson lemma, will be omitted.

Theorem 13.3.2 *Suppose that in testing $H_0 : f = f_0$ against $H_1 : f = f_1$, it is desired to find the critical region C^* in n-dimensional space, such that for any other critical region C we have*

$$A\alpha(C^*) + B\beta(C^*) \leq A\alpha(C) + B\beta(C),$$

where $A > 0, B > 0$ are given constants. Then C^* contains all points $\mathbf{x}$ such that $f_1(\mathbf{x})/f_0(\mathbf{x}) > A/B$, and its complement contains all points $\mathbf{x}$ such that $f_1(\mathbf{x})/f_0(\mathbf{x}) < A/B$. The points where $f_1(\mathbf{x})/f_0(\mathbf{x}) = A/B$ can be allocated between C^* and its complement in an arbitrary way.

Since in the case of simple hypotheses the power of the test with critical region C is $1 - \beta(C)$, the Neyman-Pearson lemma gives us in effect a rule of constructing the *most powerful* test, with the preassigned probability $\alpha(C)$ of type I error (as we will see, this probability is determined by the choice of a constant k).

We now introduce three important definitions. Let C be the critical region for testing the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$, and let $\pi_C(\theta) = P_\theta\{\mathbf{X} \in C\}$ be the power of the test.

Definition 13.3.1 The *size* of the test C is defined as

$$\overline{\alpha}(C) = \sup_{\theta \in \Theta_0} \pi_c(\theta). \qquad \square$$

Thus, the size of the test of a simple null hypothesis is the probability of a type I error, while for the composite null hypotheses the size is the least upper bound for all probabilities of type I errors.

Definition 13.3.2 Any number α satisfying the inequality $\overline{\alpha}(C) \leq \alpha$ is called the *level* of test C, and a test C satisfying $\overline{\alpha}(C) \leq \alpha$ will be called an *α-level* test. □

According to this definition a test with $\overline{\alpha}(C) = 0.01$ is a 1%-level test, as well as a 5%-level test, and so on.

Finally, we define a significance level. This is not so much a property of a test as a constraint imposed by the statistician.

Definition 13.3.3 If in testing the hypothesis H_0 the statistician chooses a number α_0 $(0 < \alpha_0 < 1)$ and decides to use only those tests C whose size satisfies the inequality $\overline{\alpha}(C) \leq \alpha_0$, then α_0 is called the *significance level* chosen by the user. □

It is often felt that a report that a given hypothesis was rejected at the significance level $\alpha = 0.05$ is not informative enough (e.g., the hypothesis might also be rejected

at the significance level $\alpha = 0.01$ or even lower). Thus it is customary to report the so called *p-value*, defined as the lowest level at which the null hypothesis would be rejected by the test. More intuitively, the p-value of the results actually observed is the probability that if the experiment were repeated, we would obtain results that give at least as strong evidence for the alternative hypothesis (or equivalently, as strong evidence against the null hypothesis) as the present result. For instance, suppose that we use statistic T for testing, and we reject the null hypothesis if T is large. The result which we observe is $T = t_0$, say. Then the p-value of this result is $P\{T \geq t_0 | H_0\}$, so that the smaller is the p-value, the stronger is the evidence that suggests rejection of the null hypothesis.[26]

In the case of two-sided tests based on statistic T with a symmetric distribution, the p-value of the result t_0 is defined as $2P(T > |t_0|)$. Here the rationale is that the observations more "in favor" of the alternative hypothesis are those above $|t_0|$ and below $-|t_0|$. In the case of tests based on a statistic that does not have a symmetric distribution, the ideas behind the concept of p-value become rather fuzzy, and there is no definition on which all statisticians agree. Consequently, one may find various "competing" definitions in different texts. For a review of these definitions, see Gibbons and Pratt (1975).

We will now illustrate the procedure of test selection by some examples. In the four initial examples we will also illustrate another aspect: how an empirical hypothesis becomes "translated" into a statistical hypothesis.

■ **EXAMPLE 13.11**

Suppose that a scientist found the exact location where memory is stored in the rats' brain and needs to confirm it by an experiment. Ten rats are trained to find food in a maze. In running through the maze, the rat makes three binary choices and as a result ends in one of eight final locations, of which only one contains food. After some surgical intervention in the part of the brain that—according to the researcher's hypothesis—stores the acquired knowledge followed by a period of healing, the rats run the maze again, and the experimenter observes the number X of rats that reach the arm with food. Without any memory left, each rat has probability $(1/2)^3 = 0.125$ of finding the way to the food on the first trial. Suppose that—according to the experimenter—if this probability is as high as 0.3, it means that part of the memory must be stored in another (undamaged) region of the brain. If the rats run through the maze independently, and each has the same probability p of choosing the path leading to food, then X has a binomial distribution BIN(10, p). We have now two hypotheses: one asserting that $p = 0.125$ (if the researcher is correct in identifying the memory storage region of the brain), and the other asserting that $p = 0.3$ (if the researcher is wrong). Suppose that from the researcher's point of view the error of rejecting the hypothesis $p = 0.125$ (i.e., no memory

[26]High p-values (close to 1) might indicate that the data were manipulated so as to make them "more conforming" to the null hypothesis. The rationale here is as follows: Suppose that the p-value of the data is 0.999. This means that the "fit" to H_0 is so perfect that only once in a 1000 times we would observe a better fit. The same principles which lie at the foundation of testing allow us to reach the conclusion that the data were fudged to fit H_0.

is left), if in fact it is true, is more serious than accepting $p = 0.125$ if in fact $p = 0.3$.

In this case we set the null hypothesis as $H_0 : p = 0.125$ and the alternative is $H_1 : p = 0.3$. Consequently $f_0(x) = \binom{10}{x}(0.125)^x(0.875)^{10-x}$ and $f_1(x) = \binom{10}{x}(0.3)^x(0.7)^{10-x}$. According to the Neyman-Pearson lemma, we should choose the critical region of the form $\{x : f_1(x)/f_0(x) \geq k\}$ for some k. But

$$\frac{f_1(x)}{f_0(x)} = \frac{(0.3)^x(0.7)^{10-x}}{(0.125)^x(0.875)^{10-x}} = 0.107 \cdot 3^x.$$

The inequality $\{x : f_1(x)/f_0(x) \geq k\}$ is equivalent to the inequality $x \geq k^*$ for some k^*.This means that the critical region C is formed of the right tail of the distribution of X. Since the possible values of X are 0, 1, ..., 10, we must simply determine the smallest value that belongs to C. Here the choice depends on the significance level α_0 of the test. Suppose that the experimenter decides to use $\alpha_0 = 0.05$. Then k^* must be such that $P\{X \geq k^* | p = 0.125\} \leq 0.05$; that is, $P\{X < k^* | p = 0.125\} \geq 0.95$. We have here $P\{X < 4 | p = 0.125\} = 0.973$ while $P\{X < 3 | p = 0.125\} = 0.881$. It follows that we must reject the null hypothesis $H_0 : p = 0.125$ if $X \geq 4$; that is, if four or more rats find their way to the food on the first try. This test has the probability of a type I error equal to $1 - 0.973 = 0.027$ and the probability of a type II error equal $P\{X < 4 | p = 0.3\} = 0.649$.

Some comments about this example appear to be in order. First, we see that the type II error has rather high probability, whereas the probability of the type I error is below the level of significance. A test with a critical region $C = \{3, 4, \ldots, 10\}$ will have $\alpha(C) = 0.119$, while $\beta(C) = 0.382$. It is clear that among tests satisfying the assertion of the Neyman-Pearson lemma, none has the probability of a type I error equal to the desired significance level 0.05. This is due to the discrete nature of the test statistic X.

A procedure with the probability of a type I error $\alpha_0 = 0.05$ exists if we allow randomized procedures. That is, suppose that we decide to reject H_0 if $X \geq 4$, accept it if $X \leq 2$, and activate some auxiliary random mechanism and reject H_0 with probability γ if $X = 3$. Then the probability of rejecting H_0 if it is true is

$$\gamma P\{X = 3 | p = 0.125\} + P\{X \geq 4 | p = 0.125\} = 0.092\gamma + 0.027,$$

which equals 0.05 if $\gamma = 0.25$.

A procedure that attains a significance level $\alpha_0 = 0.05$ is therefore as follows: If four or more rats reach food on their first trial, then reject H_0. If three rats reach food on the first trial, then toss two coins and reject H_0 if both show up heads. In all other cases accept H_0.

As was already pointed out, such a randomized procedure may be suggested and agreed upon in the process of acceptance or rejection of merchandise. However, in the case of a decision regarding scientific hypotheses, it would be disturbing to rely on the toss of a coin.

A way out from the dilemma is to abandon the concept of a significance level as a quantity imposed by the experimenter. One can then proceed to minimize a linear

combination of errors of type I and type II as in the Theorem 13.3.2. A procedure that attains it never requires randomization.

■ **EXAMPLE 13.12**

Continuing Example 13.11, if we want to find a critical region C so as to minimize the linear combination $10\alpha(C) + \beta(C)$, we must include in the critical region all x such that $f_1(x)/f_0(x) > 10$, which means that

$$\frac{(0.3)^x(0.7)^{10-x}}{(0.125)^x(0.875)^{10-x}} > 10$$

or $3^x > 93.46$; hence $x > 4.13$. Then, if the type I error is considered 10 times as serious as the type II error, the null hypothesis should be rejected only if 5 or more rats find their way to the food on the first attempt.

■ **EXAMPLE 13.13**

Suppose that we have a well with drinking water that is thought to be contaminated with bacteria. A fixed volume V of water from the well is sent to the laboratory for testing. Assume that the admissible norm is N bacteria of a certain kind in volume V, with $5N$ bacteria in volume V indicating an undesirable level of contamination. Here $N > 0$, so that some positive level of contamination is acceptable as safe.

The procedure in the laboratory is as follows: The water sent for analysis is thoroughly mixed, and then a sample of volume v is drawn from it, where $v << V$. A technician inspects the sample under the microscope and reports the number X of bacteria that she sees. It is assumed that each of the bacteria present in the observed sample is recorded by the technician with probability π (assumed known), independently of the other bacteria.

This process is repeated n times, generating reports $X_1, X_2, \ldots, X_n$ of the numbers of bacteria observed in different samples. We assume that the probability $1 - \pi$ of overlooking a bacteria is the same for all technicians.

It is desired to test the null hypothesis that the water in the well is safe against the alternative that it contains an undesirable level of bacteria. To solve this problem, it is first necessary to determine the distributions of the observable random variables and translate the empirical hypotheses stated above into statistical hypotheses.

Let $\lambda_0 = N/V$ be the density of bacteria in the well allowed by the safety standards so that the null hypothesis is $H_0 : \lambda = \lambda_0$, where λ is the actual density of bacteria per unit volume of water in the well. The alternative hypothesis is $H_1 : \lambda = 5N/V = 5\lambda_0 = \lambda_1$.

We assume now that when a sample of volume v is taken, each of the bacteria has the probability v/V of being included in the sample, independently of other bacteria. The assumption that $v << V$ makes the probability v/V small, and from Example 7.14 (about "thinning" of Poisson processes) it follows that in each sample of volume v the actual number of bacteria will have Poisson distribution with mean Mv/V, where M is the actual number of bacteria in volume V. Thus, under the null hypothesis H_0, in each sample of volume v the

number of bacteria will have Poisson distribution with mean $Nv/V = \lambda_0 v$, whereas under the alternative, the mean will be $\lambda_1 v = 5\lambda_0 v$.

We assume also that the total volume of water inspected by technicians, $v_1 + v_2 + \cdots + v_n$, is still small as compared with V so that the Poisson approximation applies to the total number of bacteria in all samples. Finally, the counts $X_1, \ldots, X_n$ (see Example 7.14) have a Poisson distribution with mean $\lambda_0 \pi$ under hypothesis H_0 and mean $\lambda_1 \pi$ under the alternative hypothesis H_1.

Letting $\mathbf{X} = (X_1, \ldots, X_n)$, we know that the test (whether most powerful, in the sense of the Neyman-Pearson lemma or minimizing the linear combination of error probabilities) is based on the likelihood ratio $f_1(\mathbf{x})/f_0(\mathbf{x})$, with values larger than a threshold leading to rejection of the null hypothesis. Denoting $S_n = X_1 + \cdots + X_n$ for $i = 0, 1$, we have

$$f_i(\mathbf{x}) = \prod_{j=1}^{n} f_i(x_j) = \prod_{j=1}^{n} \frac{(\pi\lambda_i v)^{x_j}}{x_j!} e^{-\pi\lambda_i v} = \frac{(\pi v)^{S_n}}{\prod(x_i!)} e^{-n\pi\lambda_i v} \times \lambda_i^{S_n}.$$

Consequently, remembering that $\lambda_1 = 5\lambda_0$, we obtain

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = e^{-4n\pi\lambda_0 v} \times 5^{S_n},$$

so that the null hypothesis H_0 should be rejected if the total count S_n reported by all technicians exceeds some threshold k.

The actual value of k depends on the numerical values of the parameters. For instance, suppose that we take $n = 50$ samples to be inspected by technicians, and that probability π of recording bacteria by each of them is $\pi = 0.1$ (i.e., each technician records about one bacterium out of each 10 present). Furthermore, assume that the number of bacteria allowed is 10,000 per liter (10^3 cm^3), with the volume v taken for inspection being $1/20$ of a cubic centimeter. This gives $\lambda_0 = 10^4/10^3 = 10$ with $v = 0.05$ cm^3, so

$$\pi\lambda_0 v = 0.1 \times 10 \times 0.05 = 0.05.$$

Under the null hypothesis H_0, the total count $X_1 + \cdots + X_{50}$ in all samples has a Poisson distribution with mean $0.05 \times 50 = 2.5$. Under the alternative, this count is still Poisson, but with mean 12.5.

If we want to test the null hypothesis at significance level $\alpha = 0.05$, then k is determined by

$$P\{S_{50} \geq k | \lambda = 2.5\} \leq 0.05,$$

where S_n has a Poisson distribution with mean 2.5. We have $P\{S_n \leq 5\} = 0.9580$, which means that $P\{S_n \geq 6\} = 0.042$. We may take $k = 6$, so the probability of type II error of this test is $\beta = P(S_{50} < 6 | \lambda = 12.5) = 0.015$. Thus the chances of raising a false alarm (in this case the type I error) are about 4%, whereas the chances of failing to notice a dangerous level of contamination are only about 1.5%.

■ **EXAMPLE 13.14**

Continuing Example 13.13, suppose that the laboratory is under a financial squeeze and decides to save on the cost of observations. The technicians are instructed to inspect each sample of size v and report only whether the bacteria were found. The saving here is that technicians stop searching the sample as soon as one bacterium is found (instead of continuing the search and counting). In a sample $Y_1, \ldots, Y_n$, $Y_j = 1$ or 0 depending on whether bacteria are found in the jth sample. We have, for $i = 0, 1$,

$$P\{Y_j = 1|H_i\} = 1 - P\{Y_j = 0|H_i\} = 1 - e^{-\lambda_i \pi v}. \tag{13.17}$$

Consequently, the total number of samples where bacteria were found, $S = Y_1 + \cdots + Y_n$, has a binomial distribution, with the number of trials n and probability of success given by (13.17), depending on whether H_0 or H_1 is true. Recall that $\lambda_1 = 5\lambda_0$, so the likelihood ratio is now

$$\begin{aligned} \frac{f_1(s)}{f_0(s)} &= \frac{P\{S = s|H_1\}}{P\{S = s|H_0\}} = \frac{\binom{n}{s}(1 - e^{-\pi\lambda_1 v})^s (e^{-\pi\lambda_1 v})^{n-s}}{\binom{n}{s}(1 - e^{-\pi\lambda_0 v})^s (e^{-\pi\lambda_0 v})^{n-s}} \\ &= \left[\frac{1 - e^{-5\pi\lambda_0 v}}{1 - e^{-\pi\lambda_0 v}} e^{4\pi\lambda_0 v}\right]^s e^{-4\pi\lambda_0 v n}. \end{aligned}$$

Since the fraction in brackets exceeds 1, the likelihood increases with s. Again, the critical region is the set $S \geq k$, where k has to be determined from the fact that under the null hypothesis S has a binomial distribution, BIN$(n, 1 - e^{-\pi\lambda_0 v})$. In particular, the laboratory may now ask the question: How large should n be to ensure that the same significance level and the same power as in the more expensive procedure with technicians counting the bacteria, We have now $e^{-\pi\lambda_0 v} = 1 - e^{-0.05} = 0.0488$ and $e^{-\pi\lambda_1 v} = 1 - e^{-0.2} = 0.1813$. We look for n and k such that $P(S \geq k) \leq 0.05$, where $S \sim$ BIN$(n, 0.0488)$, and $P(S' \leq k) \leq 0.015$, where $S' \sim$ BIN$(n, 0.1813)$. Based on the central limit theorem (Theorem 10.6.2), k has to satisfy

$$P\left(Z \geq \frac{k - 0.0488n}{\sqrt{(0.0488)(0.9512)n}}\right) = 0.05,$$

$$P\left(Z \leq \frac{k - 0.1813n}{\sqrt{(0.1813)(0.8187)n}}\right) = 0.015,$$

and consequently

$$\begin{cases} k &= 0.0488n + (1.645)(0.2154)\sqrt{n} \\ k &= 0.1813n - (2.17)(0.3852)\sqrt{n}. \end{cases}$$

The solution, rounded to integers, is $n = 81, k = 7$. The simplified observation procedure (recording only whether or not the bacteria are present) necessitates the increase of the number of samples from 50 to 81.

It is clear that in the case of continuous random variables we can always find a test (critical region C) such that $\alpha(C)$ equals the desired significance level α_0. Below we present examples concerning distributions other than normal, leaving the detailed presentation of various tests for the normal case to a separate section.

■ **EXAMPLE 13.15**

Suppose that the observations $X_1, \ldots, X_n$ form a random sample from the EXP(λ) distribution. We want to test hypothesis $H_0 : \lambda = 5$ against a simple alternative $H_1 : \lambda = 8$ at the significance level α_0. The likelihood ratio is then

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{\prod_{j=1}^{n}(8e^{-8x_j})}{\prod_{j=1}^{n}(5e^{-5x_j})} = \left(\frac{8}{5}\right)^n e^{-3\sum x_i},$$

and the inequality $f_1(\mathbf{x})/f_0(\mathbf{x}) > k$ is equivalent to $\sum x_i < k^*$ for some k^*.

Here $k^* = -(1/2)[n\log(5/8) + \log k]$, but the exact relation between k and k^* is not needed to determine the test. What matters most is the direction of the inequality. The way we set the solution is that we *reject* H_0 if the likelihood ratio f_1/f_0 (with alternative density on the top) is *large*. Typically this condition reduces to an inequality for some statistic (such as $x_1 + \cdots + x_n$, or equivalently, $\overline{x}$). In the present case the rejection region is the left tail (i.e., values of $x_1 + \cdots + x_n$ less than a certain threshold).

To continue, we need a value k^* such that

$$P\{X_1 + \cdots + X_n < k^* | H_0\} = \alpha_0.$$

Each X_i has exponential distribution with parameter λ_0, hence with the mean $1/\lambda_0$. Consequently, $2\lambda_0 X_i$ has an exponential distribution with mean 2, that is, a chi-square distribution with 2 degrees of freedom (see Theorem 9.4.3). Therefore $2\lambda_0(X_1 + \cdots + X_n)$ has a chi-square distribution with $2n$ degrees of freedom.

In the case under consideration, we have $\lambda_0 = 5$; hence the critical threshold k^* can be obtained from the tables of chi-square distribution with $2n$ degrees of freedom:

$$\begin{aligned} P\{X_1 + \cdots + X_n < k^* | H_0\} &= P\{10(X_1 + \cdots + X_n) < 10k^* | H_0\} \\ &= P\{\chi^2_{2n} < 10k^*\} = \alpha_0. \end{aligned}$$

Taking, as an example, $n = 10$ and $\alpha_0 = 0.1$, we obtain $10k^* = \chi^2_{0.9,20} = 12.44$ (from Table A4.), which gives $k^* = 1.244$.

The probability of a type II error is,

$$\begin{aligned} P\{X_1 + \cdots + X_n > 1.244 | H_1\} &= P\{16(X_1 + \cdots + X_n) > 16 \times 1.224 | H_1\} \\ &= P\{\chi^2_{20} > 19.904\} = 0.4637, \end{aligned}$$

as obtained from statistical software. This probability seems rather high as a probability of error. One reason is that the difference between $\lambda = 5$ and $\lambda = 8$ is equivalent to a small difference between means (1/5 and 1/8, respectively). Another reason is that the small sample size ($n = 10$) affects the quality of the inference.

Thus far we have considered testing the simple null hypothesis against a simple alternative in a parametric setup, where the distribution belongs to the same family $f(x;\theta)$, and the hypotheses are obtained by specifying $H_0 : \theta = \theta_0$ and $H_1 : \theta =$

θ_1. It is important to realize that the Neyman-Pearson lemma applies to any two simple hypotheses. For instance, suppose that we have a single observation X, that comes from the U[0, 1] distribution if H_0 is true or from the EXP(3) distribution if H_1 is true. The densities are

$$f_0(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_1(x) = \begin{cases} 3e^{-3x} & \text{if } x \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood ratio equals

$$\frac{f_1(x)}{f_0(x)} = \begin{cases} 3e^{-3x} & \text{if } 0 \le x \le 1 \\ \infty & \text{if } x > 1, \end{cases}$$

and it is undefined for $x < 0$. Each critical region should contain the set $\{x : x > 1\}$, and also a set of the form $3e^{-3x} > k$, hence a set of the form $\{x : 0 \le x \le k^*\}$ for some $k^* < 1$. To determine k^*, we must have

$$P\{X \in C | H_0\} = \int_0^{k^*} f_0(x)dx + \int_1^{\infty} f_0(x)dx = \int_0^{k^*} f_0(x)dx = k^*.$$

It follows that if we choose the significance level $\alpha_0 = 0.05$ (say), then $k^* = 0.05$, and we reject H_0 if the observation is either below 0.05 or above 1. The power of this test is

$$1 - \beta = \int_C f_1(x)dx = 1 - \int_{0.05}^{1} 3e^{-3x}dx = 1 - e^{-3} + e^{-0.15} = 0.1891.$$

PROBLEMS

13.3.1 Let $X_1, \ldots, X_{10}$ be a random sample from a POI(θ) distribution. **(i)** Find the best critical region for testing $H_0 : \theta = 0.2$ against $H_1 : \theta = 0.8$ at the significance level $\alpha = 0.05$. **(ii)** Determine the size of the test in (i).

13.3.2 A single observation X is taken from a BETA(a, b) distribution. Find the most powerful test of the null hypothesis H_0: $a = b = 1$, against the alternative H_1: **(i)** $a = b = 5$. **(ii)** $a = 2, b = 3$ **(iii)** $a = b = 1/2$. Use significance level $\alpha = 0.05$.

13.3.3 Let X have a negative binomial distribution with parameters r and p. Find the most powerful test of $H_0 : r = 2, p = 1/2$ against $H_1 : r = 4, p = 1/2$ at significance level $\alpha = 0.05$. Find probability of type II error. Use randomized test if necessary.

13.3.4 Assume that X has a N($2, \sigma^2$) distribution. Find the best critical region for testing $H_0 : \sigma^2 = 2$ against: **(i)** $H_1 : \sigma^2 = 4$. **(ii)** $H_1 : \sigma^2 = 1$.

13.3.5 Let $X_1, \ldots, X_n$ be a random sample from EXP(λ) distribution. Null hypothesis $H_0 : \lambda = \lambda_0$ is tested against the alternative $H_1 : \lambda = \lambda_1$, where $\lambda_1 > \lambda_0$. Compare the power functions of the two tests: (a) the most powerful test, and (b) the most powerful test based on the statistic $X_{1:n}$. Assume that both tests have equal probabilities of a type I error.

13.3.6 The sample space of a test statistic X has five values: a, b, c, d, e. Test the $H_0 : f = f_0$ against $H_a : f = f_1$, where distributions f_0 and f_1 are given by the table

X	a	b	c	d	e
f_0	0.2	0.2	0.0	0.1	0.5
f_1	0.2	0.4	0.3	0.0	0.1

13.3.7 An urn contains six balls, r red and $6 - r$ blue. Two balls are chosen without replacement. Find the most powerful test of $H_0 : r = 3$ against the alternative $H_1 : r = 5$, with a size as close to $\alpha = 0.05$ as possible. Find the probability of a type II error for all $r \neq 3$.

13.3.8 A multiple-choice exam gives five answers to each of its n questions, only one being correct. Assume that a student who does not know the answer chooses randomly and is correct with probability 0.2. Let θ be the number of questions to which the student knows the answers, and let X be the number of correct responses given by this student. **(i)** Determine $f(x; \theta) = P(X = x|\theta)$. **(ii)** For $n = 50$, find the most powerful test of $H_0 : \theta = 30$ against $H_1 : \theta = 40$ at the significance level $\alpha = 0.05$. **(iii)** Determine the probability of a type II error if $\theta = 40$.

13.3.9 Let $X_1, \ldots, X_n$ have a joint density $f(\mathbf{x}; \theta)$, and let U be a sufficient statistic for θ. Show that the most powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ can be expressed in terms of U.

13.4 UNIFORMLY MOST POWERFUL TESTS

From the derivation of the most powerful tests in Examples 13.13 and 13.15, notice that the final form of the test is, to a large extent, independent of the choice of a specific alternative hypothesis. For instance, in Example 13.13 observations $X_1, \ldots, X_n$ come from one of the two Poisson distributions. Omitting the details on size v of the water samples and the probability π of recording the presence of bacteria, we could test a simple hypothesis $H_0 : E(X_i) = \theta_0$ against a simple alternative $H_1 : E(X_i) = \theta_1$, where $\theta_1 > \theta_0$. The likelihood ratio, after canceling the factorials, turns out to be

$$\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} = \left(\frac{\theta_1}{\theta_0}\right)^{\sum X_i} \times \left(\frac{e^{-\theta_1}}{e^{-\theta_0}}\right)^n.$$

If $\theta_1 > \theta_0$, then the likelihood ratio is an increasing function of $\sum X_i$; hence the rejection region must be of the form

$$C = \{(x_1, \ldots, x_n) : \sum x_i \geq k\}. \tag{13.18}$$

Determination of value k involves only the value of θ_0 and a level of significance α_0; the value θ_1 plays no role, provided that $\theta_1 > \theta_0$. We know that under the null

hypothesis $S_n = \sum_{i=1}^{n} X_i$ has a Poisson distribution with mean $n\theta_0$; hence

$$k = \min\left\{ r : \sum_{j=0}^{r-1} \frac{(n\theta_0)^j}{j!} e^{-n\theta_0} \geq 1 - \alpha_0, \right\} \tag{13.19}$$

where α_0 is the desired significance level.

This means that a test with the critical region given by (13.18) and (13.19) is the most powerful for *any* alternative hypothesis if only the reasoning leading to (13.18) applies. But the only fact about θ_1 used in the derivation is that $\theta_1 > \theta_0$. This inequality causes the ratio θ_1/θ_0 to exceed 1, and hence the likelihood ratio to increase with the sum $\sum x_i$ (for $\theta_1 < \theta_0$ the likelihood ratio would decrease with an increase of $\sum x_i$ and the critical region would comprise the left tail, i.e., we would reject H_0 if $\sum x_i \leq k$).

In the continuous case, we can restrict the considerations to tests based on critical regions. In the discrete case, one can also consider randomized procedures, that is, procedures in which the rejection or acceptance of H_0 depends both on the observation of $\mathbf{X}$ and also possibly on additional randomization.

Consequently, in the definitions below we consider *procedures* for testing H_0; without much danger of confusion we will use letter C for a procedure, and define its power function as

$$\pi_C(\theta) = P_\theta\{\text{procedure } C \text{ rejects } H_0\}.$$

We will fix the significance level α_0 ($0 < \alpha_0 < 1$) and let $K(H_0, \alpha_0)$ be the class of all procedures for testing H_0 whose size is at most α_0, that is, procedures C such that

$$\sup_{\theta \in \Theta_0} \pi_C(\theta) \leq \alpha_0. \tag{13.20}$$

Definition 13.4.1 A procedure $C^* \in K(H_0, \alpha_0)$, such that

$$\pi_{C^*}(\theta) \geq \pi_C(\theta)$$

for every $\theta \in \Theta_1$ and for any $C \in K(H_0, \alpha_0)$, will be called a *uniformly most powerful* (UMP) procedure for testing H_0 against H_1 at the significance level α_0.

In the case where C^* is nonrandomized (i.e., C^* is the critical region of a test) we will speak of a UMP test of H_0 against H_1 at the significance level α_0. □

The essence of this definition lies in the fact that *the same* test (procedure) is most powerful against *all* simple hypotheses in H_1. The Neyman-Pearson lemma (or its extension to composite null hypothesis) asserts that the most powerful test exists against any simple alternative, but it sometimes happens that these most powerful tests are different for different simple alternatives.

■ **EXAMPLE 13.16**

Consider the situation of testing the simple null hypothesis $H_0 : \mu = 0$ against the composite alternative $H_1 : \mu > 0$ in case of a normal distribution $N(\mu, \sigma^2)$

with known σ^2. Let us fix $\mu_1 > 0$ and consider the likelihood ratio test of H_0 against $H_1' : \mu = \mu_1$. We have

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{\exp\{-(1/2\sigma^2)\sum(x_i - \mu_1)^2\}}{\exp\{-(1/2\sigma^2)\sum x_i^2\}} = De^{(\mu_1/\sigma^2)\sum x_i}.$$

Since $\mu_1 > 0$, this likelihood ratio is an increasing function of $\sum x_i$ (hence also of $\overline{x}$), and the critical region of the most powerful test against $\mu > 0$ is of the form $\{\mathbf{x} = (x_1, \ldots, x_n) : \overline{x} \geq k\}$.

For any given significance level α_0 we determine k from the condition

$$P\{\overline{X} \geq k | H_0\} = P\left\{Z \geq \frac{k}{\sigma/\sqrt{n}}\right\} = \alpha_0$$

so that $k = z_{\alpha_0}\sigma/\sqrt{n}$, where z_{α_0} is the upper α_0-quantile of the standard normal distribution.

The only assumption about μ_1 that is used here is $\mu_1 > 0$; consequently, the test with the critical region C,

$$\text{reject } H_0 \text{ if } \overline{X} \geq z_{1-\alpha_0}\sigma/\sqrt{n} \tag{13.21}$$

is UMP for the alternative $H_1 : \mu > 0$.

■ **EXAMPLE 13.17**

Continuing Example 13.16, observe that the test (13.21) is also UMP for the composite null hypothesis $H_0 : \mu \leq 0$ against the composite alternative $H_1 : \mu > 0$. Indeed, the power of this test is

$$\begin{aligned}\pi_C(\mu) &= P_\mu\left\{\overline{X} \geq z_{\alpha_0}\frac{\sigma}{\sqrt{n}}\right\} = P\left\{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{z_{\alpha_0}(\sigma/\sqrt{n}) - \mu}{(\sigma/\sqrt{n})}\right\} \\ &= P\left\{Z \geq z_{\alpha_0} - \frac{\mu\sqrt{n}}{\sigma}\right\}.\end{aligned}$$

Consequently, for $\mu < 0$, $\pi_C(\mu) \leq P\{Z \geq z_{\alpha_0}\} = \alpha_0 = \pi_C(0)$; hence the size of the test C is α_0.

■ **EXAMPLE 13.18**

Example 13.17 shows that there exists no UMP test for the hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu \neq 0$. Indeed, (13.21) is the most powerful test against any alternative $\mu > 0$, but it performs very poorly against the alternative $\mu < 0$: its power on such alternatives is less than α_0. On the other hand, by symmetry, the test with critical region C',

$$\text{reject } H_0 \quad \text{if} \quad \overline{X} \leq -z_{\alpha_0}\frac{\sigma}{\sqrt{n}}, \tag{13.22}$$

is most powerful against any alternative $\mu < 0$, but performs poorly against alternatives $\mu > 0$. The "compromise" test, with the critical region C'',

$$\text{reject } H_0 \quad \text{if} \quad |\overline{X}| \geq z_{\alpha_0/2}\frac{\sigma}{\sqrt{n}}, \tag{13.23}$$

performs quite well for all alternatives $\mu \neq 0$ but its power is below the power of test (13.21) for $\mu > 0$, and below the power of test (13.22) for $\mu < 0$. So it is not a UMP test.

We know from the examples above that UMP tests may not exist. On the other hand, if they do exist, they provide the best available procedures (if one takes into account only the error probabilities, and not extraneous aspects, such as computation costs, etc). It is therefore natural to ask for conditions under which UMP tests exist. The answer is given by the next theorem, which we will precede by some necessary definitions.

Definition 13.4.2 We say that the family $\{f(x;\theta), \theta \in \Theta\}$ of distributions has a *monotone likelihood ratio* in statistic T if for any two values $\theta', \theta'' \in \Theta$ with $\theta' < \theta''$, and $\mathbf{x} = (x_1, \ldots, x_n)$, the likelihood ratio

$$f_n(\mathbf{x};\theta'')/f_n(\mathbf{x};\theta')$$

depends on $\mathbf{x}$ only through the values $T(\mathbf{x})$, and this ratio is an increasing function of $T(\mathbf{x})$. □

We will now illustrate the introduced concept by some examples.

■ **EXAMPLE 13.19**

Consider the normal density depending on parameter μ, with σ^2 known, so that

$$f_n(\mathbf{x},\mu) = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right\}.$$

Let $\mu' < \mu''$, and for typographical reason, let $\mu' = \mu_1, \mu'' = \mu_2$. Then

$$\begin{aligned}\frac{f_n(\mathbf{x},\mu_2)}{f_n(\mathbf{x},\mu_1)} &= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum(x_i-\mu_2)^2 - \sum(x_i-\mu_1)^2\right]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[-2(\mu_2-\mu_1)\sum x_i + n(\mu_2^2-\mu_1^2)\right]\right\} \\ &= De^{(\mu_2-\mu_1)/\sigma^2}\sum x_i,\end{aligned}$$

where $D > 0$. Consequently, the normal distribution for a fixed σ^2 has a monotone likelihood ratio in $T = \sum X_i$, or equivalently, in $\overline{X}$.

■ **EXAMPLE 13.20**

Consider again the normal distribution, this time for known μ and unknown σ^2. Let $\sigma_1^2 < \sigma_2^2$. Then

$$\begin{aligned}\frac{f_n(\mathbf{x},\sigma_2^2)}{f_n(\mathbf{x},\sigma_1^2)} &= D\exp\left\{-\frac{1}{2\sigma_2^2}\sum(x_i-\mu)^2 + \frac{1}{2\sigma_1^2}\sum(x_i-\mu)^2\right\} \\ &= D\exp\left\{\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)\sum(x_i-\mu)^2\right\}.\end{aligned}$$

Since $D > 0$ and $1/\sigma_1^2 - 1/\sigma_2^2 > 0$, we see that now the likelihood ratio is increasing in statistic $T = \sum(X_i - \mu)^2$.

■ **EXAMPLE 13.21 Bernoulli Trials**

Let $X_1, \ldots, X_n$ be a random sample from Bernoulli distribution $f(x; p) = p^x(1-p)^{1-x}$, $x = 0, 1$, and $0 < p < 1$. For $0 < p' < p'' < 1$ we have

$$\frac{f_n(\mathbf{x}, p'')}{f_n(\mathbf{x}, p')} = \left[\frac{p''(1-p')}{p'(1-p'')}\right]^{\sum x_i} \left(\frac{1-p''}{1-p'}\right)^n.$$

Since $p''(1-p')/p'(1-p'') > 1$, the Bernoulli distribution has monotone likelihood ratio in statistic $\sum X_i$ (total number of successes in n trials).

It turns out that most of the known families of distributions have monotone likelihood ratio in some statistics. The role of families with monotone likelihood ratio for UMP tests is explained by the following theorem:

Theorem 13.4.1 *Let $\{f(x;\theta), \theta \in \Theta\}$ be a family of distributions with a monotone likelihood ratio in statistics T. Then for every α_0, $0 < \alpha_0 < 1$, there exists a test (possibly randomized) that is UMP for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ at significance level α_0. This test satisfies the following two conditions:*
(a) *There exists k such that if $T(\mathbf{x}) > k$, then H_0 is rejected and if $T(\mathbf{x}) < k$, then H_0 is accepted.*
(b) *$P_{\theta_0}\{H_0$ is rejected$\} = \alpha_0$.*
Similarly, for testing $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$, there exists k_1 such that H_0 should be rejected if $T(\mathbf{X}) < k_1$ and accepted if $T(\mathbf{X}) > k_1$.

Proof. Observe first that condition (b) may require randomization. Indeed, in view of (a) and (b), letting $\gamma = P\{H_0$ is rejected $|T(\mathbf{X}) = k\}$, we must have

$$\alpha_0 = P_{\theta_0}\{T(\mathbf{X}) > k\} + \gamma P_{\theta_0}\{T(\mathbf{X}) = k\}. \tag{13.24}$$

It follows that if $P_{\theta_0}\{T(\mathbf{X}) = k\} = 0$, then we have

$$\alpha_0 = P_{\theta_0}\{T(\mathbf{X}) > k\} = P_{\theta_0}\{T(\mathbf{X}) \geq k\},$$

and the test is nonrandomized, with critical region $C = \{\mathbf{x} : T(\mathbf{x}) \geq k\}$. If $P_{\theta_0}\{T(\mathbf{X}) > k\} < \alpha_0 \leq P_{\theta_0}\{T(\mathbf{X}) \geq k\}$, then by (13.24),

$$\gamma = \frac{\alpha_0 - P_{\theta_0}\{T(\mathbf{X}) > k\}}{P_{\theta_0}\{T(\mathbf{X}) = k\}}.$$

Clearly, we have $0 < \gamma < 1$, and an auxiliary randomization with probability of success equal to γ gives a procedure with size α_0.

Let C^* denote the procedure described by the theorem, and its power be $\pi_{C^*}(\theta) = P_\theta\{C^*$ leads to rejecting $H_0\}$. By condition (b), $\pi_{C^*}(\theta_0) = \alpha_0$.

Without loss of generality, let $\theta_1 > \theta_0$. By the Neyman-Pearson lemma, the most powerful procedure for testing the simple hypothesis $H_0^* : \theta = \theta_0$ against a

simple alternative $H_1^* : \theta = \theta_1$ is based on the likelihood ratio, and rejects H_0 if $f_n(\mathbf{x};\theta_1)/f_n(\mathbf{x};\theta_0) > k^*$ for some k^*. But by the monotonicity of the likelihood ratio, the last condition means that $T(\mathbf{X}) > k$ for some k. Since the last condition does not depend on the choice of θ_1, the procedure C^* described in the theorem is UMP for testing the simple hypothesis $H_0^* : \theta = \theta_0$ against the composite alternative $H_1 : \theta > \theta_0$.

We will also show that the power function of procedure C^* is nondecreasing. For any $\theta' < \theta''$ (regardless of their location with respect to θ_0), we have $\pi_{C^*}(\theta') \leq \pi_{C^*}(\theta'')$. This fact will be crucial for completing the proof.

Indeed, consider the class $\mathcal{K}$ of all procedures of testing the simple hypothesis $H_0' : \theta = \theta'$ against the simple alternative $H_1' : \theta = \theta''$ at a level of significance α, say. We know from the Neyman-Person lemma that C^* is the most powerful procedure in class $\mathcal{K}$, that is, for any $C \in \mathcal{K}$,

$$\pi_{C^*}(\theta'') \geq \pi_C(\theta''). \tag{13.25}$$

On the other hand, the (randomized) procedure C_0: "reject $H_0^* : \theta = \theta'$ with probability α regardless of observation" satisfies the condition $\pi_{C_0}(\theta) = \alpha$ for all θ. Clearly, $C_0 \in \mathcal{K}$. So, using (13.25), we write

$$\pi_{C^*}(\theta'') \geq \pi_{C_0}(\theta'') = \alpha = \pi_{C^*}(\theta'),$$

which shows monotonicity of $\pi_{C^*}(\theta)$.

To complete the proof we have to show that the procedure C^* is UMP not only in the class $\mathcal{B}$ of all procedures C that satisfy the condition $\pi_C(\theta_0) \leq \alpha_0$ but also in the class $\mathcal{B}'$ of all procedures C such that $\sup_{\theta \leq \theta_0} \pi_C(\theta_0) \leq \alpha_0$.

Clearly, $\mathcal{B}' \subset \mathcal{B}$, and the monotonicity of the power function $\pi_{C^*}(\theta)$ shows that C^* is an element of the smaller class $\mathcal{B}'$: Indeed, for $\theta \leq \theta_0$,

$$\pi_{C^*}(\theta) \leq \pi_{C^*}(\theta_0) \leq \alpha_0.$$

This completes the proof, since we already know that for every $\theta_1 > \theta_0$, the procedure C^* maximizes the power in the larger class $\mathcal{B}$, and hence also maximizes the power in the smaller class $\mathcal{B}'$. □

PROBLEMS

13.4.1 Check whether the following families of distributions have a monotone likelihood ratio in the parameter specified: **(i)** Poisson. **(ii)** Exponential. **(iii)** Gamma, for each parameter separately. **(iv)** Beta, for each parameter separately.

13.4.2 Let $X_1, \ldots, X_n$ be a random sample from Laplace distribution with density $f(x;\lambda) = (\lambda/2)\exp\{-\lambda|x|\}$. Find a UMP test for testing $H_0 : \lambda = 1$ against $H_1 : \lambda < 1$ at the significance level 0.01.

13.4.3 Let $X_1, \ldots, X_n$ be a random sample from a folded normal distribution with density $f(x;\theta) = \sqrt{2/\pi}\theta \exp\{-\theta^2 x^2/2\}$, for $x > 0, \theta > 0$. **(i)** Derive the UMP test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. **(ii)** Show that the power function is increasing.

13.4.4 Suppose that the number of defects in magnetic tape of length t (yards) has POI(λt) distribution. **(i)** Assume that 2 defects were found in a piece of tape of length 500 yards. Test the hypothesis $H_0 : \lambda \geq 0.02$ against the alternative $H_1 : \lambda < 0.02$. Use a UMP test at the level $\alpha \leq 0.01$. **(ii)** Find the p-value. **(iii)** Find the power of the test at $\lambda = 0.015$.

13.4.5 The effectiveness of a standard drug in treating specific illness is 60%. A new drug was tested and found to be effective in 48 out of 70 cases when it was used. Specify an appropriate alternative hypothesis and perform the test at the 0.01 level of significance. Find the p-value.

13.4.6 Suppose that $X_1, \ldots, X_n$ is a random sample from the U$[0, \theta]$ distribution. **(i)** Hypothesis $H_0 : \theta \leq \theta_0$ is to be tested against the alternative $H_1 : \theta > \theta_0$. Argue that the UMP test rejects H_0 if $X_{n:n} > c$. Find c for $\theta_0 = 5, n = 10$, and $\alpha = 0.05$. **(ii)** If the hypothesis $H_0 : \theta \geq \theta_0$ is tested against $H_1 : \theta < \theta_0$, show that the UMP test rejects H_0 if $X_{n:n} < c$. Find c if $\theta_0 = 5, n = 10$, and $\alpha = 0.05$.

13.4.7 Let $X_1, \ldots, X_n$ be a random sample from the GAM(α, λ) distribution. **(i)** Derive a UMP test for the hypothesis $H_0 : \alpha \leq \alpha_0$ against the alternative $H_1 : \alpha > \alpha_0$ if λ is known. **(ii)** Derive a UMP test for the hypothesis $H_0 : \lambda \leq \lambda_0$ against the alternative $H_1 : \lambda > \lambda_0$ if α is known.

13.4.8 Recall Problem 13.3.8. Assume that $n = 50$ and that a student with a score at most 30 will fail. Does there exist a UMP test for the hypothesis $H_0 : \theta \leq 30$? If yes, find the test; if no, justify your answer.

13.4.9 Let $X_1, \ldots, X_n$ be a random sample from the distribution $f(x; \theta) = C[\theta/(\theta + 1)]^x$, where $x = 1, 2, \ldots$ and C is the normalizing constant. Determine a UMP test of the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta > \theta_0$.

13.4.10 A reaction time to a certain stimulus (e.g., time until solving some problem) is modeled as a time of completion of r processes, running one after another in a specified order. The times $\tau_1, \ldots, \tau_r$ of completion of these processes are assumed to be iid exponential with mean $1/\lambda$. If $r = 3$ and the observed reaction times (in seconds) are 15.3, 6.1, 8.5, and 9.0, test the hypothesis $H_0 : \lambda \geq 0.8$ against the alternative $H_1 : \lambda < 0.8$. Use $\alpha = 0.05$.

13.5 UNBIASED TESTS

As we already know, there are situations where the UMP tests do not exist. One example may be a test of hypothesis $H_0 : \theta = \theta_0$ against a two-sided alternative $H_1 : \theta \neq \theta_0$ (i.e., where θ_0 is not on the boundary of the parameter space Θ). Since UMP tests are highly desirable, the theoretical efforts became directed toward a reduction of the class of tests. Such reduced class might already contain a UMP test.

We present an approach that requires unbiasedness. Consider the problem of testing the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$. Let C be a testing procedure with a power function $\pi_C(\theta)$. Here the parameter space

Θ may be multidimensional, and one or both of the hypotheses H_0 or H_1 may be composite.

Definition 13.5.1 The test C of H_0 against H_1 is called *unbiased* if

$$\sup_{\theta \in \Theta_0} \pi_C(\theta) \leq \inf_{\theta \in \Theta_1} \pi_C(\theta). \tag{13.26}$$

□

Since the left-hand side of (13.26) is the size of the test C, we say that C is unbiased if its power on the alternative hypothesis is never below its size. In particular, if the null hypothesis is simple, then the power function of an unbiased test reaches its minimum at θ_0.

It turns out that in some cases where there is no UMP test in the class of all tests, there is a UMP unbiased test.

■ **EXAMPLE 13.22**

Consider the case of testing the simple hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$, where observations $X_1, \ldots, X_n$ are a random sample from the $N(\mu, \sigma^2)$ distribution with known σ^2.

Intuition suggests that we take $T(\mathbf{X}) = \overline{X} - \mu_0$ as the test statistic, and reject H_0 if either $T(\mathbf{X}) < -k'$ or $T(\mathbf{X}) > k''$ for some suitably chosen positive numbers k' and k''. To have size equal α_0, we must choose k' and k'' so that

$$\begin{aligned} 1 - \alpha_0 &= P_{\mu_0}\{-k' < T(\mathbf{X}) < k''\} = P_{\mu_0}\{-k' < \overline{X} - \mu_0 < k''\} \\ &= P\left\{-k' \frac{\sqrt{n}}{\sigma} < Z < k'' \frac{\sqrt{n}}{\sigma}\right\}, \end{aligned}$$

that is,

$$\Phi\left(k'' \frac{\sqrt{n}}{\sigma}\right) - \Phi\left(-k' \frac{\sqrt{n}}{\sigma}\right) = 1 - \alpha_0, \tag{13.27}$$

where Φ is the cdf of standard normal random variable.

Let us investigate the power function of the test above. For the critical region $C = \{\overline{X} < \mu_0 - k'\} \cup \{\overline{X} > \mu_0 + k''\}$, we have

$$\begin{aligned} \pi_C(\mu) &= 1 - P_\mu\{-k' < \overline{X} - \mu_0 < +k''\} \\ &= 1 - P\left\{\frac{-k' + \mu_0 - \mu}{\sigma/\sqrt{n}} < Z < \frac{k'' + \mu_0 - \mu}{\sigma/\sqrt{n}}\right\} \\ &= 1 - \Phi\left(\frac{k'' + \mu_0 - \mu}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-k' + \mu_0 - \mu}{\sigma/\sqrt{n}}\right). \end{aligned} \tag{13.28}$$

Clearly, $\pi_C(\mu)$ is a continuous differentiable function of μ, and C is unbiased if $\frac{d}{d\mu}\pi_C(\mu)\Big|_{\mu=\mu_0} = 0$. We have, letting φ be the density of standard normal

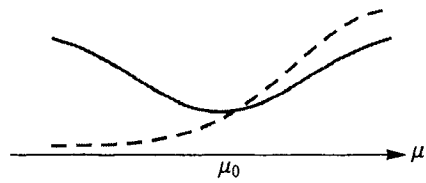


Figure 13.5 Power functions of a one-sided UMP test (dashed line) and a UMP unbiased test (solid line).

random variable,

$$\frac{d\pi_C(\mu)}{d\mu} = \varphi\left(\frac{k'' + \mu_0 - \mu}{\sigma/\sqrt{n}}\right)\frac{\sqrt{n}}{\sigma} - \varphi\left(\frac{-k' + \mu_0 - \mu}{\sigma/\sqrt{n}}\right)\frac{\sqrt{n}}{\sigma}.$$

For $\mu = \mu_0$ we obtain

$$\left.\frac{d\pi_C(\mu)}{d\mu}\right|_{\mu=\mu_0} = \left[\varphi\left(\frac{k''}{\sigma}\sqrt{n}\right) - \varphi\left(-\frac{k'}{\sigma}\sqrt{n}\right)\right]\frac{\sqrt{n}}{\sigma},$$

which equals 0 (remembering that k' and k'' are positive) only if $k' = k''$. Thus we obtain an unbiased test only if the two parts of the rejection regions are located symmetrically with respect to the null hypothesis value μ_0. Formula (13.27) gives now $k' = k'' = z_{\alpha_0/2}$.

One can show that the test obtained in Example 13.22 is actually the UMP unbiased test for hypothesis $H_0 : \mu = \mu_0$ against the two-sided alternative $H_1 : \mu \neq \mu_0$. It is not, however, a UMP test against one-sided alternatives (see Figure 13.5).

One can argue that such a specific hypothesis like $H_0 : \mu = \mu_0$ simply cannot be true. The chances that the mean is *exactly* equal to μ_0 are zero. This is true if the parameter μ is a random variable, varying from situation to situation according to some continuous prior distribution. It is also true in most cases of Bayesian priors, where μ does not vary, but the statistician's experience can be expressed in terms of a (subjective) probability distribution on μ.

Thus, the argument goes, H_0 should be rejected at once, without any testing. As explained at the beginning of this chapter, this conclusion misses the essential intention of the theory of hypotheses testing, which is to serve not as a means of establishing the truth of hypotheses, but of establishing good rules of inductive behavior. In this sense, "accepting $H_0 : \theta = \theta_0$" means simply that it is reasonable to act *as if* the parameter value were θ_0, even if in fact θ is only close to θ_0. This suggests testing a composite null hypothesis stating that θ lies in some interval, against the alternative that it lies outside it. We will illustrate this approach with an example:

■ **EXAMPLE 13.23**

Let again $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma)$ distribution with σ^2 known. We want to test the null hypothesis $H_0 : \mu_1 \leq \mu \leq \mu_2$ against the

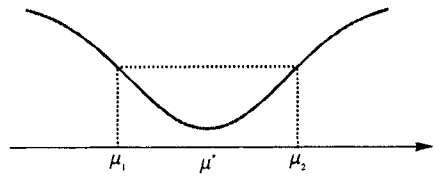


Figure 13.6 Power function of an unbiased test

alternative $H_1 : \mu < \mu_1$ or $\mu > \mu_2$. To simplify the notation, let

$$\mu^* = \frac{\mu_1 + \mu_2}{2} \tag{13.29}$$

denote the midpoint between the boundaries of the null hypothesis. It appears reasonable to use the test statistic $T(\mathbf{X}) = \overline{X} - \mu^*$, and reject H_0 if $T(\mathbf{X}) < -k'$ or $T(\mathbf{X}) > k''$, where k', k'' are some positive constants. The power of this test (call it C^*) is

$$\begin{aligned}
\pi_{C^*}(\mu) &= 1 - P_\mu\{-k' < \overline{X} - \mu^* < k''\} \\
&= 1 - P\left\{\frac{-k' + \mu^* - \mu}{\sigma/\sqrt{n}} < Z < \frac{k'' + \mu^* - \mu}{\sigma/\sqrt{n}}\right\} \\
&= 1 - \Phi\left(\frac{k'' + \mu^* - \mu}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-k' + \mu^* - \mu}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

Again, to have the test unbiased, the power curve must have a minimum at μ^*, which necessitates taking $k' = k''$. Then the size of the test (see Figure 13.6) equals to the common value of the power at the points μ_1 and μ_2, that is, after letting $k = k' = k''$ and $\Delta = (\mu_2 - \mu_1)/2$),

$$\begin{aligned}
\sup_{\mu_1 \le \mu \le \mu_2} \pi_{C^*}(\mu) &= 1 - \Phi\left(\frac{\mu^* - \mu_1 + k}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{\mu^* - \mu_1 - k}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{\Delta + k}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{\Delta - k}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

If we now require (for given μ_1, μ_2, and σ, hence given μ^* and Δ) a test with a specified size α_0 and a specified probability β_0 of a type II error at some target value μ_t in the alternative, we can determine the threshold k and sample size n from the equations

$$\Phi\left(\frac{\Delta + k}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\Delta - k}{\sigma/\sqrt{n}}\right) = 1 - \alpha_0$$

and

$$\Phi\left(\frac{\mu^* - \mu_t + k}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\mu^* - \mu_t - k}{\sigma/\sqrt{n}}\right) = \beta_0.$$

The solution has to be obtained by a numerical procedure. It can be shown that the resulting test is UMP unbiased.

PROBLEMS

13.5.1 Suppose that $X_1, \ldots, X_n$ is a random sample from the U$[0, \theta]$ distribution. Test hypothesis $H_0 : \theta = \theta_0$ against the two-sided alternative $H_1 : \theta \neq \theta_0$ using an unbiased test that rejects H_0 if $X_{n:n} < c_1$ or $X_{n:n} > c_2$. Find c_1 and c_2 if $\theta_0 = 5, n = 10$, and $\alpha = 0.05$.

13.5.2 Let X be a single observation from a distribution with density $f(x, \theta) = 1 - \theta^2(x - 0.5)^3$ for $0 < x < 1$ and zero otherwise, $-1 < \theta < 1$. Find a UMP unbiased test of $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$.

13.5.3 Let $X_1, \ldots, X_n$ be a random sample from the POI(λ) distribution. Find the (approximate) UMP unbiased test for the hypothesis $H_0 : \lambda = \lambda_0$ against the two-sided alternative $H_1 : \lambda \neq \lambda_0$, where λ_0 is assumed to be large. [*Hint:* Use the fact that if X has Poisson distribution with mean λ, then $(X - \lambda)/\sqrt{\lambda}$ converges (as $\lambda \to \infty$) in the distribution to a standard normal random variable.]

13.5.4 For testing hypothesis $H_0 : \sigma = \sigma_0^2$ against the alternative $H_1 : \sigma^2 \neq \sigma_0^2$ at level α, find an unbiased test with a critical region of the form "reject H_0 if $\sum_{i=1}^n X_i^2/\sigma_0^2 < C_1$ or $\sum_{i=1}^n X_i^2/\sigma_0^2 > C_2$." Assume that the observations $X_1, \ldots, X_n$ form a random sample from: **(i)** N(0, σ^2) distribution. **(ii)** N(μ, σ^2) distribution with both parameters unknown.

13.6 GENERALIZED LIKELIHOOD RATIO TESTS

As shown by the Neyman-Pearson lemma, analysis of the likelihood ratio is a good way of searching for test statistics. It turns out that an extension of the likelihood ratio method, originally called the "lambda principle" by Neyman (1950), often leads to tests that perform quite well. The role and importance of these tests, called generalized likelihood ratio (GLR) tests, can be compared with that of maximum likelihood estimators in the estimation problems.

A rather common situation in statistical practice occurs when we are interested in testing hypotheses about a specific parameter, but the population distribution depends also on some other parameters. Equivalently, we may say that $\underline{\theta} = (\theta_1, \ldots, \theta_r)$ is a multidimensional parameter, but we are interested in hypotheses involving one component only. One example here is testing a hypothesis about the population mean μ in the normal distribution when variance σ^2 is also unknown.

The parameter (or parameters) that is not constrained by the null hypothesis is called a *nuisance* parameter. The presence of nuisance parameters causes the null and alternative hypotheses to be composite. The generalized likelihood ratio tests that will be discussed in this section can also be applied to testing in the presence of nuisance parameters.

Suppose that we want to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta \setminus \Theta_0$. In most typical cases both hypotheses H_0 and H_1 are composite, which corresponds to the sets Θ_0 and Θ_1 containing more than one element. In analogy with most powerful tests in the case of a simple null hypothesis

and a simple alternative (which are based on the likelihood ratio), we might use the ratio

$$v(\mathbf{X}) = \frac{\sup_{\theta\in\Theta_1} f_n(\mathbf{X};\theta)}{\sup_{\theta\in\Theta_0} f_n(\mathbf{X};\theta)}, \tag{13.30}$$

and reject the null hypothesis if the observation $\mathbf{x}$ gives a high value of $v(\mathbf{x})$. The ratio here is based on the analogy with the likelihood ratio tests given by the Neyman-Pearson lemma: if the numerator in (13.30) greatly exceeds the denominator, then $\mathbf{x}$ is good evidence for the alternative hypothesis. On the other hand, small values of $v(\mathbf{x})$ constitute good evidence for the null hypothesis.

An inconvenience with the use of $v(\mathbf{x})$ is that it may be difficult to compute, especially because typically only one of the two suprema in $v(\mathbf{x})$ is attained. Consequently, it is often easier to use the statistic defined as follows:

Definition 13.6.1 The ratio

$$\lambda(\mathbf{X}) = \frac{\sup_{\theta\in\Theta_0} f_n(\mathbf{X};\theta)}{\sup_{\theta\in\Theta} f_n(\mathbf{X};\theta)}$$

will be called the *generalized likelihood ratio statistic*. □

Under some continuity assumptions, if Θ_0 is a closed set, then the suprema in $\lambda(\mathbf{x})$ are both attained. In particular, if $\hat{\theta}_0 = \hat{\theta}_0(\mathbf{x})$ is the value of the parameter that maximizes $f_n(\mathbf{X};\theta)$ over the set Θ_0, then the numerator in $\lambda(\mathbf{x})$ becomes $f_n(\mathbf{X};\hat{\theta}_0(\mathbf{X}))$. Similarly the value of θ that gives the maximum of the denominator is simply the MLE of θ, denoted $\hat{\theta} = \hat{\theta}(\mathbf{X})$. Thus a useful computational formula for $\lambda(\mathbf{X})$ is

$$\lambda(\mathbf{X}) = \frac{\max_{\theta\in\Theta_0} f_n(\mathbf{X};\theta)}{\max_{\theta\in\Theta} f_n(\mathbf{X};\theta)} = \frac{f_n(\mathbf{X};\hat{\theta}_0)}{f_n(\mathbf{X};\hat{\theta})}.$$

Clearly, $\lambda(\mathbf{X}) \leq 1$, since the denominator, being the maximum over a larger set, is at least as large as the numerator. Since $\lambda(\mathbf{X})$ does not depend on any parameter values, it is a statistic.

To see how $v(\mathbf{X})$ and $\lambda(\mathbf{X})$ are related, observe that

$$\begin{aligned} v(\mathbf{X}) &\leq \frac{\max[\sup_{\theta\in\Theta_1} f_n(\mathbf{X};\theta), \sup_{\theta\in\Theta_0} f_n(\mathbf{X};\theta)]}{\sup_{\theta\in\Theta_0} f_n(\mathbf{X};\theta)} \\ &= \frac{f_n(\mathbf{X};\hat{\theta})}{\sup_{\theta\in\Theta_0} f_n(\mathbf{X};\theta)} = \frac{1}{\lambda(\mathbf{X})}; \end{aligned}$$

hence $v(\mathbf{X})\lambda(\mathbf{X}) \leq 1$. Consequently, large values of $v(\mathbf{X})$ are associated with small values of $\lambda(\mathbf{X})$.

It happens sometimes that the distribution of $\lambda(\mathbf{X})$ under H_0 does not depend on any parameters. In such cases the α-size critical region is obtained from the condition

$$P\{\lambda(\mathbf{X}) \leq k | H_0\} = \alpha. \tag{13.31}$$

The rationale here is that small values of $\lambda(\mathbf{X})$ are a premise for the alternative: it means that the maximum of the likelihood over Θ_0 is much smaller than the overall

maximum. Hence it also means that the maximum over Θ is much higher than the maximum over Θ_0.

It may happen that exact test, using (13.31), is not available. However, it can be shown that if MLE has an asymptotically normal distribution (which is true under very general regularity conditions), then the limiting distribution of $\lambda(\mathbf{X})$ does not involve any parameters.

We have the following theorem, which we provide without proof:

Theorem 13.6.1 *Let $X_1, \ldots, X_n$ be a random sample from the $f(x, \theta)$ distribution with $\theta = (\theta_1, \ldots, \theta_m)$. If $\theta \in \Theta_0$, then statistic $-2 \log \lambda(\mathbf{X})$ has an asymptotically chi-square distribution with $m - r$ degrees of freedom, where r is the number of components of θ completely specified by the null hypothesis ($r < m$). Thus the approximate α-size test is*

$$\textit{reject } H_0 \textit{ if } -2 \log \lambda(\mathbf{X}) \geq \chi^2_{\alpha, m-r}.$$

We will now derive some important GLR tests as examples.

■ **EXAMPLE 13.24**

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from the $N(\mu, \sigma^2)$ distribution with μ and σ^2 unknown. For testing $H_0 : \mu = \mu_0, \sigma$ arbitrary, the parameter space Θ is the upper half-plane, while Θ_0 is the ray $\{(\mu, \sigma^2) : \mu = \mu_0\}$. We now have

$$f(\mathbf{X}; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2\right\}, \tag{13.32}$$

and the MLE's of μ and σ^2 are

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2. \tag{13.33}$$

After we substitute in (13.32), the denominator in $\lambda(\mathbf{X})$ becomes:

$$\begin{aligned} \max_{(\mu, \sigma^2) \in \Theta} f(\mathbf{X}; \mu, \sigma^2) &= (2\pi)^{-n/2} (\hat{\sigma^2})^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \hat{\mu})^2\right\} \\ &= \left(\frac{n}{2\pi e \sum (X_i - \overline{X})^2}\right)^{n/2}. \end{aligned}$$

It remains to find the numerator in $\lambda(\mathbf{X})$, that is,

$$\begin{aligned} \max_{(\mu, \sigma^2) \in \Theta_0} f(\mathbf{X}; \mu, \sigma^2) &= \max_{\sigma^2 > 0} f(\mathbf{X}; \mu_0, \sigma^2) \\ &= \max_{\sigma^2 > 0} (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum (X_i - \mu_0)^2\right\}. \end{aligned}$$

Here the maximum is attained at $\hat{\sigma}_0^2 = (1/n) \sum (X_i - \mu_0)^2$ and equals

$$\left(\frac{n}{2\pi e}\right)^{n/2} \left[\sum (X_i - \mu_0)^2\right]^{-n/2}.$$

Consequently,

$$\lambda(\mathbf{X}) = \left(\frac{\sum(X_i-\mu_0)^2}{\sum(X_i-\overline{X})^2}\right)^{-n/2} = \left(\frac{\sum(X_i-\overline{X})^2+n(\overline{X}-\mu_0)^2}{\sum(X_i-\overline{X})^2}\right)^{-n/2}$$
$$= \left(1+\frac{(\overline{X}-\mu_0)^2}{(1/n)\sum(X_i-\overline{X})^2}\right)^{-n/2} = \left(1+\frac{1}{n-1}t^2(\mathbf{X})\right)^{-n/2},$$

where

$$t(\mathbf{X}) = \frac{\overline{X}-\mu_0}{\sqrt{(1/n)\sum(X_i-\overline{X})^2}}\sqrt{n-1}$$

has a Student's t distribution with $n-1$ degrees of freedom; see formula (10.21). The inequality $\lambda(\mathbf{X}) \leq k$ is equivalent to the inequality $|t(\mathbf{X})| \geq k^*$.

Thus the critical region of our test is $|t(\mathbf{X})| \geq t_{\alpha/2,n-1}$. One can show (see Lehmann and Romano, 2005) that this test is UMP unbiased.

■ **EXAMPLE 13.25**

Consider the GLR test for the hypotheses $H_0 : \mu_1 = \mu_2,\ \sigma_2^2 = \sigma_1^2$ against the alternative $H_1 : \mu_1 \neq \mu_2,\ \sigma_2^2 = \sigma_1^2$ (a more general case, solved in the same way, is obtained when we assume that $\sigma_2^2 = \gamma\sigma_1^2$, where γ is a known constant).

The likelihood, letting $\sigma_1^2 = \sigma_2^2 = \sigma^2$, has the form

$$f(\mathbf{x},\mathbf{y};\mu_1,\mu_2,\sigma^2) = (2\pi\sigma^2)^{-(m+n)/2}e^{-1/2\sigma^2\left[\sum(x_i-\mu_1)^2+\sum(y_j-\mu_2)^2\right]}.$$

By maximizing the likelihood over all (μ_1,μ_2,σ^2) and on the subspace $(\mu_1 = \mu_2 = \mu,\sigma^2)$, one can derive the generalized likelihood ratio test. The exact form of the test is given in Section 13.9. We omit the details of the calculations, which are similar to those in Example 13.24.

■ **EXAMPLE 13.26 Paired Observations**

A situation deceptively similar to that in Example 13.25 occurs when we have the data obtained by observing the values of some attribute of different elements of the population, observed "before" and "after." A typical case would be to measure a certain reaction in human subjects before (X_i) and after (Y_i) some treatment. The purpose is to decide whether or not the treatment has an effect. Since we have here the situation of two samples, $X_1,\ldots,X_n$ (values "before") and $Y_1,\ldots,Y_n$ (values "after"), we cannot apply the method of Example 13.25 for $m = n$.

Such a procedure would not be correct, since in the present case the values X_i, Y_i are not independent (as observations for the same subject). Under the following assumptions, however, one can use here a one-sample test: The observations need to be such that the differences $U_i = Y_i - X_i$ have the same normal distribution with mean μ and variance σ^2. Separately, X_i's and Y_i's do not need to be normally distributed.

We may wish to test the null hypothesis $H_0 : \mu = \mu_0$ against a one- or two-sided alternative $H_1 : \mu > \mu_0$ or $H_1 : \mu \neq \mu_0$, σ^2 arbitrary. In most typical applications we take $\mu_0 = 0$ (treatment has no effect). The form of the test is the same as in Example 13.25, applied to random variables U_i, and we omit the details.

■ **EXAMPLE 13.27**

To test the efficiency of sleeping pills, a drug company uses a sample of patients with insomnia. The time (in minutes) until falling asleep is observed for each person. A few days later, the same patients are given a sleeping pill and the time until falling asleep is measured again. Suppose that the data are

Subject	No Pill (X_i)	With Pill (Y_i)
1	65	45
2	35	5
3	80	61
4	40	31
5	50	20

The proper procedure is to treat the data as paired. The differences $u_i = x_i - y_i$ are then 20, 20, 19, 9, and 30 and we want to test the hypothesis $E(U_i) = 0$ against the alternative $E(U_i) > 0$. Since $\overline{u} = 19.6$ and $s_U^2 = (1/5)\sum(u_i - \overline{u})^2 = 44.24$ we have

$$t = \frac{\overline{u}}{s_U}\sqrt{5-1} = 5.89.$$

Compared with the quantiles of the Student's t distribution with 4 degrees of freedom, the result has a p-value below 0.005.

However, if we treat the problem as a two-sample problem (which is *incorrect*), we obtain a different conclusion. The procedure is described in Section 13.9, but it is worthwhile to explain here why an analysis *of the same numerical values* can lead to two different conclusions, depending whether the values result from paired or unpaired data.

In essence, in both cases we are comparing two means, and trying to determine whether their difference is so small that it may be explained by chance variation (null hypothesis) or that it is large enough to be regarded as "significant." To make such inference we have to assess the amount of variability, to serve as a base for comparison.

Now the formulas for an estimate of variance are different in case where data are paired and in case they are not. In general, the first estimate gives a lower p-value simply because "a person is typically more similar to himself than to another person." Thus, quantitatively speaking, the same difference between means can turn out to be significant when compared with a smaller variance given by the formula for paired data, than when compared with higher variance given by the formula for independent samples.

In the present case we have $\overline{x} = 54, s_X = 16.55, \overline{y} = 32.4$, and $s_Y = 19.41$, so that

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{5s_X^2 + 5s_Y^2}} \sqrt{\frac{5+5-2}{\frac{1}{5}+\frac{1}{5}}} = \frac{21.6}{57.04}\sqrt{20} = 1.69,$$

which, for 8 degrees of freedom, is not significant at the 0.05 level.

We complete this section with a discussion of the problem of reaching a decision before the data collection is complete.

■ **EXAMPLE 13.28**

A research laboratory employs two specialists, Dr. Brown and Dr. Smith. Dr. Brown claims that he invented a certain method that is superior to the currently used method of Dr. Smith. After some debate, it is decided that Dr. Brown's method will be tested. Five experiments are to be run on five consecutive days, starting Monday, and the results $X_1, \ldots, X_5$ recorded[27] It is known that X_i's form a random sample from a normal distribution with an unknown standard deviation. It is also known that for the current method of Dr. Smith, the mean is 10, or perhaps less, so Dr. Brown's method will be declared superior if the mean of X_i's is be higher than 10; that is, if the null hypothesis $H_0 : \mu \leq 10$ (asserting that Dr. Brown's method is no better than Dr. Smith's) is rejected in favor of the alternative $H_1 : \mu > 10$. The significance level, $\alpha = 0.01$, is agreed upon during negotiations.

The observed data for the consecutive five weekdays are: 14.8, 13.6, 13.9, 10.3, and 11.4. We have here $\overline{x} = 12.8$, and $\sqrt{(1/5)\sum(x_i - \overline{x})^2} = 1.6769$, which gives

$$t = \frac{(12.8 - 10)}{1.6769}\sqrt{4} = 3.339.$$

This value is below the critical value $t_{0.01,4} = 3.747$, so Dr. Brown's method is not declared to be superior at the significance level 0.01.

Dr. Brown, however, is not someone who easily gives up. He notices that if the test were run on Wednesday (after only three observations), we would have $\overline{x} = 14.1$, and $\sqrt{(1/3)\sum(x_i - \overline{x})^2} = 0.51$, hence $t = 11.369$, which exceeds the critical value $t_{0.01,2} = 6.965$, so the p-value would be less than 0.01.

The issues involved here are serious. Generally, when the data are collected sequentially, it happens that the conclusion reached on the basis of all data values differs from a conclusion that is reached on the basis of some initial sequence of data points. Is one then justified in reaching the conclusion that is, for some reason, more convenient (e.g., in favor of one's own preferred hypothesis, and likely to get one an extension of a grant)? In particular, can one discard part of the data?

[27] One may feel that important decisions, like about the superiority of a scientific method, cannot be decided on the basis of only five observations. Of course, if possible, one should use a larger sample size to get a better quality of inference. This, however, does not change the essence of the problem.

The answer, of course, is negative, not merely because of the moral issues involved. It is equally important to realize that with modified or discarded data it may be hard, or even impossible, to assess the probabilities involved. To take the example of Dr. Brown, to assess the p-value of the result calculated with the use of the first three data points, one would have to assess the conditional probability, given the conclusion of all five data points, that is ,[28]

$$P\{(X_1, X_2, X_3) \in C_3 | (X_1, \ldots, X_5) \notin C_5\},$$

where C_3 and C_5 are critical regions for sample sizes $n = 3$ and $n = 5$.

Quite apart from the moral and computational issues involved, there exists a theory of *sequential* testing of hypotheses. At each new data point, one of the three decisions is made: "accept H_0," "accept H_1," or "take another observation." The process stops on making either of the first two decisions. The criteria for making these decisions are chosen in such a way that (1) the probabilities of making wrong decisions (type I and type II errors) are bounded by preassigned numbers, and (2) the number of observations taken (being a random variable) has finite expectation.

This theory was developed originally by A. Wald. The details can be found in many advanced textbooks on mathematical statistics.

PROBLEMS

13.6.1 Suppose that we test the null hypothesis that $\mu = 100$ against the alternative that $\mu \neq 100$. The distribution is normal with its variance unknown. We have just two observations, $X_1 = 105$ and $X_2 = 105 + a$. Find a such that the null hypothesis is rejected at the significance level $\alpha = 0.05$?

13.6.2 The following data concerning accidents on various types of highways were obtained from the Ohio Department of Transportation, September 1990 (see Al-Ghamdi, 1991):

Highway Type	Number of Accidents	Annual Million Vehicle Miles	Accident Rate
Scenic	3,621	1,021	3.55
Other 2-lane	36,752	11,452	3.21
Multi-lane	20,348	6,920	3.23
Interstate	10,460	9,412	1.11

From the table it appears that the accident rate on interstate highways is significantly lower than on other types of highways, and that on the first three types, the accident rates are essentially the same. Use the likelihood ratio test to test those claims.

[28] The analysis here is similar (but more complicated) than the ballot problem, studied in Chapter 3. There we calculated the probability that in the process of counting votes, the losing candidate will lead at least once during the counting. A moment of reflection shows that we have here a very similar situation, except that the "votes" (being random variables with values ± 1) are replaced by observations, that may "favor" one or the other hypothesis in varying degree.

13.6.3 The times for the diagnosis and repair of a car with a certain type of problem are assumed to be normally distributed with mean μ and standard deviation $\sigma = 15$ minutes. A mechanic serviced five cars in one day, and it took him a total of 340 minutes. **(i)** Test, at the level $\alpha = 0.05$, hypothesis $H_0 : \mu \leq 60$ against the alternative $H_1 : \mu > 60$. **(ii)** Suppose that you doubt the information that $\sigma = 15$, and decide to test the hypotheses in (i) without assuming anything about σ. If the sum of squares of the five diagnose/repair times is m, how small should m be to reject H_0 at the significance level $\alpha = 0.05$?

13.6.4 Let $X_1, \ldots, X_n$ be a random sample from the GAM$(3, \theta)$ distribution. Derive the GLR test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

13.6.5 A random sample of size n was selected from the EXP(θ) distribution. Perform the GLR test of $H_0 : \theta = 2$ against the alternative $H_1 : \theta \neq 2$ if the actual observations are 0.57, 0.21, 2.18, 0.85, 1.44. Use $\alpha = 0.05$.

13.6.6 Derive the GLR test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ based on a random sample $X_1, \ldots, X_n$ selected from the BETA$(1, \theta)$ distribution. Determine an approximate critical value for a size α.

13.6.7 Two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, X_n$ were selected from the EXP(θ_1) and EXP(θ_2) distributions, respectively. Derive the GLR test of size 0.05 for $H_0 : \theta_1 = \theta_2$ against $H_1 : \theta_1 \neq \theta_2$.

13.6.8 Let two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, X_n$ be selected from the N(μ_1, σ_1^2) and N(μ_2, σ_2^2) distribution, respectively. Obtain the GRL test for $H_0 : \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ against $H_1 : \mu_1 \neq \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$.

13.7 CONDITIONAL TESTS

As explained in the previous section, the parameter (or parameters) that are not constrained by the null hypothesis are called *nuisance* parameters. Besides the likelihood ratio tests, for testing in the presence of nuisance parameters one can use conditional tests.

Suppose that the distribution of X depends on a parameter $\theta = (\eta, \tau)$, where η is the parameter tested and τ is the nuisance parameter. Assume that we want to test the null hypothesis that $\eta = \eta_0$ against the alternative that $\eta < \eta_0$. These are in fact composite hypotheses:

$$H_0 : \eta = \eta_0, \tau \text{ arbitrary} \quad \text{against} \quad H_1 : \eta < \eta_0, \tau \text{ arbitrary}.$$

It may happen that there exists a sufficient statistic, say T, for τ. In such a case the conditional distribution of X given $T = t$ does not depend on τ, and it may then happen that one can find a test on a given level α for H_0 (for each t separately).

We will first illustrate such a situation by an example.

■ **EXAMPLE 13.29**

Let X and Y be independent binomial random variables: $X \sim \text{BIN}(n_1, p_1)$ and $Y \sim \text{BIN}(n_2, p_2)$. We want to test the hypothesis that $p_1 = p_2$ against one- or two-sided alternative. If the null hypothesis is true, then letting $p = p_1 = p_2$, we have

$$\begin{aligned} P\{X = x | X + Y = k\} &= \frac{P\{X = x, X + Y = k\}}{P\{X + Y = k\}} \\ &= \frac{P\{X = x\}P\{Y = k - x\}}{P\{X + Y = k\}} \\ &= \frac{\binom{n_1}{x} p^x q^{n_1 - x} \binom{n_2}{k-x} p^{k-x} q^{n_2 - k + x}}{\binom{n_1+n_2}{k} p^k q^{n_1+n_2-k}} \\ &= \frac{\binom{n_1}{x}\binom{n_2}{k-x}}{\binom{n_1+n_2}{k}}, \end{aligned} \tag{13.34}$$

which is independent of the nuisance parameter p.

The question is which values of X (given $X + Y = k$) constitute the strongest premise against the null hypothesis, and how should one assess the p-value of the result. Here the argument is as follows: We can expect X/n_1 to be close to p_1 and $Y/n_2 = (k - X)/n_2$ to be close to p_2. So, if H_0 is true, we can expect $X/n_1 \approx k/n_2 - X/n_2$. For the two-sided alternative, the "worst" cases are where X is close to 0 or close to k so that the critical region comprise two tails of the hypergeometric distribution (13.34).

Letting

$$u(j) = u(j; n_1, n_2, k) = \frac{\binom{n_1}{j}\binom{n_2}{k-j}}{\binom{n_1+n_2}{k}},$$

we may define the p-value of the result $X = x$ for two-sided test as

$$2 \min \left\{ \sum_{j \le x} u(j), \sum_{j \ge x} u(j) \right\}. \tag{13.35}$$

This is Fisher's exact test.

■ **EXAMPLE 13.30**

Let X and Y be two independent observations of two Poisson random variables, with parameters λ_1 and λ_2, respectively. We want to test the hypothesis $H_0 : \lambda_2 = \delta\lambda_1$ against the alternative $H_1 : \lambda_2 > \delta\lambda_1$, where $\delta > 0$ is some fixed constant. For instance, if $\delta = 1$, we have the hypothesis of equality of parameters in two Poisson distributions.

The joint distribution of (X, Y) is here

$$\begin{aligned} f(x, y; \lambda_1, \lambda_2) &= \frac{\lambda_1^x}{x!} \times \frac{\lambda_2^y}{y!} e^{-(\lambda_1+\lambda_2)} \\ &= \frac{1}{x!y!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^x \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^y (\lambda_1 + \lambda_2)^{x+y} e^{-(\lambda_1+\lambda_2)}. \end{aligned}$$

This suggests reparametrization with $\eta = \lambda_2/\lambda_1$ and $\tau = \lambda_1 + \lambda_2$, leading to the joint probability function

$$f(x, y; \eta, \tau) = \binom{x+y}{x} \left(\frac{1}{1+\eta}\right)^x \left(\frac{\eta}{1+\eta}\right)^y \frac{\tau^{x+y}}{(x+y)!} e^{-\tau}. \tag{13.36}$$

It is now clear that $T = X + Y$ is a sufficient statistic for τ, and that given $T = t$, the random variable X has the conditional distribution that is binomial with parameters t and $1/(1+\eta)$. Under the null hypothesis, we have $\eta = \delta$, while under the alternative hypothesis, $\eta > \delta$. Given $T = t$, we may therefore test the null hypothesis $H_0 : p = 1/(1+\delta)$ against the alternative $H_1 : p < 1/(1+\delta)$, observing the random variable X with the BIN(t, p) distribution.

■ **EXAMPLE 13.31**

Suppose that traffic engineers suggest a certain change of a traffic light sequence to reduce the number of accidents at some type of intersections. Two intersections, far apart to ensure independence but otherwise identical in all aspect (traffic intensity, road condition, etc.) are to be tested. The number X of accidents at the intersection with the new traffic light pattern is 7. Over the same period at each intersection with a previous traffic light pattern, the number of accidents is $Y = 13$. At the significance level $\alpha = 0.05$, does this indicate that the new traffic light pattern decreases the probability of an accident?

We want to test the hypothesis that $\lambda_1 = \lambda_2$ against the alternative $\lambda_1 < \lambda_2$ (so that $\delta = 1$ and $1/(1+\delta) = 1/2$). Since $t = x + y = 7 + 13 = 20$ and $X \sim$ BIN$(20, 1/2)$

$$P\{X \le 7\} = \sum_{j=0}^{7} \binom{20}{j} \left(\frac{1}{2}\right)^{20} = 0.1310.$$

The answer is clearly negative. If the new traffic lights pattern does not affect the probability of accident at all, the outcome as obtained (13 against 7), or more extreme, has more than 13% chance of occurring.

The scheme above is an application of the following theorem:

Theorem 13.7.1 *Suppose that the random sample* $\mathbf{X} = (X_1, \ldots, X_n)$ *has a joint distribution of the form*

$$f_n(\mathbf{x}, \eta, \tau_1, \ldots, \tau_r) = a(\eta, \tau_1, \ldots, \tau_r) b(\mathbf{x}) \exp\left\{\eta u(\mathbf{x}) + \sum_{j=1}^{r} \tau_j v_j(\mathbf{x})\right\} \tag{13.37}$$

for some functions $a, b, u, v_1, \ldots, v_r$. *Let* $U = u(\mathbf{X})$, *and let* $\mathbf{V} = (V_1, \ldots, V_r)$, *where* $V_j = v_j(\mathbf{X})$ *for* $j = 1, \ldots, r$. *Then for each* η, *variables* $V_1, \ldots, V_r$ *are jointly sufficient for* $(\tau_1, \ldots, \tau_r)$, *and the conditional distribution of* U *given* $V(\mathbf{x}) = \mathbf{v}$ *depends on* η *but not on* $(\tau_1, \ldots, \tau_r)$. *A test with size* α *for testing* $H_0 : \eta \le \eta_0$ *against* $H_1 : \eta > \eta_0$ *is obtained by rejecting* H_0 *if* $U(\mathbf{x}) \ge q(\mathbf{v})$, *where* $P\{U \ge q(\mathbf{v}) | \mathbf{v}\} = \alpha$ *for* $\eta = \eta_0$.

For testing $H_0 : \eta \geq \eta_0$ against $H_1 : \eta < \eta_0$ the directions of inequalities will be reversed; a two-tailed test is to be used in case of two-sided hypothesis $H_0 : \eta = \eta_0$ against $H_1 : \eta \neq \eta_0$.

Lehmann and Romano (2005) show that under some regularity conditions these tests are UMP unbiased.

■ **EXAMPLE 13.32**

To see how the situation of Example 13.30 falls under the scheme of Theorem 13.7.1, observe that the distribution of $\mathbf{X} = (X_1, X_2)$ can be reduced to the form (13.37) with $r = 1, U(\mathbf{X}) = X_1, V(\mathbf{X}) = X_1 + X_2$. This is calculated as follows: Take $s = \lambda_2/\lambda_1, t = \lambda_1 + \lambda_2$; hence $\lambda_1 = t/(1+s), \lambda_2 = st/(1+s)$). The distribution

$$f(x_1, x_2; \lambda_1, \lambda_2) = \frac{1}{x_1!}\lambda_1^{x_1} e^{-\lambda_1} \times \frac{1}{x_2!}\lambda_2^{x_2} e^{-\lambda_2}$$

reduces, after some algebra, to the form

$$f = \binom{V(\mathbf{x})}{U(\mathbf{x})} \frac{1}{V(\mathbf{x})!} e^{-t} \exp\left\{ U(\mathbf{x}) \ln \frac{1}{s} + V(\mathbf{x}) \left[\log t + \log \frac{s}{1+s} \right] \right\}.$$

Introducing new parameters $\eta = \log(1/s) = \log(\lambda_1/\lambda_2)$ and $\tau = \log t + \log s/(1+s) = \log \lambda_2$, and then expressing t in terms of τ and η, we obtain a density in the form (13.37).

PROBLEMS

13.7.1 Suppose that in a group of 10 randomly sampled Democrats only 2 favor a certain issue, whereas in a sample of 12 Republicans the same issue is favored by 5 persons. At the level $\alpha = 0.05$, does this result indicate that the fractions p_D and p_R of Democrats and Republicans favoring the issue in question are different? Find the p-value by carrying out the two-sided Fisher's test.

13.7.2 A company A that produces batteries claims that their product is "at least 50% better" than the batteries produced by company B. To test the claim, batteries A and B are used one after another in two analogue devices. That is, one device has a battery A installed and is left running until the battery becomes dead. It is then immediately replaced by another battery A, and so on. The second device runs parallel on batteries B.

Assume that the lifetimes of the batteries are exponential random variables, with densities $\lambda_A e^{-\lambda_A t}$ and $\lambda_B e^{-\lambda_B t}$. Suppose that in some time (e.g., a week) batteries A had to be replaced 5 times, whereas batteries B had to be replaced 9 times. Test the advertising claim by determining the p-value of the result. (*Hint:* If interarrival times are exponential, the process is Poisson.)

13.8 TESTS AND CONFIDENCE INTERVALS

In this section we will briefly explain how the theory of testing statistical hypotheses is related to the theory of confidence intervals, discussed in Chapter 12. This

connection was noticed by Neyman (who laid the foundations to both theories) as early as in 1938. To simplify the presentation, assume that $X_1, \ldots, X_n$ is a random sample from a distribution $f(x;\theta)$, and that θ is a one-dimensional parameter.

A confidence interval (with confidence level $1-\alpha$) is a random interval $[L, U] = [L(\mathbf{X}), U(\mathbf{X})]$ such that for every $\theta \in \Theta$,

$$P_\theta\{L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\} = 1 - \alpha. \tag{13.38}$$

Suppose that we want to test hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. The equivalence of tests and confidence intervals is based on the fact that if we can construct confidence interval (13.38), then we can also construct an α-level test of H_0 against H_1, and conversely: given a testing procedure of level α, we can construct a confidence interval.

Indeed, condition (13.38) for $\theta = \theta_0$ allows us to define the set

$$A = \{\mathbf{x} = (x_1, \ldots, x_n) : L(\mathbf{x}) \leq \theta_0 \leq U(\mathbf{x})\}.$$

Clearly, if we take the set A as the acceptance region of H_0 (equivalently, we let $C = A^c$ to be the critical region for H_0), we obtain a test with level α:

$$P\{H_0 \text{ is rejected} | H_0 \text{ is true }\} = P\{\theta_0 \notin [L(\mathbf{X}), U(\mathbf{X})] | \theta = \theta_0\} = \alpha.$$

Conversely, suppose that for every θ' we can construct an α-level test of the hypothesis $H_0 : \theta = \theta_0'$. This means that for every θ' we have a critical region $C_{\theta'}$ such that

$$P\{\mathbf{X} \notin C_{\theta'} | \theta = \theta'\} = 1 - \alpha. \tag{13.39}$$

Define now, for every $\mathbf{x}$,

$$B(\mathbf{x}) = \{\theta' : \mathbf{x} \notin C_{\theta'}\}. \tag{13.40}$$

When $\mathbf{x}$ is the observed value of random vector $\mathbf{X}$, we obtain a random set $B(\mathbf{X})$. From (13.39) and (13.40) it follows that $P_\theta\{\theta \in B(\mathbf{X})\} = 1 - \alpha$, which means that $B(\mathbf{X})$ is a confidence set for θ with confidence level $1 - \alpha$.

Except for the fact that the second part of the argument provides us with confidence sets (not necessarily intervals), the argument shows that the two theories are essentially equivalent, at least in case of one-dimensional parameters.

The main results of the theory combine optimality properties of tests and confidence intervals. For instance, confidence intervals (sets) associated with UMP tests have the property of being the shortest possible (uniformly most accurate, or UMA confidence Intervals). It is worth mentioning here that the theory extends also to the case of testing in presence of nuisance parameters, but we will not go into the details here.

13.9 REVIEW OF TESTS FOR NORMAL DISTRIBUTIONS

We now review the major testing procedures for parameters of normal distributions. This section is intended as a convenient reference for users rather than an exposition of new concepts or results. If the derivations were given in the other sections, we

refer to them. In other cases, we omit the derivations, specifying only the properties of the tests, possibly with indications of the proofs.

In all tests below α is the significance level and $1-\beta$ is the power, z_p is the upper pth quantile of standard normal random variable Z, so that

$$P\{Z > z_p\} = 1 - \Phi(z_p) = p.$$

Similarly, $t_{p,\nu}$ is the upper pth quantile of the Student's t distribution with ν degrees of freedom, $\chi^2_{p,\nu}$ is the upper pth quantile of the chi-square distribution with ν degrees of freedom, and F_{p,ν_1,ν_2} denotes an upper pth quantile of the F distribution with ν_1 and ν_2 degrees of freedom.

One-Sample Procedures

Let $X_1, \ldots, X_n$ be a random sample from distribution $\mathrm{N}(\mu, \sigma^2)$.

Hypotheses about the mean, variance known

The sufficient statistic is $\sum_{i=1}^n X_i$, hence also $\overline{X}$.

- Hypotheses:
  One-sided alternatives:
  (i) $H_0 : \mu = \mu_0$ (or $\mu \le \mu_0$) vs. $H_1 : \mu > \mu_0$.
  (ii) $H_0 : \mu = \mu_0$ (or $\mu \ge \mu_0$) vs. $H_1 : \mu < \mu_0$.

  Two-sided alternatives:
  (iii) $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \ne \mu_0$.
  (iv) $H_0 : \mu_1 \le \mu \le \mu_2$ vs. $H_1 : \mu < \mu_1$ or $\mu > \mu_2$.

  Other cases:
  (v) $H_0 : \mu \le \mu_1$ or $\mu \ge \mu_2$ vs. $H_1 : \mu_1 < \mu < \mu_2$.

- Test statistics:
  (i)–(iii) $T_1 = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$
  (iv), (v) $T_2 = \frac{\overline{X}-\mu^*}{\sigma/\sqrt{n}}$, where $\mu^* = \frac{\mu_1+\mu_2}{2}$
- Corresponding critical regions, reject H_0 if:
  (i) $T_1 \ge z_\alpha$, **(ii)** $T_1 \le z_{1-\alpha}$, **(iii)** $|T_1| \ge z_{\alpha/2}$, **(iv)** $|T_2| \ge k_\alpha$,
  (v) $|T_2| \le v_\alpha$, where k_α and v_α are determined from

$$\Phi\left(\frac{\Delta}{\sigma/\sqrt{n}} + k_\alpha\right) - \Phi\left(\frac{\Delta}{\sigma/\sqrt{n}} - k_\alpha\right) = 1 - \alpha$$

and

$$\Phi\left(\frac{\Delta}{\sigma/\sqrt{n}} + v_\alpha\right) - \Phi\left(\frac{\Delta}{\sigma/\sqrt{n}} - v_\alpha\right) = \alpha,$$

with $\Delta = (\mu_2 - \mu_1)/2$. Tests (i), (ii), and (v) are UMP tests; (iii) and (iv) are UMP unbiased tests.

- Power:

$$\text{(i) } \pi(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right)$$

$$\text{(ii) } \pi(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_\alpha\right)$$

$$\text{(iii) } \pi(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) + \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2}\right)$$

$$\text{(iv) } \pi(\mu) = 1 - \Phi\left(\frac{\mu^* - \mu}{\sigma/\sqrt{n}} + k_\alpha\right) + \Phi\left(\frac{\mu^* - \mu}{\sigma/\sqrt{n}} - k_\alpha\right)$$

$$\text{(v) } \pi(\mu) = 1 - \Phi\left(\frac{\mu^* - \mu}{\sigma/\sqrt{n}} + v_\alpha\right) + \Phi\left(\frac{\mu^* - \mu}{\sigma/\sqrt{n}} - v_\alpha\right)$$

- Sample size determination:
  Wanted: sample size n giving power at least $1 - \beta$ at μ
  (i), (ii) $n \geq \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(\mu - \mu_0)^2}\sigma^2$
  (iii) Solve numerically for n,

$$\Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} + z_{1-\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} - z_{1-\alpha/2}\right) = \beta.$$

  (iv) Solve numerically for n,

$$\Phi\left(\frac{\mu^* - \mu}{\sigma}\sqrt{n} + k_\alpha\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} - k_\alpha\right) = \beta.$$

  (v) Solve numerically for n,

$$\Phi\left(\frac{\mu^* - \mu}{\sigma}\sqrt{n} + v_\alpha\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} - v_\alpha\right) = \beta.$$

■ EXAMPLE 13.33 Generic Problem

All philogaps presently on the market have an average concentration of muzzz of at least 3.7 mg per philogap. A company claims to have discovered a new method of production that will decrease the average muzzz content to the level below 3.7 mg. To test this claim, the muzzz content of 15 philogaps of this company are analyzed, and their average muzzz content is found to be 3.52 mg.

It is known that the standard deviation of the muzzz content in a philogap does not depend on the production process and equals 0.35 mg. It is also known that the muzzz content is normally distributed. At the significance level $\alpha = 0.01$, does this finding indicate that the new production process decreases the concentration of muzzz in philogaps?

Remark You are probably curious about what are philogaps and what is muzzz (spelled with triple z). These words, to our best knowledge, mean nothing. If you so wish, substitute "objects" and "attribute A," or "cars" and

"miles per gallon," "beer" and "alcohol content," "oranges" and "sugar content in juice," and so on, and change numbers and possibly, inequality directions accordingly.

We set up this problem as being of type (ii): The null hypothesis is H_0 : $\mu \geq 3.7$ and the alternative is H_1 : $\mu < 3.7$. The value of the test statistic T_1 is $t = [(3.52-3.7)/0.35]\sqrt{15} = -1.99$. The 1% quantile of the standard normal distribution is -2.33, so the null hypothesis is not rejected. The average 3.52 or less in a sample of 15 is more likely than 0.01, if the mean is in fact 3.7. Therefore the null hypothesis cannot be rejected at the significance level 0.01. The p-value is equal to $P(Z < -1.99) = 0.0233$, or about 2.3%.

Suppose that we want not only the 1% level of significance but also at least 95% chance of detecting the improvement of the average muzzz content in philogaps by 0.15 mg. In statistical terms, this means that we want the power to be at least 0.95 at $\mu = 3.7 - 0.15 = 3.55$. Then we need to take the sample of at least

$$\frac{(z_{0.01} + z_{0.05})^2}{(3.7 - 3.55)^2}(0.35)^2 = \frac{(2.33 + 1.96)^2}{(0.15)^2}(0.35)^2 = 100.2,$$

meaning that n should be at least 101.

■ **EXAMPLE 13.34**

A food-packing company purchased a new machine to fill plastic containers with sour cream. The nominal weight, as listed on the container, is 8 oz. The dial on the machine can be set on average weight . . . , 7.98, 8.00, 8.02, 8.04, . . . oz. When it is set on 8.00 oz (say), it puts into successive containers the amounts $X_1, X_2, \ldots$, which are normally distributed with some mean μ and standard deviation (the same for all settings of the dial) $\sigma = 0.005$ oz. This standard deviation reflects the unavoidable container-to-container variability of the amounts of sour cream about their mean μ. Naturally it is impossible to build a machine that gives the average μ *exactly* equal to the setting on the dial, be it 7.995, 8.002, and so on.

A consumer protection agency may disregard the instances where the variability will occasionally lead to a container with more than the nominal amount of sour cream. However, it might strongly object if the average μ is even slightly less than 8 oz, as this constitutes a systematic theft from society. On the other hand, if μ exceeds the nominal weight even slightly, it may in time constitute a sizable free gift of the company to society.

The company decides that it will risk getting into trouble with the consumer protection agency, or absorb the loss, if the mean μ satisfies the inequality $7.995 \leq \mu \leq 8.015$ but wants to avoid both lower μ and higher μ. Careful measurements of 50 containers with sour cream give the average $\overline{x} = 8.017$.

We are now in situation (iv) where the null hypothesis states that μ lies between some bounds, and we may proceed as follows: We have here $\mu_1 = 7.995$ and $\mu_2 = 8.015$, so $\mu^* = 8.005$ oz and $\Delta = 0.01$. Consequently the observed value of T_2 is $t_2 = [(8.017 - 8.005)/0.005]\sqrt{50} = 16.97$. On the

other hand, $(\Delta/\sigma)\sqrt{n} = 14.142$, and the equation

$$\Phi(14.142 + k_\alpha) - \Phi(14.142 - k_\alpha) = 1 - \alpha$$

reduces to $\Phi(14.142 - k_\alpha) = \alpha$. For $\alpha = 0.01$ we must have $14.142 - k_\alpha = -2.33$; hence $k_\alpha = 16.472$. The observed value 16.97 exceeds k_α, and this indicates that the null hypothesis should be rejected at the level 0.01.

It should be stresses here that the conclusion is that $\mu < 7.995$ or $\mu > 8.015$ despite the fact that the result $\overline{x} = 8.017$ suggests that the second of the two inequalities holds. The point is that the null hypothesis states that $7.995 \leq \mu \leq 8.015$, and its rejection is logically equivalent to the pair of inequalities.

To conclude that $\mu > 8.015$, we should test the null hypothesis $\mu \leq 8.015$. The test statistic is then

$$t_1 = \frac{8.017 - 8.015}{0.005}\sqrt{50} = 2.82,$$

with a corresponding p-value of about 0.0025.

Hypotheses about the mean, variance unknown

The jointly sufficient statistics are $\sum X_i$ and $\sum X_i^2$, or equivalently $\overline{X}$ and $\sum(X_i - \overline{X})^2$.

- Hypotheses:

One-sided alternative:
(i) $H_0 : \mu = \mu_0, \sigma > 0$ (or $H_0 : \mu \leq \mu_0, \sigma > 0$) vs. $H_1 : \mu > \mu_0, \sigma > 0$.
(ii) $H_0 : \mu = \mu_0, \sigma > 0$ (or $H_0 : \mu \geq \mu_0, \sigma > 0$) vs. $H_1 : \mu < \mu_0, \sigma > 0$.
Two-sided alternative:
(iii) $H_0 : \mu = \mu_0, \sigma > 0$ vs. $H_1 : \mu \neq \mu_0, \sigma > 0$

- Test statistic:

$$t = \frac{\overline{X} - \mu_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}}\sqrt{n-1}.$$

Remark Many tests use the notation $t = [(\overline{X} - \mu_0)/S]\sqrt{n-1}$ or $t = [(\overline{X} - \mu_0)/S]\sqrt{n}$. These may be confusing, since one has to bear in mind that in the first case S^2 is the MLE of σ^2; that is, $S^2 = (1/n)\sum(X_i - \overline{X})^2$. In the second case, S^2 is the unbiased estimator of variance: $S^2 = [1/(n-1)]\sum(X_i - \overline{X})^2$. The notation for S^2 is *not* standardized across statistical textbooks and papers.

- Corresponding critical regions, reject H_0 if:
  (i) $t \geq t_{\alpha,n-1}$.
  (ii) $t \leq -t_{\alpha,n-1}$.
  (iii) $|t| \geq t_{\alpha/2,n-1}$.

For each of these tests the power is the function π depending on a two-dimensional variable (μ, σ^2). We have $\pi(\mu_0, \sigma^2) = \alpha$ for all σ^2. However, for $\mu \neq \mu_0$, the values $\pi(\mu, \sigma^2)$ depend on (unknown) σ^2, and are given by the noncentral Student distribu-

tion. Therefore the sample size determination requires some additional information about σ^2 (a two-stage procedure, etc.). All three tests are UMP unbiased[29] tests.

■ **EXAMPLE 13.35**

A certain make of cars is advertised as attaining gas mileage of at least 32 miles per gallon. Twelve independent tests gave the results 33, 28, 31, 28, 26, 30, 31, 28, 27, 33, 35, 29 miles per gallon. What can one say about the advertisements in light of these data?

Let us make the assumption that the observed mileages are normally distributed. We may then set the problem of evaluation of the advertising claim as that of testing the hypothesis $H_0 : \mu \geq 32$ (claim is true) against the alternative $H_1 : \mu < 32$ (claim is false). We have here $n = 12, \sum x_i = 359, \sum x_i^2 = 10,823$. Thus $\bar{x} = 29.92, \sum(x_i - \bar{x})^2 = 82.92$, so $t = -2.62$. Since the critical values for 11 degrees of freedom are $t_{0.025,11} = 2.201, t_{0.01,11} = 2.718$, the result is not significant at the 0.01 level but significant at the 0.025 level. In other words, the p-value is between 0.01 and 0.025.

Hypotheses about the variance, mean known

The sufficient statistic is $\sum(X_i - \mu)^2$. The sample size can be 1.

- Hypotheses:

  One-sided alternative:

  (i) $H_0 : \sigma^2 = \sigma_0^2$ (or $H_0 : \sigma^2 \leq \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$.

  (ii) $H_0 : \sigma^2 = \sigma_0^2$ (or $H_0 : \sigma^2 \geq \sigma_0^2$) vs. $H_1 : \sigma^2 < \sigma_0^2$.

  Two-sided alternative:

  (iii) $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$.

- Test statistic:

$$U = \sum_{i=1}^{n}(X_i - \mu)^2/\sigma_0^2.$$

- Corresponding critical regions, reject H_0 if:

  (i) $U < \chi^2_{\alpha,n}$.

  (ii) $U > \chi^2_{1-\alpha,n}$.

  (iii) $U < \chi^2_{1-\alpha/2,n}$ or $U > \chi^2_{\alpha/2,n}$.

- Power:

$$\begin{aligned}\textbf{(i)}\ \pi(\sigma^2) &= 1 - \beta(\sigma^2) = P_{\sigma^2}\{U \leq \chi^2_{\alpha,n}\} = P_{\sigma^2}\left\{\frac{\sum(X_i - \mu)^2}{\sigma_0^2} \leq \chi^2_{\alpha,n}\right\} \\ &= P_{\sigma^2}\left\{\frac{\sum(X_i - \mu)^2}{\sigma^2} \leq \sigma_0^2\frac{\chi^2_{\alpha,n}}{\sigma^2}\right\} = F_n\left(\sigma_0^2\frac{\chi^2_{\alpha,n}}{\sigma^2}\right).\end{aligned}$$

Calculations for remaining cases are similar.

Tests (i) and (ii) are UMP. Test (iii) is asymptotically (for $n \to \infty$) UMP unbiased. A UMP unbiased test in case (iii) is obtained if the thresholds are equal to the

[29] As mentioned before, two-sided UMP tests may not exist. Nevertheless, they can often be found in some restricted classes of tests, for example, in the class of unbiased tests.

endpoints of the shortest $(1-\alpha)$-level confidence interval for σ^2 (see Section 12.7).

Hypotheses about the variance, mean unknown

Jointly sufficient statistics are $\overline{X}$ and $\sum(X_i - \overline{X})^2$. Tests require $n \geq 2$.

- Hypotheses:

  One-sided alternative:

  (i) $H_0 : \sigma^2 = \sigma_0^2$ (or $H_0 : \sigma^2 \leq \sigma_0^2$) vs. $H_1 : \sigma^2 > \sigma_0^2$; $\mu \in R$.

  (ii) $H_0 : \sigma^2 = \sigma_0^2$ (or $H_0 : \sigma^2 \geq \sigma_0^2$) vs. $H_1 : \sigma^2 < \sigma_0^2$; $\mu \in R$.

  Two-sided alternative:

  (iii) $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$; $\mu \in R$.

- Test statistic:

$$V = \sum_{i=1}^{n} (X_i - \overline{X})^2 / \sigma_0^2.$$

- Corresponding critical regions, reject H_0 if:

  (i) $V > \chi^2_{\alpha,n-1}$.

  (ii) $V < \chi^2_{1-\alpha,n-1}$.

  (iii) $V < \chi^2_{1-\alpha/2,n-1}$ or $V > \chi^2_{\alpha/2,n-1}$.

- Power:

$\pi(\mu, \sigma^2) = P_{\mu,\sigma^2}\{H_0 \text{ is rejected}\}$ is obtained in the same way as in the case above.

Test (i) is unbiased. Tests (ii) and (iii) are not UMP unbiased.

The following example was taken from Larsen and Marx (1986):

■ **EXAMPLE 13.36**

The A above middle C is the note given to an orchestra, usually by the oboe, for tuning purposes. Its pitch is defined to be the sound of a tuning fork vibrating at 440 hertz (Hz). No tuning fork, of course, will always vibrate at *exactly* 440 Hz; rather, the pitch, Y, is a random variable. Suppose that Y is normally distributed with $\mu = 440$ Hz and variance σ^2 (here the parameter σ^2 is a measure of quality of the tuning fork). With the standard manufacturing process, $\sigma^2 = 1.1$. A new production technique has just been suggested, however, and its proponents claim it will yield values of σ^2 significantly less than 1.1. To test the claim, six tuning forks are made according to the new procedure. The resulting vibration frequencies are 440.8, 440.3, 439.2, 439.8, 440.6, and 441.1 Hz.

We will test the hypothesis $H_0 : \sigma^2 = 1.1$ against the alternative $H_1 : \sigma^2 < 1.1$, at the significance level $\alpha = 0.05$. First, we accept as the fact that the new production process gives the mean $\mu = 440$. Then $u = \sum(x_i - 440)^2/1.1 = 1.62$. The value $\chi^2_{0.05,6} = 1.635$, so we conclude that the new production process indeed gives a variance less than 1.1.

Suppose that we have some doubts about whether or not $\mu = 440$. We can then use the method for unknown μ. Now $v = 2.219$ and $\chi^2_{0.05,5} = 1.145$, and we do not reach the preceding conclusion.

Two-Sample Procedures

Two independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_m$ are selected from distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively.

Hypotheses about the means, variances known

Jointly sufficient statistics are $\sum_{i=1}^m X_i$ and $\sum_{i=1}^n Y_i$, or $\overline{X}$, and $\overline{Y}$.

- Hypotheses:

  One-sided alternative:

  (i) $H_0 : \mu_1 = \mu_2$ (or $H_0 : \mu_1 \leq \mu_2$) vs. $H_1 : \mu_1 > \mu_2$

  Two-sided alternative:

  (ii) $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$

Remark The opposite inequality in (i) reduces to changing the role of X_i's and Y_j's. The apparently more general hypotheses of the form $\mu_1 = \mu_2 - \Delta$, reduce to the ones above by subtracting constant Δ from all Y_j's.

- Test statistic:

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$$

- Corresponding critical regions, reject H_0 if:

  (i) $U \geq z_\alpha$.

  (ii) $|U| \geq z_{\alpha/2}$

- Power:

  (i) $\pi(\mu_1, \mu_2) = 1 - \beta(\Delta) = 1 - \Phi\left(z_\alpha + \Delta/\sqrt{\sigma_1^2/m + \sigma_2^2/n}\right)$

  (ii) $\pi(\mu_1, \mu_2) = 1 - \beta(\Delta) = 1 - \Phi\left(z_{\alpha/2} + \Delta/\sqrt{\sigma_1^2/m + \sigma_2^2/n}\right)$

$$+ \Phi\left(-z_{\alpha/2} + \Delta/\sqrt{\sigma_1^2/m + \sigma_2^2/n}\right),$$

where $\mu_2 - \mu_1 = \Delta$.

Tests (i) is UMP; test (ii) is UMP unbiased.

Hypotheses about the means, variances unknown, but $\sigma_2^2 = \gamma\sigma_1^2$.

Jointly sufficient statistics are $\overline{X}$, $\overline{Y}$, and $\sum(X_i - \overline{X})^2 + \frac{1}{\gamma}\sum(Y_j - \overline{Y})^2$.

- Hypotheses:

  One-sided alternative:

  (i) $H_0 : \mu_1 = \mu_2$ (or $H_0 : \mu_1 \leq \mu_2$) vs. $H_1 : \mu_1 > \mu_2$; $\sigma_1^2 > 0$

  Two-sided alternative:

  (ii) $H_0 : \mu_1 = \mu_2$, vs. $H_1 : \mu_1, \neq \mu_2$; $\sigma_1^2 > 0$

Remark The opposite inequality in (i) reduces to changing the role of X_i's and Y_j's. The apparently more general hypotheses of the form $\mu_1 = \mu_2 + \Delta$, reduce to the ones above by adding a constant Δ to all Y_j's.

- Test statistic:

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\sum(X_i - \overline{X})^2 + \frac{1}{\gamma}\sum(Y_j - \overline{Y})^2}} \sqrt{\frac{m+n-2}{1/m + \gamma/n}}.$$

Remark $\gamma = 1$ if $\sigma_1^2 = \sigma_2^2$.

- Corresponding critical regions, reject H_0 if:
  (i) $U > t_{\alpha,m+n-2}$.
  (ii) $|U| > t_{\alpha/2,m+n-2}$.

- Power depends on μ_1, μ_2, σ_1^2, and it is expressed through noncentral Student's t distribution. Tests (i) and (ii) are UMP unbiased.

Hypotheses about the variances, means unknown

Jointly sufficient statistics are $\overline{X}, \overline{Y}, \sum(X_i - \overline{X})^2$, and $\sum(Y_j - \overline{Y})^2$.

- Hypotheses:
  One-sided alternative:
  (i) $H_0 : \sigma_2^2 = \gamma\sigma_1^2$ (or $H_0 : \sigma_2^2 \leq \gamma\sigma_1^2$) vs. $H_1 : \sigma_2^2 > \gamma\sigma_1^2$; $\mu_1 \in R$, $\mu_2 \in R$
  Two-sided alternative:
  (ii) $H_0 : \sigma_2^2 = \gamma\sigma_1^2$ vs. $H_1 : \sigma_2^2 \neq \gamma\sigma_1^2$; $\mu_1 \in R$, $\mu_2 \in R$.

Remark $\gamma = 1$ if the hypothesis asserts equality of variances.

- Test statistic:

$$F = \frac{\sum(Y_i - \overline{Y})^2/\gamma(n-1)}{\sum(X_j - \overline{X})^2/(m-1)}.$$

- Corresponding critical regions, reject H_0 if:
  (i) $F \geq F_{\alpha,n-1,m-1}$
  (ii) $F \leq F_{1-\alpha/2,n-1,m-1} = \frac{1}{F_{\alpha/2,m-1,n-1}}$ or $F \geq F_{\alpha/2,n-1,m-1}$.

Tests (i) and (ii) are UMP unbiased.

Hypotheses about the variances; one or both means known.

Jointly sufficient statistics are $\overline{X}, \overline{Y}, \sum(X_i - \overline{X})^2, \sum(Y_j - \overline{Y})^2$. Whenever a mean is known, it replaces the sample average in ratio F, and the divisor (and number of degrees of freedom) changes into m (respectively, n). For instance, suppose that both μ_1 and μ_2 are known. Then we test $H_0 : \sigma_2^2 = \sigma_1^2$ vs. $H_1 : \sigma_2^2 > \sigma_1^2$. The testing variable is

$$F = \frac{\sum(Y_j - \mu_2)^2/n}{\sum(X_i - \mu_1)^2/m},$$

and the null hypothesis is rejected if $F \geq F_{\alpha,n,m}$.

Large Sample Tests for Binomial Distribution

If X has a BIN(n,p) distribution, then for n large, one can use either the Poisson approximation theorem or the central limit theorem to obtain testing procedures for hypotheses about p. Thus, if n is large, but np is small, then asymptotically $X \sim$

POI(np). The hypothesis $H_0 : p = p_0$ corresponds to $H_0 : \lambda = np_0$ in the Poisson distribution. We can use tests for this distribution.

If neither np nor $n(1-p)$ is small, then we have $X/n \sim \mathrm{N}(p, p(1-p)/n)$. To test the hypothesis $H_0 : p = p_0$ (against a one- or two-sided alternative), we use the fact that under H_0, the distribution of the random variable

$$Z = \frac{X/n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

is asymptotically standard normal.

For two sample situations, let $X \sim \mathrm{BIN}(n_1, p_1)$, $Y \sim \mathrm{BIN}(n_2, p_2)$, where n_1 and n_2 are both large, and assume that we can use normal approximation. Then we have

$$\frac{X}{n_1} - \frac{Y}{n_2} \sim \mathrm{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Thus under the null hypothesis $H_0 : p_1 = p_2$, we have, letting $p_1 = p_2 = p$,

$$\frac{X/n_1 - Y/n_2}{\sqrt{p(1-p)\,(1/n_1 + 1/n_2)}} \sim \mathrm{N}(0,1).$$

This statistic still involves the nuisance parameter p. However, under H_0 the MLE of p is $(X+Y)/(n_1+n_2)$, so we can use the test statistic

$$Z = \frac{\frac{X}{n_1} - \frac{Y}{n_2}}{\sqrt{\frac{X+Y}{n_1+n_2}\left(1 - \frac{X+Y}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

which is asymptotically standard normal. Whether we use the one- or two-sided test depends on the alternative hypothesis.

■ **EXAMPLE 13.37**

In primary elections, 28% of the Republicans in New Hampshire voted for candidate A. A poll of 180 Republicans in Iowa show that 41 of them will vote for candidate A. Does this result indicate that the Republican support of candidate A is lower in Iowa than in New Hampshire?

SOLUTION. We have here $p_0 = 0.28$, the known level of support for A in New Hampshire, and we want to test the hypothesis $H_0 : p \geq 0.28$ against the alternative $H_1 : p < 0.28$, where p is the fraction of Republican voters who support A in Iowa. The observed value of test statistic is

$$Z = \frac{41/180 - 0.28}{\sqrt{0.28 \times 0.72)/180}} = -1.58,$$

which corresponds to the p-value 0.059.

■ **EXAMPLE 13.38**

Continuing Example 13.37, of the 180 persons polled, 110 are women, and 20 of them of whom 21 said they would vote for A. Does this result indicate that support for A in Iowa is higher among men?

SOLUTION. We now want to test the hypothesis $p_W = p_M$ against the alternative $p_W < p_M$. The value of the test statistic is

$$z = \frac{\frac{21}{70} - \frac{20}{110}}{\sqrt{\frac{41}{180}\left(1 - \frac{41}{180}\right)\left(\frac{1}{110} + \frac{1}{70}\right)}} = 1.84.$$

The p-value here is about 0.033, and this may serve as an indication that indeed, the support for A among Republican women is lower than among Republican men.

13.10 MONTE CARLO, BOOTSTRAP, AND PERMUTATION TESTS

The tests introduced so far were based either on the known distribution of a test statistic or on its asymptotic form applicable when the sample size is large enough. There are, however, problems in which the distribution of the statistic is either unknown or difficult to handle theoretically, or the sample is not large enough to use the asymptotic properties of a test statistic. In all such cases one could apply powerful computer-intensive statistical procedures that recently became widely available with advances of computer technology.

■ **EXAMPLE 13.39**

Assume that $X_1, \ldots, X_n$ is a random sample from the Laplace distribution with a density $f(x; \mu, \lambda) = (\lambda/2)\exp\{-\lambda|x - \mu|\}$, and λ known. If we want to test the hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu > 0$ at the significance level α, then for the large sample size we could use the central limit theorem and reject H_0 if $\overline{x} > \sqrt{2}z_\alpha/\lambda$ (recall that $\mathrm{Var}(X) = 2/\lambda^2$). However, if the sample size is rather small (e.g., $n = 10$), the normal approximation of the distribution of $\overline{X}$ may not be appropriate, and other methods should be used instead.

Monte Carlo Tests

In Monte Carlo tests the critical values for given levels of significance, or conversely, the probabilities for specified thresholds, are estimated from generated samples based on the fact that relative frequencies can be used to estimate probabilities.

■ **EXAMPLE 13.40**

Continuing Example 13.39, if we are testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$ at the significance level $\alpha = 0.1$, then we need to find a critical value v_α such that $P(\overline{X} \geq v_\alpha | \mu = 0) = \alpha$. For that we generate many (at least

several thousand) samples of size $n = 10$ from the distribution with density $f(x; \mu = 0, \lambda) = (1/2)\lambda \exp\{-\lambda|x|\}$, and for each of them obtain its sample mean. Next, $\hat{v}_\alpha$ is obtained as the sample upper 0.1 quantile in the set of all means in generated samples.

In Examples 13.39 and 13.40 the sample size was small. Other situations where the use of Monte Carlo tests is recommended would be cases where distributions of statistics are difficult to track theoretically (e.g., when population distributions are mixed or contaminated) or the estimators of parameters of interest are difficult to handle theoretically (e.g., sampling distribution of the coefficient of skewness or kurtosis).

Bootstrap Tests

Bootstrap procedures are designed for the situations where the population distributions are unknown, and the actual samples provide the only available information. Consequently, the reference distribution for the statistic of interest is obtained from bootstrap samples. One may notice that bootstrap sampling is actually a Monte Carlo sampling from the distribution obtained from all n sample values, with probabilities $1/n$ assigned to each of them (the generation of bootstrap samples was described in Section 12.7). The p-values of the test are obtained as appropriate relative frequencies from the bootstrap distribution of the statistic being used.

■ **EXAMPLE 13.41**

Let us consider a problem in which hypothesis $H_0 : \mu = \mu_0$ is tested against $H_1 : \mu > \mu_0$. The test statistic that can be used here is $\overline{X}$, and its sampling distribution is imitated (as explained in Section 12.7) by the bootstrap distribution. However, we need to keep in mind that the p-values must be obtained as probabilities for the distribution determined by the null hypothesis. Since we do not know what distribution the random sample represents (with $\mu = \mu_0$ or rather $\mu > \mu_0$), before generating bootstrap samples we need to modify sample values by taking $w_i = x_i - \overline{x} + \mu_0$, where $\overline{x}$ is the mean obtained from the original sample. Then the p-value will be obtained as

$$p\text{-value} = \frac{\text{number of } \overline{w}_i^*\text{'s exceeding } \overline{x}}{B},$$

where $\overline{w}_i^*$ is the average in the ith bootstrap sample.

We can also use bootstrap tests to compare parameters in two populations. In such a case we usually have two independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. B bootstrap samples $X_1^*, \ldots, X_n^*$ and $Y_1^*, \ldots, Y_m^*$ are then generated by sampling with replacement from the same set that consists of combined $m + n$ original observations.

Permutation Tests

Permutation tests are similar to bootstrap tests as they also use only actual sample(s) to obtain a reference distribution based on which a p-value is determined. However, the sampling process is now different.

Practically all parametric tests have a corresponding permutation test that is based on the same test statistic as the parametric test. But the p-value is obtained from the sample-specific permutation distribution of that statistic rather than from the theoretical distribution derived from the parametric assumption.

We will focus here on a two-sample test that compares two population distributions; the H_0 states that both distributions (or their respective parameters) are equal. An important assumption behind a permutation test is that the observations are exchangeable under the null hypothesis. The respective samples $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ of size m and n can then be combined into one set of size $m + n$ which is then partitioned into two subsets of sizes m and n so that the value of the appropriate statistic can be obtained. This process is repeated for all possible $\binom{m+n}{m}$ partitions, eventually giving the reference distribution based on which a p-value will be determined as a measure of how "extreme" is the value obtained from the original samples.

■ **EXAMPLE 13.42**

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ be two particular samples selected randomly from distributions $f(x; \theta_0)$ and $f(y; \theta_1)$, respectively. We want to test hypothesis $H_0 : \theta_0 = \theta_1$ against the alternative $H_1 : \theta_0 < \theta_1$. Assume that the difference $\Delta = \theta_1 - \theta_0$ is estimated by some statistic $\hat{\Delta}$, and its large values provide evidence supporting the alternative hypothesis. To perform the permutation test, we first obtain $\hat{\Delta}_0 = \hat{\Delta}(\mathbf{x}, \mathbf{y})$. Then samples $\mathbf{x}$ and $\mathbf{y}$ are combined into a set $\mathbf{w} = (x_1, \ldots, x_n, y_1, \ldots, y_m)$, which is next partitioned $\binom{m+n}{m}$ times into two subsets of sizes m and n, respectively. For each partition the value of statistic $\hat{\Delta}$ is obtained and the p-value is determined as

$$p\text{-value} = \frac{\text{number of values of } \hat{\Delta} \text{ equal or exceeding } \hat{\Delta}_0}{\binom{m+n}{m}}.$$

If sample sizes m and n are such that the number of all possible partitions $\binom{m+n}{m}$ is very large, then a smaller number of partitions can be obtained in a random (rather than systematic) way. Such procedure is called a randomization test.

More information on procedures introduced in this section can be found, for example, in Good (2005). Additional examples of computer-intensive procedures presented in this section together with the R-code can be found on the book's FTP site

ftp://ftp.wiley.com/public/sc_tech_med/probability_statistical.

PROBLEMS

13.10.1 A random sample $0.38, -0.49, 0.03, 0.21, 0.12, 0.14, -0.18, -0.34, 0.46, -0.01$ was selected from a Laplace distribution with density $f(x, \mu) = \exp\{-2|x - \mu|\}$. Use Monte Carlo simulations to estimate the p-value for testing hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu > 0$.

13.10.2 A random sample $1.138, 1.103, 3.007, 1.307, 1.885, 1.153$ was obtained from a distribution with density $f(x, \theta) = 2\theta^2 x^{-3}$ for $x \geq \theta$ and 0 otherwise. Find the

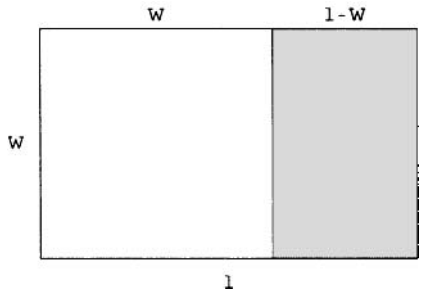


Figure 13.7 Golden rectangles

MLE of θ, and use it to test $H_0 : \theta = 1$ against $H_1 : \theta > 1$. Find the p-value and compare it with the p-value based on 2000 Monte Carlo generations.

13.10.3 Assume that the hypothesis $H_0 : \theta = 0$ is to be tested against $H_1 : \theta > 0$, where θ is a parameter in the $U[\theta, \theta + 1]$ distribution. **(i)** Derive the test of size $\alpha = 0.05$, based on one observation only, and obtain its power function for $\theta = 0.1, 0.2, \ldots, 0.9$. **(ii)** Use Monte Carlo simulations to obtain a test with a critical region $\overline{X}_3 > k$; that is, estimate k. Then again use Monte Carlo simulations to estimate the power function of this test at $\theta = 0.1, 0.2, \ldots, 0.9$. Use the same size $\alpha = 0.05$. **(iii)** Compare the power functions of both tests.

13.10.4 Use the bootstrap test for testing $H_0 : \mu = 2$ against $H_1 : \mu > 2$, based on a random sample: 3.49, 2.21, 1.07, 3.04, 2.57, 2.43, 2.18, 1.10, 1.04, 1.92, 0.99, 3.13, 0.92, 2.72, 4.03.

13.10.5 (Shoshoni Rectangles) The following problem is taken from Larsen and Marx (1986). Since antiquity, societies have expressed esthetic preferences for rectangles having a certain width (w) to length (l) ratio. For instance, Plato wrote that rectangles formed of two halves of an equilateral triangle are especially pleasing (for such rectangles $w/l = 1/\sqrt{3}$). Another standard, adopted by the Greeks, is the golden rectangle. It is defined by the condition that the ratio of its width to length must be the same to that of the part remaining after cutting off a square with the side equal to its width (i.e., the shaded area in Figure 13.7 is similar to the whole rectangle). Thus a golden rectangle must have $w/l = (l - w)/w$, which gives $w/l = \frac{1}{w/l} - 1$; hence $w/l = (\sqrt{5} - 1)/2 = 0.618$.

Both the Greeks and the Egyptians used golden rectangles in their architecture. Even today the golden rectangle remains an architectural and artistic standard (e.g., items such as drivers' licenses, business cards or picture frames often have w/l ratios close to 0.618).

The data below show width-to-length ratios of beaded rectangles used by Shoshoni Indians to decorate their leather goods. Use the bootstrap test to check whether the golden rectangle can be considered an esthetic standard for the Shoshonis.

Width-to-length ratios for 20 rectangles found on Shoshoni handicraft:

0.693	0.749	0.654	0.670	0.662	0.672	0.615	0.606	0.690	0.628
0.668	0.611	0.606	0.609	0.601	0.553	0.570	0.844	0.576	0.933

13.10.6 A study of pollution was carried out in two lakes, A and B. The level of a specific pollutant was measured using a certain instrument, and the results for lake A were 3.17, 4.22, 2.58, 4.01, and 3.79. In lake B the measurements (made with the same instrument) were 4.04, 4.32, and 4.12. Test the hypothesis that pollution levels are the same in both lakes, against the alternative that the level in lake B is higher using: **(i)** The bootstrap test. **(ii)** The permutation test. **(iii)** Compare both results.

13.10.7 Two random samples 4.49, 7.68, 5.97, 0.97, 6.88, 6.07, 3.08, 4.02, 3.83, 6.35, and 4.59, 3.39, 3.79, 6.89, 5.07, 7.41, 0.44, 2.47, 4.80, 7.23 were obtained independently from distributions with the same mean. Perform a permutation test to test the hypothesis that the variability in both populations is the same against the alternative that it is larger in the second population. As a test statistic use: **(i)** The difference of sample ranges. **(ii)** The ratio of sample variances. **(iii)** Compare both results.

CHAPTER 14

LINEAR MODELS

14.1 INTRODUCTION

This chapter is devoted to statistical problems arising in situations where the observed values (called "dependent variable," "response," etc.) are influenced by some other variable (referred to as "independent variable," "explanatory variable," "treatment," "factor," etc.). In this chapter we consider the case where only expected values are affected, while other characteristics (as well as the type of distribution) remain the same.

The theories that we present depend on the assumptions about the independent variable: Can its value be observed or not? If yes, can it be controlled by the experimenter? If no, can it be regarded as random? Depending on the answer to these questions, we have a regression analysis or different analysis of variance models (one-factor, two-factor, with or without interaction, with fixed or random effects, etc.).

Many models (theories) are available in the general setup considered in this chapter. The basic ideas and solutions we present cover a few of the most representative cases. We hope that the information we provide is enough information to motivate users of various statistical packages to try to identify conditions where a specific procedure is not the only one available. A thorough and exhaustive presentation of any of these theories can be found in any of numerous books devoted to regression

analysis or analysis of variance (e.g., see Montgomery et al., 2006; Stapleton, 1995; Myers and Milton, 1991; Myers, 1986).

In this chapter we outline some of the most common ways of analysis of data measured on a scale of at least interval type. We discuss the methods of detecting and measuring effects expressed through the mean of the observed random variable. However, we present only the most common of such methods, regression analysis and analysis of variance.

■ EXAMPLE 14.1

In a simple case of linear regression analysis we observe a random variable Y such that $Y = \alpha X + \beta + \xi$, where X is some variable (random or not, possibly controlled by the experimenter) and ξ is the "error," meaning a random variable with $E(\xi) = 0, \text{Var}(\xi) = \sigma^2 < \infty$. Both X and Y are assumed to be measured on an interval (or possibly even ratio) scale.

Specific examples may be obtained by taking Y to be the time to completion of some chemical reaction, and X being the temperature; Y may be some substance that accumulates linearly (up to random fluctuations) in human bones throughout life, and X may be the age of the person at death, and so on.

Typically we have data in the form of pairs (X_i, Y_i), $i = 1, \ldots, n$, possibly with some X_i's repeating. The problems are to estimate α, β, and σ^2 (or to test hypotheses about these parameters) and predict the value of Y to be observed for some X_0, or to estimate X corresponding to some observed value of Y.

Obvious generalizations involve a model with more than one variable X, for instance, when $Y = \alpha_1 X_1 + \alpha_2 X_2 + \beta + \xi$, where X_1 and X_2 are some variables. We may have nonlinear models, such as $Y = \alpha X^2 + \beta X + \gamma + \xi$ (quadratic regression), etc.

■ EXAMPLE 14.2 Analysis of Variance, or ANOVA

Analysis of variance applies to situations where variables X_i are nominal. We typically speak of "factors" that operate on some levels.[30] For instance, we might have data (measurements of some "response") taken from populations classified according to some criteria such as sex (male, female) and smoking status (never smoked, former smoker, current smoker). We want to find out whether any of these factors (sex and smoking status) has an effect on Y (e.g., response to some drugs).

The data can be summarized in a table:

		1 Never Smoked	2 Former Smoker	3 Current Smoker
1	Male	Y_{11}	Y_{12}	Y_{13}
2	Female	Y_{21}	Y_{22}	Y_{23}

[30]The use of the word "level" in the context of ANOVA does not imply any specific ordering of the levels of factors such as "sex," F (female) and M (male).

Here Y_{ij} is the response of a subject from group (i, j), where we have one observation in each cell. The model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \xi_{ij},$$

for which we can assume $\alpha_1 + \alpha_2 = 0, \beta_1 + \beta_2 + \beta_3 = 0$, and ξ_{ij} to be the error term satisfying $\xi_{ij} \sim \mathrm{N}(0, \sigma^2)$. We may wish to test the hypotheses $H_0 : \alpha_1 = \alpha_2 = 0$ and $H_0' : \beta_1 = \beta_2 = \beta_3 = 0$ (no sex effect, and no smoking effect) against the alternatives $H_1 : \alpha_1 \neq \alpha_2$ and $H_1' : \beta_k \neq \beta_l$ for some $k, l = 1, 2, 3, k \neq l$.

The scheme of Example 14.2 can be modified in a number of ways, of which we will discuss some. The main issue we note here is that comparing the results within each pair separately (until either a pair of "significantly different" results is found or until all pairs are checked and found not to differ significantly) is *not a correct method.* Indeed, for many pairs the probability of finding a pair with a large difference becomes quite likely because of chance fluctuations, even where the null hypothesis is in fact true. The correct method requires testing all pairs *at once*, which is accomplished by the analysis of variance methodology.

14.2 REGRESSION OF THE FIRST AND SECOND KIND

Let Y denote a dependent variable (assumed to be one-dimensional) and X (or $\mathbf{X}$) denote an independent variable. In most cases X will be one-dimensional, but the considerations can usually be extended to m-dimensional vectors $\mathbf{X} = (X_1, \ldots, X_m)$, where $m > 1$.

A few comments about the nature of $\mathbf{X}$ and some illustrative examples are in order here. We start with the case where the values of the independent variable (or variables) can be observed. Depending on how the values of $\mathbf{X}$ are chosen, $\mathbf{X}$ may be random, either one- or multidimensional, nonrandom, or even under the experimenter's control. Typical cases of such situations are exemplified below.

■ EXAMPLE 14.3

In some genetic theories one studies a characteristic Y of offspring, such as height, which depends on characteristics $\mathbf{X} = (X_1, X_2)$ of a father and a mother (X_1, X_2 may be heights, but it is also possible to study some other features affecting the offspring's height Y).

■ EXAMPLE 14.4

Sometimes the randomness of Y for a given $X = x$ has to be postulated, since it is not possible to observe more than one Y for the same x. This happens, for instance, if x is the calendar time, and we observe the stock market index Y_x at the end of the day x. Here we treat Y_x as random, since we cannot predict its value with complete precision. We only have the conviction that "if such and such events had occurred, the value Y_x would be different from that being

observed." But for each x only one single value is recorded, and we have no empirical access to the distribution unless we make some assumptions about the nature of randomness of Y_x across different times x.

■ **EXAMPLE 14.5**

It may happen that X is not random. For instance (in the case of one dimension), we may study the relation between some developmental characteristic Y of a child, say height or size of vocabulary, as dependent on age X. The randomness concerns values of Y for the same $X = x$, since in the population of children of age x there is some variability of values of Y.

■ **EXAMPLE 14.6**

It may even happen that the values of X are totally under the experimenter's control. For instance, a chemist may be interested in some characteristic Y of a chemical reaction (e.g., its duration) depending on temperature X. Then X can be determined arbitrarily by the experimenter, and for a given temperature $X = x$, the randomness of Y may be due to measurement error, or to some other factors.

In each of the cases under consideration the data have the form of a set of pairs (x_i, y_i), $i = 1, 2, \ldots n$, where values of x_i can repeat. Even if the x_i's are not random, the formulas are identical with those obtained under the assumption that X has discrete uniform distribution over the set $\{x_1, \ldots, x_n\}$, with probabilities appropriately increased in the case of repeated values. Thus we will proceed as if X (respectively, $\mathbf{X}$) were a random variable. We will assume throughout this chapter that X and Y have finite variances, hence also finite expectations.

Let us begin by recalling some facts from Chapter 8. Suppose that we know the value of X, $X = x$, and we want to predict the value of Y. The best prediction of Y (in the sense of mean square error) is given by the conditional expectation

$$u(x) = E(Y|X = x). \tag{14.1}$$

In other words, we have

$$\min_{\xi} E[(Y - \xi)^2|X = x] = E[(Y - u(x))^2|X = x].$$

Accordingly we introduce the following definition:

Definition 14.2.1 The conditional expectation of Y given $X = x$, that is, the function $u(x)$ given by (14.1), will be called the *true regression*, or regression of the *first kind*, of Y on X. □

In the case of a continuous random variable X, it may happen that the regression function $u(x)$ is defined only for almost all x, that is, on a set A such that $P\{X \in A\} = 1$. However, in all cases that are encountered in statistics, there is typically a function $u(x)$ that is "natural" in a given problem. We will select such a function and comment on the nonuniqueness of a regression function only when such a comment is essential for the problem.

The usefulness of the regression function $u(x)$ is not restricted to the prediction of Y. To mention just one other use of it, suppose that we do not know how X and Y are related; we only formulated the hypothesis that the true regression function is equal to some function u_0. In other words, the null hypothesis is $H_0 : E(Y|X = x) = u_0(x)$ for all x. If we know something about the conditional distribution of Y given X, we may test this hypothesis, rejecting it if the value of Y observed for given $X = x$ is "far" from $u_0(x)$.

Finding the true regression $u(x)$ requires the knowledge of the joint distribution of (X, Y) or, more precisely, the conditional distribution of Y given $X = x$. Such knowledge can come from a sufficiently deep understanding of the stochastic mechanisms that connect X and Y, or from very extensive data on pairs (X_i, Y_i). There are many practical cases when neither is available. Moreover, even if we do know the joint distribution of (X, Y), determining $u(x)$ may present formidable difficulties.

To cover such situations, another type of regression has been introduced. Starting again from problem of prediction, suppose that we want to find the best predictor of Y that is *linear* in X. It will have the form $Y_p = a + bX$, and we will have to find such a and b that Y_p is the best predictor of Y–in the sense of minimizing the mean square error $E(Y - Y_p)^2$. We have here

$$\begin{aligned} E(Y - Y_p)^2 &= E[Y - (a + bX)]^2 \\ &= E(Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX) \\ &= E(Y^2) + a^2 + b^2E(X^2) - 2aE(Y) - 2bE(XY) + 2abE(X). \end{aligned}$$

By differentiating with respect to a and b and setting the derivatives to be 0, we obtain the *normal equations*:

$$\begin{aligned} a + bE(X) &= E(Y), \\ aE(X) + bE(X^2) &= E(XY). \end{aligned}$$

Multiplying the first equation by $E(X)$ and subtracting it from the second, we obtain

$$b = \frac{E(XY) - E(X)E(Y)}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}, \tag{14.2}$$

and consequently

$$a = E(Y) - \rho\frac{\sigma_Y}{\sigma_X}E(X). \tag{14.3}$$

Thus coefficients a and b require only means, variances, and covariances of X, Y. We now introduce the following definition:

Definition 14.2.2 The relation

$$Y = a + bX = E(Y) + \rho\frac{\sigma_Y}{\sigma_X}[X - E(X)] \tag{14.4}$$

is called the *linear* regression, or regression of the *second kind*, of Y on X. □

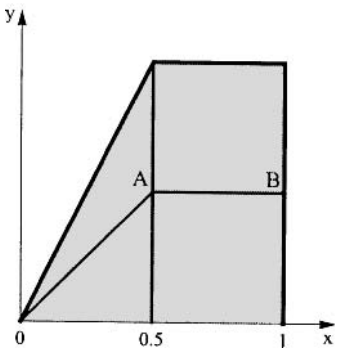


Figure 14.1 True regression

■ EXAMPLE 14.7

Let us consider the situation where the true regression is the function

$$u(x) = \begin{cases} x & \text{for } 0 \le x \le \frac{1}{2} \\ \frac{1}{2} & \text{for } \frac{1}{2} \le x \le 1. \end{cases}$$

Such a regression can occur for various conditional distributions of Y. To fix the idea, suppose that given x, the random variable Y is distributed uniformly on an interval whose leftmost point is 0. Thus, for $0 \le x \le \frac{1}{2}$, random variable Y_x is uniform on $[0, 2x]$, whereas for $\frac{1}{2} \le x \le 1$, random variable Y_x is uniform on $[0, 1]$. The situation is presented in Figure 14.1. The shaded area shows all possible points (X, Y). The polygonal line OAB is the graph of the true regression $u(x)$. For use in the next example, note that we have here

$$\mathrm{Var}(Y|X = x) = \begin{cases} \frac{x^2}{3} & \text{for } 0 \le x \le \frac{1}{2} \\ \frac{1}{12} & \text{for } \frac{1}{2} \le x \le 1. \end{cases} \tag{14.5}$$

■ EXAMPLE 14.8

In Example 14.7 the distribution of X did not play a role in determining regression. In fact X could have been nonrandom. On the other hand, to determine the linear regression of Y on X, it is necessary to know the marginal distribution of X. The intuitive justification here is that in linear regression the objective is to find the best approximation of Y by a straight line. Thus it matters which values of X occur more often and which occur less often. Suppose that the joint distribution of the vector (X, Y) is uniform on the shaded area in Figure 14.2. To determine the linear regression, we need the expectations and variances of X and Y as well as their covariance. We have here, using simple

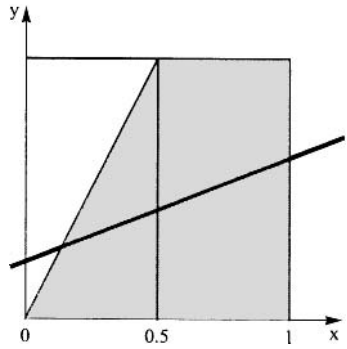


Figure 14.2 Linear regression for uniform distribution of (X, Y)

geometry, for the marginals of X and Y:

$$f_X(x) = \begin{cases} cx & \text{for } 0 \le x \le \frac{1}{2} \\ \frac{c}{2} & \text{for } \frac{1}{2} \le x \le 1. \end{cases} \tag{14.6}$$

The value of c is determined from the condition

$$1 = \int_0^1 f_X(x)dx = \frac{3}{8}c;$$

hence $c = 8/3$. Consequently

$$E(X) = \int_0^1 x f_x(x)dx = \frac{11}{18}.$$

After simple integration we obtain $E(X^2) = 31/72$, hence $\mathrm{Var}(X) = 37/648$. Similarly, $f_Y(y) = (4/3)(1 - y/2)$ for $0 \le y \le 1$, and consequently $E(Y) = 4/9$, and $\mathrm{Var}(Y) = 13/162$.

Finally, the joint density $f(x, y)$ is constant, equal to $4/3$ on the shaded area in Figure 14.2. Thus $E(XY) = 7/24$, and using (14.4), we obtain the equation for linear regression (see Figure 14.2):

$$Y = \frac{X}{4} + \frac{7}{24}.$$

It is of some interest to compare the average square error of prediction of Y on the basis of X if we use true and linear regressions. For prediction under true regression, we find the average square error, using (14.5) and (14.6):

$$\begin{aligned} E_X[\mathrm{Var}(Y|X)] &= \int_0^1 \mathrm{Var}(Y|X = x) f_X(x)dx \\ &= \int_0^{1/2} \frac{x^2}{3} \times \frac{8}{3}x\,dx + \int_{1/2}^1 \frac{1}{12} \times \frac{4}{3}x\,dx = \frac{5}{72} = 0.0694. \end{aligned}$$

For prediction with the use of linear regression, we have the error

$$E\left(Y - \frac{X}{4} - \frac{7}{24}\right)^2 = E(Y^2) + \frac{1}{16}E(X^2) + \left(\frac{7}{24}\right)^2 - \frac{1}{2}E(XY) - \frac{7}{12}E(Y) + \frac{1}{48}E(X) = 0.0738.$$

Thus, in this case, the mean square error of the linear prediction is about 5.5% higher than the corresponding error of prediction based on the true regression. Whether a 5.5% difference is important or negligible depends on the context. The point is that the mean square error of a linear prediction is *always* at least as large as the error of prediction based on a true regression. Also one can easily construct examples where the ratio of these two errors is as large as one wishes. In the extreme case, if $Y = g(X)$, where g is a deterministic nonlinear function, g is the true regression and there is no error involved in a prediction of Y using $g(X)$ as the predictor. But there is an error involved if a nonlinear function g is approximated by a linear function $a + bX$.

Example 14.8 shows that the issue of prediction errors in comparison of two types of regression is only partially statistical. One component is due to the randomness of Y for a given $X = x$, and the other is due to replacing the true regression $u(x)$ by a straight line.

The situation may be improved by considering a regression of the second order, where the predictor is of the form $Y_p = a + bX + cX^2$, or a regression of some other special form, and as $Y_p = a + b\sin(cX)$, and so on. The point is that none of these regressions can give the mean square error lower than that based on true regression.

PROBLEMS

14.2.1 Let X and Y have a joint distribution uniform on a parallelogram with vertices at points $(-1, -1)$, $(0, -1)$, $(1, 1)$ and $(0, 1)$. Find the true and linear regression of: bf(i) X on Y. **(ii)** Y on X.

14.2.2 Find an example of random variables (X, Y) such that their true and linear regressions of Y on X coincide, but (X, Y) do not have a joint normal distribution.

14.2.3 Suppose that the true regression of Y on X is not linear in X. Is it possible that the marginal distribution of X is such that the expected square error of the best linear predictor of Y is the same as the expected square error of predictor based on the true regression u?

14.2.4 Let (x_i, y_j), $i = 1, \ldots, n$ be the data where at least one value of x_i is not zero. Find an estimate of the slope parameter b if it is known that $a = 0$ [i.e., find the best fit of the model $E(Y|x) = bx$].

14.2.5 Suppose that $\mathbf{X} = (X_1, X_2)$. Find coefficients a, b_1, b_2 such that $Y_p = a + b_1X_1 + b_2X_2$ is the best linear predictor of Y given $\mathbf{X}$.

14.2.6 Find true regression Y on X if X and Y have a joint trinomial distribution

$$P\{X = x, Y = y\} = \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y (1 - p_1 - p_2)^{n-x-y}$$

for $x = 0, 1 \ldots n, y = 0, 1, \ldots, n$, and $0 \leq x + y \leq n$.

14.2.7 The number of eggs in nests of a certain species of birds is one, two, or three, with two eggs found in about 80% of nests, and one or three eggs in about 10% of nests each. In one-egg nests, the egg hatches successfully in 75% of cases. In two-egg nests the probabilities for the number of offspring are 0–20%, 1–30%, 2–50%, whereas in three-egg nests, these probabilities are 0–10%, 1–20%, 2–60%, 3–10%. Find the best linear predictor of Y being the number of offspring based on the observation of X—the number of eggs in the nest.

14.3 DISTRIBUTIONAL ASSUMPTIONS

In addition to the classification of regression models with respect to the two types of regression, a meaningful and useful classification is obtained when one considers typical assumptions about the distribution of the response Y for a given value of X.

The oldest methods that can be included in regression theory dates back to the beginning of the nineteenth century and the works of Legendre. He found a method of best approximation of a set of points by a straight line. The word "best" was understood in the sense of least squares. In other words, Legendre found, for a set of data points (x_i, y_i), the coefficients a and b such that the sum $\sum_i [y_i - (a + bx_i)]^2$ attained its minimum. In this formulation no assumptions about randomness are needed: the best-fitting line always exists, even if the "best" fit does not mean a "good" fit. The extension to the best linear fit in two (or k) dimensions—finding a_1, a_2, b to minimize the sum $\sum [y_i - (a + b_1 x_i^{(1)} + b_2 x_i^{(2)})]^2$—is now straightforward. Similarly the theory extends naturally to other forms of relations, such as quadratic or periodic functions obtained by minimizing the sums $\sum_i [y_i - (a + bx_i + cx_i^2)]^2$ or $\sum_i [y_i - a\cos(bx_i + c)]^2$.

Viewed in this way, regression theory belongs properly to the domain of numerical analysis. However, to allow statistical inference, one usually makes some assumptions about the randomness inherent in the model. One of the standard assumptions is that for every x we have $Y = \varphi(x) + \epsilon$, where φ is some deterministic function and ϵ is the "error" random variable, such that:

(i) $E(\epsilon) = 0$.
(ii) $\mathrm{Var}(\epsilon) = \sigma^2 > 0$.
(iii) The errors $\epsilon, \epsilon', \epsilon'', \ldots$, corresponding to different observations of Y (for the same, or for distinct values of x are uncorrelated.[31]

Under these assumptions, $\varphi(x)$ is the true regression of Y on X. If one now imposes some parametric model on φ, one can set up the least square equations for parameters of φ and the parameter σ^2.

[31] The assumption that variance is independent of x is often called *homoscedasticity*, as opposed to *heteroscedasticity*.

If the assumptions (i)–(ii) are replaced by

(iv) The errors $\epsilon, \epsilon', \ldots,$ have a N(0,σ^2) distribution,

then the errors are independent. The independence of errors and knowledge of the distribution, have profound theoretical consequences: we can write the likelihood of the data and find the maximum likelihood estimators of the parameters.

One can easily see that in this case the MLE's coincide with least square estimators, so it might appear that nothing is gained by replacing (i)–(iii) with (iii) and (iv). However, under (iv) we know the sampling distribution of the estimators, and we are therefore able to use the results from estimation and testing theory to build confidence intervals, tests of various hypotheses, and so on.

Finally, let us mention, that if in addition to (i), (ii), and (iv) we can assume that

(v) $\varphi(x) = a + bx$,

(vi) X has a normal distribution,

then the joint distribution of (X, Y) is bivariate normal. The converse is also true, as shown in Theorem 9.5.4. We may summarize these facts as the following:

Theorem 14.3.1 *In case of bivariate normal distribution (X, Y), the true and linear regression coincide.*

A comment appears necessary here in regards to assumptions (iii) and (iv), and their implementation in practical situations. To illustrate potential difficulties, suppose that we collect data on a regression where x is the age of a child, and Y is some response, such as reaction time to a specific stimulus. Suppose that we need two observations for the same age x. Then it is not correct to measure the reaction time Y twice for the same child of age x, even if such observations can be regarded as independent. The correct procedure is to take two children of the same age x and observe their reaction times. The reason is that variability of Y has two components: the between-children variability and the within-child variability. Taking two measurements for the same child will involve only the second component, whereas other observations will involve both components, violating (among others) the assumption of homoscedasticity.

Assumptions (i)–(vi) are by no means the only sensible assumptions one can make in regression analysis. The linearity assumption $E(Y_x) = a + bx$ is often acceptable as an approximation or in some narrow range. A typical case of this kind is illustrated by the following example:

■ **EXAMPLE 14.9 Logistic Regression**

Assume that the response variable Y_x is of the binary character (e.g., success or failure). We can always take the possible values of Y_x as 0 and 1 so that $E(Y_x) = P\{Y_x = 1\} = \pi(x)$. We will assume that x is a numerical variable (random or not, depending on the situation under study), and that $\pi(x)$ is the true regression of Y on x. The assumption $\pi(x) = a + bx$ can be realistic only in a narrow range of values of x, since we must have $0 \leq \pi(x) \leq 1$ for all x. Still the inference about the shape of function $\pi(x)$ is of considerable interest, and linearity is a powerful assumption (in the sense of allowing many analytical results). To realistically utilize such an assumption, we consider the odds ratio $\pi(x)/[1 - \pi(x)]$, with range $(0, \infty)$, and its logarithm, with range

$(-\infty, \infty)$. Thus we consider the model

$$\log \frac{\pi(x)}{1-\pi(x)} = a + bx,$$

or equivalently,

$$\pi(x) = \frac{e^{a+bx}}{1+e^{a+bx}}. \tag{14.7}$$

The right-hand side of (14.7) is called the *logistic function* (and is a cdf of a logistic distribution if $b > 0$), which explains the name logistic regression. The problem now is to estimate parameters a and b given the data $(x_i, n_i, N_i), i = 1, \ldots, m$, where x_i's are the values of the independent variable at which observations are taken, while n_i and N_i are the number of successes and the number of trials at value x_i, respectively. The likelihood is therefore

$$\begin{aligned} L &= \prod_{i=1}^{m} \binom{N_i}{n_i} [\pi(x_i)]^{n_i} [1-\pi(x_i)]^{N_i - n_i} \\ &= \left[\prod_{i=1}^{m} \binom{N_i}{n_i}\right] \frac{e^{a\sum n_i + b\sum x_i n_i}}{\prod_{i=1}^{m}(1+e^{a+bx_i})^{N_i}}. \end{aligned}$$

Determining a and b that maximize L requires using numerical iterations.

PROBLEMS

14.3.1 The random variable Y_x has the density

$$f(y|x) = \begin{cases} 1 & \text{for } a+bx-\frac{1}{2} < y < a+bx+\frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Find the MLE's of a and b given the sample: $(-1, 1.3), (0, 1.4), (1, 0.1), (2, -0.4), (3, -0.1$ Compare the MLE's with the LS estimates.

14.3.2 Let (X, Y) have the distribution uniform on the quadrangle with vertices $(0,0), (1,1), (0,2)$, and $(0.5, 1)$. Find: **(i)** The true regression of Y on X and of X on Y. **(ii)** $\mathrm{Var}(X|Y)$ and $\mathrm{Var}(Y|X)$.

14.3.3 Assume that the total number X of eggs laid by a bird has a Poisson distribution with parameter λ, and that each egg hatches with probability p, independent of other eggs. Find the regression of the number Y of eggs hatched on X. Let X_i, Y_i be the number of eggs laid and hatched by the ith bird. Derive the likelihood of the data for n birds, and the equation for estimating λ and p. Find the MLE's of λ and p.

14.3.4 Answer Problem 14.3.3 assuming a more realistic model, where X_i's have a Poisson distribution conditional on a positive value, that is, for $k = 1, 2, \ldots,$

$$P\{X_i = k\} = \frac{(\lambda^k/k!)e^{-\lambda}}{1-e^{-\lambda}}.$$

14.3.5 A still more realistic model is that when all eggs are laid, a disaster occurs, and with probability α all eggs are destroyed. If there is no disaster, eggs hatch independently, each with probability p. Find the regression $E(Y|X)$, the likelihood of the data, and (where possible) the estimators of λ, α, and p under assumptions of Problems 14.3.3 (Poisson distribution of X) and 14.3.4 (conditional Poisson). State carefully the assumptions you make.

14.4 LINEAR REGRESSION IN THE NORMAL CASE

In this section we will present the main results in the case where for any given x, the response Y is of the form

$$Y = a + bx + \epsilon$$

for $\epsilon \sim \mathrm{N}(0, \sigma^2)$. Moreover, we will assume that the errors ϵ for different observations of Y (for the same as well as for a distinct x) are independent.

We will assume that the data have the form

$$(x_i, y_i), \qquad i = 1, \ldots, n, \tag{14.8}$$

where $y_1, \ldots, y_n$ are independent observations. We will use the notation

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and assume that not all x_i's are identical.

The estimates of parameters a, b, and σ^2 are easy to obtain. The likelihood function of the sample (14.8) is

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i - a - bx_i)^2/2\sigma^2} \\ &= \sigma^{-n}(2\pi)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \right\}; \end{aligned}$$

hence

$$\log L = C - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

Differentiating with respect to a and b, we obtain a pair of equations (in which σ^2 cancels out) that are the same as the equations for the least squares estimates of a and b. The solutions can be obtained from formulas (14.2) and (14.3) by treating X as a random variable uniformly distributed on $x_1, \ldots, x_n$. Thus

$$\hat{b} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \qquad \text{and} \qquad \hat{a} = \overline{y} - \hat{b}\overline{x}. \tag{14.9}$$

The denominator in the expression for $\hat{b}$ is not zero in view of the assumption that not all x_i's are equal.

Differentiating $\log L$ with respect to σ^2 and setting the derivative equal to 0, we obtain the MLE of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2. \tag{14.10}$$

To determine the sampling distribution of estimators (under variability of y_j's only, for fixed x_j's), let us regard y_j as a value of the random variable Y_{x_j}. Then the estimators of b, a, and σ^2 are, respectively,

$$\begin{aligned} U &= \frac{\sum(x_i - \overline{x})(Y_{x_i} - \overline{Y})}{\sum(x_i - \overline{x})^2} \\ V &= \overline{Y} - U\overline{x} \\ T^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_{x_i} - V - Ux_i)^2. \end{aligned} \tag{14.11}$$

Notice that the randomness of U, V, and T^2 is related only to the variability of the dependent variable Y_x about its mean $a + bx$. The values x_i may be nonrandom, and even if they are arising from the sampling, U, V, and T^2 involve only a conditional distribution of Y given $X = x$.

It follows from (14.11) that U, as a linear combination of normally distributed random variables, has itself a normal distribution. To determine the mean, observe that

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}E(Y_{x_i}) = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + b\overline{x}.$$

Consequently we have

$$E(U) = \frac{\sum(x_i - \overline{x})E(Y_{x_i} - \overline{Y})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})[a + bx_i - (a + b\overline{x})]}{\sum(x_i - \overline{x})^2} = b.$$

To find the variance of U, we write

$$U = \frac{\sum(x_i - \overline{x})Y_{x_i} - \overline{Y}\sum(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})Y_{x_i}}{\sum(x_i - \overline{x})^2}.$$

Consequently, by assumption of homoscedasticity,

$$\sigma_U^2 = \mathrm{Var}\left\{\frac{\sum(x_i - \overline{x})Y_{x_i}}{\sum(x_i - \overline{x})^2}\right\} = \frac{\sum(x_i - \overline{x})^2\mathrm{Var}(Y_{x_i})}{[\sum(x_i - \overline{x})^2]^2} = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}. \tag{14.12}$$

Next the estimator V given by (14.11) is also a linear combination of the normal random variables, so it has a normal distribution. Here

$$E(V) = E(\overline{Y}) - E(U)\overline{x} = a + b\overline{x} - b\overline{x} = a$$

so that U and V are unbiased estimators of b and a.

The derivation of the variance of estimator V is somewhat messy. We have

$$\begin{aligned}\sigma_V^2 &= \mathrm{Var}(\overline{Y}) + \overline{x}^2\,\mathrm{Var}(U) - 2\overline{x}\,\mathrm{Cov}(U,\overline{Y}) \\ &= \mathrm{Var}(\frac{1}{n}\sum Y_{x_i}) + \overline{x}^2\sigma_U^2 - 2\overline{x}\,\mathrm{Cov}\left[\frac{\sum(x_i-\overline{x})Y_{x_i}}{\sum(x_i-\overline{x})^2}, \frac{1}{n}\sum Y_{x_j}\right] \\ &= \frac{\sigma^2}{n} + \overline{x}^2\frac{\sigma^2}{\sum(x_i-\overline{x})^2} - \frac{2\overline{x}}{n\sum(x_i-\overline{x})^2}\,\mathrm{Cov}\left[\sum(x_i-\overline{x})Y_{x_i}, \sum Y_{x_j}\right].\end{aligned}$$

In the last term, all covariances corresponding to $i \neq j$ vanish while all others are equal σ^2. Consequently

$$\mathrm{Cov}\left[\sum(x_i-\overline{x})Y_{x_i}, \sum Y_{x_j}\right] = \sum(x_i-\overline{x})\sigma^2 = 0.$$

After some algebra we obtain

$$\sigma_V^2 = \sigma^2\frac{\sum x_i^2}{\sum(x_i-\overline{x})^2}. \tag{14.13}$$

In a similar manner, we can show that

$$\mathrm{Cov}(U,V) = -\sigma^2\frac{\overline{x}}{\sum(x_i-\overline{x})^2} \tag{14.14}$$

(we leave the proof as an exercise).

These results suggest that (U, V) has a bivariate normal distribution. In fact the following theorem holds:

Theorem 14.4.1 *The estimators (V, U, T^2) are jointly sufficient for the parameter $\theta = (a, b, \sigma^2)$. Moreover, (U, V) have a bivariate normal distribution with means, variances, and covariance given by* (14.12) through (14.14). *Finally, T^2 is independent of (U, V), and nT^2/σ^2 has a chi-square distribution with $n-2$ degrees of freedom.*

Proof. We can use the joint moment generating function

$$m(t_1,t_2,t_3) = E(e^{t_1U+t_2V+t_3T^2}).$$

Substituting the expression for U, V and T^2, and integrating with respect to the joint density of $(Y_1, \ldots, Y_n)$, we can show (after considerable algebra) that $m(t_1, t_2, t_3) = m_1(t_1, t_2) \times m_2(t_3)$, which proves the independence of (U, V) and T^2. The form of functions m_1 and m_2 will then show that claims about the distributions of (U, V) and T^2 are also valid. We omit the details. □

We will now illustrate the applications of Theorem 14.4.1 with some examples on the construction of confidence intervals and tests for parameters of the regression model.

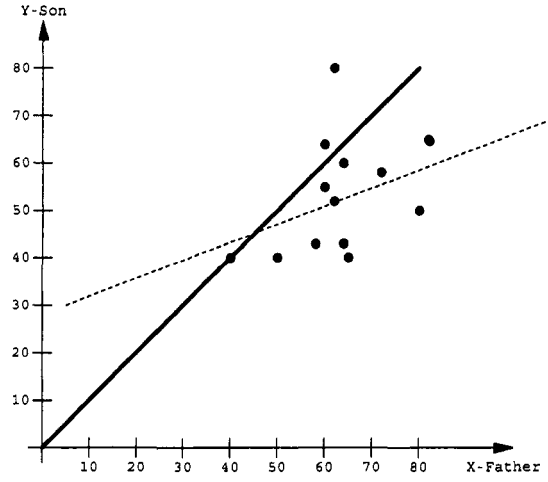


Figure 14.3 Ages of Polish kings and their heirs at death

■ EXAMPLE 14.10 Is It Good to Be a Royal Prince?

Poland had altogether 13 kings whose fathers were also Polish kings (starting from the fifteenth century, Polish kings were elected, and the election of a late king's son, although often likely, was by no means ensured). In one pair, the son died young in battle; deaths in the remaining 12 pairs came from natural causes. The ages at death of fathers (x_i) and sons (y_i) are listed below.

Father	x_i	Son	y_i
Mieszko I	62	Bolesław the Brave	59
Bolesław the Brave	59	Mieszko II	44
Casimir I the Restorer	42	Bolesław the Bold	42
Władysław I the Short	73	Casimir III the Great	60
Władysław II Jagiełło	83	Casimir IV Jagiellonian	65
Casimir IV Jagiellonian	65	John I Albert	42
Casimir IV Jagiellonian	65	Alexander Jagiellonian	45
Casimir IV Jagiellonian	65	Sigismund the Old	81
Sigismund I the Old	81	Sigismund II Augustus	52
Sigismund III Vasa	66	Władysław IV Vasa	53
Sigismund III Vasa	66	John II Casimir Vasa	63
Augustus II the Strong	63	Augustus III	67

As can be seen, only in 2 out of 12 pairs did the son live longer than his father. Can this be attributed to chance, or does it indicate some systematic trend (e.g., a "pampered" heir to the crown is not capable of dealing with the stress and additional duties demanded of him upon his accession to the crown). We have here

$$\overline{x} = 65.83, \qquad \overline{y} = 56.08$$

which means that on average, sons lived about 10 fewer years than their fathers. For the regression coefficients, and their variances and covariances, we obtain

$$\hat{a} = 29.214, \qquad \hat{b} = 0.408,$$

and also $\sum x_i^2 = 53,224, \sum(x_i - \overline{x})^2 = 1215.67$. Finally, $T^2 = 115.034$. The individual points, as well as the regression line $y = 29.214 + 0.408x$, are presented in Figure 14.3 (broken line). We will return to the analysis in Example 14.11.

We begin by constructing confidence intervals for the regression intercept a and slope b. It is known that $(U - b)/\sigma_U$ has a standard normal distribution. This quantity, however, involves the nuisance parameter σ. Since nT^2/σ^2 is independent of U and has a chi-square distribution with $n - 2$ degrees of freedom, the ratio

$$t = \frac{(U - b)/\sigma_U}{\sqrt{nT^2/[\sigma^2(n-2)]}} = \frac{(U - b)\sqrt{(n-2)\sum(x_i - \overline{x})^2}}{\sqrt{\sum(Y_{x_i} - Ux_i - V)^2}} \tag{14.15}$$

has a Student's t distribution with $n-2$ degrees of freedom. Consequently, (14.15) is a pivotal quantity for b (see Definition 12.7.2). Letting $\hat{a}$ and $\hat{b}$ denote the observed values of estimators V and U of intercept and slope, we obtain the $(1 - \gamma)$-level confidence interval for slope b as

$$\hat{b} \pm \frac{t_{\gamma/2,n-2}}{\sqrt{(n-2)\sum(x_i - \overline{x})^2}} \sqrt{\sum(y_j - \hat{a} - \hat{b}x_j)^2}.$$

In a similar way we can derive the $(1 - \gamma)$-level confidence interval for the intercept a in regression line:

$$\hat{a} \pm t_{\gamma/2,n-2} \sqrt{\frac{\sum x_i^2}{(n-2)\sum(x_i - \overline{x})^2}} \sqrt{\sum(y_i - \hat{a} - \hat{b}x_i)^2}.$$

■ **EXAMPLE 14.11**

To continue Example 14.10, the 95% confidence intervals for the regression slope and the intercept are now, respectively,

$$[0.408 - 0.751; 0.408 + 0.751] = [-0.343, 1.159]$$

and

$$[29.214 - 50.000; 29.214 + 50.000] = [-20.786; 79.214].$$

Since the confidence interval for regression slope b covers the value 0, we cannot exclude the possibility that the true regression is a constant a. This means that the age of the son at death does not depend on the father's age at death. Thus the evidence is not conclusive with regard to the effect of being born to a royal family on duration of life.

Obviously this example should not be taken too seriously. There are many factors affecting the length of one's life that should also be taken into account in a real study.

We can even construct a simultaneous confidence set for both regression parameters using the F distribution. We want to show (we omit the proof) that the random variable

$$Q = \frac{1}{\sigma^2}[n(U-b) + 2n\overline{x}(U-b)(V-a) + \sum x_i^2(V-a)^2]$$

has the chi-square distribution with two degrees of freedom. Consequently, the random variable

$$\frac{Q/2}{nT^2/(n-2)} = \frac{n-2}{2} \times \frac{[n(U-b) + 2n\overline{x}(U-b)(V-a) + \sum x_i^2(V-a)^2]}{\sum(Y_{x_i} - Ux_i - V)^2}$$

has the F distribution with $(2, n-2)$ degrees of freedom. Thus, the ellipsoid in the (a, b)-plane,

$$n(\hat{b}-b)^2 + 2n\overline{x}(\hat{b}-b)(\hat{a}-a) + \sum x_i^2(\hat{a}-a)^2 \leq \frac{2}{n-2}F_{\gamma,2,n-2}\sum(y_i - \hat{a} - \hat{b}x_i)^2,$$

is a $(1-\gamma)$-level confidence set for the regression parameters.
Notice that the results obtained thus far can be used to build estimators and to construct tests of hypotheses about linear combinations of the regression coefficients a and b. We will illustrate the situation with an example.

■ **EXAMPLE 14.12**

Suppose that we need to estimate the parameter $\theta = Aa + Bb$, where A and B are given constants. A special case is obtained here if $A = 1, B = x_0$, so that the objective is to estimate the mean response at $X = x_0$, namely $E(Y_{x_0}) = a + bx_0$.

Clearly, the unbiased estimator of θ is $W = AV + BU$, whose value for the sample is $\hat{\theta} = A\hat{a} + B\hat{b}$. The distribution of W is normal since W is a sum of two normally distributed random variables. So we have

$$\begin{aligned}\sigma_W^2 &= A^2\sigma_V^2 + B^2\sigma_U^2 + 2AB\mathrm{Cov}(U, V) \\ &= \frac{\sigma^2}{\sum(x_i - \overline{x})^2}[A^2\sum x_i^2 - 2AB\overline{x} + B^2].\end{aligned}$$

Proceeding as before, we can show that the random variable

$$\frac{[A(V-a) + B(U-b)]\sqrt{(n-2)\sum(x_i - \overline{x})^2}}{\sqrt{[A^2\sum x_i^2 - 2AB\overline{x} + B^2][\sum(Y_{x_i} - Ux_i - V)^2]}}$$

has the Student's t distribution with $n-2$ degrees of freedom. This gives the $(1-\gamma)$-level confidence interval for θ as

$$\hat{\theta} \pm t_{\gamma/2,n-2}\sqrt{\frac{A^2\sum x_i^2 - 2AB\overline{x} + B^2}{(n-2)\sum(x_i - \overline{x})^2}}\sqrt{\sum(y_i - \hat{a} - \hat{b}x_i)}.$$

These results can also be used to test hypotheses about the regression coefficients a and b. The testing procedures (likelihood ratio tests) use the same Student-type ratios as the confidence intervals above. We will give the results for the slope coefficient b, leaving the derivation of the tests for intercept a as exercises.

Suppose that we want to test the null hypothesis

$$H_0 : b = b_0$$

against a one- or two-sided alternative. Then, given H_0, the random variable $(U - b_0)/\sigma_U$ has a standard normal distribution, and consequently the test statistic obtained upon division by $\sqrt{nT^2/[\sigma^2(n-2)]}$, that is

$$t = \frac{U - b_0}{\sqrt{\sum (Y_{x_i} - Ux_i - V)^2}} \sqrt{(n-2)\sum (x_i - \overline{x})^2}$$

has t distribution with $n - 2$ degrees of freedom. A one- or two-sided rejection region is then used depending on the alternative hypothesis.

PROBLEMS

14.4.1 Derive the test for the null hypothesis $H_0 : a = a_0$ against the one- or two-sided alternative.

14.4.2 The scores on an entrance exam (x) and the GPA's upon graduation (y) for 10 randomly selected students of a certain university are

x	355	361	402	365	375	404	349	380	420	395
y	3.66	3.49	3.86	3.24	3.55	3.92	3.11	3.19	3.76	3.75

Assume normality and homoscedasticity. **(i)** Compute the MLE's of a, b, and σ^2. **(ii)** Test the hypothesis that there is no relation between the grade on the entrance exam and the GPA, against the alternative that higher scores on the entrance exam tend to be associated with higher GPA's. **(iii)** Find the shortest 95% confidence intervals for a and for b. **(iv)** Find the joint confidence set for (a, b) with confidence level 0.95; sketch it and compare with your answers to (iii).

14.4.3 Suppose that observations are taken only at two values, x_1 and x_2, of an independent variable. Let $\overline{y}_1$ and $\overline{y}_2$ be the average observed responses for $x = x_1$ and for $x = x_2$, respectively. Show that the estimated regression line passes through points $(x_1, \overline{y}_1)$ and $(x_2, \overline{y}_2)$.

14.4.4 Using the ideas given on deriving the confidence set for (a, b), derive the testing procedure for the null hypothesis $H_0 : a = a_0, b = b_0$ against the general alternative $H_1 :$ H_0 is false. Consider two cases: **(i)** σ known. **(ii)** σ unknown.

14.4.5 Suppose it is known that the true regression (assuming a normal case) has the form $E(Y_x) = bx$. Derive the MLE's for b and for σ^2.

14.5 TESTING LINEARITY

In this section we make the following two additional assumptions regarding the design of the experiment for collecting the data on regression parameters:

(i) There are at least three distinct values among $x_1, x_2, \ldots, x_n$.

(ii) There exist at least two repeated values among $x_1, \ldots, x_n$.

We will show that under (i) and (ii) it is possible to construct a test for linearity of regression, that is, a test of the null hypothesis

$$H_0 : E(Y_x) = a + bx \text{ for some } a, b \qquad \text{against} \qquad H_1 : E(Y_x) = u(x),$$

where $u(x)$ is not a linear function of x.

The test for linearity will be based on construction of two independent unbiased estimators of σ^2. One of them will estimate σ^2 regardless of whether or not H_0 is true, and the other will be an unbiased estimator of σ^2 only under the null hypothesis.

For the considerations of the present section only, it will be convenient to change the labeling of the sample (x_i, y_i), $i = 1, \ldots, n$, as follows: Let $x_1', x_2', \ldots, x_r'$ be all those x_i at which multiple observations were made, and let $n_1, \ldots, n_r$ be the numbers of observations made for those values. Furthermore, let the observations made for x_j' be $y_{j,1}', \ldots, y_{j,n_j}'$, and regarded as values of iid random variables $Y_{j,1}', \ldots, Y_{j,n_j}'$. Finally, the remaining values of independent variable will be denoted by $x_{r+1}', \ldots, x_m'$, with the corresponding observations $y_{r+1}', \ldots, y_m'$ being the values of random variables $Y_{r+1}', \ldots, Y_m'$.

We have therefore

$$n_1 + \cdots + n_r + (m - r) = n, \tag{14.16}$$

with $n_i \geq 2, i = 1, \ldots, r$.

Consider now the following decomposition of the sum of the squared deviations:

$$\begin{aligned}
s^2 &= \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 = \sum_{j=1}^{r}\sum_{t=1}^{n_j}(y_{j,t}' - \hat{a} - \hat{b}x_j')^2 + \sum_{k=r+1}^{m}(y_k' - \hat{a} - \hat{b}x_k')^2 \\
&= \sum_{j=1}^{r}\sum_{t=1}^{n_j}(y_{j,t}' - \overline{y}_j' + \overline{y}_j' - \hat{a} - \hat{b}x_j')^2 + \sum_{k=r+1}^{m}(y_k' - \hat{a} - \hat{b}x_k')^2 \\
&= \sum_{j=1}^{r}\sum_{t=1}^{n_j}(y_{j,t}' - \overline{y}_j')^2 + \sum_{j=1}^{r} n_j(\overline{y}_j' - \hat{a} - \hat{b}x_j')^2 + \sum_{k=r+1}^{m}(y_k' - \hat{a} - \hat{b}x_k')^2 \\
&= s_1^2 + s_2^2 + s_3^2.
\end{aligned}$$

In passing from the third to the fourth expression, the cross-products were omitted. One can check that indeed they are equal to zero.

Now the sums s_1^2, s_2^2, and s_3^2 are observed values of random variables

$$\begin{aligned} S_1^2 &= \sum_{j=1}^{r}\sum_{t=1}^{n_j}(Y_{j,t}' - \overline{Y}_j')^2, \\ S_2^2 &= \sum_{j=1}^{r} n_j(\overline{Y}_j' - V - Ux_j')^2, \\ S_3^2 &= \sum_{k=r+1}^{m} (Y_k' - V - Ux_k')^2. \end{aligned}$$

Under the assumption of normality of the distributions of Y_x, the random variables S_1^2 and S_2^2 are independent; random variables S_3^2 and S_1^2 are independent as well. The proof of these facts is similar to the proof of Theorem 10.2.1, and it will be omitted.

Finally, in view of (14.16), S_1^2/σ^2 has a chi-square distribution with the number of degrees of freedom equal to

$$\sum_{j=1}^{r}(n_j - 1) = \sum_{j=1}^{r} n_j - r = n - m,$$

where $n - m > 0$, since $r \geq 1$ and $n_j \geq 2$ for all j. It is important to recognize that this statement about the distribution of S_1^2 holds *regardless of whether the null hypothesis H_0 about linearity of regression is true*.

On the other hand, if H_0 *is true*, then also

$$S_2^2/\sigma^2 + S_3^2/\sigma^2 \sim \chi_{m-2}^2.$$

Again, the number of degrees of freedom is positive, in view of the assumption that m, the number of distinct values of x_i, is at least 3.

To obtain the testing procedure, observe finally that any violations of the null hypothesis will tend to increase the expected value of the sum $S_2^2 + S_3^2$, since $E(Y - \xi)^2$ is minimized for $\xi = E(Y)$.

Consequently,

$$F = \frac{(S_2^2 + S_3^2)/(m-2)}{S_1^2/(n-m)}$$

has the F-distribution with $(m-2, n-m)$ degrees of freedom, provided that H_0 is true, while any lack of fit to the linear model will tend to inflate the value of F. Thus the testing procedure with size α is as follows: reject the hypothesis of linearity of regression if

$$\frac{[\sum_{j=1}^{r} n_j(\overline{y}_j' - \hat{a} - \hat{b}x_j')^2 + \sum_{k=r+1}^{m}(y_k' - \hat{a} - \hat{b}x_k')^2]/(m-2)}{\sum_{j=1}^{r}\sum_{t=1}^{n_j}(y_{j,t}' - \overline{y}_j')^2/(n-m)} > F_{\alpha,m-2,n-m}.$$

■ **EXAMPLE 14.13**

To develop some sort of intuition concerning the linearity test, we will analyze the situation in a deliberately oversimplified case: two observations for $x = 0$ are d and $-d$, an observation for $x = 1$ is c, and an observation for $x = 2$ is 0. We will find the range of values c (for fixed d), and the range of values d (for given c) when the linearity hypothesis should be rejected, on a level of significance $\alpha = 0.05$, say.

Intuitively, for a fixed d linearity will be rejected if the middle point deviates too far from the x-axis in any direction, and (for fixed $c \neq 0$) when d is close to 0.

We have here $\overline{x} = (0 + 0 + 1 + 2)/4 = 3/4$ and $\sum(x_i - \overline{x})^2 = 11/4$. Moreover $\overline{y} = (d - d + c + 0)/4 = c/4$. This gives

$$\hat{b} = \frac{c}{11}, \qquad \hat{a} = \frac{2c}{11}.$$

The estimated regression line is therefore $\hat{a} + \hat{b}x = (c/11)(2+x)$. To compute the F ratio, we have $s_1^2 = 2d^2$ with $n - m = 1$. For the numerator we find

$$s_2^2 + s_3^2 = 2\left[0 - \frac{c(2+0)}{11}\right]^2 + \left[c - \frac{c(2+1)}{11}\right]^2 + \left[0 - \frac{c(2+2)}{11}\right]^2 = \frac{84}{121}c^2,$$

with $m - 2 = 1$ degree of freedom. Since $F_{\alpha,1,1} = 161.45$, the hypothesis of linearity should be rejected if

$$\frac{\frac{84}{121}c^2}{2d^2} > 161.45 \quad \text{or} \quad \frac{c^2}{d^2} > 443.99.$$

Since $d > 0$, this is equivalent to the inequality $|c| > 21.56d$.

PROBLEMS

14.5.1 The output of a certain device is suspected to decrease linearly with the temperature. Two observations were taken for each temperature, and the data (in appropriate units) are as follows:

Temperature	55	65	75	85	95	105
Observation 1	2.01	2.01	2.02	1.48	1.93	1.90
Observation 2	2.03	2.02	2.00	1.48	1.95	1.94

At the significance level 0.05, test the hypothesis that the output is a linear function of temperature.

14.5.2 Suppose that we have six data points in addition to those in Problem 14.4.2:

x	355	402	402	309	375	375
y	3.44	3.91	3.95	3.24	3.52	3.31

Test the hypothesis (using all 17 data values) that the regression of GPA's on the scores from the entrance exam is linear.

14.5.3 Check the identity (14.17).

14.5.4 Derive, if possible, a test of linearity of regression under the assumptions of this section, if the individual observations for values $x'_{r+1}, \ldots, x'_m$ are now unknown, and instead we have the data on averages $y'_{r+1}, \ldots, y'_m$, and the corresponding numbers of observations $n'_{r+1}, \ldots, n'_m$.

14.6 PREDICTION

Consider now the problem of prediction. As before, the data have the form of a set of pairs $(x_i, y_i), i = 1, \ldots, n$, where y_i is the observed value of a random variable Y_{x_i} assumed to be normal with mean $a + bx_i$ and standard deviation σ. The random variables $Y_{x_i}, \ldots, Y_{x_n}$ are independent, and at least two among $x_1, \ldots, x_n$ are distinct. The problem is to predict Y_{x_0} as precisely as possible.

More generally, we want to predict the average of k independent observations of Y_{x_0} . By prediction, we mean here providing an interval, as short as possible, such that the value of the predicted random variable will fall into this interval with a preassigned probability, say $1 - \gamma$. Let $\overline{Y}_{x_0}$ denote the average of k observations to be taken at the value x_0 of the independent variable. We are looking for an interval $[c_1, c_2]$ such that

$$P\{c_1 \leq \overline{Y}_{x_0} \leq c_2\} = 1 - \gamma.$$

The solution will be obtained as follows. Note first that the average $\overline{Y}_{x_0}$ has distribution $\mathrm{N}(a + bx_0, \sigma^2/k)$. Consequently,

$$Z = \frac{\overline{Y}_{x_0} - a - bx_0}{\sigma}\sqrt{k}$$

is a standard normal random variable, and we have

$$\{-z_{\gamma/2} \leq Z \leq z_{\gamma/2}\} = 1 - \gamma.$$

A simple argument based on symmetry around 0 of normal density shows that $(-z_{\gamma/2}, z_{\gamma/2})$ is the shortest prediction interval for Z. Thus the corresponding shortest prediction interval for $\overline{Y}_{x_0}$ (given that the regression parameters a, b and σ^2 are known) is

$$a + bx_0 \pm z_{\gamma/2}\frac{\sigma}{\sqrt{k}},$$

and its length is $2(\sigma/\sqrt{k})z_{\gamma/2}$. The actual prediction interval has to take into account the fact that the regression parameters are estimated. The construction is based on an analogue to the pivotal quantity. Thus the random variable

$$L = \overline{Y}_{x_0} - Ux_0 - V$$

has a normal distribution (being a linear combination of normal random variables), and $E(L) = 0$ in view of the fact that U and V are unbiased estimators of b and a.

We have, using the fact that $\overline{Y}_{x_0}$ is independent of (U, V),

$$\begin{aligned}\sigma_L^2 &= \operatorname{Var}(L) = \operatorname{Var}(\overline{Y}_{x_0}) + x_0^2 \operatorname{Var}(U) + \operatorname{Var}(V) + 2x_0 \operatorname{Cov}(U, V) \\ &= \frac{\sigma^2}{k} + x_0^2\sigma_U^2 + \sigma_V^2 + 2x_0 \operatorname{Cov}(U, V) \\ &= \frac{\sigma^2}{k} + x_0^2 \frac{\sigma^2}{\sum(x_i - \overline{x})^2} + \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \overline{x})^2} - 2x_0 \frac{\sigma^2 \overline{x}}{\sum(x_i - \overline{x})^2} \\ &= \sigma^2 \left[\frac{1}{k} + \frac{x_0^2 + \frac{1}{n}\sum x_i^2 - 2x_0\overline{x}}{\sum(x_i - \overline{x})^2}\right] = \sigma^2 \left[\frac{1}{k} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right].\end{aligned}$$

Consequently, the random variable

$$\frac{L}{\sigma_L} = \frac{\overline{Y}_{x_0} - Ux_0 - V}{\sigma\sqrt{1/k + 1/n + (x_0 - \overline{x})^2/\sum(x_i - \overline{x})^2}}$$

has a standard normal distribution. Dividing by $\sqrt{nT^2/[\sigma^2(n-2)]}$, we obtain

$$t = \frac{(\overline{Y}_{x_0} - Ux_0 - V)\sqrt{n-2}}{\sqrt{1/k + 1/n + (x_0 - \overline{x})^2/\sum(x_i - \overline{x})^2}\sqrt{\sum(\overline{Y}_{x_0} - Ux_i - V)^2}},$$

which has the t–distribution with $n-2$ degrees of freedom. Substituting the observed values $\hat{a}$ and $\hat{b}$ of V and U, we obtain the prediction interval for $\overline{Y}_{x_0}$ with prediction probability $1 - \gamma$ as

$$\hat{a} + \hat{b}x_0 \pm \frac{t_{\gamma/2n-2}}{\sqrt{n-2}}\sqrt{\left[\frac{1}{k} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right]\left[\sum(y_{x_i} - \hat{a} - \hat{b}x_i)^2\right]}. \quad (14.17)$$

Let us remark here that the prediction interval (for fixed n and k) is shortest if $x_0 = \overline{x}$ (i.e., it is "easier" to predict values of dependent variable for x_0 close to $\overline{x}$).

Observe also that as $k \to \infty$, the length for prediction interval for *known* a, b, and σ tends to 0. In the present case the increase of k has much less effect, and as $k \to \infty$, the length of the prediction interval tends to a positive quantity depending on n and on the location of observations $x_1, x_2, \ldots, x_n$. This is consistent with intuition, according to which in the present case the uncertainty of prediction has two sources: randomness of the dependent variable about its estimated mean, and uncertainty as to the exact location of the true mean.

PROBLEMS

14.6.1 At harvest, the weight of a certain fruit grown in a greenhouse has the $N(a + bt, \sigma^2)$ distribution, where t is the average temperature. Weights in a sample of five fruits for $t = 80°F$ are 1.02, 1.03, 0.98, 1.05, 1.02, while a sample of seven fruits for $t = 86°F$ (other conditions being equal) are 1.03, 1.03, 1.09, 1.07, 1.04, 1.02, 1.08. Give a 95% prediction interval for the average weight of four fruits grown: **(i)** In the first greenhouse. **(ii)** In the second greenhouse. **(iii)** In a greenhouse for $t = 84°F$.

14.6.2 Combine data points in Problem 14.4.2 and 14.5.2, and find the 95% prediction interval for the GPA of a student who scored 400 on an entrance exam.

14.6.3 Find the prediction interval with a probability $1-\gamma$ of coverage for an observation to be taken at the value x_0 by an independent variable, given the data $(x_i, y_i), i = 1, \ldots, n$, and assuming the normal model of the form $Y_x = bx + \epsilon$ with $\epsilon \sim \mathrm{N}(0, \sigma^2)$.

14.6.4 Under condition of Problem 14.6.3 find the prediction interval for the mean of k observations taken for value x_0 of the independent variable.

14.7 INVERSE REGRESSION

Inverse regression is the problem of inference about an unknown value x_0 of an independent variable on the basis of a number of observations of a response for this value. The data consist of two groups of observations. One group is, as before, the sample

$$(x_i, y_i), \quad i = 1, \ldots, n,$$

where y_i is the observed value of random variable $Y_i \sim \mathrm{N}(a + bx_i, \sigma^2)$, with the usual assumption of independence. The second group, $(y_1', \ldots, y_m')$, is a random sample from an $\mathrm{N}(a + bx_0, \sigma^2)$ distribution, where x_0 is unknown. The objective is to estimate x_0, with a, b, and σ^2 being unknown.

The likelihood of the data is

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\sigma^2(y_i - a - bx_i)^2} \prod_{j=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\sigma^2(y_i' - a - bx_0)^2} \\ &= C(\sigma^2)^{-(m+n)/2} \exp\left\{ - \frac{\sum_{i=1}^{n}(y_i - a - bx_i)^2 + \sum_{j=1}^{m}(y_j' - a - bx_0)^2}{2\sigma^2} \right\}. \end{aligned}$$

Taking the logarithm and differentiating with respect to a, b, σ^2, and x_0, we obtain the equations

$$\begin{aligned} &\sum_{i=1}^{n}(y_i - a - bx_i) + \sum_{j=1}^{m}(y_j' - a - bx_0) = 0, \\ &\sum_{i=1}^{n}(y_i - a - bx_i)x_i + x_0 \sum_{j=1}^{m}(y_j' - a - bx_0) = 0, \\ &\sum_{i=1}^{n}(y_i - a - bx_i)^2 \sum_{j=1}^{m}(y_j' - a - bx_0)^2 - (m+n)\sigma^2 = 0, \\ &\sum_{j=1}^{m}(y_j' - a - bx_0) = 0. \end{aligned}$$

Letting $\overline{y}' = (1/m)\sum_{j=1}^{m} y_j'$, we obtain from the last equation the estimate

$$\hat{x}_0 = \frac{\overline{y}' - \hat{a}}{\hat{b}}, \tag{14.18}$$

assuming, of course, that $\hat{b} \neq 0$. The third equation gives

$$\hat{\sigma}^2 = \frac{1}{m+n}\left[\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 + \sum_{j=1}^{m}(y_j - \overline{y}')^2\right].$$

Using (14.18) in the first two equations, we can easily check that in each case the sum involving y_j' equals 0, which means that the expressions for $\hat{a}$ and $\hat{b}$ are given by (14.9).

These results are consistent with our intuition. Indeed, since we do not know x_0, the observations $y_1', \ldots, y_m'$ cannot provide any information about the slope and intercept of the regression line. On the other hand, if only $m > 1$, the values $y_1', \ldots, y_m'$ provide additional information about σ^2.

To set a confidence interval for $\hat{x}_0$, we may proceed as in the case of a prediction. The random variable $W = \overline{Y}' - Ux_0 - V$ has the normal distribution with mean 0 and variance

$$\sigma_W^2 = \sigma^2\left(\frac{1}{m} + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum(x_0 - \overline{x})^2}\right).$$

Consequently, W/σ_W has a standard normal distribution. To eliminate σ, we note that the random variable

$$\frac{(m+n)\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(Y_{x_i} - Ux_i - V)^2 + \sum_{j=1}^{m}(Y'_{x_{0,j}} - \overline{Y}')^2\right]$$

has the chi-square distribution with $m + n - 3$ degrees of freedom. Indeed, the two sums are independent, with $n - 2$ and $m - 1$ degrees of freedom, respectively. Consequently the random variable

$$\frac{W/\sigma_W}{\sqrt{(m+n)\hat{\sigma}^2/\sigma^2(m+n-3)}} \tag{14.19}$$

has the Student's t distribution with $m + n - 3$ degrees of freedom. The observed value of the random variable given by (14.19) is

$$t = \frac{(\overline{y}' - \hat{a} - \hat{b}x_0)\sqrt{m+n-3}}{\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum(x_0-\overline{x})^2}} \times \sqrt{\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_0)^2 + \sum_{j=1}^{m}(y_j' - \overline{y}')^2}},$$

and the confidence interval is obtained by converting the inequality

$$-t_{\gamma/2,m+n-3} < t < t_{\gamma/2,m+n-3} \tag{14.20}$$

into an inequality for x_0. Observe, however, that (14.20) is now a *quadratic* inequality.

■ EXAMPLE 14.14

The amounts of a chemical compound (y) that dissolve in a given amount of water at different temperatures (x) are given in the following table:

x (°C)	y (grams)		
5	3	4	2
10	7	7	6
15	10	13	11
20	15	18	17
25	21	18	19

Two measurements for an unknown temperature x_0 are 14 and 16. What can one say about x_0?

SOLUTION. We have here $n = 15$ and $m = 2$. The relevant quantities are

$$\hat{a} = -1.14, \quad \hat{b} = 0.853, \quad \overline{x} = 15, \quad \sum (x_i - \overline{x})^2 = 750,$$

$$\sum (y_i - \hat{a} - bx_i)^2 = 21.46 \, \overline{y}' = 15, \quad \sum (y_i' - \overline{y}')^2 = 2.$$

Thus the point estimate of x_0 is $\hat{x}_0 = 19.23$. Since $t_{0.05,14} = 2.145$ ($n + m - 3 = 14$), to obtain a 95% confidence interval for x_0, we must solve the inequality (14.20), which in our case takes the form

$$\frac{|15 - 0.853x_0 + 1.4|}{\sqrt{\frac{1}{2} + \frac{1}{15} + (x_0 - 15)^2/750}} \frac{\sqrt{2 + 15 - 3}}{\sqrt{21.46 + 2}} < 2.145$$

or

$$\frac{|346.98 - 18.05x_0|}{\sqrt{x_0^2 - 30x_0 + 650}} < 2.145.$$

After some algebra, we obtain $x_0^2 - 38.567 + 365.499 < 0$, which gives the confidence interval

$$16.96 < x_0 < 21.80.$$

PROBLEMS

14.7.1 Five measurements of Y at $x = 10$ are 10.5, 10.6, 9.7, 11.1, and 12.3. Six measurements of Y at $x = 20$ are 3.1, 3.6, 3.1, 4.0, 5.2, and 2.9. Assuming that the regression of Y on X is linear, estimate the value of x if two observations made at this value are 6.3 and 7.1.

14.7.2 Using data from Problems 14.4.2 and 14.5.2, estimate the score on an entrance exam of a student who graduated with a GPA equal to 3.95.

14.8 BLUE

Most of the results presented thus far rely on the assumptions that the random variable Y_x is normally distributed with constant variance and that the observations are independent. These assumptions allow us to use the likelihood and provide access

to the distributions of MLE's. The natural question is what to do if the assumptions above are not satisfied.

First, with regard to homoscedasticity, there exist numerous variance-stabilizing transformations. These transformations have been suggested by statisticians as an ad hoc remedy against heteroscedasticity: instead of data of the form (x_i, y_i), one can use the data (x_i, y_i^*), with $y_i^* = g(y_i)$, where g is some suitably selected function.

One can also use transformation of y's that depend on x's, that is, by replacing the pair (x_i, y_i) with $(x_i, g_{x_i}(y_i))$, where g_x is some function [e.g., replacing (x_i, y_i) by $(x_i, y_i/x_i)$]. Which transformation should be used in a given situation can be hard to resolve, especially when little is known about the distribution of Y_x. It is rather statistical intuition and experience that can serve as a guide.

A question interesting from a both theoretical and practical point of view is what to do if the distribution of Y_x is not normal. Suppose, for instance, that the model analyzed is that of linear regression $Y_x = a + bx + \epsilon_x$, where the errors ϵ_x can be assumed to satisfy the conditions;

(i) $E(\epsilon_x) = 0$ (unbiasedness).
(ii) $E(\epsilon_x^2) = \sigma^2$ (homoscedasticity).
(iii) $E(\epsilon_x \epsilon_{x'}) = 0$ (orthogonality).

The least squares estimators of regression coefficients a and b (which are also the MLE's in the normal case) are

$$\hat{b} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2} \qquad \text{and} \qquad \hat{a} = \overline{y} - \hat{b}\overline{x}.$$

The estimator of σ^2 may be based on residuals, where

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{j=1}^{n}(y_j - \hat{a} - \hat{b}x_j)^2$$

is the observed value of the estimator

$$S^2 = \frac{nT^2}{(n-2)} = \frac{1}{n-2}\sum_{j=1}^{n}(Y_j - V - U_{x_i})^2,$$

with U, V, and T given by (14.11). Clearly, U, V, and S^2 are unbiased estimators of b, a, and σ^2, with variances and covariances obtained as before. This is true because the calculations of moments did not rely on the assumption of normality.

It is possible to show that under the assumption of normality, U and V are also minimum variance estimators of b and a (i.e., that their variances coincide with the bounds given by the Rao-Cramèr inequality).

The question is: Is there a reason to use U and V as estimators of b and a when the normality assumption does not hold? The mere availability of the computational formulas in statistical packages is hardly a justification. Unbiasedness of U and V is a desirable property only if it can be related to MSE. As we have seen, there are situations where the use of biased estimators is recommended, as leading to smaller mean square error.

The answer is positive and is given by the following theorem (which we state here without a proof).

Theorem 14.8.1 (Gauss-Markov) *Consider the observations* $(x_i, Y_j), i = 1, 2, \ldots, n$, *with* $Y_i = a + bx_i + \epsilon_i$, *where the errors* ϵ_i *satisfy conditions (i)–(iii) of unbiasedness, homoscedasticity, and orthogonality. Let* $\mathcal{L}$ *be the class of all statistics of the form*

$$r_1 Y_1 + r_2 Y_2 + \cdots + r_n Y_n,$$

where r_i *are constants—depending possibly on the vector* $(x_1, \ldots, x_n)$. *Furthermore, let* $\mathcal{L}_a \subset \mathcal{L}, \mathcal{L}_b \subset \mathcal{L}$ *be the subsets of* $\mathcal{L}$ *consisting of statistics that are unbiased estimators of* a *and of* b. *Then the statistics* V *and* U *given by* (14.11) *have minimal variances in classes* $\mathcal{L}_a$ *and* $\mathcal{L}_b$.

The acronym used here is BLUE, which stands for the "best linear unbiased estimator." Thus V is BLUE for a, and U is BLUE for b.

Note that in view of unbiasedness, "best" estimators mean those with a minimum mean square error. Recall that in normal case U and V are best (in the sense of MSE) estimators of regression parameters in the class of *all* estimators, linear or not. In the present case, under the weaker assumption the conclusion is also weaker: U and V are best estimators in a more restricted class of estimators, namely those that depend linearly on the data.

PROBLEMS

14.8.1 Suppose that the number of errors in a text of length x is known to be a Poisson random variable with unknown mean λ. We observe n texts of lengths $x_1, x_2, \ldots, x_n$ and find the numbers of errors they contain, $Y_1, \ldots, Y_n$, satisfy $E(Y_j) = \mathrm{Var}(Y_j) = \lambda x_j$. Find the BLUE of λ.

14.8.2 Carry out the calculations in the following direct proof of the Gauss-Markov theorem showing that the LS estimators of a and b are BLUE. For a you need to determine the constants $a_i, i = 1, \ldots, n$ such that the statistic $T = \sum_{i=1}^{n} a_i Y_i$ satisfies the conditions: (1) $E(T) = a$. (2) Variance of T is the smallest among all linear estimators for which condition (1) holds. Show first that condition (1) implies

$$\sum a_j = 1, \qquad \sum a_j x_j = 0. \tag{14.21}$$

Next use the fact that Y_j's are uncorrelated and homoscedastic, and show that $\mathrm{Var}(T) = \sigma^2 \sum a_j^2$, that is, minimize $\sum a_j^2$ subject to constraints (14.21). Using Lagrange multipliers, this means that one must minimize

$$\sum a_j^2 - \lambda_1 \sum (a_j - 1) - \lambda_2 \sum a_j x_j.$$

Take the derivatives with respect to $a_1, \ldots, a_n, \lambda_1$, and λ_2, solve the resulting $n + 2$ equations, and check that the solution a_j gives the LS estimator of a.

14.8.3 Provide the same argument as in Problem 14.8.2 for the LS estimator of b.

14.8.4 Given data $(x_i, y_i), i = 1, \ldots, n$, find the LS estimators of the quadratic regression $a + bx + cx^2$.

14.9 REGRESSION TOWARD THE MEAN

The phenomenon known as *regression toward the mean* was discovered in the nineteenth century by Galton, who studied various hereditary traits. He noticed that (e.g., in using height) tall fathers tend to have tall sons, but their sons tend to be closer to the average than the fathers. Similarly short fathers tend to have short sons, but their sons tend to be closer to the average than their fathers. Galton called it a "tendency toward mediocrity." Galton's choice of the word "mediocrity" may also explain why he chose the term "regression," a word with somewhat derogatory connotations.

It should be realized that regression, understood as "affecting the mean of dependent variable Y by independent variable X," need not imply any casual relationship between the values of X and Y. One of the more common types of relations between X and Y, that leads to the regression phenomenon and yet does not involve any casual effects of X on Y, is exemplified by the following situation: Imagine some attribute of objects, for example, length. Suppose that a person takes two observations, typically differing somewhat because of a measurement error, and calls the first and second measurement of the ith object x_i and y_i.

Imagine now that points (x_i, y_i) are plotted, resulting in a scatter diagram. If the measurement errors are small, and/or the objects measured differ in their true lengths, we will observe that the points (x_i, y_i) have a strong linear relationship with a slope close to 1.

Such a relationship will appear stronger when the variability of lengths of measured objects is higher. This effect is utilized in designing some psychological questionnaires.

■ EXAMPLE 14.15 Psychological Test Scores

In the areas of psychology dealing with personality or motivation (as opposed to areas such as memory studies, with more quantifiable experiments) a researcher typically introduces some *construct* ("neuroticism," "self-esteem," etc.). Those constructs are then used to explain and/or predict some behavior. The explanation has the form of specific hypotheses, such as "persons with low self-esteem are more likely to be aggressive," etc. In addition to a theoretical justification of such hypotheses, there arises a problem of testing them empirically. Clearly, one needs here a tool for measuring the level of the construct (a tool to measure the level of neuroticism, self-esteem, etc.).

A typical tool has the form of a questionnaire and a scoring rule. For some questions, it is the answer "yes" that contributes to the total score; for some other questions it is the answer "no." This is done in order to eliminate any bias arising from a possible tendency toward some type of answers.

When the questionnaire is applied to a subject s, one obtains the score X_s. Upon repetition, the score may be different, say X_s'. One of the central assumptions of the theory of psychological tests is that the expected scores X_s and X_s' are equal:

$$E(X_s) = E(X_s') = T_s.$$

Moreover, the deviations $\epsilon_s = X_s - T_s$ and $\epsilon_s' = X_s' - T_s$ are expected to satisfy the conditions

$$\text{Var}(\epsilon_s) = \text{Var}(\epsilon_s') = \sigma_s^2, \qquad E(\epsilon_s \epsilon_s') = 0.$$

For fixed s, the value T_s (called the *true value* of the measured construct for person s) is a constant, while ϵ_s is a random variable (reflecting intra-person variability of response to the questionnaire, upon hypothetical repetitions of measurement).

Assume now that person s is sampled from some population according to a certain probability distribution. Using Theorem 8.6.4, we have

$$\begin{aligned} \sigma_X^2 &= \text{Var}(X) = E_s\{\text{Var}(X|T)\} + \text{Var}\{E_s(X|T)\} \\ &= E(\sigma_s^2) + \text{Var}(T_s) = \sigma_\epsilon^2 + \sigma_T^2. \end{aligned}$$

Clearly, $\sigma_{X'}^2 = \sigma_\epsilon^2 + \sigma_T^2$. Similarly

$$\begin{aligned} \text{Cov}(X, X') &= E(XX') - E(X)E(X') = E\{(T+\epsilon)(T+\epsilon')\} - [E(T)]^2 \\ &= E(T^2) - [E(T)]^2 = \sigma_T^2, \end{aligned}$$

since $E(\epsilon T) = E_s\{E(\epsilon T)\} = E_s\{T_s E(\epsilon)\} = E(0) = 0$, and similarly for the other products. Thus

$$\rho(X, X') = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\epsilon^2}. \tag{14.22}$$

The last ratio is called the *reliability* of the test, and formula (14.22) shows that reliability is equal to the test–retest correlation. This correlation approaches 1 with an increase of σ_T^2, that is, with an increase of variance of the true scores of the test in the population under study.

The reliability of a psychological test is (unlike of the instruments for physical measurements) not intrinsic for the test only but depends also on how diverse the population is to which the test is applied.

14.10 ANALYSIS OF VARIANCE

In the remainder of this chapter we deal with testing for the existence of the effects of an independent variable X (often measured on the nominal scale) on a response Y.

The standard terminology of the analysis of variance (ANOVA) is that of *levels of factors*. These factors cross each other in the sense that every level of one factor can be combined with every level of the other factor.

The central assumptions of analysis of variance models are very much the same as in regression models:

(i) The response variable Y has, for each level of factors, a normal distribution with the same (unknown) variance σ^2 (*homoscedasticity*).
(ii) The factors may affect only the mean of the response.

(iii) Distinct observations for the same or different levels of factors are independent.

We show the tests for the hypothesis that a given factor has no effect on the response variable against the alternative that it has some effect.

The main issue here is that these tests can be carried out for various factors *on the same data*. In fact ANOVA models originated from questions arising in agriculture, where one is interested in the response variable (e.g., size of harvest Y) as dependent on combinations of various factors (e.g., type of soil, time of planting, time of harvesting, type of cultivation, and use of various fertilizers). Since a typical experiment lasts for one season, it is imperative to find a design that will allow us to study the effects of various factors using the same data. We begin with the simplest case of one factor only.

14.11 ONE-WAY LAYOUT

Consider the situation where the data are partitioned into groups, each corresponding to one level of a factor. Alternatively, the same setup may be described as "independent samples from different populations."

We let n_i denote the number of observations from the ith group, where $i = 1, \ldots, I$. We have here $I \geq 2$ and $n_i \geq 2$ for at least one i.

Let $n_1 + n_2 + \cdots + n_I = N$ be the total number of observations, and let $y_{i1}, y_{i2}, \ldots, y_{in_i}$ be the observations corresponding to the ith level of the factor. These observations are regarded as the recorded values of the random variables $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$.

According to the assumptions stated at the beginning of this section, all random variables Y_{ij} are independent, normally distributed, with $\text{Var}(Y_{ij}) = \sigma^2$. Moreover, since the effect of a factor is expressed only through the mean, we must have

$$E(Y_{ij}) = \mu_i, \qquad j = 1, \ldots, n_i.$$

The objective is to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I \quad \text{against} \quad H_1 : \mu_i \neq \mu_{i'} \text{ for some } i, i'.$$

It will be convenient to let

$$\mu_i = \mu + \alpha_i, \qquad i = 1, \ldots, I,$$

where

$$\mu_i = \frac{\mu_1 + \cdots + \mu_I}{I}. \tag{14.23}$$

We have therefore

$$\sum_{i=1}^{I} \alpha_i = 0,$$

and the hypotheses tested can be formulated as

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0 \quad \text{against} \quad H_1 : \alpha_i \neq 0 \text{ for at least one } i.$$

Before we develop the testing procedure, let us observe that if $I = 2$, we have the problem of comparing the means of two normal populations with the same (unknown) variance σ^2. This problem was solved in Chapter 13 (Section 13.9). The testing procedure used the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

It would seem that if $I > 2$, we can use this result for the present case by comparing pairs of levels of the factor until either we find a pair where the difference is significant (and then we reject H_0) or we find no significant difference in all pairs tested (in which case we accept H_0). The reason why such a procedure is unacceptable lies in the fact that it is impossible to determine its significance level because (i) the procedures for overlapping pairs of factor levels are not independent and (ii) even for nonoverlapping pairs, if the null hypothesis is true, the chances of at least one incorrect rejection of null hypothesis increases quickly with the number of tested pairs.

Consequently, the objective is to find a procedure that can test the null hypothesis of no effect of the factor with a preassigned level of significance. The construction here will be based on a partition of the sum of squared deviations from the mean, which is very much similar to the technique used for testing linearity of regression in Section 14.6.

In the derivation below, the subscript + will stand for averaging over the values of the index replaced by +. Thus

$$\overline{Y}_{i+} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \qquad \overline{y}_{i+} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},$$

and

$$\overline{Y}_{++} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N}\sum_{i=1}^{I} n_i \overline{Y}_{i+},$$

and similarly for $\overline{y}_{++}$.

In the identities below we omit the cross-products. We encourage the reader to verify that all the cross-products are indeed zero. Using the fact that by (14.23) we have $E(Y_{ij}) = \mu + \alpha_i$, we decompose the sum of squared deviations of the variables from their means as follows:

$$\begin{aligned} S^2 &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \mu - \alpha_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+} + \overline{Y}_{i+} - \mu - \alpha_i)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 + \sum_{i=1}^{I}(\overline{Y}_{i+} - \mu - \alpha_i)^2 \qquad (14.24) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 + \sum_{i=1}^{I} n_i(\overline{Y}_{i+} - \overline{Y}_{++} + \overline{Y}_{++} - \mu - \alpha_i)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 + \sum_{i=1}^{I} n_i(\overline{Y}_{i+} - \overline{Y}_{++} - \alpha_i)^2 + N(\overline{Y}_{++} - \mu)^2. \end{aligned}$$

Under the assumption that Y_{ij} are normally distributed, the three terms in the last row are independent random variables. Moreover, the first term, upon division by

σ^2, has the chi-square distribution with the number of degrees of freedom equal $(n_1 - 1) + \cdots + (n_I - 1) = N - I$ *regardless of whether or not the null hypothesis is true*. The second sum, again upon division by σ^2, has the chi-square distribution with $I - 1$ degrees of freedom, provided that $E(\overline{Y}_{i+} - \overline{Y}_{++}) = \alpha_i$ for every i. Thus, under H_0, the sum

$$\frac{1}{\sigma^2} \sum_{i=1}^{I} n_i (\overline{Y}_{i+} - \overline{Y}_{++})^2 \tag{14.25}$$

has a chi-square distribution with $I - 1$ degrees of freedom, and any violation of H_0 will increase the expectation of the sum (14.25). This suggests the use of an appropriate F ratio to test the hypothesis H_0.

Finally, if $\mu = 0$, then $N(\overline{Y}_{++})^2/\sigma^2$ has the chi-square distribution with one degree of freedom, which allows us to test hypothesis $H_0' : \mu = 0$ against the alternative $H_1' : \mu \neq 0$. In terms of observations, the testing procedure is most often displayed in the form of the following ANOVA table:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F Ratio
Mean	1	SSM	$MSM = \frac{SSM}{1}$	$F_M = \frac{MSM}{MSR}$
Factor	$I - 1$	SSA	$MSA = \frac{SSA}{I-1}$	$F_A = \frac{MSA}{MSR}$
Residual	$N - I$	SSR	$MSR = \frac{SSR}{N-I}$	
Total	N	SST		

where

$$SSM = N(\overline{y}_{++})^2, \qquad SSA = \sum_{i=1}^{I} n_i (\overline{y}_{i+} - \overline{y}_{++})^2,$$

$$SSR = \sum_{i=1}^{I} \sum_{j=1}^{n_i} n_i (y_{ij} - \overline{y}_{i+})^2, \qquad SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}^2.$$

We reject the hypothesis $H_0 : \alpha_1 = \cdots = \alpha_I = 0$ (at level γ) if the ratio F_A exceeds the upper quantile $F_{\gamma, I-1, N-I}$. It is important to remember that in ANOVA *one always uses a one-sided critical region* (since any violation of H_0 tends to increase the numerator without affecting the denominator).

If, for some reason, one is interested in the hypothesis $H_0' : \mu = 0$, then one should reject it in favor of $H_1' : \mu \neq 0$ if the ratio F_M exceeds $F_{\gamma, 1, N-I}$. Such a test is useful if observations can be positive as well as negative (e.g., deviations from the required standards in a technological process).

PROBLEMS

14.11.1 Verify that the cross-products in (14.24) do indeed equal zero.

14.11.2 To test the milage achieved by cars produced by different companies, but of comparable price, size, and so on, one make of cars was selected from among the three major American companies and two foreign companies. For each make selected, a number of new cars was chosen and their mpg (miles per gallon) recorded. The data (in mpg) are as follows:

$n_1 = 5$	$n_2 = 4$	$n_3 = 5$	$n_4 = 3$	$n_5 = 6$
25.1	27.1	39.9	25.4	29.2
26.2	26.4	21.4	28.2	29.3
24.9	26.8	22.2	27.1	30.4
25.3	27.2	22.5		28.5
23.9		20.8		28.9
				29.2

At the significance level 0.05, test the hypothesis that the average mpg is the same for all makes of cars tested.

14.11.3 Suppose that we take a random sample of size n from a normal distribution $\mathrm{N}(\mu, \sigma^2)$. We divide the observations into k groups of sizes $n_1, \ldots, n_k$, where $n_i \geq 2$ for $i = 1, \ldots, k$ and $n_1 + \cdots + n_k = n$. Let S_i^2 be the sample variance in the ith group. Find: **(i)** The distribution of $[(n_1 - 1)S_1^2 + \cdots + (n_k - 1)S_k^2]/\sigma^2$. **(ii)** The distribution of S_k^2/S_1^2.

14.11.4 Show that the test developed in this chapter is equivalent to the Student's t test for the case when the factor operates at two levels only.

14.12 TWO-WAY LAYOUT

Assume now that we have data concerning possible effects of two factors, A and B, with I and J levels, respectively. Let y_{ij} be the observation for the ith level of A and the jth level of B. Such data can be arranged into a matrix $[y_{ij}]$ with I rows and J columns. For the moment assume that we have one observation in each of the IJ cells formed by crossing factors A and B. As before, we regard y_{ij} as the recorded value of a random variable Y_{ij}, assumed to have normal distribution $Y_{ij} \sim \mathrm{N}(\mu + \alpha_i + \beta_j, \sigma^2)$, where α_i and β_j represent the effects of factors A and B. Without loss of generality, we can write

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0. \tag{14.26}$$

We want to construct a test for the hypothesis

$$H_0^{(A)} : \alpha_1 = \cdots = \alpha_I = 0 \qquad \text{against} \qquad H_1^{(A)} : \alpha_i \neq 0 \text{ for some } i,$$

as well as a test for

$$H_0^{(B)} : \beta_1 = \cdots = \beta_J = 0 \qquad \text{against} \qquad H_1^{(B)} : \beta_j \neq 0 \text{ for some } j.$$

The tests are built on a partition of the sum of squares, as in the case of one factor. Omitting again the cross–products (which are zero), we have

$$\begin{aligned} S^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\mu-\alpha_i-\beta_j)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{++}+\overline{Y}_{++}-\mu-\alpha_i-\beta_j)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{++}-\alpha_i-\beta_j)^2+IJ(\overline{Y}_{++}-\mu)^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i+}-\overline{Y}_{+j}+\overline{Y}_{++})^2+\sum_{i=1}^{I}J(\overline{Y}_{i+}-\overline{Y}_{++}-\alpha_i)^2 \\ &\quad +\sum_{j=1}^{J}I(Y_{+j}-\overline{Y}_{++}-\beta_j)^2+IJ(\overline{Y}_{++}-\mu)^2. \end{aligned}$$

Substituting the observed data for random variables and letting $\mu=0, \alpha_i=\beta_j=0$ for all i, j, we obtain the following sums of squares:

$$SSM = IJ(\overline{y}_{++})^2, \qquad SSA = \sum_{i=1}^{I}J(\overline{y}_{i+}-\overline{y}_{++})^2,$$

$$SSB = \sum_{i=1}^{J}I(\overline{y}_{+j}-\overline{y}_{++})^2, \qquad SSR = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\overline{y}_{i+}-\overline{y}_{+j}+\overline{y}_{++})^2,$$

which add up to the total sum of squares

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}y_{ij}^2.$$

The four terms in the last decomposition of S^2 are independent random variables. The SSR does not involve any parameters and (upon division by σ^2) has a chi-square distribution with the number of degrees of freedom equal to

$$IJ-(I-1)-(J-1)-1=(I-1)(J-1),$$

regardless of whether the hypotheses are true or false. The SSA sum has a chi-square distribution if $\alpha_1=\cdots=\alpha_I=0$, and a similar statement holds for SSB if $\beta_1=\cdots=\beta_J=0$. The numbers of degrees of freedom are $I-1$ and $J-1$, respectively. Any deviation from the null hypothesis tends to increase the corresponding sum of squares.

The ANOVA table is now:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F Ratio
Mean	1	SSM	$MSM = \frac{SSM}{1}$	$F_M = \frac{MSM}{MSR}$
A	$I-1$	SSA	$MSA = \frac{SSA}{I-1}$	$F_A = \frac{MSA}{MSR}$
B	$J-1$	SSB	$MSB = \frac{SSB}{J-1}$	$F_B = \frac{MSB}{MSR}$
Residual	$(I-1)(J-1)$	SSR	$MSR = \frac{SSR}{(I-1)(J-1)}$	
Total	$N = IJ$	SST		

The hypothesis $H_0^{(A)}$ (that factor A has no effect) is rejected whenever $F_A > F_{\gamma,I-1,(I-1)(J-1)}$, and $H_0^{(B)}$ is rejected if $F_B > F_{\gamma,J-1,(I-1)(J-1)}$, where γ is the desired level of significance.

PROBLEMS

14.12.1 Twelve overweight subjects participated in a study comparing the effectiveness of three weight-reducing diets. The subjects were grouped according to their initial weight, and each of three subjects from each initial weight group was randomly assigned to a diet. The weight loss (in pounds) at the end of the experimental period is given below:

	Diet		
Initial Weight	A	B	C
150–174	10	23	24
175–199	12	21	26
200–224	12	31	21
Over 224	20	28	33

(i) Do these data provide sufficient evidence that (after eliminating the effect of initial weight) the diets are different in their effectiveness? Use $\alpha = 0.01$. **(ii)** Does the initial weight affect the loss of weight?

14.12.2 The nutritional value of a certain vegetable is measured on 18 specimens grown in two varieties in three geographical regions. The data are as follows:

	Geographical Region		
Variety	A	B	C
	6.3	9.2	6.8
1	11.5	5.1	7.2
	9.2	8.1	5.5
	11.0	5.4	7.1
2	7.3	5.0	7.8
	8.2	6.1	8.4

Study the effect of variety and geographical location on the nutritional value, taking the average of three observations in each cell as a response. Use the significance level $\alpha = 0.05$.

14.12.3 The model for three factors is $E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$, where $\sum \alpha_i = \sum \beta_j = \sum \gamma_k$. Assume that Y_{ijk} is normally distributed with variance σ^2. Derive the tests for the hypotheses analogous to those in the case of two factors.

14.13 ANOVA MODELS WITH INTERACTION

The model considered in Section 14.12 was of the form $E(Y_{ij}) = \mu + \alpha_i + \beta_j$, by which we assume that the effects of the factors A and B are additive. The hypotheses $H_0^{(A)}$ and $H_0^{(B)}$ are tested within this model.

In general, the effects of A and B need not be additive, and one might wish to consider the case where $E(Y_{ij})$ is an entirely arbitrary function of i and j. Such a function can always be represented in the form

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

where $\sum_i \alpha_i = \sum_j \beta_j = \sum_{ij} \gamma_{ij} = 0$ for all i, j (we leave the proof as an exercise). The constants γ_{ij} are referred to as *interaction* terms. Testing for the presence of interaction is based on an idea similar to that used in testing linearity of regression. It requires a model-independent estimate of σ^2, which in turn can be achieved if we have more than one observation for each combination of levels of factors A and B. We assume therefore that the data now have the form of a three-dimensional array $y_{ijk}, i = 1, \ldots, I; j = 1, \ldots J; k = 1, \ldots, K$, where $I \geq 2, J \geq 2, K \geq 2$. Leaving the details of the derivation as an exercise, we can write the decomposition of the sum of squared deviations of the random variables Y_{ijk} from their means as

$$\begin{aligned} S^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2 \qquad (14.27) \\ &= IJK(\overline{Y}_{+++} - \mu)^2 \\ &\quad + \sum_{i=1}^{I} JK(\overline{Y}_{i++} - \overline{Y}_{+++} - \alpha_i)^2 + \sum_{j=1}^{J} IK(\overline{Y}_{+j+} - \overline{Y}_{+++} - \beta_j)^2 \\ &\quad + \sum_{i=1}^{I}\sum_{j=1}^{J} K(\overline{Y}_{ij+} - \overline{Y}_{i++} - \overline{Y}_{+j+} + \overline{Y}_{+++} - \gamma_{ij})^2 \\ &\quad + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij+})^2. \end{aligned}$$

As before, the last sum (upon division by σ^2) has the chi-square distribution with $IJ(K-1)$ degrees of freedom, given only normality and homoscedasticity (regardless of any other hypotheses tested). Thus, it can serve as a denominator in all F ratios used for testing, while the two single sums and the double sum can be used to

test the hypotheses about the effects of A, of B, and of their interaction. To put it differently, each of the sums above, divided by its number of degrees of freedom, is an unbiased estimator of σ^2 if the appropriate hypothesis is true (except for the last sum, for which this holds regardless of any hypothesis). Letting

$$
\begin{aligned}
SSM &= IJK(\overline{y}_{+++})^2, \\
SSA &= \sum_{i=1}^{I} JK(\overline{y}_{i++} - \overline{y}_{+++})^2, \\
SSB &= \sum_{j=1}^{J} IK(\overline{y}_{+j+} - \overline{y}_{+++})^2, \\
SSAB &= \sum_{i=1}^{I}\sum_{j=1}^{J} K(\overline{y}_{ij+} - \overline{y}_{i++} - \overline{y}_{+j+} + \overline{y}_{+++})^2, \\
SSR &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \overline{y}_{ij+})^2, \\
SST &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2,
\end{aligned}
$$

the corresponding ANOVA table is as follows:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F Ratio
Mean	1	SSM	$MSM = \frac{SSM}{1}$	$F_M = \frac{MSM}{MSR}$
A	$I-1$	SSA	$MSA = \frac{SSA}{I-1}$	$F_A = \frac{MSA}{MSR}$
B	$J-1$	SSB	$MSB = \frac{SSB}{J-1}$	$F_B = \frac{MSB}{MSR}$
AB	$(I-1)(J-1)$	$SSAB$	$MSAB = \frac{SSI}{(I-1)(J-1)}$	$F_{AB} = \frac{MSI}{MSR}$
Residual	$IJ(K-1)$	SSR	$MSR = \frac{SSR}{IJ(K-1)}$	
Total	$N = IJK$	SST		

■ EXAMPLE 14.16

A researcher studies the effects of sex and type of stimulus ("soothing" or "exciting") on the aggressive behavior of parrots. Six male and six female birds of a given species are each placed in a separate cage isolated from other cages. The six birds of a given gender are randomly divided into two groups. The cages are covered for the night. Before uncovering the cage in the morning, the birds hear a tape. One is a "soothing" tape, with the voice of the experimenter talking quietly to the birds. The other tape contains angry voice of the experimenter. The observed value Y is the number of times the bird attacks the experimenter's hand when she uncovers the cage and puts food into the

plate. The data are as follows (the three numbers in each cell represent the data for three birds):

	Soothing (S)	Angry (A)
M	8 6 13	22 28 33
F	5 10 6	12 14 9

The within-cell averages $\overline{y}_{ij+}$ and row and column averages Y_{i+}, y_{+j}, as well as y_{++} are

	S	A	
M	9	27.67	18.33
F	7	11.67	9.33
	8	19.67	13.83

The sum SSR of the squared deviations from cell means is $SSR = 1113.33$. We therefore have the next table:

	df	SS	MS	F
Mean	1	2296.33	2296.33	162.10
Gender	1	243.00	243.00	17.15
Stimulus	1	408.33	408.33	28.82
Interaction	1	147.00	147.00	10.38
Residuals	8	113.33	14.17	

Since $F_{0.05,1,8} = 5.318$, we can conclude that both the gender of the parrot and the type of stimulus have an effect, and there is interaction between gender and type of stimulus at the 0.05 significance level. The mean response is significantly different than zero.

To better grasp the meaning of interaction, we will use a graphical representation. Let us arrange the categories of one factor, say A, along the horizontal axis (there may be no numerical values attached, and the categories need not have any "natural" order). Then along the horizontal axis we can plot average responses $\overline{y}_{ij+}$ for various j. The effects of A, B and of their interaction can now be interpreted as follows (see Figure 14.4). The effect of A means that at least one of the curves differs significantly from the horizontal line (last two figures). The effect of B, but not of any interaction, means that the curves for various levels of B are parallel to one another but significantly different from one another (Figure 14.4b and c). Finally, the interaction reveals itself by the lack of parallelism of the curves (Figure 14.4d).

PROBLEMS

14.13.1 The scores measuring the level of emotional maturity of young adult males classified by age and extent of use of marijuana are given below.

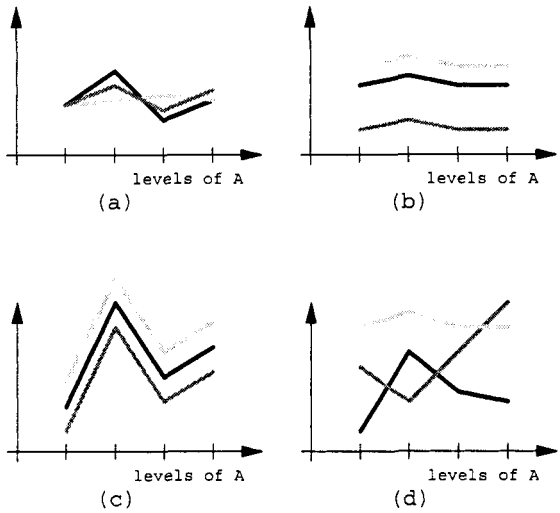


Figure 14.4 (a) No effects of A, B or AB. (b) No effect of A, effect of B, no interaction. (c) Effect of A, no effect of B, no interaction. (d) Effect of A, effect of B, effect of interaction.

Age	Marijuana Use: Never	Occasionally	Daily
	25	18	17
15–19	28	23	24
	22	19	19
	28	16	18
20–24	32	24	22
	30	20	20
	25	14	10
25–29	35	16	8
	30	15	12

Test for effects of age, extent of use, and their interaction. Use $\alpha = 0.05$

14.13.2 Use data given in Problem 14.12.2 to test for the existence of interaction between variety and geographical notation.

14.13.3 Show that if $E(Y_{ij}) = c(i,j) = \mu + \alpha_i + \beta_j + \gamma_{ij}$, where $i = 1, \ldots, I, j = 1, \ldots, J$, then one can always find constants μ, α_i, β_j, and γ_{ij} such that

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij} = 0.$$

14.14 FURTHER EXTENSIONS

The ANOVA models presented in the last three sections can easily be extended to the case of more than two factors, with or without interaction, provided that we consider completely balanced designs. Each level of a factor can be combined with all combinations of levels of other factors, and in each cell we have the same number of observations. Testing for interactions is possible only if the number of observations per cell is at least two. The number of observations needed very quickly becomes unattainable practically. For instance, with three factors, each on five levels, we need $5^3 = 125$ observations, and twice that if we want to test for interactions.

This situation led to research in two major directions. The first was to invent experimental schemes that allow testing for the presence of effects (as well as estimation of those effects) with as small a number of experiments as possible. If one resigns from the stringent requirement that every level of a factor has to appear with every combination of levels of other factors, there are many possibilities of experiments (e.g., forming the so-called Latin or Greco-Latin squares). To use a simple example, imagine that we have three factors, each appearing on five levels. Representing one factor as a row, the other as a column, and the third as letter (with levels a, b, c, d, e), we can arrange the experiment as follows:

a	b	c	d	e
b	c	d	e	a
c	d	e	a	b
d	e	a	b	c
e	a	b	c	d

Each level of the first factor (row) combines exactly once with each level of the second factor (column), and exactly once with the third factor (letters). The same is true for the other two factors. However, of $5^3 = 125$ possible combinations, only 25 actually appear.

It is possible to include here the fourth factor (e.g., Greek letters, $\alpha, \beta, \gamma, \delta, \epsilon$) as follows:

$a\alpha$	$b\beta$	$c\gamma$	$d\delta$	$e\epsilon$
$b\epsilon$	$c\alpha$	$d\beta$	$e\gamma$	$a\delta$
$c\delta$	$d\epsilon$	$e\alpha$	$a\beta$	$b\gamma$
$d\gamma$	$e\delta$	$a\epsilon$	$b\alpha$	$c\beta$
$e\beta$	$a\gamma$	$b\delta$	$c\epsilon$	$d\alpha$

Now each row and each column has exactly one of the Roman letters and exactly one of the Greek letters, and each Roman letter is combined exactly once with each Greek letter.

One can therefore plan an agricultural experiment (say) in which in each row are plants of one of five varieties of seed, and in each column one of the five varieties

of fertilizers. Then each Roman letter would correspond to one of the five different amounts of watering, and each Greek letter to one of the five different times of planting.

There are seemingly countless variety of experimental designs to cover all contingencies that occur in practice, each design with its own testing or estimation procedures. By introducing appropriate criteria, one can search for designs that optimize these criteria.

CHAPTER 15

RANK METHODS

15.1 INTRODUCTION

This chapter is devoted to statistical tests applicable to data measured on ordinal scales only and more precisely, when only ordinal relations are taken into account. These methods[32] fall under the general heading "nonparametric statistics," and they involve techniques based on ranks, which offer several advantages. While the data sampled must be from a continuous distribution, the type of the distribution is irrelevant. Nonparametric methods are almost as powerful as normal theory methods. They can therefore be applied when a population distribution is not normal (e.g., it is skewed) or if the sample size is not large enough for testing normality.

■ **EXAMPLE 15.1**

Suppose that four observations randomly selected from some population are $x_1 = 8, x_2 = 7, x_3 = 3, x_4 = 15$, and five observations from another population are $y_1 = 5, y_2 = 4, y_3 = 0, y_4 = 6, y_5 = 1$. If the data are expressed on the ordinal (but not any higher) scale, these data contain the same information as such data as $x_1 = 1000, x_2 = 32, x_3 = 30, x_4 = 1001, y_1 = 35, y_2 =$

[32] See Hollander and Wolfe (1999) for an exhaustive presentation.

$31, y_3 = -5, y_4 = 36, y_5 = 0$. In effect all relevant information is contained in the sequence

$$YYXYYYXXX,$$

which tells us that the observations from the first sample occupy places 3, 7, 8, and 9 in the ordered combined sample.

Methods based on ranks *should* be used for the data measured on an interval scale. They, however, *can* be used for the data measured on an interval or ratio scale. In other words, one can always use methods for a weaker scale.

The assumption of continuity is what distinguishes the methods of this chapter from the methods of categorical data that sometimes are also of an ordinal character (discussed in Chapter 16) and exemplified by questionnaire responses such as "strongly agree," "agree," "neutral," "disagree," "strongly disagree." The essential difference is that under the assumption of a population distribution being continuous, we can disregard ties in our theoretical considerations. Practically, it means that ties in the data are rather exceptional.

We begin with the study of the behavior of empirical cdf's. The empirical cdf converges almost surely and uniformly to the cdf of the underlying random variable. This is a general fact, and it is true regardless of whether or not the random variable is continuous. If we restrict the analysis to the continuous case, we can construct a test for the hypothesis that the random sample $X_1, \ldots, X_n$ comes from a population with a specific continuous cdf.

Next we will consider the two-sample problem, tests of the hypothesis that two random samples $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$ come from the same continuous distribution. We will also introduce tests for randomness and present procedures for testing hypotheses about population medians in one, two, and more than two populations. Last, we will present an ANOVA type of inference in case of response distribution that is other than normal.

15.2 GLIVENKO-CANTELLI THEOREM

Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be a random sample from some population. Specifically, this means that $X_1, \ldots, X_n$ is the initial fragment of a sequence $\mathbf{X}_\infty = \{X_i, i = 1, 2, \ldots\}$ of iid random variables. We let F denote the common cdf of X_i's so that $F(t) = P\{X_i \leq t\}$. At the moment, we can make no assumptions about the nature of F. We only know that F is an arbitrary nondecreasing right-continuous function, satisfying

$$\lim_{t \to \infty} F(t) = 1, \qquad \lim_{t \to -\infty} F(t) = 0$$

(see Chapter 6). For any n and t, we define the empirical cdf of the sample $X_1, \ldots, X_n$ as

$$F_n(t) = \frac{\text{number of data values that do not exceed } t}{n}.$$

Thus $F_n(t)$ is a step function that increases by $1/n$ (or by a multiple of $1/n$) at each point of the sample. It is important to keep in mind that $F_n(t)$ is a random function. It depends formally on the sequence $\mathbf{X}_\infty$ but in fact only on the first n observations.

If we fix the value t, then $nF_n(t)$ is a random variable, equal to the number of X_i's among the first n observations, satisfying the condition $X_i \leq t$. Thus $nF_n(t)$ has a binomial distribution with parameters n and $p = P\{X_i \leq t\} = F(t)$. Consequently, by Theorem 10.5.19, for every t,

$$F_n(t) = \frac{nF_n(t)}{n} \to F(t) \tag{15.1}$$

with probability 1.

To appreciate the meaning of this result, and the meaning of its extension below, let us write explicitly $F_n(t, \mathbf{X})$, where $\mathbf{X} = \mathbf{X}_\infty$. Since $\mathbf{X}$ is the element of the sample space corresponding to sampling an infinite sequence of values of iid random variables, each with cdf F, the phrase "with probability 1" or "almost certainly" means "for all sequences $\mathbf{X}$, except sequences in a set $\mathcal{N}$ with $P(\mathcal{N}) = 0$."

Specifically, (15.1) means that for every t there exists a set $\mathcal{N}_t$ of sequences $\mathbf{X}$ such that $P(\mathcal{N}_t) = 0$ and if $\mathbf{X} \notin \mathcal{N}_t$, then

$$\lim_{n\to\infty} F_n(t, \mathbf{X}) = F(t). \tag{15.2}$$

The following theorem is an extension of this result to the convergence for all t at once, and uniform:

Theorem 15.2.1 (Glivenko-Cantelli) *As $n \to \infty$, we have*

$$\sup_t |F_n(t) - F(t)| \to 0$$

with probability 1.

Proof. The theorem means that one can find a set $\mathcal{N}$ of sequences $\mathbf{X}$ such that $P(\mathcal{N}) = 0$ and $\sup_t |F_n(t, \mathbf{X}) - F(t)| \to 0$ as $n \to \infty$ for all $\mathbf{X} \notin \mathcal{N}$. We show first that such a common set $\mathcal{N}$ exists for convergence $F_n(t)$ to $F(t)$ for all t (convergence being not necessarily uniform). Let A be the set of all points at which F is discontinuous. Since F is nondecreasing and continuous on the right, the condition $t \in A$ means that $F(t) - F(t-) > 0$, where

$$F(t-) = \lim_{\tau\to t^-} F(\tau) = \sup_{\tau<t} F(\tau).$$

If $t \in A$, we define the random variables

$$U_n(t) = U_n(t, \mathbf{X}) = \begin{cases} 1 & \text{if } X_n = t \\ 0 & \text{if } X_n \neq t, \end{cases} \tag{15.3}$$

so that $E[U_n(t)] = P\{U_n(t) = 1\} = F(t) - F(t-)$. Clearly, the random variables $U_1(t), U_2(t), \ldots$ are iid, and by Theorem 10.5.19, we have

$$\frac{1}{n}\sum_{j=1}^{n} U_j(t, \mathbf{X}) \to F(t) - F(t-)$$

for all $\mathbf{X} \notin \mathcal{N}_t^*$, where $P(\mathcal{N}_t^*) = 0$. Here $\mathcal{N}_t^*$ is the "exceptional" set chosen for specific t.

Clearly, the set A is at most countable. Indeed, if $A_k \subset A$ is the set of points $t \in A$ with $F(t) - F(t-) \geq 1/k$, then A_k has at most k elements because of monotonicity of F and the condition $F(\infty) - F(-\infty) = 1$. Since $A = \bigcup_k A_k$, the set A is at most countable, and the condition (15.3) holds for all $\mathbf{X} \notin \bigcup_{t\in A} \mathcal{N}_t^*$, with

$$P(\bigcup_{t\in A} \mathcal{N}_t^*) \leq \sum_{t\in A} P(\mathcal{N}_t^*) = 0.$$

Now let Q be the set of all rational t (or any other countable dense subset of real line). We may assert that $F_n(t) \to F(t)$ for all $t \in Q$ with probability 1, since by (15.2) we have $F_n(t, \mathbf{X}) \to F(t)$ for all t if $\mathbf{X} \notin \bigcup_{t\in Q} \mathcal{N}_t$, and again $P(\bigcup_{t\in Q} \mathcal{N}_t) \leq \sum_{t\in Q} P(\mathcal{N}_t) = 0$.

Thus we showed that with probability 1 we have $F_n(t) \to F(t)$ for all $t \in Q$ and $F_n(t) - F_n(t-) \to F(t) - F(t-)$ for all $t \in A$. We will next show that with probability 1 we also have $F_n(t) \to F(t)$ for all $t \notin A$, and that the convergence is uniform. These last statements do not involve any probability considerations: They are true for any $\mathbf{X} \notin (\bigcup_{t\in A} \mathcal{N}_t^*) \cup \left(\bigcup_{t\in Q} \mathcal{N}_t\right)$, so one may suppress the dependence on $\mathbf{X}$ and consider a fixed sequence of cdf's $F_n(\cdot)$.

Thus let t be a continuity point of F, and let $t \notin Q$. We want to show that $F_n(t) \to F(t)$. Let $\epsilon > 0$, and let $t_1, t_2 \in Q$ satisfy the relations $t_1 < t < t_2$ and

$$F(t_2) - F(t_1) < \frac{\epsilon}{2} \tag{15.4}$$

(which is possible because t is a continuity point of F). Next we choose N such that for $n \geq N$,

$$|F_n(t_1) - F(t_1)| < \frac{\epsilon}{2} \quad \text{and} \quad |F_n(t_2) - F(t_2)| < \frac{\epsilon}{2}. \tag{15.5}$$

This is possible because $F_n(t_i) \to F(t_i)$ as $n \to \infty$ for $i = 1, 2$. By monotonicity of F we have $F(t_1) \leq F(t) \leq F(t_2)$; hence by (15.4),

$$F(t_2) - \frac{\epsilon}{2} < F(t) < F(t_1) + \frac{\epsilon}{2}.$$

Using (15.4), (15.5), and monotonicity of $F_n(\cdot)$, we can write for $n \geq N$,

$$F(t) \leq F(t_2) \leq F(t_1) + \frac{\epsilon}{2} \leq \left(F_n(t_1) + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2} \leq F_n(t) + \epsilon$$

and also

$$F(t) \geq F(t_2) - \frac{\epsilon}{2} \geq \left(F_n(t_2) - \frac{\epsilon}{2}\right) - \frac{\epsilon}{2} \geq F_n(t) - \epsilon,$$

which gives $|F_n(t) - F(t)| \leq \epsilon$. It remains to prove that the convergence $F_n(t) \to F(t)$ is uniform, that is, $\sup_t |F_n(t) - F(t)| \to 0$ as $n \to \infty$. The proof is after Chung (2001).

Now assume the contrary. There exists $\epsilon_0 > 0$, and sequences n_k and t_k such that $|F_{n_k}(t_k) - F(t_k)| > \epsilon_0$ for all t. Since $F_n(\cdot)$ and $F(\cdot)$ are cdf's, we cannot have $t_k \to +\infty$ or $t_k \to -\infty$; hence the sequence $\{t_k\}$ is bounded. Without loss of generality we can assume that $t_k \to t^*$.

Moreover, there exists either a subsequence of $\{t_k\}$ which converges to t^* monotonically from below or monotonically from above. Similarly, since $|F_{n_k}(t_k) - F(t_k)| > \epsilon_0$, there exists either a subsequence at which $F_{n_k}(t_k) > F(t_k) + \epsilon_0$ or a subsequence at which $F_{n_k}(t_k) < F(t_k) - \epsilon_0$. Restricting the analysis to these subsequences, we distinguish four cases:

1. $t_k \uparrow t^*, t_k < t^*, F_{n_k}(t_k) > F(t_k) + \epsilon_0$.

2. $t_k \uparrow t^*, t_k < t^*, F_{n_k}(t_k) < F(t_k) - \epsilon_0$.

3. $t_k \downarrow t^*, F_{n_k}(t_k) > F(t_k) + \epsilon_0$.

4. $t_k \downarrow t^*, F_{n_k}(t_k) < F(t_k) - \epsilon_0$.

We will select $t', t'' \in Q$ such that $t' < t^* < t''$.

In case 1, using the monotonicity of F_{n_k} and F, we can write, for all k sufficiently large,:

$$\begin{aligned} \epsilon_0 &< F_{n_k}(t_k) - F(t_k) \le F_{n_k}(t^*-) - F(t') \\ &\le F_{n_k}(t^*-) - F_{n_k}(t^*) + F_{n_k}(t'') - F(t'') + F(t'') - F(t'). \end{aligned}$$

If we let $k \to \infty$, the difference $F_{n_k}(t^*-) - F_{n_k}(t^*)$ converges to $-(F(t^*) - F(t^*-))$, and the difference $F_{n_k}(t'') - F(t'')$ converges to 0. Letting $t'' \downarrow t^*, t' \uparrow t^*$ along values in Q, the last difference converges to $F(t^*) - F(t^*-)$. Thus the right-hand side can be made as small as possible, which gives a contradiction.

In case 2, we have

$$\epsilon_0 < F(t_k) - F_{n_k}(t_k) < F(t^*-) - F_{n_k}(t') \le F(t^*-) - F(t') + F(t') - F_{n_k}(t').$$

Letting $k \to \infty$, we obtain $F(t') - F_{n_k}(t') \to 0$, and letting next $t' \uparrow t^*$, we obtain $F(t^*-) - F(t') \to 0$, which gives another contradiction.

In case 3, we have

$$\begin{aligned} \epsilon_0 &< F_{n_k}(t_k) - F(t_k) < F_{n_k}(t'') - F(t^*) \\ &\le F_{n_k}(t'') - F_{n_k}(t') + F_{n_k}(t') - F(t') + F(t') - F(t^*). \end{aligned}$$

Now $F_{n_k}(t'') - F_{n_k}(t') \to F(t'') - F(t')$, while $F_{n_k}(t') - F(t') \to 0$. We let $t'' \downarrow t^*$ and use the continuity of F on the right, which results in the term $F(t'') - F(t')$ converging to $F(t^*) - F(t')$. So we again have a contradiction.

Finally, in case 4, we can write

$$\begin{aligned} \epsilon_0 &\le F(t_k) - F_{n_k}(t_k) \le F(t'') - F_{n_k}(t^*) \\ &\le F(t'') - F(t') + F(t') - F_{n_k}(t') + F_{n_k}(t^*-) - F_{n_k}(t^*). \end{aligned}$$

As $k \to \infty$ we have $F(t') - F_{n_k}(t') \to 0$, while $F_{n_k}(t^*-) - F_{n_k}(t^*) \to F(t^*-) - F(t^*)$. Letting now $t' \uparrow t^*$ and $t'' \downarrow t^*$, we have $F(t'') - F(t') \to F(t^*) - F(t^*-)$, which also leads to contradiction.

PROBLEMS

15.2.1 Let $F(t) = 0$ for $t < 0$, $F(t) = p$ for $0 \le t < 1$, and $F(t) = 1$ otherwise. Use the central limit theorem to evaluate directly the distribution of $\sup_t |F_n(t) - F(t)|$, and show that $F_n(t)$ tends to $F(t)$ almost surely and uniformly in t.

15.2.2 Let X be a random variable such that $P\{X = a\} = 1/2, P\{X = b\} = 1/3$ and $P\{X = c\} = 1/6$. Let $X_1, X_2, \ldots, X_{200}$ be a random sample from the distribution of X. Suppose that among the first 100 observation of X_i, 55 were equal to a and 38 were equal to b. Among the next 100 observations, 51 are equal to a, and 30 are equal to b. Determine $\sup_t |F_{200}(t) - F(t)|$ and $\sup_t |F_{100}(t) - F(t)|$ in all six cases $a < b < c$, $a < c < b$, $b < a < c$, and so on, as well as in all cases such as $b = c < a, a = b < c, \ldots$, and $a = b = c$.

15.3 KOLMOGOROV-SMIRNOV TESTS

One-Sample Kolmogorov-Smirnov Test

Let

$$D_n = \sup_t |F_n(t) - F(t)|. \tag{15.6}$$

The distance D_n converges a.s. to zero if the F_n's are empirical cdf's of random samples drawn from distribution F. Research on the rate of this convergence, that is, finding constants $c_n \to \infty$ such that the sequence of random variables $\{c_n D_n, n \ge 1\}$ has a limiting distribution, led to a remarkable discovery, due to Kolmogorov and Smirnov: if F is continuous, then this limiting distribution exists for $c_n = \sqrt{n}$, and moreover, it does not depend on F.

The following theorem specifies this limiting distribution, and serves as a foundation for a test of the hypothesis that a random sample comes from a specific distribution. We will omit the proof here.

Theorem 15.3.1 (Kolmogorov and Smirnov) *Let $X_1, X_2, \ldots$ be iid random variables, with a continuous cdf F. Then for every $z > 0$,*

$$\lim_{n\to\infty} P\{\sqrt{n} D_n \le z\} = Q(z),$$

where

$$Q(z) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 z^2}. \tag{15.7}$$

The function $Q(z)$ is a cdf of a continuous distribution called a *Kolmogorov distribution*. The values of this cdf are given in Table A7.

Suppose now that we want to test the hypothesis

$$H_0 : F = F_0 \qquad \text{against the alternative} \qquad H_1 : F \ne F_0,$$

where the hypothetical cdf F_0 is continuous. Under the null hypothesis, for large n, the distribution of $\sqrt{n} D_n$ is given by $Q(x)$. If the true distribution of X_i's is F^*, we

can write

$$\begin{aligned}\sup_t |F_n(t) - F_0(t)| &\leq \sup_t \{|F_n(t) - F^*(t)| + |F^*(t) - F_0(t)|\} \\ &\leq \sup_t |F_n(t) - F^*(t)| + \sup_t |F^*(t) - F_0(t)| \\ &= D_n + \sup_t |F^*(t) - F_0(t)|.\end{aligned}$$

Upon multiplication by $\sqrt{n}$, the first term, $\sqrt{n}D_n$, has the limiting distribution (15.7), while the second term $\sqrt{n}\sup_t |F^*(t) - F_0(t)|$ tends to infinity if $F^*(t) \not\equiv F_0(t)$.

The last property means that the test should reject H_0 if the observed value of the statistic $\sqrt{n}D_n$ exceeds the critical value determined from the right tail of the distribution (15.7). This test has—in the limit—power 1 against *any* alternative.

■ **EXAMPLE 15.2**

A small town had 30 fires last year: on January 5 and 18, February 3, 4, 21, and 26, March 5, 10, and 13, April 6, May 16 and 25, June 19, July 10 and 21, August 12 and 15, September 1, 8, and 21, November 2, 6, 7, 19, and 29, December 3, 9, 12, 17, and 24. Are these data consistent, on the 0.05 level, with the hypothesis that the occurrences of fires follow a Poisson process?

SOLUTION. One of the solutions may be based on the fact that if the fires form a Poisson process, then they occur throughout the year according to the uniform distribution. The data are summarized in Table 15.1

Table 15.1

Fire i	Day x_i	$F_n(x_i) = i/30$	$F(x_i) = x_i/365$	Fire i	Day x_i	$F_n(x_i) = i/30$	$F(x_i) = x_i/365$
1	5	0.033	0.014	16	224	0.533	0.614
2	18	0.067	0.049	17	227	0.567	0.622
3	34	0.100	0.093	18	244	0.600	0.668
4	35	0.133	0.096	19	251	0.633	0.688
5	52	0.167	0.142	20	264	0.667	0.723
6	57	0.200	0.156	21	306	0.700	0.838
7	64	0.233	0.175	22	310	0.733	0.849
8	69	0.267	0.189	23	311	0.767	0.852
9	72	0.300	0.197	24	323	0.800	0.885
10	96	0.333	0.263	25	333	0.833	0.912
11	136	0.367	0.373	26	337	0.867	0.923
12	145	0.400	0.397	27	343	0.900	0.940
13	170	0.433	0.466	28	346	0.933	0.948
14	191	0.467	0.523	29	351	0.967	0.962
15	202	0.500	0.553	30	363	1.000	0.995

Since $F_n(\cdot)$ is a step function, it suffices to inspect only the differences $|F(x_i) - F_n(x_i)|$ and $|F(x_i) - F_n(x_{i-1})|$. The largest difference in Table 15.1

is $F(x_{21}) - F_{30}(x_{20}) = 0.838 - 0.667 = 0.171$, and $\sqrt{n}D_n = \sqrt{30} \times 0.171 = 0.937$. Since (see Table A7.) $P\{\sqrt{n}D_n > 0.9\} = 1 - 0.6073 = 0.3927$, there is no reason to reject the null hypothesis that the fires form a Poisson process.

It is important to realize that the Kolmogorov-Smirnov test applies only when the null hypothesis specifies completely the distribution F. The test cannot be used in cases where the null hypothesis specifies only the distribution type but does not provide the values of the parameters. If one estimates parameters from the sample to obtain a specific distribution F^*, $F^*(x)$ is then a *random* cdf that depends on the sample, and consequently the limiting distribution of $\sqrt{n}D_n^* = \sqrt{n}\sup_x |F_n(x) - F^*(x)|$ will *not* given by formula (15.7).

■ EXAMPLE 15.3

In this application of the Kolmogorov-Smirnov test we will regard points distributed uniformly on a circle. The quality of the estimation depends heavily on how good the method is that is used for generating random points in the circle.

Without loss of generality, we can assume that the circle is $x^2 + y^2 \leq 1$. Letting ξ_1, ξ_2 be a pair of independent random variables distributed uniformly on $[0, 1]$, we let $x = 2\xi_1 - 1, y = 2\xi_2 - 1$ and then check whether $x^2 + y^2 \leq 1$. If yes, the pair (x, y) is accepted. If not, we sample a new pair (ξ_1, ξ_2), and proceed in this way until we obtain an accepted pair (x, y). One can continue this process to generate as many accepted pairs as needed. The distribution of (x, y) is easily seen to be uniform on the circle with radius 1 centered at the origin, as long as ξ_1, ξ_2 are iid uniform on $[0, 1]$.

To "save computer time" it was proposed that when $x^2 + y^2 > 1$, rather than rejecting both x and y, the value of x is kept, and the computer generates a new ξ_2, transforms it to a new $y = 2\xi_2 - 1$, and tests if $x^2 + y^2 \leq 1$ (i.e., once x is sampled, it is retained, and only y is added to it).

For an estimation of the savings in the computer time, and for the distribution of the resulting points (x, y), see Problem 15.3.3. To test the uniformity of the distribution, $n = 100$ points are generated on the circle according to a "time-saving" scheme. Two tests are performed, both reducing the problem to a one-dimensional Kolmogorov-Smirnov test. First, the angles are measured from an arbitrary direction; the angles between this direction and the line connecting point (x, y) with the origin are uniformly distributed on $[0, 2\pi]$. Next, the positive x-axis is the direction chosen, and, counting counterclockwise, the values $\varphi_i = \arctan(x_i/y_i) + (\pi/2)[1 - \text{sign}(y_i)]$ take on a $U[0, 2\pi]$ distribution. The ordered values φ_i are shown in the left part of Table 15.2.

The empirical cdf of these values increases by 0.01 at each $\varphi_{i:100}$, while $F(t) = t/2\pi$. For our data, $\sqrt{n}D_n = \sqrt{n}\sup|F_n(t) - F(t)|$ equals 1.136, which corresponds to a p-value of about 0.15 (see Table A7.).

On the other hand, if the points have distribution uniform on the circle with radius 1, the distance R of a random point from the center has a cdf $F_R(t) = P\{R \leq t\} = \pi t^2/\pi = t^2$ for $0 \leq t \leq 1$. Thus the theoretical cdf of distance R is a parabola t^2. The right side of Table 15.2 gives the ordered dis-

tances. We can then compute the values of the Kolmogorov-Smirnov statistic $\sqrt{n}D_n = \sqrt{n}\sup|F_n(t) - t^2|$, which equals 1.737; the corresponding p-value (see Table A7.) is less than 0.005.

Table 15.2

$\varphi_{i:100}$			$R_{i:100}$		
0.0537	1.7784	3.7217	0.0432	0.6757	0.8811
0.0570	1.8854	3.8933	0.1238	0.6969	0.8827
0.0571	1.8967	3.9110	0.1411	0.7045	0.8834
0.1170	1.9798	3.9215	0.1872	0.7145	0.8846
0.1242	2.0339	4.0757	0.2365	0.7168	0.9027
0.1525	2.0721	4.0778	0.2774	0.7268	0.9104
0.3353	2.1705	4.1081	0.3865	0.7312	0.9134
0.4496	2.1833	4.1494	0.4014	0.7326	0.9169
0.4866	2.2170	4.1922	0.4036	0.7365	0.9200
0.5360	2.3099	4.3407	0.4125	0.7462	0.9223
0.5655	2.3482	4.5311	0.4262	0.7547	0.9254
0.5957	2.4588	4.5437	0.4457	0.7557	0.9278
0.6237	2.4618	4.5571	0.4905	0.7833	0.9333
0.6343	2.5279	4.7786	0.4966	0.7836	0.9349
0.7253	2.5874	4.8246	0.5178	0.7924	0.9421
0.8785	2.5840	5.0184	0.5298	0.7967	0.9443
0.8945	2.8752	5.2162	0.5351	0.8175	0.9460
0.9055	2.9165	5.2770	0.5371	0.8268	0.9468
0.9839	2.9489	5.3605	0.5701	0.8290	0.9494
1.0182	2.9534	5.3661	0.5729	0.8311	0.9496
1.1167	2.9606	5.4251	0.5734	0.8337	0.9503
1.1671	2.9669	5.6105	0.5788	0.8412	0.9536
1.2113	2.9774	5.7050	0.5865	0.8453	0.9566
1.3437	3.0322	5.8654	0.5888	0.8472	0.9591
1.4048	3.0663	5.8959	0.6185	0.8476	0.9616
1.4319	3.1596	6.0020	0.6232	0.8485	0.9621
1.4435	3.2035	6.0355	0.6251	0.8491	0.9687
1.4962	3.3165	6.1418	0.6261	0.8521	0.9709
1.5223	3.3226	6.1496	0.6279	0.8632	0.9795
1.5231	3.4023	6.1781	0.6355	0.8656	0.9815
1.5716	3.4204	6.2117	0.6509	0.8669	0.9880
1.5774	3.4407	6.2705	0.6556	0.8692	0.9988
1.6020	3.4981		0.6578	0.8713	
1.7543	3.5595		0.6717	0.8724	

To find the correct p-value of the observed result, we can perform two tests, one giving nonsignificant result corresponding to a p-value 0.15 and the other giving a highly significant result with a p-value of 0.005.

To see what is involved here, suppose that instead of the p-value we just carry out two tests on significance levels α_1 and α_2, respectively. We decide to *reject* H_0 *if either test 1 or test 2 rejects it*. Consequently the probability of

type 1 error equals

$$\begin{aligned}\alpha &= P\{\text{test 1 rejects or test 2 rejects}|H_0\} \\ &= P\{\text{test 1 rejects}|H_0\} + P\{\text{test 2 rejects}|H_0\} - P\{\text{both tests reject}|H_0\} \\ &= \alpha_1 + \alpha_2 - P\{\text{both tests reject}|H_0\}.\end{aligned}$$

If test 1 and test 2 are performed on two independent samples, then the attained significance level is

$$\alpha = \alpha_1 + \alpha_2 - \alpha_1\alpha_2. \tag{15.8}$$

In the present case, the testing was performed on the same data set. Normally the determination of a significance level in such case is difficult because it requires knowledge of the joint distribution of the statistics used in both tests 1 and 2.

In the case under consideration, however, the situation is simple. Since φ and R are independent (under null hypothesis), we can apply (15.8), which gives the result 0.154.

Two-Sample Kolmogorov-Smirnov Test

In a two-sample problem we need to determine whether or not two samples come from populations with the same distribution. We have already encountered its special cases (e.g., in considering the Student's t test for equality of two normal distributions with the same variance). We now consider this problem in its generality, under the only assumption that the samples are drawn from continuous distributions. Thus we let $X_1, \ldots, X_m$ and $Y_1, \ldots, X_n$ be two independent random samples from distributions with continuous cdf's F and G, respectively. The objective is to test the hypothesis

$$H_0 : F = G \qquad \text{against} \qquad H_1 : F \neq G.$$

Unlike in the one-sample case of Section 15.3, the null hypothesis is now composite. The following theorem, also due to Kolmogorov and Smirnov, is presented without proof.

Theorem 15.3.2 *Let* $\mathbf{X}_m = (X_1, \ldots, X_m)$ *and* $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ *be random samples from distributions with continuous cdf's* F *and* G*, respectively, and let* $F_m(t)$ *and* $G_n(t)$ *be the respective empirical cdf's. Furthermore let*

$$D_{n,m} = \sup_t |F_m(t) - G_n(t)|.$$

Then

$$\lim_{m,n\to\infty} P\left\{\sqrt{\frac{mn}{m+n}} D_{m,n} \leq t\right\} = Q(t),$$

with $Q(t)$ *given by* (15.7).

If H_0 is true, then the statistic $\sqrt{\frac{mn}{m+n}} D_{m,n}$ has a limiting distribution given by the cdf $Q(t)$, regardless of the particular cdf that governs the sampling of X_i's and Y_j's. Suppose now that H_0 is false. As $m, n \to \infty$, we have $F_m(t) \to F(t)$ and

$G_n(t) \to G(t)$ almost surely and uniformly in t, by the Glivenko-Cantelli theorem. Then

$$\begin{aligned} D_{m,n} &= \sup_t |F_m(t) - G_n(t)| \\ &= \sup_t |F_m(t) - F(t) + F(t) - G(t) + G(t) - G_n(t)| \\ &\leq \sup_t |F_m(t) - F(t)| + \sup_t |F(t) - G(t)| + \sup_t |G_n(t) - G(t)|, \end{aligned}$$

and we can write

$$\begin{aligned} \sqrt{\frac{mn}{m+n}} D_{m,n} &\leq \sqrt{\frac{m}{1+\frac{m}{n}}} \sup_t |F_m(t) - F(t)| + \sqrt{\frac{mn}{m+n}} \sup_t |F(t) - G(t)| \\ &\quad + \sqrt{\frac{n}{1+\frac{n}{m}}} \sup_t |G_n(t) - G(t)| \\ &\leq \sqrt{m} \sup_t |F_m(t) - F(t)| + \sqrt{\frac{mn}{m+n}} \sup_t |F(t) - G(t)| \\ &\quad + \sqrt{n} \sup_t |G_n(t) - G(t)|. \end{aligned}$$

The two extreme terms in the last expression have limiting distributions, while the middle term tends to infinity for any single hypothesis contained in the alternative. Thus, again, the test rejects the null hypothesis when the value of the statistic $\sqrt{mn/(m+n)}D_{m,n}$ is large enough.

■ **EXAMPLE 15.4**

Is a Poisson process observed at every other event also a Poisson process?

SOLUTION. We know, of course, that the answer is negative: If we observe every other event in a Poisson process, then the inter-event times are sums of two exponential random variables and hence are not exponential. Let us verify this fact empirically. Table 15.3 gives $m = 12$ interarrival times in a Poisson process with mean 1 (where every event is observed) and $n = 16$ interarrival times for every other event in another Poisson process with mean $\frac{1}{2}$ (so the mean interarrival times are the same).

Column T_i gives the observed interarrival times in both samples, jointly ordered, while the next two columns give the values of $F_m(t)$ and $G_n(t)$ at the observed points—$F_m(t)$ increases at points from the first sample while $G_n(t)$ increases at points from the second sample. It may be seen therefore that the six shortest interarrival times are all in a Poisson process observed at every event. Here the value $D_{m,n}$ is 0.5 and the statistic $\sqrt{mn/(m+n)}D_{m,n}$ equals 1.309, corresponding to a p-value of about 0.065.

Table 15.3

T_i	$F_m(\cdot)$	$G_n(\cdot)$	T_i	$F_m(\cdot)$	$G_n(\cdot)$
0.049	0.083	0.000	0.942	0.750	0.375
0.198	0.166	0.000	0.969	0.750	0.437
0.237	0.250	0.000	1.033	0.833	0.437
0.259	0.333	0.000	1.094	0.833	0.500
0.310	0.416	0.000	1.375	0.833	0.562
0.352	0.500	0.000	1.392	0.833	0.625
0.381	0.500	0.062	1.555	0.833	0.687
0.546	0.500	0.125	1.625	0.833	0.750
0.547	0.500	0.187	1.697	0.833	0.812
0.569	0.583	0.187	2.019	0.916	0.812
0.801	0.583	0.250	2.065	1.000	0.812
0.803	0.666	0.250	2.114	1.000	0.875
0.878	0.750	0.250	2.244	1.000	0.937
0.895	0.750	0.312	2.534	1.000	1.000

PROBLEMS

15.3.1 Let the observed values $x_1, \ldots, x_n$ be such that $1/3 \leq x_i \leq 2/3$ for all i. What can be said about n, if the null hypothesis that the underlying distribution is U[0, 1] cannot be rejected by the Kolmogorov-Smirnov test at $\alpha = 0.05$ level?

15.3.2 Suppose that the data are as in Example 15.2, except that there were only 25 fires, none of them in November. Test that the fires occur according to the Poisson process.

15.3.3 Find the joint density of (X, Y) resulting from the "time-saving scheme" of Example 15.3. Find also the expected number of random variables ξ_i necessary to sample in order to obtain one pair (X, Y) under both schemes.

15.3.4 What is the minimal possible value of the statistic $D_{m,n}$ if k values of X_i precede the third in the magnitude value Y_j? If $m = 100, k = 30$, and $n = 200$, is there enough evidence to reject (at the level $\alpha = 0.05$) the null hypothesis that the distributions of X's and Y's are the same?

15.3.5 Suppose that out of 30 fires in Example 15.2 those on January 5 and 18, February 3 and 21, March 10, April 6, May 25, June 19, December 3 were caused by arson, and in the remaining cases arson was excluded. Use the Kolmogorov-Smirnov statistic to test the hypothesis that the occurrences of "arson" and "nonarson" fires within a year follow the same distribution.

15.3.6 Suppose that one sample contains $2m$ data points while the other contains $2m + k$ data points. The first $2m$ and the last $2m$ data points in the joint sample alternate between samples. Thus the ordered data has the form

$$\underbrace{YX \cdots YX}_{2m}\underbrace{YY \cdots Y}_{k}\underbrace{XY \cdots XY}_{2m}. \tag{15.9}$$

(i) For given m find k such that the Kolmogorov-Smirnov test will reject the hypothesis that both samples are drawn from the same population ($\alpha = 0.05$). **(ii)** Solve this problem if the string of k consecutive Y's occurs at the beginning of the joint ordering.

15.3.7 Assume that each of two samples contains $2m + k$ elements with the following ordering:

$$\underbrace{YX\cdots YX}_{2m}\underbrace{YY\cdots Y}_{k}\underbrace{XY\cdots XY}_{2m}\underbrace{XX\cdots X}_{k} \tag{15.10}$$

For given m find k such that the Kolmogorov-Smirnov test will reject the hypothesis that both samples are drawn from the same population ($\alpha = 0.05$).

15.4 ONE-SAMPLE RANK TESTS

We begin this section with the Wilcoxon signed rank test. This test, used for testing hypotheses about a location parameter in symmetric distributions (median), is an excellent example of the simplicity and versatility of nonparametric methods.

Assume that we have a random sample $X_1, \ldots, X_n$ from a continuous, symmetric distribution with a cdf F and a density f. Then there exists θ such that for any x,

$$F(\theta - x) = 1 - F(\theta + x).$$

Equivalently we can say that $f(\theta + x) = f(\theta - x)$ for any x.

The null hypothesis $H_0 : \theta = \theta_0$ will be tested against either the one-sided alternative $H_1 : \theta > \theta_0$ or the two-sided alternative $H_1 : \theta \neq \theta_0$. We can use this test also for $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. The case of null and alternative hypotheses involving opposite inequalities can be obtained by an obvious change of signs. To define the test statistic, consider the absolute differences $V_1 = |X_1 - \theta_0|, V_2 = |X_2 - \theta_0|, \ldots, V_n = |X_n - \theta_0|$. Since the underlying distribution is continuous, we can assume that all V_i's are distinct and that none equals 0.

Let us arrange the V_i's in increasing order and assign ranks $R_1, R_2, \ldots, R_n$ to them, with rank 1 assigned to the smallest V_i. Furthermore let

$$\eta_i = \begin{cases} +1 & \text{if } X_i > \theta_0 \\ -1 & \text{if } X_i < \theta_0. \end{cases}$$

The *Wilcoxon signed rank statistic* is defined as

$$S_n = \sum_{i=1}^{n} \eta_i R_i.$$

We construct a test for any of the hypotheses mentioned above, we need to:

1. Find the distribution (or at least, limiting distribution as n becomes large) of the statistic S_n under the null hypothesis.

2. Study the effect of values of θ in the alternative hypothesis on the values of S_n.

Suppose therefore that the true value of θ is θ_0. In this case the signs η_i are equally likely to be positive or negative:

$$P\{\eta_i = 1\} = P\{\eta_i = -1\} = \frac{1}{2}.$$

Moreover, the random variables $\eta_1, \ldots, \eta_n$ are independent (since each is determined by a different X_i), and also η_i is independent of R_i, by symmetry of the distribution of X_i about θ_0. The values $R_1, \ldots, R_n$ form a permutation of numbers $1, \ldots, n$, so we write

$$S_n = \sum_{i=1}^{n} i\eta_i.$$

Since $E(\eta_i) = 0$ and $\text{Var}(\eta_i) = E(\eta_i^2) = 1$, we also have $E(S_n) = 0$ and

$$\text{Var}(S_n) = \sum_{i=1}^{n} i^2 \text{Var}(\eta_i) = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

It is possible, though tedious, to determine the distribution of S_n for small values of n; the exact distribution of S_n can be found in almost any sufficiently large collection of statistical tables. For large n one can prove that $S_n/\sqrt{\text{Var}(S_n)}$ has a limiting standard normal distribution.

Indeed, S_n is the sum of n independent random variables $\eta_1 + 2\eta_2 + \cdots + n\eta_n$, where $i\eta_i = \pm i$ with probability $1/2$ each. To apply the Lindeberg-Feller theorem (10.6.5), we have $\text{Var}(i\eta_i) = E[(i\eta_i)^2] = i^2$, so $s_n^2 = \text{Var}(S_n) = \sum_{i=1}^{n} i^2 \sim n^3$. Since $E(i\eta_i) = 0$ for $i = 1, 2, \ldots$, we have to show that

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \int_{|x| \geq \epsilon s_n} x^2 dF_i(x) \to 0$$

for every $\epsilon > 0$, where F_i is the cdf of $i\eta_i$. Since $s_n \sim n^{3/2}$, we have $\epsilon s_n > n$ for n large enough, and each integral equal to 0; since $i\eta_i$ is either i or $-i$, we have $|i\eta_i| \leq n$. This shows that Lindeberg-Feller condition is satisfied, and the proof is complete.

To determine now whether to use the right tail, left tail, or both tails of the distribution of S_n (limiting or exact), observe that if the true value of θ exceeds θ_0, then

$$P\{\eta_i = +1\} = P\{X_i > \theta_0\} > P\{X_i > \theta\} = \frac{1}{2}.$$

Consequently, the positive signs are more likely than negative signs, and this will tend to increase the value of S_n. Thus, in tests of $H_0 : \theta = \theta_0$ or $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, large values of S_n provide evidence for the alternative, and the right tail should be used as the critical region.

■ EXAMPLE 15.5

Assume that f is a density symmetric about 0; that is, it satisfies the condition $f(x) = f(-x)$ for all x. Let $X_1, \ldots, X_n$ be a random sample from the distribution with density f, and suppose that we observe the values $Y_i = \theta + X_i, i =$

$1, \ldots, n$. Thus θ is the median of Y_i (and also its mean, if $E|X_i| < \infty$). Let the observed values of $n = 20$ observations of Y_i be 315, 493, 366, 291, 501, 503, 388, 526, 308, 410, 418, 540, 285, 360, 426, 475, 336, 455, 301, 359. We want to test the hypothesis that the median θ satisfies the inequality $\theta \leq 350$, against the alternative $\theta > 350$. The consecutive values $|Y_i - 350|$, the signs η_i of the differences $Y_i - 350$, and ranks R_i are listed in Table 15.4.

Table 15.4

i	$\lvert Y_i - 350\rvert$	η_i	Rank	i	$\lvert Y_i - 350\rvert$	η_i	Rank
1	35	−1	5	11	68	+1	12
2	143	+1	16	12	190	+1	20
3	16	+1	4	13	65	−1	11
4	59	−1	9	4	10	+1	2
5	151	+1	17	15	76	+1	13
6	153	+1	18	16	525	+1	15
7	38	+1	6	17	14	−1	3
8	176	+1	19	18	105	+1	14
9	42	−1	7	19	49	−1	8
10	60	+1	10	20	9	+1	1

The value of statistic S_{20} equals 124. The asymptotic variance of S_{20} is $20 \times 21 \times 41/6 = 2870$, and therefore the observed value of statistic $S_{20}/\sqrt{\text{Var}(S_{20})}$ is 2.31. The corresponding p-value is 0.0129, so we have strong evidence in the data that the median of the population exceeds 350.

The Wilcoxon one-sample test can also be used to compare two related populations. This setup was already introduced in Chapter 13, in the test we developed for the mean difference in paired data.

■ EXAMPLE 15.6 Paired Data

When comparing the effect of two treatments A and B, part of the variation in the data is caused by other factors—for example in medical experiments this could be patient's age, gender, health status, and so on. To eliminate, or at least significantly reduce that variability, the experimenter should apply treatments A and B to the same, or at least very similar ("almost the same") subjects. Therefore our data will concern pairs, either formed naturally (e.g., twins) or carefully matched as closely as possible (age, gender, etc.). It is important to realize that the pairing process is not random. In fact, the strength of the final conclusion depends largely on how well the members of each pair are matched by as many factors as possible.

Next within each pair the experimenter allocates one member to treatment A, and the other to treatment B, this allocation being random. Let X_i and Y_i denote the results of treatments A and B, respectively, for members of the ith pair of subjects, and let $Z_i = X_i - Y_i$. If the treatments do not differ, then Z_i is as likely to be positive as negative (i.e., its median is zero). It will have a symmetric distribution because members of the pair are matched and

treatments are allocated at random to members of the pair. Thus the hypothesis "treatments do not differ" and "treatment A is superior" are now expressed as $H_0 : \theta = 0$ and $H_1 : \theta > 0$. The Wilcoxon signed rank procedure can be used to test these hypotheses.

We will present here one more procedure, called the runs test. In the case of a single sample, it can be used to test the hypothesis that the sample elements were selected randomly. In the case of two samples, it can be used to test the equality of two underlying distributions. For example, suppose that in a sample of size $n = 10$, we observe that the first 5 values are all negative while the last 5 values are all positive. Is this an indication that the process of taking the sample was not random?

The general idea of the runs test is to partition the observations $X_1, \ldots, X_n$ into two classes, say A and $B = A^c$, in such a way that the partition is induced by the values of X_i only, not by their order (A may be the set of observations that are positive, observations that exceed the hypothetical median, etc.). Formally this means that if the vector $(X_1, \ldots, X_n)$ leads to a choice of observations with indices $i_1, \ldots, i_k$ to form the set A and $\pi(1), \ldots, \pi(n)$ is a permutation of indices $(1, \ldots, n)$, then the choice of set A from the permuted vector of observations $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ is the set with indices $\pi(i_1), \ldots \pi(i_k)$. The randomness of the sample, combined with the fact that with probability 1 there are no ties among the sample values, imply that each set of indices of appropriate size is equally likely to be the set A of the partition.

In the special case we could have two samples, $X_1, \ldots, X_m$ from a distribution F and $Y_1, \ldots, Y_n$ from a distribution G, and we would test the null hypothesis that $F \equiv G$ (i.e., samples come from the same distribution). We would then consider the joint sample of X's and Y's and the partition would correspond to the two constituent samples. In this case, if the null hypothesis is true, each arrangement of X's and Y's would be equally likely.

The test in now based on the intuitive idea that if the sampling is really random (or, in the two-sample case, if samples come from the same distribution), then the partition into sets A and $B = A^c$ is "random." One of the possible measures of deviations from randomness is to observe the number of runs. To fix the idea, imagine that we have $m = 5$ symbols A and $n = 7$ symbols B arranged in some order, such as

$$\underline{BB}A\underline{B}AAA\underline{BB}A\underline{BB}.$$

A run is a string of elements of one kind, bordered either by elements of the other kind or by the end of the string. For instance, the string above has four runs of elements B, as underlined, and three runs of elements A, not underlined. The test is based on the intuitive expectation that a too small number of runs, such as

$$AAAAABBBBBBB,$$

or a too large number of runs, such as

$$BABABABABABB,$$

indicates lack of randomness.

To develop the test, we need the distribution of the number of runs under the null hypothesis. Let R be the total number of runs. In the sequel we will assume that

$m \leq n$. Clearly, the smallest value of R is 2, while the largest possible value of R is obtained for alternating runs of A's of length 1. If $m < n$, then the maximal value of R is $2m + 1$, while if $m = n$, then it is $2m$.

It remains to determine the probabilities $P\{R = r\}$ for all possible values of r. Consider first the case where r is even ($r = 2k$). The sequence must then contain k runs of each kind that alternate, starting either with a run of A's or with a run of B's. Let us imagine m elements A arranged in a string. Dividing the string into k runs means choosing $k - 1$ out of $m - 1$ places separating consecutive A's. This can be done in $\binom{m-1}{k-1}$ ways. In a similar way the string of B's of length n can be divided into k runs in $\binom{n-1}{k-1}$ distinct ways.

A joint string with $2k$ runs is now formed by dividing A's and B's into k runs each, as described above, and joining them by taking alternating runs. For instance, suppose $k = 3, m = 5$, and $n = 7$. The string of five A's can be partitioned into three runs in $\binom{4}{2} = 6$ ways:

$$AAA|A|A \quad AA|AA|A \quad AA|A|AA \quad A|A|AAA \quad A|AAA|A \quad A|A|AAA.$$

In a similar way, a string of seven B's can be divided into 3 strings in $\binom{6}{2} = 15$ ways. Taking one such partition for A's and one for B's, we obtain two arrangements giving $r = 2 \times 3 = 6$ runs. For example, taking the partition $AAA|A|A$ and $B|BBBB|BB$ we obtain two arrangements:

$$AAABABBBBABB \quad \text{and} \quad BAAABBBBABBA.$$

Consequently

$$P(R = 2k) = \frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{m+n}{m}}, \tag{15.11}$$

where the denominator gives the total number of arrangements of m objects A and n objects B. For $r = 2k + 1$ we must have either k runs of A's and $k + 1$ runs of B's, or vice versa. Reasoning analogous to that used in obtaining (15.11) leads to

$$P(R = 2k + 1) = \frac{\binom{m-1}{k-1}\binom{n-1}{k} + \binom{m-1}{k}\binom{n-1}{k-1}}{\binom{m+n}{m}}. \tag{15.12}$$

Formulas (15.11) and (15.12) therefore give the distribution of the number of runs R.

■ **EXAMPLE 15.7**

A machine produces items whose nominal diameter is c. Because of inherent variability the diameters of the items produced are random, sometimes above c and sometimes below it. The machine was designed in such a way that the diameter of the each item produced has no effect on the diameter of the next one. The diameters of 12 consecutively produced items were recorded and classified as "above c" (A) or "below c" (B). The resulting sequence was $BAAAAABBBBBB$, so $m = 5, n = 7$, and $R = 3$. The small number of runs led to the suspicion that there may be some systematic low-frequency

oscillation in operation of the machine, which tends to produce long runs of items with dimensions above c, followed by long runs of items with dimensions below c.

To test the hypothesis on "randomness" of dimensions of items against the alternative of "low-frequency oscillation," we must choose the left tail of the distribution of R (low number of runs) as the critical region, or equivalently, determine the p-value of the observed result, that is, $P(R \leq 3)$. In this case the calculations are straightforward: using formulas (15.11) and (15.12), we obtain, for $m = 5, n = 7$, and $k = 1$,

$$\begin{aligned} P(R \leq 3) &= P(R = 2) + P(R = 3) \\ &= \frac{1}{\binom{12}{5}} \left\{ 2\binom{4}{0}\binom{6}{0} + \left[\binom{4}{0}\binom{6}{1} + \binom{4}{1}\binom{6}{0} \right] \right\} \\ &= \frac{2+6+4}{\binom{12}{5}} = \frac{12}{792} = 0.0152. \end{aligned}$$

Thus observing only three runs in this situation is a strong indication of the low-frequency oscillation effect.

Calculations such as in the example above are cumbersome for large m and n. Fortunately, we can use the normal approximation showing first that

$$E(R) = \frac{2mn}{m+n} + 1 \qquad \text{and} \qquad \text{Var}(R) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}.$$

We also have the following theorem:

Theorem 15.4.1 *If $m \to \infty, n \to \infty$ in such a way that $m/n \to \eta$ with $0 < \eta < \infty$, then*

$$\frac{R - 2m/(1+\eta)}{\sqrt{4\eta m/((1+\eta)^3}}$$

converges in distribution to $\text{N}(0, 1)$.

Thus, replacing η by m/n, we can expect that the random variable

$$\frac{R - 2mn/(m+n)}{2mn} \sqrt{(m+n)^3} \tag{15.13}$$

has approximate standard normal distribution, provided that m and n are large.

We can show that the approximation is very good for $m, n \geq 20$, and quite acceptable when $m, n \geq 10$.

■ **EXAMPLE 15.8**

Recall Example 15.7. Using (15.13) we have an approximation

$$\begin{aligned} P\{R \le 3\} &= P\left\{ \frac{R - 2\times 5\times 7/(5+7)}{2\times 5\times 7}\sqrt{(5+7)^3} \right. \\ &\le \left. \frac{3 - 2\times 5\times 7/(5+7)}{2\times 5\times 7}\sqrt{(5+7)^3} \right\} \\ &\approx P\left\{ Z \le \frac{3-5.83}{70} \times 41.57 \right\} = P\{Z \le -1.68\} = 0.0465. \end{aligned}$$

As compared with the exact p-value 0.0152, the approximation is not good.

This shows that for small sample sizes one should try to determine the exact p-value (by direct evaluation, use of special tables, or an appropriate statistical package).

PROBLEMS

15.4.1 Prove the asymptotic normality of the Wilcoxon signed rank statistic S_n using the Liapunov theorem.

15.4.2 Out of 15 data points one is between 0 and 1, two are between -2 and -1, three are between 2 and 3, four are between -4 and -3, and five are between 4 and 5. Use the Wilcoxon signed rank statistic to test the hypothesis that the median is: **(i)** 0. **(ii)** 1.

15.4.3 Twelve pairs of subjects, matched within each pair with respect to age, gender, health status, and initial weight, were put on two types of diets. The data on pounds lost after five weeks are as follows:

Pair	1	2	3	4	5	6	7	8	9	10	11	12
Diet A	15	33	21	17	14	25	25	31	18	5	46	11
Diet B	18	17	10	10	32	11	8	26	-3	19	5	8

Use the Wilcoxon signed ranked test to test the hypothesis that both diets have the same effect, against the alternative that diet B is more efficient than diet A. Use $\alpha = 0.05$.

15.4.4 Some texts define the Wilcoxon signed rank statistic as $\tilde{S}_n = \sum \tilde{\eta}_i R_i$, where $\tilde{\eta}_i = 1$ if $X_i > \theta_0$ and 0 otherwise. Determine the mean and variance of $\tilde{S}_n$, and show that tests based on S_n and on $\tilde{S}_n$ are equivalent in the following sense: under a null hypothesis, $S_n = \tilde{S}_n - \tilde{S}_n^*$, where $\tilde{S}_n$ and $\tilde{S}^*$ have the same distribution.

15.4.5 Solve Problems 15.3.6 and 15.3.7 using a runs test. Compare the results obtained by different methods (use the same $\alpha = 0.05$). Explain the differences, if they exist, for $m = 10$ and $k = 5$.

15.4.6 A machine is set up to produce items, each with a diameter above 1 inch. The diameters of 15 consecutive items produced are 1.11, 1.15, 0.98, 1.11, 1.08, 1.06, 0.97, 0.97, 1.05, 1.02, 0.98, 0.99, 0.96, 1.03, 1.01. Use a runs test, taking 1 inch as a threshold to test the hypothesis that the measurements represent random deviations from the required standard.

15.5 TWO-SAMPLE RANK TESTS

Let $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be random samples from two continuous distributions with cdf's F and G, respectively. We want to test the null hypothesis

$$H_0 : F \equiv G$$

against some alternatives, whose form will be discussed later.

We present one of the most important of the rank tests, the Wilcoxon-Mann-Whitney two-sample test. The underlying idea of the test is as follows:

First, we combine both samples and then arrange all observations in an increasing order:

$$U_1 \leq U_2 \leq \cdots \leq U_{m+n},$$

where each U_i belongs to one of the samples. Because of the assumed continuity of F and G, we can disregard the possibility of ties in the joint sample, so we assume that all inequalities among U_i's are strict.

Next we assign ranks from 1 to $m + n$ to consecutive elements U_i. If the null hypothesis is true, then the m ranks of elements of the first sample and the n ranks of elements of the second sample are mixed randomly, in the sense that each of the $\binom{m+n}{m}$ allocations of the m ranks of elements of the first sample has the same probability $1/\binom{m+n}{m}$.

Wilcoxon suggested to use the statistic W_X, defined as the sum of ranks of elements of the sample $(X_1, \ldots, X_m)$, in the joint ordering of both samples. Formally, we can write

$$W_X = \sum_{i=1}^{m+n} i I_X(U_i),$$

where $I_X(U_i) = 1$ if U_i comes from sample $\mathbf{X}$ and is 0 otherwise.

Equivalently we can use statistic W_Y, being the sum of ranks of elements of the second sample in the joint ordering. Statistics W_X and W_Y carry the same information. Indeed we have

$$\begin{aligned} W_X + W_Y &= \sum_{i=1}^{m+n} i I_X(U_i) + \sum_{i=1}^{m+n} i I_Y(U_i) = \sum_{i=1}^{m+n} i(I_X(U_i) + I_Y(U_i)) \\ &= \sum_{i=1}^{m+n} i = \frac{(m+n)(m+n-1)}{2}, \end{aligned}$$

so

$$W_Y = \frac{(m+n)(m+n-1)}{2} - W_X.$$

Let us begin with finding the expectation and variance of W_X. We have

$$E(W_X) = E\left(\sum_{i=1}^{m+n} i I_X(U_i)\right) = \sum_{i=1}^{m+n} i E[I_X(U_i)] = \sum_{i=1}^{m+n} i P\{I_X(U_i) = 1\}.$$

Since all allocations of the m ranks of elements of the first sample among $m + n$ elements of both samples are equally likely, the probability that the ith ranking

element comes from the first sample is $P(I_X(U_i) = 1) = m/(m+n)$. We have therefore

$$E(W_X) = \sum_{i=1}^{m+n} i\frac{m}{n+m} = \frac{m(m+n+1)}{2}.$$

The calculations of variance is somewhat more tedious (we leave it as exercise). The result is

$$\mathrm{Var}(W_X) = \frac{mn(m+n+1)}{12}. \tag{15.14}$$

One could also show (we omit the proof) that as $m \to \infty, n \to \infty$, the random variable

$$Z_{m,n} = \frac{W_X - E(W_X)}{\sqrt{\mathrm{Var}W_X}} = \frac{W_X - m(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}$$

converges in distribution to the standard normal random variable.

To design a testing procedure, it is now necessary to specify the alternative hypothesis. In other words, we have to determine the class of hypotheses such that if (F, G) belongs to this class, then the values of $Z_{m,n}$ will tend to be relatively large (or small, or large in absolute value).

One such class of alternative hypotheses is obtained by taking

$$H_1 : G(t) = F(t-\theta) \tag{15.15}$$

for all t and some θ. To grasp the meaning of this hypothesis and its consequence for W_X, let us consider the case $\theta > 0$. Here $G(t) = F(t-\theta) \leq F(t)$, which means that $P\{Y \leq t\} \leq P\{X \leq t\}$; hence the values of X tend to be smaller than values of Y (since whatever the value t, the random variable X is more likely to be below t than the random variable Y). Consequently the observations $X_1, \ldots, X_m$ will tend to be located closer to the left end and so have smaller ranks. Thus small values of W_X support the alternative hypothesis $H_1 : \theta > 0$. The case $\theta < 0$ is analogous, whereas for the alternative $H_1 : G(t) = F(t-\theta), \theta \neq 0$, one should take the two-sided test.

Actually the class of alternatives against which the Wilcoxon statistic W_X can be used is larger. Recall that the random variable X is *stochastically larger* than Y if $P\{X \leq t\} \leq P\{Y \leq t\}$ for all t. In the present notation, X is stochastically larger than Y if $F(t) \leq G(t)$ for all t. Thus stochastic dominance of Y by X (or X by Y) is an alternative that will tend to inflate (or decrease) the statistic W_X.

About the same time as Wilcoxon introduced his statistic W_X, Mann and Whitney introduced another statistic, pertaining to the same two-sample problem:

$$R = \text{number of pairs } (X_i, Y_j) \text{ with } X_i > Y_j.$$

The statistic W_X can be written as

$$W_X = \sum_{j=1}^{m} R_j,$$

where R_j is the rank of $X_{j:m}$. This means that R_j equals the number of elements in the combined sample that do not exceed $X_{j:m}$.

By definition, there are j elements in the sample $X_1, \ldots, X_m$ that are less or equal to $X_{j:m}$, so we can write

$$R_j = j + \text{ number of } Y_i \text{ with } Y_i < X_{j:m}.$$

Consequently

$$W_X = \sum_{j=1}^{m} [j + \text{ number of } i\text{'s with } Y_i < X_{j:m}] = \frac{m(m+1)}{2} + R.$$

It follows that

$$E(R) = E(W_X) - \frac{m(m+1)}{2} = \frac{m(m+n+1)}{2} - \frac{m(m+1)}{2} = \frac{mn}{2}$$

and

$$\text{Var}(R) = \text{Var}(W_X) = \frac{mn(m+n+1)}{2}.$$

It also follows, from the asymptotic normality of W_X, that as $m \to \infty, n \to \infty$, the statistic

$$Z_{m,n} = \frac{R - mn/2}{\sqrt{mn(m+n+1)/2}} \tag{15.16}$$

converges in distribution to $\text{N}(0,1)$. Since the tests based on W_X and on R are equivalent, they became known as Wilcoxon-Mann-Whitney tests.

Observe now that there are mn possible pairs with elements of the pair coming from different samples. Consequently, R/mn is a consistent estimator of the probability $P\{Y < X\}$, so as $m \to \infty, n \to \infty$, we have

$$\frac{R}{mn} \xrightarrow{p} P\{Y < X\} = \int_{-\infty}^{+\infty} G(t) f(t) dt = \int_{-\infty}^{+\infty} [1 - F(t)] g(t) dt.$$

Under the null hypothesis $F \equiv G$, we have $P\{Y < X\} = 1/2$. If $P\{Y < X\} = \xi \neq 1/2$, then, using (15.16), we can write

$$\begin{aligned} Z_{m,n} &= \sqrt{\frac{2mn}{m+n+1}} \left(\frac{R}{mn} - \frac{1}{2} \right) \\ &= \sqrt{\frac{2mn}{m+n+1}} \left(\frac{R}{mn} - \xi \right) + \sqrt{\frac{2mn}{m+n+1}} \left(\xi - \frac{1}{2} \right). \end{aligned}$$

The first term converges in distribution to a standard normal random variable, while the second diverges to $+\infty$ or $-\infty$ if only $\xi \neq 1/2$. This shows that in the limit as $m \to \infty, n \to \infty$, the Wilcoxon-Mann-Whitney test (based on W_X or R) has asymptotic power 1 for all alternatives (F, G) with $\xi \neq 1/2$.

■ EXAMPLE 15.9

Consider two athletes, A and B; one of them is to be selected to represent the country in some competition. Assume that both athletes have attained some stable level of proficiency in their discipline. Their results (in competitions,

their best daily training results, etc.) may be taken as random samples of some random variables, say $X^{(A)}$ and $X^{(B)}$. Furthermore, assume that the discipline is such that tied results are unlikely. Let $n = 15$ be the results of A and $n = 20$ the results of B, with none of the results repeating. Suppose that after arranging these results jointly from worst to the best, we obtain the sequence

$$A\,A\,A\,B\,A\,B\,B\,A\,A\,A\,B\,B\,B\,B\,A\,B\,A\,B\,B\,A\,B\,B\,A\,A\,B\,B\,A\,B\,B\,B\,A\,B\,B\,B\,A$$

Here the three lowest results are of athlete A, the fourth is of athlete B, and so on. The best result belongs also to A.

We can define the "better" athlete (of a pair) as the one who has better than even chances of defeating the other one. Thus A is better than B if

$$\xi = P(A \text{ beats } B) > \frac{1}{2}.$$

The idea that lies at the foundation of the US system of selecting Olympic representatives is the same as that in the case under consideration: A is better than B because the best result of A is better than the best result of B. Let us therefore test the hypothesis that

$$\xi = P(A \text{ beats } B) = P(X > Y) \leq \frac{1}{2}$$

(A is equal to B, or inferior to B) against the alternative that $\xi > 1/2$. The Wilcoxon statistic (sum of ranks of A) is

$$W_X = 1+2+3+5+8+9+10+15+17+20+23+24+27+31+35 = 230.$$

For $m = 15, n = 20$ we have $E(W_X) = 270, \mathrm{Var}(W_X) = 900$, so the p-value is $P\{Z > (230 - 270)/30\} = P\{Z > -1.33\} = 0.9082$. There is therefore no reason to reject the null hypothesis that A is no better than B, despite the fact that the best result is attained by athlete A.

PROBLEMS

15.5.1 Prove formula (15.14) showing first that for $i \neq j$

$$\mathrm{Var}(I_{A_i}) = \frac{mn}{(m+n)^2} \text{ and } \mathrm{Cov}(I_{A_i}, I_{A_j}) = -\frac{mn}{(m+n)^2(m+n-1)}.$$

15.5.2 Two samples of sizes $m = 20$ and $n = 10$, respectively, are selected from two populations. Let $r_1, \ldots, r_{10}$, denoting the numbers of observations from the first sample that do not exceed the kth ($k = 1, \ldots, 10$) element in the ordered second sample, be $1, 1, 2, 4, 4, 6, 8, 9, 11, d$. **(i)** Find the value of Wilcoxon and Mann-Whitney statistic as a function of d. **(ii)** Find the value of Kolmogorov-Smirnov statistic as a function of d. **(iii)** Suggest the appropriate alternative hypothesis, and determine d for which the hypothesis about the same median can be rejected based on the Mann-Whitney or Wilcoxon test. **(iv)** Answer (iii) using the Kolmogorov-Smirnov test.

15.5.3 Use a runs test for the data of Problem 15.5.2.

15.5.4 Assume that n is odd and that all m elements of the first sample are below the median of the second sample. Find the range of the test statistics and the rejection region for the appropriate alternative hypothesis using: **(i)** The Wilcoxon test. **(ii)** The Mann-Whitney test. **(iii)** The runs test. **(iv)** The Kolmogorov-Smirnov test for the appropriate alternative hypothesis. ($\alpha = 0.05, n = 51, m = 20$).

15.6 KRUSKAL-WALLIS TEST

Finally, we present the Kruskal–Wallis test, a rank-based counterpart of the one-way analysis of variance test. We consider k random samples, of sizes $n_1, n_2, \ldots n_k$, where $k \geq 2$. We let X_{ij} be the jth element ($j = 1, \ldots, n_i$) in ith sample, and we assume that $X_{i,1}, \ldots, X_{i,n_i}$ are iid random variables with a cdf given by

$$F_i(x) = G(x - \theta_i)$$

for all x. Here G is assumed to be a cdf of some continuous random variable. We want to test the null hypothesis

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k \qquad \text{against} \qquad H_1 : \text{ not all } \theta_i's \text{ are equal.}$$

If G is a cdf of a distribution symmetric around 0—satisfying $G(-x) = 1 - G(x)$—then the median and the mean (if it exists) of F_i is easily seen to be θ_i, and the null hypothesis asserts that the medians (or means) of all populations are the same. The variance need not exist, but the fact that all populations have the same distribution up to a location parameter corresponds to the assumption of homoscedasticity of Chapter 14.

Let us order all $n = n_1 + \cdots + n_k$ observations X_{ij} from smallest to largest, and let R_{ij} be the rank of observation X_{ij} in the joint ordering. Furthermore let $R_{i+} = \sum_{j=1}^{n_i} R_{ij}$ be the sum of ranks corresponding to elements in the ith sample (if we have two populations, i.e., $k = 2$, then R_1 and R_2 correspond to Wilcoxon statistics W_X and W_Y from the preceding section).

The test of null hypothesis H_0 is based on Kruskal-Wallis statistic

$$B = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{R_{i+}^2}{n_i} - 3(n+1). \tag{15.17}$$

We have the following theorem:

Theorem 15.6.1 *If H_0 is true, and all sample sizes $n_1, n_2, \ldots, n_r$ increase to infinity in such a way that $n_i/n \to p_i > 0$ for all i, then the distribution of statistic B converges to the chi-square distribution with $r - 1$ degrees of freedom.*

We will not give a full proof, but outline the argument to explain why the large values of the statistic B are compatible more with the alternative than with the null hypothesis.

First, under the null hypothesis,

$$E\left(\frac{R_{i+}}{n_i}\right) = \frac{n+1}{2} \qquad \text{and} \qquad \text{Var}\left(\frac{R_{i+}}{n_i}\right) = \frac{(n+1)(n-n_i)}{12n_i}. \tag{15.18}$$

The proof is left as an exercise. Now letting

$$V_i = \frac{R_{i+}/n - E(R_{i+}/n)}{\sqrt{\text{Var}(R_{i+}/n)}},$$

we write

$$\begin{aligned}\sum_{i=1}^{r}(1-\frac{n_i}{n})V_i^2 &= \sum_{i=1}^{r}(1-\frac{n_i}{n})\left[\frac{R_{i+}/n_i - E(R_{i+}/n_i)}{\sqrt{\text{Var}(R_{i+}/n_i)}}\right]^2 \\ &= \sum_{i=1}^{r}\frac{n-n_i}{n}\times\frac{(R_{i+}/n_i-(n+1)/2)^2}{(n+1)(n-n_i)/12n_i} \qquad (15.19) \\ &= \frac{12}{n(n+1)}\sum_{i=1}^{r}\frac{R_{i+}^2}{n_i} - 3(n+1).\end{aligned}$$

First, from expression (15.19) we have that any deviation from the null hypothesis will tend to increase B. Thus the critical region is always the right tail of the appropriate chi-square distribution, as in ANOVA tests under normal assumptions.

Second, we have to show that V_i converges in the distribution to a standard normal variable, or equivalently that the distribution of R_{i+} is asymptotically normal. Once this fact is established, the proof of the theorem can be completed by observing that the random variables $R_{1+}, \ldots, R_{r+}$ are constrained by the condition $\sum_{i=1}^{r} R_{i+} = n(n+1)/2$, which reduces the number of degrees of freedom to $r-1$. It remains to check, for instance, that the asymptotic mean and variance of B agree of those of chi-square distribution with $r-1$ degrees of freedom.

The proof of asymptotic normality of R_{i+} lies beyond the scope of this book; it relies on one of the central limit theorems for the sum of exchangeable random variables.

PROBLEMS

15.6.1 Prove relations (15.18).

15.6.2 STAT 102 can only taken be taken by students who passed STAT 101. Among 15 students in STAT 102, five took STAT 101 from instructor X, four took it from instructor Y, and the rest took the course from instructor Z. Ordered according to their performance, the students of the three instructors are

$$ZXZZZYXZYXYXZXY.$$

(this means that the best student was taught by instructor Z, second best by instructor X, etc.). At the significance level $\alpha = 0.05$, test the hypothesis that the performance of students in STAT 102 class does not depend on who taught them STAT 101.

This Page Intentionally Left Blank

CHAPTER 16

ANALYSIS OF CATEGORICAL DATA

16.1 INTRODUCTION

The term *categorical data* refers to observations recorded either on a nominal scale or on a discrete ordinal scale (so ties are expected to occur often). Typical examples of the nominal scale occur when data represent frequencies of categories of some qualitative attribute (e.g., responses in a questionnaire about the state of residence or religious affiliation). The discrete ordinal scale consists of a set of naturally ordered categories. For example, opinion on a specific issue may be classified as "favorable," "neutral," or "unfavorable"; education achieved may be classified as "high school," "junior college," "four-year college or university," "graduate school"; and so on.

A special case is played by a binary data, for example male vs. female, smoker vs. nonsmoker. There is no "natural" ordering here, but one can always assign values 0 and 1 (or any other two values) to the categories.

The following examples illustrate some of the possible problems in the analysis of categorical data.

■ **EXAMPLE 16.1**

A typical case of data on a nominal scale occur in genetic experiments. Suppose that we have a gene with two forms, A and a, so that an individual (e.g.,

plant) belongs to one of the three categories AA, Aa, or aa. If none of the forms is dominant, the three genotypes can be identified (they coincide with the phenotypes). According to genetic theory, the probabilities of three genotypes are θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$, respectively, where θ is the unknown frequency of allele A. Suppose that out of 200 plants we have 100 of type AA, 89 of type Aa, and 11 of type aa. Are these data in agreement with genetic theory?

■ **EXAMPLE 16.2**

Suppose that we have joint data on ethnicity and incidence of various types of cancer. In studying the association between these two attributes, we might search for genetic, dietary, or other reasons for a lack of independence.

In the last example, both variables (ethnicity and type of cancer) were of a nominal character. When both variables are measured on an ordinal scale, we may additionally be interested in the strength of association between variables. Such questions are discussed later in this chapter, where we present selected methods of analysis of ordinal aspects of categorical data.

■ **EXAMPLE 16.3**

For data measured originally on an ordinal scale and later grouped into classes, the methods of Chapter 15 are not applicable. Suppose that we want to study a relationship between level of education (classified as "I, no college"; "II, some college"; and "III, at least four-year college/university") and the frequency of changing jobs (again classified as high, moderate, or low). The data take the form

	Education		
Frequency of Job Change	I	II	II
Low	N_{11}	N_{12}	N_{13}
Moderate	N_{21}	N_{22}	N_{23}
High	N_{31}	N_{32}	N_{33}

Here N_{ij} is the frequency of occurrence of a given category among randomly selected study participants. We want to test the hypothesis that there exists an association (positive or negative) between the frequency of changing jobs and the level of education.

The theory of categorical data is a domain with a long tradition, the chi-square test being one of the oldest examples of statistical procedures that use only frequencies. At the same time this domain has been developing rapidly over the last four decades. We will introduce some of the basic information on problems and methods in analysis of categorical data. An exhaustive presentation of the field can be found in Agresti (2002).

16.2 CHI-SQUARE TESTS

In this section we present one of the oldest and best known statistical tests. Beginning with the case of a discrete distribution with a finite number of values, we let X be a random variable with r possible values $x_1, \ldots, x_r$ and the corresponding probabilities $p_i = P\{X = x_i\}, i = 1, \ldots, r$. For a random sample $X_1, \ldots, X_n$, counts $N_1, \ldots, N_r$, where $N_1 + \cdots + N_r = n$, are respective frequencies of values $x_1, \ldots, x_r$. The vector $(N_1, \ldots, N_{r-1})$ (notice that one of the coordinates N_i is redundant), called the *count vector* of the sample, carries all the necessary information about the distribution of X. Formally, the likelihood of the data is

$$L = Cp_1^{N_1} p_2^{N_2} \cdots p_r^{N_r} = Cp_1^{N_1} p_2^{N_2} \cdots p_{r-1}^{N_{r-1}} (1 - \sum_{i=1}^{r-1} p_i)^{n - \sum_{i=1}^{r-1} N_i},$$

where $C = n!/(N_1!N_2! \cdots N_r!)$. Consequently, the count vector is a set of jointly sufficient statistics for the vector $(p_1, p_2, \ldots p_{r-1})$.

An important remark here is that procedures for the inference about the distribution $(p_1, p_2, \ldots, p_{r-1})$ of X on the basis of the counts $(N_1, \ldots, N_{r-1})$ do not depend on $x_1, \ldots, x_r$.

Suppose that we want to test the hypothesis

$$H_0 : p_i = p_i^0, \quad i = 1, \ldots, r \tag{16.1}$$

against the general alternative

$$H_1 : \text{ hypothesis } H_0 \text{ is not true.}$$

Here $(p_1^0, \ldots, p_r^0)$ is some fixed probability distribution, and we assume that $p_i^0 > 0$ for $i = 1, \ldots, r$.

Under the null hypothesis, for each i, the marginal distribution of the count N_i is BIN(n, p_i^0). The test will be based on the following theorem:

Theorem 16.2.1 *Let $(N_1, \ldots, N_r)$ be the count vector of random sample of size n from multinomial distribution with probabilities $(p_1^0, \ldots, p_r^0)$, so that $p_1^0 + \cdots + p_r^0 = 1$ and $N_1 + \cdots + N_r = n$. Then the statistic*

$$Q^2 = \sum_{j=1}^{r} \frac{(N_j - np_j^0)^2}{np_j^0} \tag{16.2}$$

has the limiting (as $n \to \infty$) chi-square distribution with $r - 1$ degrees of freedom. The statistic Q^2 is often referred to as Person's chi-square.

Proof. A simple argument shows that the theorem is true for $r = 2$. In this case we have, remembering that $p_2^0 = 1 - p_1^0$ and $N_2 = n - N_1$,

$$\begin{aligned} Q^2 &= \frac{(N_1 - np_1^0)^2}{np_1^0} + \frac{(N_2 - np_2^0)^2}{np_2^0} = \frac{(N_1 - np_1^0)^2}{np_1^0} + \frac{(n - N_1 - n(1 - p_1^0))^2}{n(1 - p_1^0)} \\ &= \frac{(N_1 - np_1^0)^2}{np_1^0} + \frac{(N_1 - np_1^0)^2}{n(1 - p_1^0)} = \left[\frac{N_1 - np_1^0}{\sqrt{np_1^0(1 - p_1^0)}} \right]^2 \approx Z^2 \sim \chi_1^2. \end{aligned}$$

For an arbitrary r the algebra is more complicated: the statistic Q^2 is represented as a sum of $r-1$ squares of random variables, each converging in distribution to a standard normal random variable, and such that their coefficients of correlation tend to 0 as n increases. We omit the details. □

As a practical rule, one obtains a reasonable approximation of the distribution of Q^2 if expected counts np_i^0, $i = 1, \ldots, r$, are at least 5, and the approximation is good if the expected counts exceed 10.

To test the null hypothesis $H_0 : p_j = p_j^0, j = 1, \ldots, r$ against the alternative H_1 : "hypothesis H_0 is false," we need to determine the critical region. Since any violations of the null hypothesis in the chi-square test will tend to increase the value of statistic Q^2, we should take the right tail as the critical region, with $\chi^2_{\alpha,r-1}$ as a critical value.

■ **EXAMPLE 16.4**

According to genetic theory, the seeds collected from a field of pink pea should produce plants with white, pink, and red flowers, in the proportion $1 : 2 : 1$. Of 400 plants grown from such seeds, 93 are white, 211 are pink and 96 are red. Does this result contradict genetic theory?

SOLUTION. We have here 400 observations of a three-valued random variable. According to the null hypothesis we have $p_1^0 = 1/4, p_2^0 = 1/2$, and $p_3^0 = 1/4$, so the expected counts np_i^0 are 100, 200, and 100. The observed value of the test statistic Q^2 is

$$Q^2 = \frac{(93-100)^2}{100} + \frac{(211-200)^2}{200} + \frac{(96-100)^2}{100} = 1.255.$$

Since we have here 2 degrees of freedom, the p-value is 0.534, and consequently, the data do not provide enough evidence against the null hypothesis.

As already mentioned, the chi-squared test can be used for testing the hypothesis that the data of a continuous type follow a specific distribution. In this case, to obtain the count vector $(N_1, \ldots, N_r)$ we partition the range of values of the observed random variable X into r sets, say $C_1, C_2, \ldots, C_r$. Typically the sets C_j are intervals, but this is not necessary; the sets C_j need not to be connected, and may consist of a number of non-contiguous intervals. The count N_j, given the sample of n observations of random variable X, is defined as a frequency of observations in set C_j. Clearly, $N_1 + \cdots + N_r = n$. If f is the density of X specified by the null hypothesis, then we have

$$p_j^0 = \int_{C_j} f(x)dx, \quad j = 1, \ldots, r.$$

The test statistic Q^2 depends not only on the sample but also on the choice of the partition into sets C_j. Sometimes the choice of the partition is natural, while in other cases the partition is chosen in a rather arbitrary way.

■ EXAMPLE 16.5

Fox and James (1987) give the following data about the birth signs of 851 prominent chess players:

Capricorn	63	Lasker, Keres, Chiburdanidze
Aquarius	79	Spassky, Bronstein
Pisces	101	Fischer, Tarrasch, Geller, Larsen
Aries	76	Smyslov, Kasparov, Korchnoi, Portisch
Taurus	77	Miles, Nunn, Steinitz
Gemini	67	Petrosian, Karpov, Short, Euwe
Cancer	54	Morphy, Anderssen
Leo	67	Botvinnik
Virgo	63	Philidor
Libra	69	Rubinstein, Fine
Scorpio	71	Capablanca, Alekhine, Nimzowitsch, Tal
Sagitarius	64	Reshevsky, Pillsbury.

It seems that Pisces have a significantly higher number of prominent chess players.

We want to test the null hypothesis that the birthday is not related to chess talent. In this case, we can expect that the birthdays of $n = 851$ prominent chess players form a random sample from a distribution uniform on the year (which we may conveniently regard as a continuous distribution). The expected count for each sign will be $851 \times (1/12) = 70.917$, and

$$Q^2 = \frac{(63 - 70.917)^2}{70.917} + \cdots + \frac{(64 - 70.917)^2}{70.917} = 21.53.$$

The p-value of the observed result is 0.028 which means that such a result will occur, on average, about once in 35 times, if the null hypothesis is true. Someone with a firm belief that there is some truth in astrology may use this result to support an argument. A sceptic who is convinced that the configuration of stars and planets at the time of one's birth cannot affect this person's talents will regard the observed result as an example of a type I error.

■ EXAMPLE 16.6

The left tail (indicating a good fit) of the chi-square test can be used for detecting whether the data were tampered with. The point is that the p-value gives the probability that in a repetition of the experiment, one would observe worse (in discrete case, no better) fit than the one actually observed. Thus if the p-value is close to 1, say 0.99, it means that on average, only once in 100 repetitions one can expect to observe a better fit. That strongly suggests that the data were "improved" to make them conform better to the null hypothesis. An interesting example of such a type of inference is provided by Fisher's analysis of data on heredity by Gregor Mendel. For a detailed explanation, see Freedman et al. (1992). Here we briefly sketch the idea and results.

Mendel studied the laws of inheritance of various characteristics, eventually introducing the concept of a gene. In a typical experiment of Mendel,

plants of genotype AA are crossed with plants of genotype aa. All seeds are then hybrids Aa. A number of such seeds are grown and the plants are cross-pollinated. According to genetic theory, the ratio of plants of genotypes AA, Aa, and aa are $1:2:1$. If AA and Aa cannot be distinguished, then the ratio of genotype aa to all others is $1:3$, and so on. For example, in the last case, of 800 plants, about 200 can be expected to be aa. In all of Mendel's experiments, the observed numbers differ suspiciously little from the expected. For example, suppose that in the last case Mendel reported 205 plants with an aa genotype out of 800. The chi-square fit of such a result is

$$\frac{(205-200)^2}{200}+\frac{(595-600)^2}{600}=0.167,$$

which corresponds to a p-value of about 90%.

A single result with so high a p-value is not unusual, but when Fisher combined all of Mendel's data using the combined chi-square test, the p-value was 0.99996 (i.e., only 4 times out of 100,000 one can expect a better fit). Thus either Mendel had some extraordinary luck, or his data were "beautified" to conform better to his theory.

In the cases considered above, the null hypothesis is simple; that is, the hypothetical distribution is specified completely. More often the null hypothesis comprises a class of distributions. We will now consider cases where the distribution of a random variable X depends on some parameter θ (possibly vector valued) in a specified family of distributions. Thus we will assume that X is discrete random variable, with r possible values $x_1,\ldots,x_r$, and such that

$$P\{X=x_j\}=p_j(\theta),$$

where $\theta=(\theta_1,\ldots,\theta_k)\in H_k$. Moreover we will assume that the number of classes r satisfies the inequality $r\geq k+2$ and that $p_j(\theta)>0$ for all j and all $\theta\in H_k$.

As before, we assume that we have a random sample of values of X, leading to the count vector $(N_1,\ldots,N_r)$, where $N_1+\cdots+N_r=n$. This time we cannot form the test statistic (16.2) because the expected class frequencies $np_j(\theta)$ depend now on the unknown parameter θ. According to Theorem 16.2.2, we must use instead the estimated expected frequencies obtained as functions of the MLE of θ. Given the count $(N_1,N_2,\ldots,N_r)$ the likelihood of the data is

$$L(\theta;N_1,\ldots,N_r)=[p_1(\theta)]^{N_1}\cdots[p_r(\theta)]^{N_r}. \tag{16.3}$$

Let $\hat{\theta}=(\hat{\theta}_1,\ldots,\hat{\theta}_k)$ denote the value of the parameter θ that maximizes the likelihood (16.3). We have then the following theorem:

Theorem 16.2.2 *The statistic*

$$Q^2=\sum_{j=1}^{r}\frac{[N_j-np_j(\hat{\theta})]^2}{np_j(\hat{\theta})} \tag{16.4}$$

has, as $n\to\infty$, *the limiting chi-square distribution with* $r-1-k$ *degrees of freedom.*

We omit the proof, which can be found in Cramér (1946).

■ **EXAMPLE 16.7**

Consider a gene with two alleles, A and a. Let the frequency of gene A in population be θ. Under random mating, the frequencies of individuals of genotypes AA, Aa, and aa are $\theta^2, 2\theta(1-\theta)$, and $(1-\theta)^2$. To test the theory, n individuals are randomly selected, and the count of the three genotypes is N_1, N_2, N_3. Then the likelihood of the data is

$$L = [\theta^2]^{N_1}[2\theta(1-\theta)]^{N_2}[(1-\theta)^2]^{N_3} = 2^{N_2}\theta^{2N_1+N_2}(1-\theta)^{N_2+2N_3}.$$

Differentiating $\log L$, we obtain easily the MLE of θ:

$$\hat{\theta} = \frac{2N_1 + N_2}{2n}. \tag{16.5}$$

For a numerical example, suppose that $n = 200, N_1 = 25, N_2 = 10$, and $N_3 = 165$. We have then $\hat{\theta} = 60/400 = 0.15$, and using (16.5), we get

$$Q^2 = \frac{(25-30)^2}{30} + \frac{(10-51)^2}{51} + \frac{(165-144.5)^2}{144.5} = 36.70.$$

Since this result exceeds $\chi^2_{0.0005,1} = 7.879$, the p-value is less than 0.005, so the evidence against the null hypothesis of random mating provided by such data is very strong.

Theorem 16.2.1 asserts that if the null hypothesis is simple (specifies completely the distribution), then the limiting distribution of Q^2 (under null hypothesis) is chi-square with $r-1$ degrees of freedom. On the other hand, if the null hypothesis is composite, and we have to estimate the expected counts by finding the MLE's of the parameters $\theta_1, \ldots, \theta_k$ given the counts, we lose k degrees of freedom.

■ **EXAMPLE 16.8**

Assume that we have raw data for the numbers $x_{1905}, x_{1906}, \ldots, x_{1991}$ of cloudless nights in the last $n = 87$ years at some prospective telescope site. Suppose that it is known that $x_{1905} + x_{1906} + \cdots + x_{1991} = 21,163$ and $x^2_{1905} + \cdots + x^2_{1991} = 5,226,819$. However, the actual data are not available, and instead we have the following counts N_i of years with given numbers of cloudless nights:

Interval	N_i
160 or below	1
161 to 180	3
181 to 200	7
201 to 220	17
221 to 240	18
241 to 260	26
261 to 280	9
281 to 300	4
301 or above	2

We want to test the hypothesis that the number of the cloudless nights at the site is normally distributed.

The MLE's of μ and σ^2 computed from original data are

$$\bar{x} = \frac{21,163}{87} = 243.25 \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{87}(5,226,819) - (243.25)^2 = 907.82;$$

hence $\hat{\sigma} = 30.13$. The estimated class probabilities are, letting X denote the number of cloudless nights,

$$\begin{aligned} p_1 &= P(X \leq 160) = \Phi\left(\frac{160 - 243.25}{30.13}\right) = \Phi(-2.76) = 0.0029 \\ p_2 &= P(160 \leq X \leq 180) = \Phi(-2.10) - \Phi(-2.76) = 0.015, \end{aligned}$$

and similarly for subsequent intervals.

The actual and expected counts for the consecutive classes are therefore

Interval	$87p_i$	N_i
160 or below	0.252	1
161 to 180	1.305	3
181 to 200	4.959	7
201 to 220	12.676	17
221 to 240	19.192	18
241 to 260	22.281	26
261 to 280	15.356	9
281 to 300	7.056	4
301 or above	2.612	2

To avoid too small expected class sizes, we combine the first three classes together as well as the last two classes and obtain:

Interval	$87p_i$	N_i
200 or below	6.516	11
201 to 220	12.676	17
221 to 240	19.192	18
241 to 260	22.281	26
261 to 280	15.356	9
281 or above	9.668	6

The observed value of the statistic Q^2 is now 9.278. This corresponds to the p-value equal 0.026 (3 df).

PROBLEMS

16.2.1 Ladislaus von Bortkiewicz, a Russian economist and statistician, is known for the data he collected on the number of Prussian cavalryman being killed by the kick of a horse. He observed 10 army corps for 20 years, obtaining 200 observations. The total of 122 deaths was distributed as follows:

0	1	2	3	4	5	6
109	65	22	3	1	0	0

Test the hypothesis that the number of deaths from horse kicks has Poisson distribution.

16.2.2 A certain type of toy is sold with three batteries included. The number of defective batteries (X) in a random sample of 200 toys are as follows:

X	0	1	2	3
Count	51	92	40	17

Test the hypothesis that the number of defective batteries in a toy has a binomial distribution.

16.2.3 Assume that the genders of children in a family are independent. In a human population, the probability that a child is a male is very close to 0.5. Numbers of boys (X) in a random sample of 100 families with four children are:

X	0	1	2	3	4
Count	7	21	40	27	5

Test the hypothesis that the distribution of the number of boys in a family of four children is indeed BIN(4, 0.5).

16.2.4 Suppose that counts of female offsprings in a certain animal species with four offsprings are 3, 8, 28, 40, 21. Test the hypothesis that the corresponding distribution is binomial.

16.2.5 Given the data on numbers of hits of various areas of London by V2 rockets (see Example 9.14), test the hypothesis that the numbers of hits have a Poisson distribution.

16.3 HOMOGENEITY AND INDEPENDENCE

Chi-square tests can also be used for testing hypotheses about jointly distributed categorical variables. The data obtained in a random sample are summarized in a *contingency table* with r rows and c columns representing levels of respective classification variables. If n observations are taken, then N_{ij} is the number of observations classified at the ith level of the first classification variable and at the jth level of the second. The counts can be arranged into a matrix $[N_{ij}]$, where we have $\sum_{ij} N_{ij} = n$.

We will postpone a more detailed analysis of types of contingency tables to subsequent sections. Here we will consider only the simplest case, corresponding to random sampling from a population whose elements are categorized according to some classification system (gender vs. educational level, smoking status vs. cause of death, etc.).

Let p_{ij} be the probability that a single observation belongs to the ith category in first classification, and jth category in the second classification. The marginal

probabilities here are

$$p_{i+} = \sum_j p_{ij}, \qquad p_{+j} = \sum_i p_{ij},$$

and the most obvious null hypothesis is that of independence of the two classifications:

$$H_0 : p_{ij} = p_{i+}p_{+j} \quad \text{for all } i, j.$$

The alternative hypothesis is

$$H_a : H_0 \text{ is not true.}$$

Let now $\hat{p}_{ij}$ denote the MLE of the probability p_{ij} based on the count matrix $[N_{ij}]$. Then the test statistic (16.4) takes on the form

$$Q^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}. \tag{16.6}$$

This statistic, according to Theorem 16.2.2, has a limiting (as $n \to \infty$) chi-square distribution with $rc - 1 - k$ degrees of freedom, where k is the number of estimated parameters.

To determine k, and also find the estimators $\hat{p}_{ij}$, observe that under the null hypothesis H_0 we have (by the invariance property of MLE's) $\hat{p}_{ij} = \hat{p}_{i+}\hat{p}_{+j}$. Now there are r values of the marginal probabilities p_{i+}, and c values of marginal probabilities p_{+j}, but in each of these marginal distributions one value is a function of others, for instance,

$$p_{r+} = 1 - p_{1+} - p_{2+} - \cdots - p_{r-1,+}$$

and similarly

$$p_{+c} = 1 - p_{+1} - p_{+2} - \cdots - p_{+,c-1}.$$

Thus the number of estimated parameters is $k = (r-1) + (c-1)$, and consequently the number of degrees of freedom of the limiting distribution of (16.6) is

$$rs - 1 - (r-1) - (c-1) = (r-1)(c-1). \tag{16.7}$$

Finally, we know that the MLE of a probability in multinomial distribution is the relative frequency, so

$$\hat{p}_{i+} = \frac{N_{i+}}{n}, \qquad \hat{p}_{+j} = \frac{N_{+j}}{n},$$

where

$$N_{i+} = \sum_j N_{ij}, \qquad N_{+j} = \sum_i N_{ij}. \tag{16.8}$$

Thus $\hat{p}_{ij} = N_{i+}N_{+j}/n^2$, and substitution into (16.6) gives the following theorem:

Theorem 16.3.1 *Assume that n elements of a random sample are classified according to two attributes with r and c categories, respectively, producing a contingency table (count matrix) $[N_{ij}]$. For testing the hypothesis*

$$H_0 : \textit{row and column classifications are independent}$$

against the general alternative

$$H_a: \text{ hypothesis } H_0 \text{ is false,}$$

one can use the statistic

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n}, \tag{16.9}$$

which (under H_0) has limiting chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

Any deviations from H_0 tend to increase Q^2. So the critical region contains values of Q^2 that exceed $\chi^2_{\alpha,(r-1)(c-1)}$.

■ **EXAMPLE 16.9**

The Special Election Issue of *Newsweek* (November/December 1992) gives the exit poll results for the 1992 presidential election. One of the tables is the following:

		Clinton	Bush	Perot
White	87%	41%	38%	21%
Black	9%	82%	11%	7%

Although the conclusion seems quite clear, let us try to analyze this table and test the hypothesis that the preferences for the three candidates are independent of race. The total sample size given is $n = 15,241$ voters, and the margin of error is given as 1.1 percentage points.

Observe first that the marginal percentages of whites and blacks do not add up to 100%, which means that the data for 4% of the voters (e.g., Hispanics) were not taken into account. Thus we can estimate the sample size for our test to be about $0.96 \times 15,241 = 14,631$. Now the marginal totals for whites and blacks can be estimated as

$$N_{W+} = 14,631 \times \frac{0.87}{0.87 + 0.09} = 13,259,$$

and therefore $N_{B+} = 1,327$.

Next the percentages in each row add to 100%; this allows us to estimate the counts $N_{W,Clinton}$, $N_{W,Bush}$, and so on. The whole contingency table takes on the form

	Clinton	Bush	Perot	
White	5,436	5,039	2,784	13,259
Black	1,125	151	96	1,372
	6,561	5,190	2,880	14,631

We can now compute the observed value of statistic Q^2 given by formula (16.9). We obtain here $Q^2 = 845.2$, which exceeds the critical value for a chi-square distribution with 2 degrees of freedom chosen for any reasonable level of significance α. The race and voting preference are definitely dependent.

Continuing this analysis we could ask what confidence level the pollsters use in announcing their "margin of error." Here it is given as ± 1.1 percentage points. We are estimating here the true proportion p on the basis of sample of size $n = 14,631$. If X is the number of observations of a given category, then X has a binomial distribution with parameters n and p; hence X is approximately normal $\mathrm{N}(np, np(1-p))$. Consequently, the estimated percentage, $100\hat{p} = 100X/n$ has an approximate normal distribution with mean $E(100\hat{p}) = 100p$ and $\mathrm{Var}(100\hat{p}) = \mathrm{Var}((100/n)X) = (100^2/n^2)\mathrm{Var}(X) = 100^2 p(1-p)/n$. The $(1-\alpha)$-confidence interval for the mean $100p$ is $100\hat{p} \pm z_{\alpha/2}\sqrt{\mathrm{Var}100\hat{p}}$.

The fact that the error given is the same for all data (instead of depending on the observed percentage) suggests that an upper bound $p(1-p) \leq 1/4$ is used. We have therefore the inequality

$$z_{\alpha/2}\sqrt{\frac{100^2 p(1-p)}{n}} \leq z_{\alpha/2}\frac{100}{\sqrt{4 \times 14,631}}.$$

The right-hand side equals 1.1 for $z_{\alpha/2} = 2.575$, which suggests that the pollsters used a 99% confidence level.

The chi-square test for independence described in Theorem 16.3.1 concerns the case where the counts N_{ij} arise from a cross-classification of independent and identically distributed observations. In experiments where one of the classification variables is not random but controlled by the experimenter, the data cannot be treated as a sample from a bivariate distributions.

■ EXAMPLE 16.10 Prospective and Retrospective Studies

Data summarized in contingency tables are usually used to analyze the hypothesized relationship between *cause* and *effect* (also referred to as *stimulus* and *response*, *explanatory* and *response* variable, *independent* and *dependent* variable, etc.). Let X and Y denote these variables, with X having r categories (also referred to as *levels* or *treatments*, depending on the context) and Y having c categories.

In *prospective studies* one selects groups of subjects corresponding to various levels of X and then classifies each group separately according to levels of Y. For instance, in the social sciences, one may be interested in productivity (Y) and its dependence on stress level (X). The data may result from selecting r groups of subjects, exposing the ith group to the ith stress level, and then observing levels of productivity. Consequently the totals N_{i+} (sizes of groups exposed to levels of stress) are not random. Similarly in medical research, one may select two groups of patients and then administer the treatment being studied to one group and a placebo to the other group. The response Y can be observed in both groups. Again, the sizes of the groups are not random but are under the control of the experimenter.

In *retrospective studies* the situation is similar, except that now the marginal counts N_{+j} of the response categories are controlled by the experimenter. An example might be provided by typical data on smoking habits and lung cancer.

A sample of subjects who died from lung cancer is selected and compared with a sample (possibly matched with respect to various attributes such as sex, age, etc.) of subjects who died from other causes. The sizes of these two samples are chosen largely at will. The two samples are then classified according to categories related to smoking.

For the analysis of contingency tables in which one of the marginal frequency vectors is fixed, let us first introduce the appropriate notation and then formulate the hypothesis to be tested.

The data form, as before, a count matrix $[N_{ij}]$, with marginal counts $N_{i+} = \sum_j N_{ij}$ and $N_{+j} = \sum_i N_{ij}$. The total number of observations is $n = \sum_{ij} N_{ij}$. For the sake of argument, assume that the counts $N_{i+}, i = 1, \ldots, r$, are not random. For each i, the vector $(N_{i1}, N_{i2}, \ldots, N_{ic})$ is assumed to represent the counts from N_{i+} iid observations, sampled from the distribution corresponding to the ith level of the first attribute. The probabilities in this distribution will be denoted by $(p_{1|i}, p_{2|i}, \ldots, p_{c|i})$, where $p_{1|i} + p_{2|i} + \cdots + p_{c|i} = 1$. Here $p_{j|i}$ stands for the probability that the observation will fall into the jth class of the second attribute, if the sample is taken from the population of objects with the ith level of the first attribute. Note that $p_{j|i}$ is not a conditional probability as long as i is not random. We use, however, the symbols appropriate for conditional probabilities, since in the special case where i is random, we have the obvious relation

$$p_{ij} = p_{j|i} p_{i+}. \tag{16.10}$$

The hypothesis of interest here is the hypothesis of homogeneity, which may be stated as follows:

$$H_0: \text{ distributions } (p_{1|i}, \ldots, p_{c|i}) \text{ do not depend on } i; \tag{16.11}$$

that is, for each j we have $p_{j|1} = \cdots = p_{j|r} = p_j$.

As before, we will obtain a test for H_0 against the general alternative

$$H_1: \text{ hypothesis } H_0 \text{ is not true.}$$

Despite the differences between the independence hypothesis and the homogeneity hypothesis above, they are tested by the same statistic. Indeed, for any fixed i, the component of the chi-square sum is

$$X_i^2 = \sum_{j=1}^{c} \frac{(N_{ij} - N_{i+}\hat{p}_{j|i})^2}{N_{i+}\hat{p}_{j|i}},$$

where $\hat{p}_{j|i}$ is the MLE of probability $p_{j|i}$. Clearly, under the null hypothesis of homogeneity we have

$$X_i^2 = \sum_{j=1}^{c} \frac{(N_{ij} - N_{i+}\hat{p}_j)^2}{N_{i+}\hat{p}_j}, \tag{16.12}$$

and $\hat{p}_j = N_{+j}/n$ (i.e., the MLE of p_j is the relative frequency of the jth category). Adding over i, we obtain the test statistic

$$Q^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n}. \tag{16.13}$$

The number of degrees of freedom equals $r(c-1)-k$, where k is the number of estimated parameters. Indeed, each Q_i^2 will have $c-1$ degrees of freedom if the value p_j is known. The parameters estimated are $p_1, \ldots, p_{c-1}$, so $k = c-1$, and we obtain the following theorem:

Theorem 16.3.2 *If the null hypothesis H_0 (16.11) is true, and if $N_{i+} \to \infty$ for $i = 1, \ldots, r$, then the statistic Q^2 given by* (16.13) *has the limiting chi-square distribution with $r(c-1)-(c-1) = (r-1)(c-1)$ degrees of freedom.*

Any violation of the null hypothesis will tend to increase the value of Q^2. So, again, the null hypothesis will be rejected if the observed value of the statistic Q^2 exceeds $\chi^2_{\alpha,(r-1)(c-1)}$.

PROBLEMS

16.3.1 Show that in the case of a 2×2 contingency table, the statistic Q^2 given by (16.9) is proportional to $(N_{11}N_{22} - N_{21}N_{12})^2$, and find the proportionality constant.

16.3.2 For each 3×4 contingency table below find such k that the hypothesis about independence of two classification variables is rejected at the 0.05 significance level.

(i)
$$\begin{matrix} k & k & k & k \\ k & k & 0 & k \\ k & k & k & k \end{matrix}$$

(ii)
$$\begin{matrix} 5 & 5 & 5 & 5 \\ 5 & 5 & k & 5 \\ 5 & 5 & 5 & 5. \end{matrix}$$

16.3.3 Mrs. Smith, who teaches an elementary statistics course, classified each student in the class according to whether the grade on the first exam was below or above the median for this exam, and then did the same for the second exam. The results obtained are:

	Second Exam	
First Exam	Below	Above
Below	30	5
Above	5	30

Compute Q^2 and find the p-value. What legitimate conclusion can be made?

16.3.4 Professionals from various disciplines participated in a study on job-related stress. A random sample of size 100 was selected from each group of professionals (physicians, engineers, and lawyers) and each person was asked to evaluate the level of job-related stress as low, moderate, or high. The results of a study are given below.

	L	M	H
Physicians	5	25	70
Engineers	25	25	50
Lawyers	10	30	60

Specify the hypothesis to be tested, perform the test, and make the appropriate conclusions.

16.3.5 A random sample of 29 university students was selected and each student was then classified according to their high school GPA and college GPA. Both classifications had the same two categories: "I, below 3.0," "II, at least 3.0." Formulate a hypothesis to be tested. Perform the test and make appropriate conclusions.

High School GPA	College GPA I	II
I	5	3
II	12	9

16.4 CONSISTENCY AND POWER

Chi-square tests of either independence or homogeneity serve as tests against a general alternative, asserting simply that the null hypothesis is false (such tests are called *omnibus tests*). If the independence hypothesis is not true, then (as the sample size increases) the probability of rejection of the null hypothesis tends to 1. We may rephrase this property by stating that power of the chi-square test for independence tends to 1 for any simple hypothesis contained in the alternative as the sample size increases. This property of the test is called *consistency*.

The situation is similar in the case of tests for homogeneity, except that now the sample sizes refer to rows (levels of explanatory variable), and are determined not by chance but by the experimenter. Again, if the null hypothesis is not true, then at least two of the rows of the matrix $[p_{i|j}]$ are different. The power of the chi-square test will tend to 1 on a simple alternative in which the rows labeled i_0 and i_0' are different, if the sample sizes N_{i_0+} and $N_{i_0'+}$ both tend to infinity.

In general, it is difficult to analyze the power of chi-square tests for independence or homogeneity, since it involves determining the exact or limiting distribution of a statistic under some distribution in the underlying population.

However, if one looks at the main motivation of an analysis of the power of a test, namely to decide "which hypothesis should be accepted if one rejects null hypothesis," then one can suggest the following approach.

When the null hypothesis is not valid, it typically is due to the fact that there is a strong association between some specific values of the two attributes analyzed, while for other values the null hypothesis is, at least approximately, satisfied.

In symbols, the null hypothesis (of independence) asserts that all absolute differences

$$|p_{ij} - p_{i+}p_{+j}| \tag{16.14}$$

are zero. If the null hypothesis is not valid, then there exist absolute differences that are positive. What typically happens in such cases is that a few of those differences are high; other may be close to zero (rather than all of these differences being small). In these cases one would like to identify the cells for which the differences (16.14) are high.

To present the solution, let us first derive an alternative to the chi-square test, namely the generalized likelihood ratio (GLR) test (see Section 13.6). Consider the case of testing the hypothesis of independence, which states that $p_{ij} = p_{i+}p_{+j}$ for all i, j. The union of H_0 and H_1 allows the probabilities p_{ij} to be arbitrary (subject only to the constraint that $\sum_{i,j} p_{ij} = 1$). The data form the contingency table $[N_{ij}]$.

The likelihood of the data is (up to a multiplicative constant) equal to $L = \prod_i \prod_j (p_{ij})^{N_{ij}}$. It is maximized at $\hat{p}_{ij} = N_{ij}/n$, where $n = \sum_{ij} N_{ij}$ is the total sample size. Consequently the denominator in the GLR, equal to the maximum over all parameter space, is

$$\sup_{H_0 \cup H_1} L = \prod_i \prod_j \frac{(N_{ij})^{N_{ij}}}{n^{N_{ij}}} = n^{-n} \prod_i \prod_j (N_{ij})^{N_{ij}}.$$

On the other hand, the likelihood over the null hypothesis equals (up to the same multiplicative constant)

$$L = \prod_i \prod_j (p_{i+})^{N_{ij}} (p_{+j})^{N_{ij}}.$$

This is maximized at $\hat{p}_{i+} = N_{i+}/n$ and $\hat{p}_{+j} = N_{+j}/n$, and we obtain

$$\begin{aligned} \max_{H_0} L &= \prod_i \left(\frac{N_{i+}}{n}\right)^{N_{i+}} \prod_j \left(\frac{N_{+j}}{n}\right)^{N_{+j}} \\ &= n^{-2n} \prod_i \prod_j (N_{i+})^{N_{+i}} (N_{+j})^{N_{+j}}. \end{aligned}$$

The generalized likelihood ratio equals therefore

$$\lambda = \frac{\prod_i (N_{i+})^{N_{i+}} \prod_j (N_{+j})^{N_{+j}}}{n^n \prod_i \prod_j (N_{ij})^{N_{ij}}} = \prod_i \prod_j \left(\frac{N_{i+} N_{+j}}{n N_{ij}}\right)^{N_{ij}}. \tag{16.15}$$

Under the null hypothesis H_0, the statistic

$$G^2 = -2 \log \lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \log \frac{N_{ij}}{(N_{i+} N_{+j})/n} \tag{16.16}$$

has a limiting (as $n \to \infty$) chi-square distribution with a number of degrees of freedom equal to the difference in the number of estimated parameters in the denominator and in the numerator. Thus the number of degrees of freedom is

$$(rc - 1) - [(r - 1) + (c - 1)] = (r - 1)(c - 1).$$

Symbolically we may write $Q^2 = \sum (N_{ij} - E(N_{ij}))^2 / E(N_{ij})$, where N_{ij} and $E(N_{ij})$ stand for observed and expected (under null hypothesis) counts. On the other hand, using the fact that

$$\log(1 + x) = x + \frac{x^2}{2} + \cdots,$$

we can write

$$\begin{aligned} G^2 &= 2 \sum_{i,j} N_{ij} \log \left(\frac{N_{ij}}{E(N_{ij})}\right) = 2 \sum_{i,j} N_{ij} \log \left(1 + \frac{N_{ij} - E(N_{ij})}{E(N_{ij})}\right) \\ &\approx 2 \sum_{i,j} N_{ij} \frac{N_{ij} - E(N_{ij})}{E(N_{ij})} + \sum_{i,j} \frac{(N_{ij} - E(N_{ij}))^2}{E(N_{ij})} \times \frac{N_{ij}}{E(N_{ij})}. \end{aligned}$$

Now, as $n \to \infty$, we have $N_{ij}/E(N_{ij}) \to 1$ in probability, and also $E[\sum(N_{ij} - E(N_{ij}))] = 0$. This suggests that the first sum is close to 0, and the second sum is close to Q^2.

Although the argument above falls short of being a proof, it suggests that Q^2 and G^2 are close to one another for large samples. In fact one can show that $Q^2 - G^2 \to 0$ in probability as $n \to \infty$. The idea of using the statistic G^2 to investigate the power lies in the additivity property of the chi-square distribution. If W has χ^2_k distribution, and $k = k_1 + k_2 + \cdots + k_s$ is a sum of positive integers, then there exist independent random variables $Y_1, \ldots, Y_s$ such that $W = Y_1 + \cdots + Y_s$ and $Y_i \sim \chi^2_{k_i}$.

In the case under consideration, we have the random variable G^2 given by (16.16), which (under the null hypothesis) has a limiting chi-square distribution with $(r-1)(c-1)$ degrees of freedom. It may be shown that G^2 can be represented as a sum of independent random variables $G_1^2, \ldots, G_s^2$, each corresponding to a subtable of the original contingency table.

The subtables are obtained by taking a part of the original table and then collapsing some of the categories. Goodman (1969, 1971) and Lancaster (1949, 1969) formulated following necessary conditions for components G_i^2 in the sum

$$G^2 = G_1^2 + G_2^2 + \cdots + G_s^2 \tag{16.17}$$

are independent:

1. The degrees of freedom for the components $G_1^2, \ldots, G_s^2$ must sum to the number $(r-1)(c-1)$ of the degrees of freedom of G^2.
2. Each cell count N_{ij} of the original table must appear in exactly one subtable.
3. Each of the marginal counts N_{i+} and N_{+j} of the original table must appear as marginal count in exactly one subtable.

If all three conditions are satisfied, then the values of statistic G^2 computed for the whole table and for the subtables satisfy (16.17).

■ **EXAMPLE 16.11**

Consider a 3×3 contingency table

	Y_1	Y_2	Y_3	Marginals
X_1	N_{11}	N_{12}	N_{13}	N_{1+}
X_2	N_{21}	N_{22}	N_{23}	N_{2+}
X_3	N_{31}	N_{32}	N_{33}	N_{3+}
Marginals	N_{+1}	N_{+2}	N_{+3}	

An example of a decomposition satisfying the three conditions of independence is as follows:

Subtable 1:

	Y_1	Y_2	
X_1	$\mathbf{N_{11}}$	$\mathbf{N_{12}}$	$N_{11} + N_{12}$
X_2	$\mathbf{N_{21}}$	$\mathbf{N_{22}}$	$N_{21} + N_{22}$
	$N_{11} + N_{21}$	$N_{12} + N_{22}$	

Subtable 2:

	Y_1 or Y_2	Y_3	
X_1	$N_{11} + N_{12}$	$\mathbf{N_{13}}$	$\mathbf{N_{1+}}$
X_2	$N_{21} + N_{22}$	$\mathbf{N_{23}}$	$\mathbf{N_{2+}}$
	$N_{11} + N_{12} + N_{21} + N_{22}$	$N_{13} + N_{23}$	

Subtable 3:

	Y_1	Y_2	
X_1 or X_2	$N_{11} + N_{21}$	$N_{12} + N_{22}$	$N_{11} + N_{12} + N_{21} + N_{22}$
X_3	$\mathbf{N_{31}}$	$\mathbf{N_{32}}$	$N_{31} + N_{32}$
	$\mathbf{N_{+1}}$	$\mathbf{N_{+2}}$	

Subtable 4:

	Y_1 or Y_2	Y_3	
X_1 or X_2	$N_{11} + N_{12} + N_{21} + N_{22}$	$N_{13} + N_{23}$	$N_{1+} + N_{2+}$
X_3	$N_{31} + N_{32}$	$\mathbf{N_{33}}$	$\mathbf{N_{3+}}$
	$N_{+1} + N_{+2}$	$\mathbf{N_{+3}}$	

The number of degrees of freedom of the original table is 4, and for each of the subtables it is 1, so the first condition is met.

Regarding conditions 2 and 3, we can check that they are met by simple inspection. In the four subtables, all entries and marginals from the original table are identified by boldface.

The algebraic verification that the condition (16.17) holds can be found in Lancaster (1949). Here we observe only that the partition of this example is not the only one possible for a 3×3 table: There are eight other partitions.

■ EXAMPLE 16.12

In a study of marijuana use in colleges, 445 students were sampled and classified according to a response variable (use of marijuana or other drugs) into three categories: "never," "occasionally," "regularly." As a possible explanatory variable the experimenters selected the number of parents who were alcohol or drug users ("neither one," "exactly one," "both"). The data that follow are from Devore (1991):

	Student's Drug Use			
Parents' Drug Use	Never (N)	Occasionally (O)	Regularly (R)	
None	141	54	40	235
One	68	44	51	163
Both	17	11	19	47
	226	109	110	445

For 3×3 tables, the statistic G^2 has a limiting chi-square distribution with 4 degrees of freedom. The value for the table above is $G^2 = 22.254$, which is highly significant (p-value equals 0.00018). Incidentally, for this table we have $Q^2 = 22.394$, which illustrates the closeness of G^2 and Q^2 for large samples. Thus we may conclude that there is a relationship between parental and student use of drugs.

One of the questions we could ask here is whether a positive or a negative example is stronger. In other words, taking for granted that the frequency of marijuana use tends to increase with the number of parents who use alcohol or drugs, the question is whether the effect of a bad example of one parent tends to outweigh the good example of the other parent. To get an insight into this question, we consider two decompositions of G^2. The first decomposition correspond to the following four subtables:

	1		2			3		4	
	N	O	$N+0$	R		N	O	$N+O$	R
None	141	54	195	40	At most one	209	98	307	91
One	68	44	112	51	Both	17	11	28	19

The values of G^2 four these for tables are, respectively, $G_1^2 = 4.344$, $G_2^2 = 10.957$, $G_3^2 = 0.616$, and $G_4^2 = 6.336$, and we check that $G^2 = G_1^2 + G_2^2 + G_3^2 + G_4^2$. The corresponding p-values for 1 degree of freedom are, respectively, 0.037, 0.001, 0.432, and 0.012.

An alternative decomposition is as follows:

	1*		2*			3*		4*	
	O	R	N	$O+R$		N	O	$N+O$	R
One	44	51	68	95	Neither	54	40	141	94
Both	11	19	17	30	At least one	55	70	85	125

Now the values are $G_{1\cdot}^2 = 0.871$, $G_{2\cdot}^2 = 0.470$, $G_{3\cdot}^2 = 3.893$, and $G_{4\cdot}^2 = 17.019$. The corresponding p-values are 0.351, 0.493, 0.048, and 0.000.

These results indicate that a bad example, of just one parent, has a big influence on a student's marijuana use. As subtables 1* and 2* show, if at least one parent uses drugs or alcohol, then it does not really matter whether the other parent does as well. On the other hand, as subtables 3* and 4* show, there is a significant difference where none of the parents use drugs or alcohol. This effect increases greatly the likelihood that a student will never use the marijuana, and—if he uses it—at a decreasing frequency of use.

The conclusions are strengthened if one analyzes also a first decomposition. Here the category "one" is either compared with "neither," or category "at most one" is compared with category "both." As can be seen, subtables 1, 2, and 4 show the significant effect of a parent using alcohol or drugs. Subtable 3 shows that any positive effect of the other parent is limited. The presence or absence of a positive model does not have any significant effect on the frequency of marijuana use as long as this frequency is low (or zero). Subtable 4 suggests that the presence of a positive parent role model has only the effect of lowering the probability of a regular use of marijuana during the college years.

PROBLEMS

16.4.1 Find a decomposition of G^2 into independent components by decomposing: **(i)** A $2 \times k$ table. **(ii)** A 3×4 table.

16.4.2 The data on the incidence of a certain disease, classified by age and gender, are as follows:

	< 20	20–39	40–59	60 +
Men	10	15	30	40
Women	20	30	60	300

Find the values of statistics Q^2 and G^2. Find the decomposition of G^2, and verify that the "source" for lack of independence is the very high incidence of this disease among women over 60.

16.4.3 At the beginning of the semester a random sample of 104 students was selected out of students in all introductory statistics classes. Students were then classified according to their GPA (I, "below 3.0"; II, "between 3.0 and 3.5"; and III, "above 3.5") and their attitude towards the statistics course (i, "I hate to take this class but I have to"; ii, "I do not mind taking this class but it is not one of my favorites"; iii, "I look forward to taking this class"). The results of classifications are given below.

	i	ii	iii
I	14	8	2
II	10	14	12
III	6	16	22

Test the hypothesis of independence of both classification variables. In the case of lack of independence, use decomposition to identify the cause of dependence.

16.5 2×2 CONTINGENCY TABLES

We begin with the simplest, but very common situation, where both variables have only two categories (i.e., are treated as binary variables). Denoting those levels by 1 and 2; we have then two distributions of the response, corresponding to the values of the explanatory variable, namely

$$(p_{1|1}, p_{2|1}) \quad \text{and} \quad (p_{1|2}, p_{2|2}),$$

where $p_{2|1} = 1 - p_{1|1}$ and $p_{2|2} = 1 - p_{1|2}$.

The null hypothesis of homogeneity reduces to

$$H_0 : \frac{p_{1|1}}{p_{1|2}} = 1.$$

When response 1 is in some sense undesirable (death, relapse of disease, etc.) the ratio $p_{1|1}/p_{1|2}$ was often called the *relative risk*. This term is now used generally regardless of the context (similarly as "success" and "failure" in a binomial distribution).

The alternative hypothesis of a positive association asserts that a higher value of explanatory variable gives a higher probability of a higher value of the response variable, that is, $p_{2|1} < p_{2|2}$. The latter inequality is equivalent to $1-p_{1|1} < 1-p_{1|2}$; hence $p_{1|1}/p_{1|2} > 1$. Similarly a negative association means that the relative risk is less than 1.

An alternative formulation, which also suggests a testing procedure, is as follows: Consider a binary distribution $(\pi, 1-\pi)$, where π is the probability of some event A. Then the ratio

$$\eta = \frac{\pi}{1-\pi} \tag{16.18}$$

is called the *odds* (for the event A). It is clear that η determines π uniquely, namely $\pi = \eta/(1+\eta)$. As π increases from 0 to 1, the odds η increase from zero to infinity. The odds for the complement of the event A are $1/\eta$. In the case of the distributions $(p_{1|1}, p_{2|1})$ and $(p_{1|2}, p_{2|2})$, the odds (for the response 1) are

$$\eta_1 = \frac{p_{1|1}}{p_{2|1}} \quad \text{and} \quad \eta_2 = \frac{p_{1|2}}{p_{2|2}}.$$

To formulate the null and alternative hypotheses, it appears natural to consider the *odds ratio* :

$$\theta = \frac{\eta_1}{\eta_2} = \frac{p_{1|1}p_{2|2}}{p_{1|2}p_{2|1}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}, \tag{16.19}$$

where the latter expression is meaningful in the case where the explanatory variable is random.

The null hypothesis, of either homogeneity or independence, has the form

$$H_0 : \theta = 1. \tag{16.20}$$

The alternative H_1^+ of positive and H_1^- of negative association are, respectively,

$$H_1^+ : \theta > 1 \quad \text{and} \quad H_1^- : \theta < 1. \tag{16.21}$$

Suppose now that the null hypothesis $H_0 : \theta = 1$ is to be tested against the alternative of positive association $H_a^+ : \theta > 1$. The data have the form of a 2×2 count matrix (contingency table):

$$\begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}.$$

The question is: How to determine the p-value of the observed contingency table, that is, the *probability* (calculated under the assumption that the null hypothesis is true) *of observing—if the experiment were to be repeated—a contingency table that will at least be as much in favor of the alternative as the contingency table actually observed.*

To implement the idea of determining the p-value of the observed result, one needs to proceed as follows:

1. Explicate the ordering of the contingency tables according to the relation of "being more in favor of the alternative hypothesis."

2. Specify the probabilities of occurrence of the various contingency tables under the null hypothesis. (Note that the null hypothesis of independence is composite; hence it does not lead directly to numerical values of probabilities.)

Regarding step 1, as a statistic, one can take a sample counterpart of the odds ratio θ, replacing the probabilities p_{ij} by the corresponding relative frequencies N_{ij}/n, where n is the total sample size. This leads to the statistic U, defined as

$$U = \frac{N_{11}N_{22}}{N_{12}N_{21}}$$

(and $U = \infty$ if N_{12} or N_{21} is zero).

Thus, if the observed contingency table is $[n_{ij}]$, then the p-value of this result is defined formally as

$$P_0\left\{[N_{ij}] : \frac{N_{11}N_{22}}{N_{12}N_{21}} \geq \frac{n_{11}n_{22}}{n_{12}n_{21}}\right\}. \tag{16.22}$$

It remains to specify the probability distribution P_0 on a suitably selected class of 2×2 contingency tables. The main requirement is that P_0 should not depend on any parameters so that its numerical value can be determined for every contingency table for which P_0 is defined.

Now, if we take a random sample of size n from a population whose elements are classified according to two dichotomous classifications, then the probability of a particular contingency table with its sum of entries equal n is given by the multinomial distribution

$$P\{N_{ij} = n_{ij}, i, j = 1, 2\} = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}}. \tag{16.23}$$

If the null hypothesis H_0 is true, then $p_{ij} = p_{i+}p_{+j}$, and substitution into (16.23) gives

$$P_{H_0}\{N_{ij} = n_{ij}, i, j = 1, 2\} = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} p_{1+}^{n_{1+}} p_{2+}^{n_{2+}} p_{+1}^{n_{+1}} p_{+2}^{n_{+2}}, \tag{16.24}$$

where $n_{i+} = n_{i1} + n_{i2}$ and $n_{+j} = n_{1j} + n_{2j}$ for $i, j = 1, 2$.

It should be clear that in order to obtain the probabilities of the contingency tables that do not depend on the parameter, it is necessary to restrict the definition of P_0 to the class of tables with given marginals n_{i+} and n_{+j} $(i, j = 1, 2)$. This suggests taking as P_0 the conditional probabilities given both marginals, since the product $p_{1+}^{n_{1+}} p_{2+}^{n_{2+}} p_{+1}^{n_{+1}} p_{+2}^{n_{+2}}$ will then cancel.

We have, for the probability of marginals being (n_{1+}, n_{2+}) and (n_{+1}, n_{+2}), the product of two binomial probabilities:

$$P\{N_{1+} = n_{1+}\} = \binom{n}{n_{1+}} p_{1+}^{n_{1+}} (1 - p_{1+})^{n - n_{1+}} = \binom{n}{n_{1+}} p_{1+}^{n_{1+}} p_{2+}^{n_{2+}}$$

and

$$P\{N_{+1} = n_{+1}\} = \binom{n}{n_{+1}} p_{+1}^{n_{+1}} (1 - p_{+1})^{n - n_{+1}} = \binom{n}{n_{+1}} p_{+1}^{n_{+1}} p_{+2}^{n_{+2}}.$$

Consequently

$$P\{N_{1+} = n_{1+}, N_{+1} = n_{+1}\} = P\{N_{1+} = n_{1+}\}P\{N_{+1} = n_{+1}\} \quad (16.25)$$
$$= \binom{n}{n_{1+}}\binom{n}{n_{+1}} p_{1+}^{n_{1+}} p_{2+}^{n_{2+}} p_{+1}^{n_{+1}} p_{+2}^{n_{+2}}.$$

Dividing (16.24) by (16.25), we obtain, after rearranging the multinomial coefficients, the following theorem in which $N_{ij}(x)$ is the 2×2 table with given marginals and $N_{11} = x$:

Theorem 16.5.1 *Under the null hypothesis* $H_0 : \theta = 1$, *for any integers* n, n_{1+}, n_{+1} *satisfying the conditions* $n > 0, 0 \leq n_{1+} \leq n, 0 \leq n_{+1} \leq n$,

$$P\{[N_{ij}] = [N_{ij}(x)] | N_{1+} = n_{1+}, N_{+1} = n_{+1}\} = \frac{\binom{n_{+1}}{x}\binom{n-n_{+1}}{n_{1+}-x}}{\binom{n}{n_{1+}}}$$
$$= \frac{\binom{n_{1+}}{x}\binom{n-n_{1+}}{n_{+1}-x}}{\binom{n}{n_{+1}}}. \quad (16.26)$$

Observe now that we have

$$\frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{x(n - n_{1+} - n_{+1} + x)}{(n_{1+} - x)(n_{+1} - x)},$$

which is an increasing function of x. This means that the ordering of contingency tables with the same marginal totals coincides with an ordering with respect to the element $x = n_{11}$. Consequently, in view of Theorem 16.5.1, the p-value defined by (16.22) becomes

$$\sum_{x \geq n_{11}} \frac{\binom{n_{1+}}{x}\binom{n-n_{1+}}{n_{+1}-x}}{\binom{n}{n_{+1}}}.$$

■ **EXAMPLE 16.13**

To study whether or not there exists a positive association between musical and mathematical abilities, a group of 12 fourth graders was classified according to their scores (high or low) in these subject areas. The results are as follows:

	Mathematics		
Music	High	Low	
High	4	1	5
Low	2	5	7
	6	6	12

Do the observed data lead to rejecting the null hypothesis of independence of musical and mathematical abilities in favor of the alternative of a positive association?

SOLUTION. There is only one contingency table with the same marginals as original one, which is more in favor of the alternative, namely

$$\begin{array}{cc|c} 5 & 0 & 5 \\ 1 & 6 & 7 \\ \hline 6 & 6 & 12 \end{array}$$

The p-value of the observed result is therefore the sum of probabilities of both tables; hence it is

$$\frac{\binom{5}{4}\binom{7}{2}}{\binom{12}{6}} + \frac{\binom{5}{5}\binom{7}{1}}{\binom{12}{6}} = 0.121.$$

Clearly, at level 0.1 (and therefore also at any lower level) the null hypothesis should not be rejected: There is about a 12% chance of observing a result at least as much in favor of the alternative hypothesis if only due to random fluctuations (i.e., if in fact the null hypothesis is true).

Let us mention here that the test as described above is applicable also to the cases of prospective or retrospective studies where one of the marginals is not random. Indeed, if the marginals n_{1+} and n_{2+} are selected by the experimenter, then the contingency table is

$$\begin{array}{cc|c} n_{11} & n_{12} & a \\ n_{21} & n_{22} & b \\ \hline n_{+1} & n_{+2} & n \end{array}$$

(where we let a and b denote the nonrandom marginals to distinguish them from the random marginals). The corresponding table of conditional probabilities, under the null hypothesis of homogeneity, is

$$\begin{bmatrix} p_{1|1} & p_{2|1} \\ p_{1|2} & p_{2|2} \end{bmatrix} = \begin{bmatrix} \gamma & 1-\gamma \\ \gamma & 1-\gamma \end{bmatrix}.$$

The likelihood of the data, under the null hypothesis, is therefore a product of two binomial probabilities:

$$\begin{aligned} &\binom{a}{n_{11}}\gamma^{n_{11}}(1-\gamma)^{n_{12}}\binom{b}{n_{21}}\gamma^{n_{21}}(1-\gamma)^{n_{22}} \\ &\quad = \binom{a}{n_{11}}\binom{n-a}{n_{+1}-n_{11}}\gamma^{n_{+1}}(1-\gamma)^{n_{+2}}. \end{aligned}$$

On the other hand, the probability of the observed column marginals is, again under H_0,

$$\binom{n}{n_{+1}}\gamma^{n_{+1}}(1-\gamma)^{n_{+2}}.$$

Thus the conditional probability, given the column marginal, is free of γ, and it equals

$$\frac{\binom{a}{n_{11}}\binom{n-a}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}},$$

which agrees with (16.26).

■ **EXAMPLE 16.14**

Returning to Example 16.13, the p-value of the observed contingency table would be the same as calculated there (i.e., about 0.12) if we selected six students with high and six with low math scores, and then classified them according to their music scores.

■ **EXAMPLE 16.15**

Twenty occasional headache sufferers participated in testing a newly developed headache remedy. They were given the drug and later asked whether or not it was significantly better than the drug they usually took. In fact, however, every fourth subject tested received not a drug but a placebo. The data are as follows:

	Significant Improvement	No Significant Improvement	
Drug	11	4	15
Placebo	3	2	5
	14	6	20

Do these data indicate that the new drug is better than the drugs usually taken?

SOLUTION. We are testing here the hypothesis of homogeneity against the alternative of a positive association between taking the new drug and beneficial effects for patients. To determine the p-value observe that the count n_{11}, given the marginals, may be only 11, 12, 13, or 14 if the table is to be at least as much in favor of the alternative as the observed one. Tables

$$\begin{bmatrix} 11 & 4 \\ 3 & 2 \end{bmatrix}, \begin{bmatrix} 12 & 3 \\ 2 & 3 \end{bmatrix}, \begin{bmatrix} 13 & 2 \\ 1 & 4 \end{bmatrix}, \begin{bmatrix} 14 & 1 \\ 0 & 5 \end{bmatrix},$$

have probabilities

$$\frac{\binom{15}{11}\binom{5}{3}}{\binom{20}{14}} = 0.352, \quad \frac{\binom{15}{12}\binom{5}{2}}{\binom{20}{14}} = 0.117, \quad \frac{\binom{15}{13}\binom{5}{1}}{\binom{20}{14}} = 0.014, \quad \frac{\binom{15}{14}\binom{5}{0}}{\binom{20}{14}} = 0.0004,$$

respectively. Hence the p-value is about 0.48, so there in no evidence to back the claim that the new drug is superior to the drugs used so far.

The testing procedure described above is known as the Fisher's exact test (see Section 13.7). In testing the null hypothesis $H_0 : \theta = 1$ against a two-sided alternative $H_1 : \theta \neq 1$, one encounters the usual problem of defining p-values in the case of two-sided alternatives. There seems to be no agreement among statisticians as to what is the proper procedure in such situations. In the case of continuous and symmetric distributions of a test statistic (e.g., Student's t), the p-value for the observed result is usually taken as a doubled p-value for the one-sided alternative. In the case of an asymmetric distribution, there is little justification of doubling the one-sided p-value. In the case of discrete distributions, an additional source of difficulty is

that by such a procedure one can obtain a value exceeding 1. Some authors suggest taking as the p-value the sum of all probabilities of tables that are at most as likely as the observed one (see Freeman and Halton, 1951).

In all situations already considered in this section, observations were classified according to two arbitrary variables. There are, however, cases when classification variables are dependent, very often it is the same classification variable applied to data collected on two occasions. For example one might want to study the effect of a certain medical treatment by comparing data from some tests collected from the same patients before and after the treatment was administered. If the classification variables have dichotomous response (such as 0 and 1, "Yes" and "No", etc.) one can use the so-called *McNemar's test* to compare marginal distributions. The McNemar's test is considered a counterpart of a matched-pair t test discussed in Chapter 13. The data to be analyzed are summarized as in the table below:

	Variable 2		
Variable 1	Yes	No	
Yes	N_{11}	N_{12}	N_{1+}
No	N_{21}	N_{22}	N_{2+}
	N_{+1}	N_{+2}	n

Marginal distributions are (p_{1+}, p_{2+}) and (p_{+1}, p_{+2}) for variable 1 and variable 2, respectively. To test if marginal distributions of classification variables are the same (marginal homogeneity) we will be testing

$$H_0 : p_{1+} = p_{+1} \quad \text{against} \quad H_1 : p_{1+} \neq p_{+1}. \tag{16.27}$$

Hypotheses in (16.27) reduce to hypotheses

$$H_0 : p_{12} = p_{21} \quad \text{and} \quad H_1 : p_{12} \neq p_{21}. \tag{16.28}$$

Of course we could perform the McNemar's test against one-sided alternatives, either

$$H_0 : p_{1+} = p_{+1} \quad \text{against} \quad H_1 : p_{1+} > p_{+1}, \tag{16.29}$$

or

$$H_0 : p_{1+} = p_{+1} \quad \text{against} \quad H_1 : p_{1+} < p_{+1}. \tag{16.30}$$

The measure of the difference between marginal probabilities $p_{1+} - p_{+1}$ reduces to $p_{12} - p_{21}$ and its MLE is the difference of relative frequencies $N_{12}/n - N_{21}/n$. Under the H_0 the distribution of N_{12} is BIN$(N^*, 1/2)$, where $N^* = N_{12} + N_{21}$. Since it is a symmetric distribution, its normal approximation N$(N^*/2, N^*/4)$ can be applied for $N > 10$. Consequently, the test statistic we then take

$$Z = \frac{N_{12} - N^*/2}{\sqrt{N^*/4}} = \frac{N_{12} - N_{21}}{\sqrt{N_{12} + N_{21}}} \tag{16.31}$$

has approximately N$(0, 1)$ distribution.

■ **EXAMPLE 16.16**

A students who was not doing well in his statistics class was offered an online tutorial to enhance his algebra skills. To assess the effectiveness of an online tutorial, the student was required to take a pre-test at the beginning and a post-test at the of the last tutorial session. Both, pre- and post-tests consisted of the same 30 multiple-choice questions so each answer was either correct or incorrect. Student's results are summarized in the table below.

	Post-test		
Pre-test	Correct	Incorrect	
Correct	6	2	8
Incorrect	18	4	22
	24	6	30

Did the student improve his algebra skills by taking the online tutorial?

SOLUTION. Here we will be testing null hypothesis $H_0 : p_{12} = p_{21}$ of no change against one-sided alternative $H_1 : p_{12} < p_{21}$. The value of the test statistic (16.31) is $(2-18)/\sqrt{2+18} = -3.578$, what corresponds to p-value equal 0.0002. The data provide strong evidence that student's algebra skills improved by taking online tutorial.

At the end of this section we would like to mention that similarly to all tests that were discussed we could obtain confidence intervals for various statistics. Unlike the McNemar's test statistic that uses only the information about the counts in different categories for two variables (N_{12} and N_{21}), while the counts N_{11} and N_{22} are irrelevant, the formula for the confidence interval depends on all four counts. Interested readers are advised to check the texts on the analysis of categorical data, for example Agresti (2002).

PROBLEMS

16.5.1 Use the data from Problem 16.3.5 to test if high school GPA's are positively related to college GPA's.

16.5.2 Two out of ten randomly selected men, and four out of ten randomly selected women were found to be allergic to some drug. **(i)** Does these data indicate that there is a difference between men and women in their propensity to develop an allergic reaction to the drug in question? **(ii)** Does these data indicate that men are less likely than women to develop an allergic reaction to the drug in question? Find corresponding p-values for (i) and (ii) and explain why are they different.

16.5.3 A study of change in employment status was conducted among residents of a certain county. A random sample of 100 adult residents was surveyed on their employment status, and their responses (Yes/No) were recorded. The same people were contacted after 12 months and asked the same question. Their answers are summarized in the table below. Does the employment status in the county changed significantly within these 12 months?

Survey 1: Employed	Survey 2: Employed Yes	No	
Yes	80	10	90
No	2	8	10
	82	18	100

16.5.4 Fifty randomly selected college students that take courses with on-line quizzes were classified according to their answer on two questions: "Do you cheat taking on-line quizzes?" and "Do other students that you know well cheat taking on-line quizzes?" Explain what can be said about student's cheating. Formulate appropriate hypotheses and perform the test using the data in the table below.

Student	Other students cheat Yes	No
Yes	5	2
No	13	30

16.6 $R \times C$ CONTINGENCY TABLES

At the end, we will extend the results for 2×2 to the general case of $r \times c$ tables, in which the categories corresponding to rows and those corresponding to columns are ordered. The assumption of Theorem 16.5.1 carries over to the present case. By conditioning on both marginals, we obtain (under the null hypothesis of independence or homogeneity) a probability distribution defined on the class of all $r \times c$ tables with the given marginals, which does not involve any unknown parameters.

For the case where both marginals are random, the situation is as follows: Suppose that we take n independent observations, each with the same distribution, and classify them according to two systems with r and c categories, respectively. The cell counts are N_{ij}, and the marginal counts are N_{i+} and N_{+j}, where $i = 1, \ldots, r$ and $j = 1, \ldots, c$. We have then the following theorem:

Theorem 16.6.1 *Under the assumption of independence, for any contingency table* $[n_{ij}], i = 1, \ldots, r, j = 1, \ldots, s$, *with* $\sum_{i,j} n_{ij} = n$ *and with marginal counts* n_{i+} *and* n_{+j}, *we have*

$$P\{[N_{ij}] = [n_{ij}] | N_{i+} = n_{i+}, N_{+j} = n_{+j}\} = \frac{\prod_{i=1}^{r}(n_{i+})! \prod_{j=1}^{c}(n_{+j})!}{n! \prod_{i=1}^{r} \prod_{i=1}^{c}(n_{ij})!}. \quad (16.32)$$

In order to implement Definition 16.22 for the p-value of an observed contingency table, say A_0, it remains to order all $r \times c$ tables with marginals the same as those of A_0, according to their "strength of support" for the alternative hypothesis of (say) a positive association.

In presenting the solution, we will use an approach different than that which we used for 2×2 tables. Instead of expressing the null hypothesis in terms of a single parameter and then finding its empirical counterpart (as we did in the case of 2×2 tables), we will introduce several indices (statistics, i.e., functions of the observed counts N_{ij}). Each of these indices will provide, on an intuitive ground, an ordering of contingency tables according to the strength of support for the alternative.

Symbolically, let A_0 be the observed contingency table, and let A be any contingency table with the same marginals as A_0. Furthermore, let $t = t(A)$ be a real-valued function defined on the considered class of contingency tables. Then the p-value of the observed table is defined as

$$\sum_{A:t(A)\geq t(A_0)} P_0(A). \tag{16.33}$$

The only problem remaining is the choice of statistic t so that its values could reflect the order of "strength of support" of the alternative hypothesis. The choices here are based on the following idea: Consider a pair of observations. Since each observation is classified as belonging to one of the rows and one of the columns, such a pair determines two pairs of coordinates, say (r', c') and (r'', c'').

Definition 16.6.1 The pair of observations is called *concordant* if

$$(r'' - r')(c'' - c') > 0;$$

it is called *discordant* if

$$(r'' - r')(c'' - c') < 0,$$

and it is called *tied* if $(r'' - r')(c'' - c') = 0$. □

It should be obvious that every concordant pair provides support in favor of the alternative of positive association: the differences $r'' - r'$ and $c'' - c'$ are both nonzero and of the same sign. This means that a higher evaluation on one variable is accompanied by a higher evaluation on the other variable. By the same logic, every discordant pair provides a support for the alternative of a negative association: a higher classification on one variable is accompanied by a lower evaluation on the other variable.

Let C = number of concordant pairs, and let D = number of discordant pairs. If we can assume that any concordant (or discordant) pair equally supports the alternative, then the overall support for the alternative of (say) a positive association is a function of the difference $C - D$. Standardizing this difference, we let

$$\hat{\gamma} = \frac{C - D}{C + D}. \tag{16.34}$$

The choice of the symbol reflects the fact that $\hat{\gamma}$ (the estimator of some parameter γ) depends on the observed contingency table.

We can now define the p-value of the observed contingency table (in testing the alternative of positive association) as the sum of probabilities of all the tables with the same marginals as the one observed, and with the value of index $\hat{\gamma}$ at least as high as for the original table.

■ **EXAMPLE 16.17**

For 2×2 tables, the numbers of concordant and discordant pairs of observations are $C = n_{11}n_{22}$ and $D = n_{12}n_{21}$, respectively; hence

$$\hat{\gamma} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = \frac{\theta - 1}{\theta + 1}, \qquad \text{where} \qquad \theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Notice that the ordering of the values of $\hat{\gamma}$ coincides with the ordering according to the values of θ.

■ **EXAMPLE 16.18**

For the case of marijuana use in colleges (recall Example 16.12), the categories represented by the rows and those represented by the columns were ordered by a natural way: by the number of parents who use drugs or alcohol and by the frequency of marijuana use by the student.

A concordant pair is formed by observations such that one of them is below and to the right of the other:

$$\begin{aligned} C &= 141 \times (44 + 51 + 11 + 19) + 54 \times (51 + 19) \\ &\quad + 68 \times (11 + 19) + 44 \times 19 = 24,281, \end{aligned}$$

and similarly

$$\begin{aligned} D &= 40 \times (68 + 44 + 17 + 11) + 54 \times (68 + 17) \\ &\quad + 51 \times (17 + 11) + 44 \times 17 = 12,366. \end{aligned}$$

Thus $\hat{\gamma} = 0.325$.

To evaluate the p-value of an observed contingency table, one should use a statistical package such as SAS.

The index $\hat{\gamma}$ utilizes only the information contained in the numbers C and D of concordant and discordant pairs. The numbers of tied pairs are not used, since a pair of observations tied on one or both variables provides evidence neither in favor of the null hypothesis nor in favor of the alternative. However, one could argue that a large number of tied pairs (as compared with $C + D$) is an indication that the difference $C - D$ might be insignificant. This argument has led to the introduction of two indices that take tied pairs into account, called Kendall's tau-b and Somers' d.

The total number of all possible pairs of observations (disregarding the order) is

$$\binom{n}{2} = \frac{n(n-1)}{2}.$$

Therefore the total number of tied observations is $n(n-1)/2 - C - D$. Let T_X, T_Y, and T_{XY} denote numbers of pairs of observations tied on the first coordinate (row), on the second coordinate (column), and on both coordinates (falling to the same cell), respectively.

Observe that the categories of ties are not disjoint: any pair tied on both coordinates is counted twice, in T_X and in T_Y.

Definition 16.6.2 The statistic

$$\tau_b = \frac{C - D}{\sqrt{[n(n-1)/2 - T_X]\,[n(n-1)/2 - T_Y]}} \tag{16.35}$$

is called *Kendall's tau-b*. □

Definition 16.6.3 The statistic

$$d = \frac{C - D}{n(n-1)/2 - T_X} \tag{16.36}$$

is called *Somers' d.* □

In the formulas above we have

$$T_X = \sum_{i=1}^{r} \frac{n_{i+}(n_{i+} - 1)}{2} \quad \text{and} \quad T_Y = \sum_{j=1}^{c} \frac{n_{+j}(n_{+j} - 1)}{2}. \tag{16.37}$$

■ **EXAMPLE 16.19**

Returning to the marijuana data (Examples 16.12 and 16.18), we have $T_X = 41,779$ and $T_Y = 37,306$. The total number of all possible pairs is $445 \times 444/2 = 98,790$. We have here $\tau_b = 0.201$ and $d = 0.209$ (while $\hat{\gamma} = 0.325$).

The calculation of the p-value based on τ_b or d for the alternative of (say) a positive association again uses the sum of all probabilities of tables with the same marginals as those observed, and the index τ_b, or d, at least as high as that for the observed table.

It ought to be mentioned that the very fact of an introduction of an index is—to a certain extent—an attempt to reduce the problem to one dimension. That enables one to compare objects (in this case, contingency tables) and select the best. This may be, however, a deceptive comfort, since not all things are comparable in such a simple way. In the present case, just thinking in terms of an index (e.g., $\hat{\gamma}$) leads to the danger of attaching significance to certain values of that index. For instance, one may tend to take $\hat{\gamma} = 1$ first as an indication of a very strong positive association (which it is), and then tend to attach to it a fixed significance level. To see this, observe that tables

$$\begin{bmatrix} 2 & 2 & 0 \\ 0 & 2 & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 3 & 3 & 0 \\ 0 & 2 & 2 \end{bmatrix}$$

both have $D = 0$; hence $\hat{\gamma} = 1$. The p-value for the first table equals

$$\frac{(4!4!)(2!4!2!}{8!2!2!2!2!} = 0.086,$$

whereas the p-value for the second table is about half of that for the first table:

$$\frac{(6!6!)(3!5!2!)}{10!3!2!3!2!} = 0.008.$$

This is simply the effect of regarding as equivalent two situations characterized by the same value of $\hat{\gamma}$. This example shows that the values of $\hat{\gamma}$ are not comparable for experiments with different marginal counts.

Computations of the p-value become cumbersome especially when the number of row and/or column categories increases. When cell counts are sufficiently large, normal approximation of sampling distribution may be used (e.g., see Agresti, 2002).

PROBLEMS

16.6.1 A study reported in *Science* magazine investigated the relationship between sex, handedness (right- or left-handed), and relative foot size (left foot bigger than right foot, left foot within one-half shoe size of right foot, or right foot bigger). A random sample of 150 adults gave the following data:

	Right-handed		Left-handed	
Relative foot size	Male	Female	Male	Female
$L > R$	2	55	6	0
$L \approx R$	10	18	6	2
$L < R$	28	14	0	9

Test the association between gender and handedness considering three groups of people: these whose left foot is bigger than their right foot, these whose both feet are almost the same, and these whose left foot is bigger. Compare the results obtained.

16.6.2 Measure the association of GPA and attitude toward statistics courses based on the data from Problem 16.3.5 using $\hat{\gamma}$, τ_b, and d coefficients.

Statistical Tables

Table A1 Binomial Distribution

Table A2 Standard Normal Distribution

Table A3 Student's t Distribution

Table A4 Chi-square Distribution

Table A5 Quantiles for the shortest CI for σ.

Table A6 F Distribution

Table A7 Kolmogorov Distribution

Table A1. Cumulative binomial probabilities $\sum_{k=0}^{x} b(k; n, p) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$.

		p										
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.33	0.35	0.40	0.45	0.50
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4489	0.4225	0.3600	0.3025	0.2500
	1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8911	0.8775	0.8400	0.7975	0.7500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.3008	0.2746	0.2160	0.1664	0.1250
	1	0.9927	0.9720	0.9392	0.8560	0.8437	0.7840	0.7452	0.7182	0.6480	0.5747	0.5000
	2	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9641	0.9571	0.9360	0.9089	0.8750
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.2015	0.1785	0.1296	0.0915	0.0625
	1	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5985	0.5630	0.4752	0.3916	0.3125
	2	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8918	0.8735	0.8208	0.7585	0.6875
	3	1.0000	0.9999	0.9995	0.9984	0.9961	0.9919	0.9881	0.9850	0.9744	0.9590	0.9375
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1350	0.1160	0.0778	0.0503	0.0313
	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4675	0.4284	0.3370	0.2562	0.1875
	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7950	0.7648	0.6826	0.5931	0.5000
	3	1.0000	0.9995	0.9978	0.9933	0.9844	0.9692	0.9564	0.9460	0.9130	0.8688	0.8125
	4	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9961	0.9947	0.9898	0.9815	0.9687
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0905	0.0754	0.0467	0.0277	0.0156
	1	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3578	0.3191	0.2333	0.1636	0.1094
	2	0.9978	0.9841	0.9527	0.9011	0.8306	0.7443	0.6870	0.6471	0.5443	0.4415	0.3438
	3	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.9031	0.8826	0.8208	0.7447	0.6562
	4	1.0000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9830	0.9777	0.9590	0.9308	0.8906
	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9987	0.9982	0.9959	0.9917	0.9844
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0606	0.0490	0.0280	0.0152	0.0078
	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2696	0.2338	0.1586	0.1024	0.0625
	2	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5783	0.5323	0.4199	0.3164	0.2266
	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8318	0.8002	0.7102	0.6083	0.5000
	4	1.0000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9566	0.9444	0.9037	0.8471	0.7734
	5	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9935	0.9910	0.9812	0.9643	0.9375
	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9994	0.9984	0.9963	0.9922
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0406	0.0319	0.0168	0.0084	0.0039
	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.2006	0.1691	0.1064	0.0632	0.0352
	2	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4764	0.4278	0.3154	0.2201	0.1445
	3	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7481	0.7064	0.5941	0.4770	0.3633
	4	1.0000	0.9996	0.9971	0.9896	0.9727	0.9420	0.9154	0.8939	0.8263	0.7396	0.6367
	5	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9813	0.9747	0.9502	0.9115	0.8555
	6	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9976	0.9964	0.9915	0.9819	0.9648
	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9993	0.9983	0.9961
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0272	0.0207	0.0101	0.0046	0.0020
	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1478	0.1211	0.0705	0.0385	0.0195
	2	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3854	0.3373	0.2318	0.1495	0.0898
	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6585	0.6089	0.4826	0.3614	0.2539
	4	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8602	0.8283	0.7334	0.6214	0.5000
	5	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9596	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9922	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9991	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980

Table A1. *(Continued)*

						p					
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.33	0.35	0.40	0.45	0.50
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0182	0.0135	0.0060	0.0025	0.0010
	1	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.1080	0.0860	0.0464	0.0233	0.0107
	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.3070	0.2616	0.1673	0.0996	0.0547
	3	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5684	0.5138	0.3823	0.2660	0.1719
	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7936	0.7515	0.6331	0.5044	0.3770
	5	1.0000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9268	0.9051	0.8338	0.7384	0.6230
	6	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9815	0.9740	0.9452	0.8980	0.8281
	7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9968	0.9952	0.9877	0.9726	0.9453
	8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9983	0.9955	0.9893
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
15	0	0.4632	0.2059	0.0874	0.0352	0.0134	0.0047	0.0025	0.0016	0.0005	0.0001	0.0000
	1	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0206	0.0142	0.0052	0.0017	0.0005
	2	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0833	0.0617	0.0271	0.0107	0.0037
	3	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.2171	0.1727	0.0905	0.0424	0.0176
	4	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.4148	0.3519	0.2173	0.1204	0.0592
	5	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.6291	0.5643	0.4032	0.2608	0.1509
	6	1.0000	0.9997	0.9964	0.9819	0.9434	0.8689	0.8049	0.7548	0.6098	0.4522	0.3036
	7	1.0000	1.0000	0.9994	0.9958	0.9827	0.9500	0.9163	0.8868	0.7869	0.6535	0.5000
	8	1.0000	1.0000	0.9999	0.9992	0.9958	0.9848	0.9711	0.9578	0.9050	0.8182	0.6964
	9	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9921	0.9876	0.9662	0.9231	0.8491
	10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9984	0.9972	0.9907	0.9745	0.9408
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9981	0.9937	0.9824
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9963
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
20	0	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0003	0.0002	0.0000	0.0000	0.0000
	1	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0036	0.0021	0.0005	0.0001	0.0000
	2	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0189	0.0121	0.0036	0.0009	0.0002
	3	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0642	0.0444	0.0160	0.0049	0.0013
	4	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1589	0.1182	0.0510	0.0189	0.0059
	5	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.3083	0.2454	0.1256	0.0553	0.0207
	6	1.0000	0.9976	0.9781	0.9133	0.7858	0.6080	0.4921	0.4166	0.2500	0.1299	0.0577
	7	1.0000	0.9996	0.9941	0.9679	0.8982	0.7723	0.6732	0.6010	0.4159	0.2520	0.1316
	8	1.0000	0.9999	0.9987	0.9900	0.9591	0.8867	0.8182	0.7624	0.5956	0.4143	0.2517
	9	1.0000	1.0000	0.9998	0.9974	0.9861	0.9520	0.9134	0.8782	0.7553	0.5914	0.4119
	10	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.9650	0.9468	0.8725	0.7507	0.5881
	11	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9881	0.9804	0.9435	0.8692	0.7483
	12	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9966	0.9940	0.9790	0.9420	0.8684
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9992	0.9985	0.9935	0.9786	0.9423
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9984	0.9936	0.9793
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9941
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998

Table A2. Standard normal cumulative distribution function $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8314	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table A3. Upper quantiles $t_{\alpha,\nu}$ of Student's t distribution with ν degrees of freedom.

	α									
ν	0.400	0.300	0.200	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.614	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.334	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
100	0.254	0.526	0.845	1.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table A4. Upper quantiles $\chi^2_{\alpha,\nu}$ of the chi-square distribution with ν degrees of freedom.

	α									
ν	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.000	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
120	83.852	86.923	91.573	95.705	100.624	140.233	146.567	152.211	158.950	163.648

Table A5. Quantiles of the chi-square distribution for determining the shortest CI for σ.

	$\alpha = 0.10$				$\alpha = 0.05$			
ν	α_1	$\chi^2_{\alpha_1,\nu}$	$\chi^2_{\alpha_2,\nu}$	%	α_1	$\chi^2_{\alpha_1,\nu}$	$\chi^2_{\alpha_2,\nu}$	%
1	0.09998	0.016	18.189	24.92	0.04999	0.004	19.511	24.99
2	0.09988	0.211	18.056	49.02	0.04998	0.103	21.640	49.48
3	0.09948	0.582	17.647	61.21	0.04988	0.351	20.726	61.82
4	0.09882	1.056	18.100	68.50	0.04969	0.708	21.047	69.10
5	0.09800	1.594	18.907	73.41	0.04943	1.139	21.806	73.94
6	0.09708	2.175	19.871	76.95	0.04911	1.623	22.736	77.41
7	0.09612	2.788	20.927	79.65	0.04876	2.147	23.791	80.04
8	0.09516	3.426	22.041	81.77	0.04839	2.703	24.910	82.11
9	0.09420	4.084	23.182	83.48	0.04802	3.284	26.083	83.78
10	0.09328	4.758	24.352	84.90	0.04764	3.886	27.270	85.16
11	0.09238	5.447	25.530	86.09	0.04726	4.505	28.472	86.32
12	0.09152	6.147	26.719	87.11	0.04689	5.141	29.689	87.31
13	0.09070	6.858	27.915	87.98	0.04653	5.790	30.915	88.16
14	0.08990	7.579	29.109	88.75	0.04619	6.451	32.153	88.91
15	0.08916	8.308	30.313	89.42	0.04585	7.123	33.386	89.56
16	0.08844	9.045	31.514	90.02	0.04552	7.804	34.619	90.14
17	0.08774	9.788	32.711	90.55	0.04521	8.495	35.859	90.67
18	0.08710	10.539	33.916	91.03	0.04490	9.193	37.091	91.13
19	0.08648	11.295	35.117	91.46	0.04461	9.899	38.328	91.56
20	0.08588	12.056	36.315	91.85	0.04433	10.612	39.563	91.94
21	0.08530	12.823	37.510	92.21	0.04405	11.331	40.791	92.29
22	0.08476	13.595	38.708	92.54	0.04379	12.056	42.024	92.61
23	0.08424	14.371	39.903	92.84	0.04354	12.787	43.254	92.91
24	0.08374	15.151	41.095	93.12	0.04329	13.523	44.478	93.18
25	0.08326	15.935	42.286	93.37	0.04306	14.264	45.706	93.43
26	0.08278	16.723	43.471	93.61	0.04283	15.009	46.928	93.67
27	0.08234	17.514	44.659	93.83	0.04261	15.759	48.149	93.89
28	0.08192	18.310	45.846	94.04	0.04240	16.513	49.369	94.09
29	0.08150	19.108	47.027	94.23	0.04219	17.271	50.583	94.28
30	0.08110	19.909	48.208	94.41	0.04199	18.032	51.797	94.46
40	0.07780	28.063	59.927	95.75	0.04031	25.823	63.834	95.78
50	0.07536	36.418	71.499	96.57	0.03903	33.855	75.694	96.59
60	0.07346	44.918	82.949	97.13	0.03803	42.064	87.418	97.14
80	0.07066	62.232	105.571	97.83	0.03653	58.859	110.535	97.84
100	0.06866	79.847	127.915	98.25	0.03544	76.015	133.321	98.26

Table A6(a). Upper quantiles $F_{0.05,\nu_1,\nu_2}$ of the F distribution.

	ν_1									
ν_2	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Table A6(a). (*Continued*)

ν_2	ν_1 12	15	20	24	30	40	60	120	∞
1	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.23	254.30
2	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Table A6(b). Upper quantiles $F_{0.01,\nu_1,\nu_2}$ of the F distribution.

	ν_1									
ν_2	1	2	3	4	5	6	7	8	9	10
1	4052.2	4999.5	5403.4	5624.6	5763.7	5859.0	5928.4	5981.1	6022.5	6055.9
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.47	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Table A6(b). (*Continued*)

ν_2	ν_1 12	15	20	24	30	40	60	120	∞
1	6106.3	6157.3	6208.7	6234.6	6260.7	6286.8	6313.1	6339.4	6366.0
2	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Table A7. Tail probabilities $1 - Q(z)$ of the Kolmogorov distribution.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.3	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9995	0.9992	0.9987	0.9981
0.4	0.9972	0.9960	0.9945	0.9926	0.9903	0.9874	0.9840	0.9800	0.9753	0.9700
0.5	0.9639	0.9572	0.9497	0.9415	0.9325	0.9228	0.9124	0.9013	0.8896	0.8772
0.6	0.8643	0.8508	0.8367	0.8222	0.8073	0.7920	0.7764	0.7604	0.7442	0.7278
0.7	0.7112	0.6945	0.6777	0.6609	0.6440	0.6272	0.6104	0.5936	0.5770	0.5605
0.8	0.5441	0.5280	0.5120	0.4962	0.4806	0.4653	0.4503	0.4355	0.4209	0.4067
0.9	0.3927	0.3791	0.3657	0.3527	0.3399	0.3275	0.3154	0.3036	0.2921	0.2809
1.0	0.2700	0.2594	0.2492	0.2392	0.2296	0.2202	0.2111	0.2024	0.1939	0.1857
1.1	0.1777	0.1700	0.1626	0.1555	0.1486	0.1420	0.1356	0.1294	0.1235	0.1177
1.2	0.1122	0.1070	0.1019	0.0970	0.0924	0.0879	0.0834	0.0794	0.0755	0.0717
1.3	0.0681	0.0646	0.0613	0.0582	0.0551	0.0522	0.0495	0.0469	0.0443	0.0420
1.4	0.0397	0.0375	0.0354	0.0335	0.0316	0.0298	0.0282	0.0266	0.0250	0.0236
1.5	0.0222	0.0209	0.0197	0.0185	0.0174	0.0164	0.0154	0.0145	0.0136	0.0127
1.6	0.0120	0.0112	0.0105	0.0098	0.0092	0.0086	0.0081	0.0076	0.0071	0.0066
1.7	0.0062	0.0058	0.0054	0.0050	0.0047	0.0044	0.0041	0.0038	0.0035	0.0033
1.8	0.0031	0.0029	0.0027	0.0025	0.0023	0.0021	0.0020	0.0018	0.0017	0.0016
1.9	0.0015	0.0014	0.0013	0.0012	0.0011	0.0010	0.0009	0.0009	0.0008	0.0007
2.0	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0004	0.0003	0.0003
2.1	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
2.2	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
2.3	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Bibliography

1. Agresti, A. 1984. *Analysis of Ordinal Categorical Data.* Wiley, New York.

2. Agresti, A. 2002. *Categorical Data Analysis*, 2nd ed. Wiley, New York.

3. Al–Ghamdi, A. S. 1991. The application of mathematical statistics to traffic accidents data. MS dissertation. Ohio State University, Columbus, OH.

4. Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: Critique des postulats de l'école Americaine. *Econometrica* **21**: 503–546.

5. Antle, C. E., and L. J. Bain. 1969. A property of maximum likelihood estimators of location and scale parameters. *SIAM Review* **11**: 251.

6. Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja. 1992. *A First Course in Order Statistics*. Wiley, New York.

7. Bartoszyński, R. 1974. On certain combinatorial identities. *Colloquium Mathematicum* **30** (2): 289–293.

8. Bartoszyński, R., and M. L. Puri. 1981. Some remarks on strategy in playing tennis. *Behavioral Science* **26**: 379–387.

9. Blackwell, D., and M. A. Girshick. 1954. *Theory of Games and Statistical Decisions*. Wiley, New York.

10. Casella, G. and R. L. Berger. 2002. *Statistical Inference*. Duxbury Press, Pacific Grove, CA.

11. Chernoff, H. and E. L. Lehmann. 1954. The use of the maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics* **25**: 579–586.

12. Chow, Y. S., H. Robbins, and D. Siegmund. 1972. *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.

13. Chow, Y. S., and H. Teicher. 1997. *Probability Theory*, 3rd ed. Springer-Verlag, New York.

14. Chung, K. L. 2001. *A Course in Probability Theory Revised*, 3rd ed. Academic Press, New York.

15. Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

16. Davidson, D., P. Suppes, and S. Siegel. 1957. *Decision Making: An Experimental Approach*. Stanford University Press, Palo Alto, CA.

17. De Groot, M. 1986. *Probability and Mathematical Statistics*. Addison-Wesley. Reading, MA.

18. Devore, J. L. 1991. *Probability and Statistics for Engineering and the Sciences*, 3rd ed. Brooks and Cole, Pacific Grove, CA.

19. Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

20. Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

21. Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*, vol. 1. Wiley, New York.

22. Feller, W. 1971. *An Introduction to Probability Theory and Its Applications*, vol. 2. Wiley, New York.

23. Fox, M., and R. James. 1987. *The Complete Chess Addict.* Faber and Faber, London.

24. Freedman, D., R. Pisani, R. Purves, and A. Adhikari. 1992. *Statistics*. 2nd ed. Norton, New York.

25. Freeman, G. H., and J. H. Halton. 1951. Note on an exact treatment of contingency, goodness-of-fit and other problems of significance. *Biometrika* **38**: 141–149.

26. Gafrikova, V., and M. Niewiadomska-Bugaj. 1992. Ordering and comparing estimators by means of Gini separation index. *Probability and Mathematical Statistics* **13**: 215–228.

27. Genest, C. 1987. Frank's family of bivariate distributions. *Biometrika* bf 74: 545–555.

28. Gibbons, J. D., J. W. and Pratt. 1975. P-values: Interpretation and methodology. *American Statistician* **29** (1): 20–25.

29. Good, P. I. 2005. *Introduction to Statistics Through Resampling Methods and R/S-Plus*. Wiley, New York.

30. Goodman, L. A. 1969. On partitioning chi-square and detecting partial association in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **31**: 486–498.

31. Goodman, L. A. 1971. The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association* **66**: 339–344.

32. Green, R. F. 1974. A note on outlier-prone families of distributions. *Annals of Statistics* **2**: 1293–1295.

33. Green. R. F. 1976. Outlier-prone and outlier-resistant families of distributions. *Journal of the American Statistical Association* **71**: 502–505.

34. Green, R. F. 1980. Bayesian birds: A simple example of Oaten's stochastic model of optimal foraging. *Theoretical Population Biology* **18** (2): 244–256.

35. Hollander, M., and D. A. Wolfe. 1999. *Nonparametric Statistical Methods*. Wiley, New York.

36. Hotelling, H., and L. M. Solomons. 1932. The limits of a measure of skewness. *Annals of Mathematical Statistics* **3**: 141–142.

37. Jordan, W. 1869. Ueber die Bestimmung der Genauigkeit mehrfach wiederholter Beobachtungen einer Unbekannten. *Astronom. Nachr.* **74**: 209–226.

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

39. Kennedy, W. J., and J. E. Gentle. 1980. *Statistical Computing*. Marcel Dekker, New York.

40. Kolmogorov, A. N. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin (English transl. by N. Morrison, *Foundations of the Theory of Probability*, Chelsea, New York, 1956).

41. Kotz, S., T. Kozubowski, and K. Podgorski. 2001. *The Laplace Distribution and Generalizations*. Birkhäuser, Boston.

42. Krantz, D. H., R. D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of Measurement*, vol. 1. Academic Press, New York.

43. Lancaster, H. O. 1949. The derivation of and partition of χ^2 in certain discrete distributions. *Biometrika* **36**: 117–129.

44. Lancaster, H. O. 1969. *The Chi-Squared Distribution*. Wiley, New York.

45. Larsen, R. L., and M. L. Marx. 1986. *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, Englewood Cliffs, NJ.

46. Lehmann, E. L., and J. P. Romano. 2005. *Testing Statistical Hypotheses*. Springer-Verlag, New York.

47. Luce, R. D., and D. H. Krantz, 1971. Conditional expected utility. *Econometrica* **39**: 253–271.

48. Marsaglia, G., and T. A. Bray. 1964. A convenient method for generating normal variables. *SIAM Review* **6**: 260–264.

49. Montgomery, D. C., E. A. Peck, and G. G. Vining. 2006. *Introduction to Linear Regression Analysis*, 4th ed. Wiley, New York.

50. Myers, R. H. 1986. *Classical and Modern Regression with Applications*, Duxbury Press, Boston.

51. Myers, R. H. and J. S. Milton. 1991. *A First Course in the Theory of Linear Statistical Models*. PWS-Kent, Boston.

52. Neyman, J. 1950. *A First Course in Probability and Statistics.* Holt, New York.

53. Neyman, J., and E. L. Scott. 1971. Outlier proneness of phenomena and of related distributions. In: J. S. Rustagi (ed.), *Optimizing Methods in Statistics*. Academic Press, New York, pp. 413–430.

54. O'Cinneide, C. A. 1990. The mean is within one standard deviation of any median. *American Statistician* **44**: 292–293.

55. Oehlert, G. W. 1992. A note on the delta method. *American Statistician* **46**: 27–29.

56. Ramsay, F. P. 1926. Truth and probability. [Reprinted in *Studies in Subjective Probability*, H. E. Kyberg and H. E. Smokler, eds. Wiley, New York, 1964, pp. 61–92].

57. Roberts, F. S. 1979. *Measurement Theory with Applications to Decisionmaking, Utility and the Social Sciences.* Addison-Wesley, Reading, MA.

58. Ross, S. 2006. *Simulation*. Academic Press, New York.

59. Savage, L. J. 1954. *The Foundations of Statistics*. Wiley, New York.

60. Scott, D. W. 1979. On optimal and data-based histograms. *Biometrika* **66**: 605–610.

61. Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

62. Shapley, L. S., and M. Shubik. 1954. A method of evaluating the distribution of power in a committee system. *American Political Science Review* **48**: 787–792.

63. Srivastava, R. C. 1981. On some characterization of the geometric distribution. In: C. Taillie, G. P. Patil, and B. A. Baldessari (eds.), *Statistical Distribution in Scientific Work*. Reidel Publishers, Dordrecht, pp. 349–355.

64. Stapleton, J. 1995. *Linear Statistical Models*. Wiley, New York.

65. Stevens, S. S. 1946. On the theory of scales of measurement. *Science* **103**: 677–680.

66. Stigler, S. M. 1986. *The History of Statistics.* Harvard University Press, Cambridge, MA.

67. Stigler, S. M. 1996. Statistics and the question of standards. *J. Res. Natl. Inst. Stand. Technol* **101**: 779–789.

68. Taylor, H. M., and S. Karlin. 1984. *An Introduction to Stochastic Modeling*. Academic Press, New York.

69. Valleman, P. P., and L. Wilkinson. 1993. Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician* **47** (1): 65–72.

70. von Neumann, J., and O. Morgenstern. 1944. *Theory of Games and Economic Behavior.* Princetown University Press, Princeton, NJ.

Answers to Odd-Numbered Problems

CHAPTER 1

1.2.1 (i) H1, H2, ..., H6, TTT, TTH, THT, THH; (ii) TT, HTT, THT, THHT, HTHT, HHTT, HHHH, THHH, HTHH, HHTH, HHHT.

1.2.3 134, 234, 135, 235.

1.2.5 (i) 600, 510, 501, 420, 411, 402, 330, 321, 312, 303, 240, 231, 222, 213, 204, 150, 141, 132, 123, 114, 105, 060, 051, 042, 033, 024, 015, 006; (ii) 600, 510, 420, 411, 330, 321, 222; (iii) Six allocations for two white balls: 200, 110, 101, 020, 011, 002. Fifteen allocations for four red balls: 400, 310, 301, 220, 211, 202, 130, 121, 112, 103, 040, 031, 022, 013, 004. Each allocation of two white and four red balls is a combination of two allocations ($6 \times 15 = 90$).

1.2.7 (i) THT; (ii) A_2, A_4, A_6; (iii) 6 (MG+SP, MG+SJ, MP+SJ, MP+SG, MJ+SG, MJ+SP).

1.2.9 Yes. If the answer is "yes" and the interviewer manages to find out that the repondent was not born in April, the privacy is not mantained.

1.3.1 (i) False; (ii) False; (iii) False; (iv) True; (v) True; (vi) True.

1.3.3 (i) $X = A$; (ii) $X = \emptyset$; (iii) $X = A^c$; (iv) $X = B \div A$.

1.3.5 $D_1 = E_{11}$, $D_2 = E_2$, $D_3 = E_8$, $D_4 = E_6$, $D_5 = E_4 = E_5$, $D_6 = E_3$, $D_7 = E_1$, $D_8 = E_9$, $D_9 = E_{10}$, $D_{10} = E_7$.

1.3.7 (i) $x = 0, y = 4$; (ii) Either $x = 0$ or $y \geq 3$; (iii) $x = 0$, no inference about y possible; (iv) $x \leq 4$, no inference about y possible .

1.4.1 $\lim A_n = \emptyset$.

1.4.3 (i) 2^{2^n} ; (ii) $2^{3 \times 2^{n-2}}$; (iii) 2^{n+1} ; (iv) 2 ; (v) Answers are the same as for the field.

CHAPTER 2

2.3.1 True:(i), (iii), (v); False: (iii), (iv), (vi), (vii), (viii).
2.3.3 (i) 13/24.
2.4.1 5/18.
2.4.3 5/6.
2.4.5 (i) 0.6; (ii) 0.4; (iii) 0.5.
2.4.7 (i) 0.1; (ii) 0.3; (iii) 1; (iv) 0.6.
2.4.9 0.1.
2.4.11 0.95.
2.4.13 3/7.
2.5.1 0.25.
2.5.3 25.

CHAPTER 3

3.2.1 (i) 10; (ii) 12.
3.2.3 (i) 10; (ii) 13; (iii) 14.
3.2.5 (i) $2/n$; (ii) $1/n$; (iii) 0.5; (iv) $2(n-3)/[n(n-1)]$.
3.2.7 (i) 28,800; (ii) 86,400.
3.2.9 (i) 153; (ii) a hat; (iii)138.
3.2.11 $(2n+1)!/(2^n n!)$.
3.2.13 320.
3.2.15 $k \geq n$.
3.2.17 (i) $v(A) = 0.6$; (ii) $v(A) = 0.4, v(B) = 0.25, v(x) = 7/60$ for other members.
3.2.19 0.016 for each non-permanent member, 0.168 for each permanent member.
3.3.1 (i) $\binom{1000}{51}$; (ii) $\binom{1000}{949}$; (iii) $\binom{1001}{50}$.
3.3.3 $k \geq 30$.
3.3.5 4!49!
3.3.7 (i) 4; (ii) 36; (iii) 5108; (iv) 624; (v) 3744; (vi) 54,912.
3.3.9 $p = 0.0444$. If Queen is replaced by Jack then $p = 0.0204$.
3.3.11 (i) 0.8964; (ii) 0.04255; (iii) 0.0611.
3.3.13 $A(n,k)/k^n$.
3.3.15 (i) 252; (ii) 0; (iii) 5.
3.3.17 $2\binom{a+b}{b+r+k}/\binom{a+b}{a}$.
3.6.1 1.11×10^{-14}

CHAPTER 4

4.1.1 (ii) (a) $P(Z=2)$, (b) $P(X+Y=0|Z>0)$, (c) $P[A \cap B|(A \cup B) \cap (Z>0)]$.
4.2.1 (i) False; (ii) False; (iii) True; (iv) True; (v) False.
4.2.3 $1/(1+a)$.
4.2.5 1/3.
4.2.7 1/4
4.2.9 (i) 0.492; (ii) 0.123; (iii) 1; (iv) 1/4; (v) 0.5, 0.105, 1, 0.210; (vi) 0.25 for n odd, less than 0.25 for n even; (vii) $P(\text{sum odd}) \rightarrow 0.5$, $P(\text{product odd}) \rightarrow 0.125$, $P(\text{product odd}|\text{sum odd}) \rightarrow 0.25$.
4.2.11 $3/4^n$.
4.2.13 (i) 0.145; (ii) 0.5; (iii) 0.005.
4.3.1 1/11.

4.3.3 (i) 0.2113; (ii) 0.2510; (iii) 0.2526; (iv) 0.2091.
4.3.5 (i) $\sum_{i=1} \pi_0 \cdots \pi_{i-1}(1-\pi_i)$; (ii) $1-(1-\mu)/(1-\lambda\mu)$.
4.4.1 (i) 0.5; (ii) 0.00276.
4.4.3 (i) 23/45; (ii) 13/45; (iii) 14/23.
4.4.5 15/29.
4.4.7 15/22.
4.4.9 1/3.
4.5.1 (i) False; (ii) True; (iii) False; (iv) True.
4.5.3 0 for $k \geq 1$ and $1-k$ for $k < 1$.
4.5.5 A and B are not independent–all other pairs are.
4.5.7 $0 < P(A) < 1$.
4.5.9 (i) True; (ii) True.
4.5.11 6/11.
4.5.13 6/11.
4.5.15 $(6!)^2/21^6 = 0.006$.
4.5.17 (i) No; (ii) No.
4.5.19 $P_A(Y = h_{2k}) = \prod_{j=1}^{2k}[1-(1-p(h_j))^3][1-p()h_{2k+1}]^3$, $P_B(Y = h_k) = \prod_{j=1}^{2k}[1-(1-p(h_{2j}))^3][1-p()h_{2k+2}]^3$.
4.6.1 (ii) $-1/(N-1)$.

CHAPTER 5

5.1.1 $P(X_{n+1} = j|X_n = i) = p_j$ for $i = 0, \ldots, m$. If $i > m$ then $P(X_{n+1} = j|X_n = i) = p_{j-i+m}$ for $j = 0, 1, \ldots$.
5.2.1 (i) $p_{jk} = p$ if $j = k-1$, r if $j = k$, and q if $j = k+1$. (ii) The formulas remain valid if we replace p by $p/(1-r)$ and q by $q/(1-r)$.
5.2.3 Rows of the matric are: (0, 1/2, 1/2), (1/4, 0, 3/4), (1/8, 7/8, 0).
5.2.5 $p_{i,0} = r_i/(r_i + r_{i+1} + \cdots), p_{i,i+1} = 1 - r_i/(r_i + r_{i+1} + \cdots)$.
5.3.3 $p_{up}^{(2)} = r^2(100-C)[2A + (100-B) + 2C]/40000$.
5.4.1 2.
5.4.5 $\binom{N}{j}\binom{N}{N-j}/\binom{2N}{N}$.
5.5.1 $p_{30-30} = \sum_{n=0}^{\infty}(2pq)^n p^2 = p^2/(1-2pq)$.

CHAPTER 6

6.2.1 (i) False; (ii) True; (iii) True; (iv) False; (v) True; (vi) True; (vii) False; (viii) False; (ix) True.
6.2.3 (i) 0.7; (ii) 0.0608; (iii) 0.7786.
6.2.5 0 for $t < 0$, $\pi(t/a)^2$ for $0 \leq t < a/2$, $\sqrt{(2t/a)^2-1} + 2(t/a)^2 - [\pi/2 - 2\arctan\sqrt{(2t/a)^2-1}]$ for $a/2 \leq t < a\sqrt{2}/2$, and 1 for $t \geq a\sqrt{2}/2$.
6.3.1 Even (56/91).
6.3.3 (i) 0.7753; (ii) 0.3162.
6.3.5 (i) $(1-e^{-\lambda/2})/(1-e^{-\lambda})$; (iii) $(e^{-\lambda/2} - e^{-\lambda})/(1-e^{-\lambda})$.
6.3.7 (i) $P(X_4 = -4) = P(X_4 = 4 = 1/16, P(X_4 = -2) = P(X_4 = -2) = 4/16, P(X_4 = 0) = 6/16$; (ii) 5/16; (iii) $P(X_4 > 0|X_4 \geq 4) = 5/11$ for $n = 4$, and $P(X_5 > 0|X_5 \geq 0) = 1$ for $n = 5$.
6.3.9 (i) $a = 5, b = 22$; (ii) 0.3195; (iii) 0.2065.
6.3.11 (i) 5/6; (ii) 0.115; (iii) 0.5.
6.4.1 (i) $P(Y = y) = 1/6$ for $y = 0, 4, 9, 16$ and $P(Y = 1) = 1/3$; $P(Z = 0.5) =$

$P(Z = 1.5) = 1/3$, $P(Z = 2.5) = P(Z = 3.5) = 1/6$.
6.4.3 (i)$2tf_X(t^2)$; (ii) $e^t f_X(e^t)$; (iii) $(1/t^2)f_X(1/t)$; (iv) $(1/t)f_X(\log t)$.
6.4.5 $g(u) = 2u\lambda e^{-\lambda u^2}$ for $u \geq 0$ and 0 for $u < 0$.
6.4.7 $f_Y(y) = 2/\sqrt{1-4y}$ for $0 \leq y < 0.25$, and 0 otherwise.
6.4.9 (i) U[0, 1]; (ii) U[$-1, 1$].
6.4.11 (i) $g(y) = 2/\sqrt{1-4y}$ for $0 \leq y \leq 0.25$ and 0 otherwise; (ii) $h(w) = 2$ for $0.5 \leq w \leq 1$, and 0 otherwise.
6.4.13 $g(u) = m^{-3/2}k\sqrt{2u}e^{-2bu/m}, u > 0$.
6.5.1 $h(t) = 1/(1-t), 0 < t < 1$.
6.5.3 (i) $q = (\alpha/2)e^{\alpha}$; (ii) For $0 < t < 1, S(t) = 0.5(1-t) + 0.5$ and $h(t) = 1/(2-t)$. For $0 < t < 1$, $S(t) = 0.5e^{-\alpha(t-1)}$ and $h(t) = \alpha$.
6.5.5 0.1876.

CHAPTER 7

7.1.1 (ii) 22/36, 0, 1/6; (iii) 2/3, 2/3, 10/26.
7.1.3 (ii) -0.1048.
7.1.5 (i) $k = 2$; (ii) 5/16.
7.1.7 (i) 6; (ii) 3/20; (iii) 0 for $u < 0, v < 0$, u^2v^3 for $0 \leq u \leq 1, 0 \leq v \leq 1$, v^3 for $u > 1, 0 \leq v \leq 1$, u^2 for $v > 1, 0 \leq u \leq 1$, and 1 for $u > 1, v > 1$.
7.1.9 (i) 0.6534; (ii) 0.5.
7.2.1 (i) 1/14; (ii) 0; (iii) 5/7; (iv) 4/7; (v) 11/14.
7.2.5 (i) 1; (ii) POI(λ); (iii) Yes.
7.2.7 1/18, 1/9, 2/9, 3/9.
7.2.9 (ii) $\int_0^{\infty} h_1(x)\exp\{-\int_0^x [h_1(u) + h_2(u)]du\}dx$.
7.2.11 (i) e^{-5}; (ii) 0.4678.
7.2.13 $(z/\sigma^2)\exp\{-z^2/(2\sigma^2)\}$ for $z > 0$.
7.2.17 (4) for X and (7) for Y.
7.3.3 (i) $1 - \exp\{-(\lambda_1 + \lambda_2)\}$; (ii) POI($\lambda_1 + \lambda_2$); (iii) BIN($n, \lambda_1/(\lambda_1 + \lambda_2)$).
7.3.7 (i) 1; (ii) $2x^2 + 1/3$ and $y^2 + 2/3$; (iii) 0.3011;
7.4.1 (i) $\{\pi(z^2+1)\}^{-1}$; (ii) $e^{u/2}/\sqrt{2\pi u}$; (iii) same as (i).
7.4.3 $kw^{\alpha-1}(1-w)^{\beta-1}$.
7.4.5 (i) $k = 1/[4(a+b)]$; (ii) $[(a-b)\sqrt{z} + b]/[4(a+b)z^2]$.
7.4.9 $b(1-\sqrt{w})^2$ for $0 \leq w \leq 1$.
7.5.1 (ii) 19/34.
7.5.3 (i) 4; (ii) 5/48; (iii) 10/18.

CHAPTER 8

8.2.1 (i) $a = 0.25, b = 0.35$; (ii) $a = 1 - m/2, b = m/2 - 0.4, 0.8 \leq m \leq 2$.
8.2.3 (i) 3.36; (ii) approximately 22.97.
8.2.5 (i) $1/(n\lambda)$; (ii) $a + (b-a)/(n+1)$.
8.4.3 511/8, 255/2048, $(1/8)\sum_{k=1}^{8} 2^{2^k}$.
8.4.5 (i) 1 ; (ii) 1 ; (iii) 1.5 ; (iv) $2\log 2$.
8.4.7 15.
8.5.1 $(e^{nt} - 1)/[n(e^t - 1)]$ for $t \neq 1$, and 1 for $t = 0$.
8.5.3 λ^4.
8.5.5 $\alpha_n = \sqrt{n}(\sqrt{n} - 1)\mu, \beta_n = \sqrt{n}$
8.5.7 (i) 0; (ii) $-2\sqrt{\alpha(\alpha+2)}(\alpha-1)/[\alpha(\alpha+3)]$; (iii) $(1-2p)/\sqrt{p(1-p)}$; (iv)

$\lambda^{-1/2}$.
8.5.9 (i) $e^{\lambda(e^{it}-1)}$; (ii) $p/(1-q^{it})$; (iii) $\lambda/(\lambda-it)$.
8.6.1 0.1633.
8.6.3 (i) $164-30c$; (ii) $(c-2)/s\sqrt{3}$; (iii) $-1.464 \le c \le 5.464$.
8.6.5 $0, 1-\rho^2, 0$.
8.7.1 (i) 2; (ii) $(13x^2+13x+4)/[6(4x+1)]$.

CHAPTER 9

9.1.1 (i) True; (ii) False.
9.1.3 (i) r/n; (ii) $-r(n-r)/[n^2(n-1)]$.
9.1.5 0.5.
9.1.9 $E(U)=6^5, \mathrm{Var}(U)=6^5(6^5-1)$.
9.1.11 $\pi=M/(1+M), \lambda_2=100M^2/[C(M+1)^2], \gamma=1-100M/[C(M+1)]$.
9.1.13 (i) $3(p^3+q^3)+12pq(p^2+q^2)+30p^2q^2$; (ii) $2(p^2+q^2)+6pq$.
9.2.1 $P(X=j)=\binom{5}{j}\binom{4}{3-j}/\binom{9}{3}, j=0,1,2,3$.
9.3.1 (i) $[\lambda]$; (ii) $0.5(1+e^{-2\lambda})$.
9.3.3 (i) 0.1094; (ii) 0.0042.
9.3.5 (i) 0.4164; (ii) 0.1677.
9.3.7 (i) 0.0588 ;(ii) 0.1567.
9.3.11 (i) $\sum_{k=7}^{1} 0\binom{10}{k}(9/30)^k(21/30)^{10-k}$; (ii) $7n/30$.
9.4.3 (i) 0.5665; (ii) $E(T)=(m+1)/\lambda, \mathrm{Var}(T)=(m+1)/\lambda^2$.
9.4.5 WEI$(k, 1/\theta)$.
9.4.7 (i) 2.5 years; (ii) 0.189.
9.5.1 (i) 0.4099; (ii) 0.4221 ; (iii) 0.0672 ; (iv) 0.6807 ; (v) 0.3078 ; (vi) 0.1335.
9.5.3 0.1747
9.5.5 About 48 feet above the average level.
9.5.7 (i) 0.3520; (ii) 0.0793.
9.6.1 (i) BETA(β,α); (ii) $\Gamma(\alpha+\beta)\Gamma(k+\alpha)\Gamma(m+\beta)/[\Gamma(\alpha)\Gamma(\beta)\Gamma(k+m+\alpha+\beta)]$.
9.6.3 $\alpha=\beta=(1/k^2-1)/2$, provided $|k|<1$.

CHAPTER 10

10.1.3 $\mathrm{N}(\mu,(\sigma^2/9)(1/n_1+1/n_2+1/n^3))$ and $\mathrm{N}(\mu,\sigma^2/(n_1+n_2+n_3))$.
10.2.3 0.9975.
10.3.1 EXP$(n\lambda)$
10.3.3 (i) 1/2; (ii) $1-(1/2)^{n-1}$.
10.3.5 7.
10.4.1 0, 1, 0, 1, 2, 4.
10.4.3 0.336517, -0.037361.
10.4.5 0.146930, 0.115247, 0.420803.
10.4.9 (ii) (0.548291, 0.304935).
10.5.3 The cdf's of the limiting distributions are: (i) $1-y^{-2n}$ for $y \ge 1$ and 0 otherwise; (ii) $1-y^{-2}$ for $y \ge 1$, and 0 otherwise; (iii) e^{-1/y^2} for $y>0$ and 0 otherwise.
10.5.5 0.3679, 0.6262.
10.6.1 $c=1.939$.
10.6.3 $a=11.16, b=1.24$.
10.6.5 0.36.

10.6.7 $n \geq 32$.
10.6.9 0.9875.
10.6.11 $\sqrt{n}(X_n(1-X_n) - p(1-p)) \to N(0, p(1-p)(1-2p)^2)$.

CHAPTER 12

12.2.5 $T = NX_{1:n}$.
12.3.1 (i) $a+b=1$; (ii) $b = 1-a, a = \sigma_2^2/(\sigma_1^2+\sigma_2^2)$; (iii) $a = (\sigma_2^2 - C)/(\sigma_1^2 + \sigma_2^2 - 2C)$.
12.3.3 $\theta^2/n, \theta^2/(n+1)$.
12.3.5 $(1-\theta)/(n\theta^2), (1-\theta)(n+1-\theta)/[\theta^2(n+1)^2]$.
12.3.7 (ii) $1/(3n), 2/[(n+1)(n+2)]$.
12.4.1 0.5.
12.4.3 $1/(\lambda^2 n)$.
12.4.5 (i) $2/\sigma^2, 1/(2\sigma^4)$; (ii) 0, $1/(2\sigma^4)$.
12.4.7 (i) $\gamma_1 = \sigma^2/n, \gamma_2 = \mu$; (ii) $\gamma_1 = p(1-p)/n, \gamma_2 = p$.
12.5.1 (i) $T_a = (\overline{X})^2/[(1/n)\sum X_i^2 - (\overline{X})^2], T_b = \overline{X}/[(1/n)\sum X_i^2 - (\overline{X})^2]$; (ii) $T_a = b\overline{X}, T_b = a/\overline{X}$.
12.5.3 (i) $T = \overline{X} + \sqrt{\overline{X}^2 + 2\overline{X} + 3}$; (ii) $X_{n:n}$; (iii) same as in (ii) except that it cannot be observed when $X_{n:n} \leq -1$; (iv) $\hat{\theta} = (n - U_n)/U_n$ if $U_n > 0$.
12.5.7 (i) $k/(k+1)$; (ii) $(k-1)/k$.
12.5.11 0.5
12.5.13 (i)$\overline{X}$. (ii) MLE does not exist.
12.5.15 $\sqrt{(1/n)\sum(X_i - \mu)^2}$.
12.5.17 $(X_{1:n} + X_{n:n})/2$
12.5.19 0.0587.
12.5.21 (i) $\hat{\mu_1} = \overline{X}, \hat{\mu_2} = \overline{Y}, \hat{\sigma}^2 = [\sum(X_i - \overline{X})^2 + \sum(Y_j - \overline{Y})^2]/(m+n)$.
12.6.1 (i) $\sum X_i^2$; (ii) $\sum |X_i|^2$; (iii) $\prod X_i(1-X_i)$; (iv) $X_{1:n}$, and $X_{n:n}$.
12.6.3 $X_{n:n}$.
12.6.5 $\sum X_i$ and $X_{1:n}$.
12.6.7 $T_1 = \prod X_i, T_2 = \sum X_i$ are minimal jointly sufficient.
12.6.9 $\sum_{i=1}^{k-1} X_{i:n} + (n-k+1)X_{k:n}$.
12.7.1 BETA(5, 3), 0.227.
12.7.3 (i) $\prod X_i$; (ii) $\hat{\theta}_n(1 \pm z_{\alpha/2}/\sqrt{n})$.
12.7.5 (i) 1.117 ± 0.035 and 1.117 ± 0.053; (ii) 1.117 ± 0.016 and 1.117 ± 0.020 ; (iii) [0.00068, 0.00520] and [0.026, 0.072].
12.7.7 (i) $E(L_\alpha^2) = 4\sigma^2[t_{\alpha/2,n-1}]^2/n$; (ii) $n \geq 8[t_{\alpha/2,n-1}]^2$.
12.7.9 $P\{\chi_{24}^2 < 35.179.\}$
12.7.11 (i) $\overline{x} \pm z_{\alpha/2}\sqrt{\overline{x}/n}$; (ii) CI from part (i) is [3.955, 5.825], the approximated CI is 4.80 ± 0.9306.
12.7.13 (i) [0.00058, 0.00365] ; (ii) [274.05, 1728.62] ; (iii) same as in (ii); (iv) [0.0049, 0.4294].
12.7.15 (i) [0.428, 1.081]; (ii) [0.856, 5.459].

CHAPTER 13

13.2.1 $\pi(\theta) = 0.5(2\theta + \sqrt{0.1})^2$ for $0 < \theta < 0.5(1-\sqrt{0.1})$, $\pi(\theta) = 1 - 0.5(2 - 2\theta - \sqrt{0.1})^2$ for $0.5(1-\sqrt{0.1}) \leq \theta < 1 - 0.5\sqrt{0.1}$, and $\pi(\theta) = 1$ for $\theta \geq 1 - 0.5\sqrt{0.1}$.
13.2.3 (i) $\pi_a(\theta) = \pi_1(\theta)\pi_2(\theta)\pi_3(\theta)$, $\pi_b(\theta) = \pi_1(\theta)\pi_2(\theta)\pi_3(\theta) + [1-\pi_1(\theta)]\pi_2(\theta)\pi_3(\theta) +$

$\pi_1(\theta)[1-\pi_2(\theta)]\pi_3(\theta)+\pi_1(\theta)\pi_2(\theta)[1-\pi_3(\theta)]$, $\pi_c(\theta)=1-[1-\pi_1(\theta)][1-\pi_2(\theta)][1-\pi_3(\theta)]$; (ii) A is always better than any test C_i, B is better than any test C_i if $0<\pi(\theta)<0.5$, C is always worse.
13.2.5 (i)$r(5-r)/10$; (ii) $2r(5-r)/25$; (iii) P(Type I error)=0 in each case.
13.3.1 (i) Reject H_0 if $\sum X_i \geq 6$; (ii) 0.017.
13.3.3 X-the number of failures preceding the rth success. Reject H_0 if $X \geq 7$, accept H_0 if $X \leq 5$. For $X=6$ reject H_0 with prob. 0.5413, $\beta=0.7837$.
13.3.5 Reject H_0 if $2\lambda_0 \sum X_i \leq \chi^2_{0.05,2n}$, $\pi(\lambda)=P\{\chi^2_{2n} \leq (\lambda/\lambda_0)\chi^2_{0.05,2n}\}$. Reject H_0 if $X_{1:n}<-(\log 0.95)/(\lambda_0 n)$, $\pi_{X_{1:n}}=1-(0.95)^{\lambda/\lambda_0}$.
13.3.7 $1-r(r-1)/30$.
13.4.3 Reject H_0 is $\sum X_i^2 < k$.
13.4.5 $H_1: p>0.6$, p-value=0.072.
13.4.7 (i) Reject H_0 for large values of $T_1=\prod X_i$; (ii) Reject H_0 for small values of $T_2=\sum X_i$.
13.4.9 Reject H_0 if $\sum X_i \geq r$, where r is the smallest integer such that $\sum_{x=1}^{r} C_0[\theta_0/(\theta_0+1)^x \geq \alpha$.
13.5.1 $c_1=3.71, c_2=5$.
13.5.3 Reject H_0 if $|X-\lambda_0|/\sqrt{\lambda_0} > z_{\alpha/2}$.
13.6.1 $a<0.8543$.
13.6.3 (i) Not significant at 0.05 level; (ii) $m<24,940$.
13.6.5 $-2\log\lambda=3.58<\chi^2_{0.05,1}=3.841$ Do not reject H_0.
13.6.7 $\lambda=(m+n)^{m+n}/[(m+n\overline{Y}/\overline{X})^m(m\overline{X}/\overline{Y}+n)^n]$.
13.7.1 0.536.

CHAPTER 14

14.2.1 (i) $E(X|Y=y)=y/2$ and $X=Y$;(ii) $E(Y|X=x)=x$ and $Y=X$.
14.2.5 $a=E(Y)-b_1E(X_1)-b_2E(X_2), b_1=(\rho_Y/\rho_X)(\rho_{X_1,Y}-\rho_{X_1,X_2}\times\rho_{X_2,Y})/(1-\rho^2_{X_1,X_2})$, $b_2=(\rho_Y/\rho_X)(\rho_{X_2,Y}-\rho_{X_1,X_2}\times\rho_{X_1,Y})/(1-\rho^2_{X_1,X_2})$.
14.2.7 $0.335+0.475X$.
14.3.1 Any point (a,b) such that $a>0.9, b<-a/2-0.05, b<-a+0.6, b>a-1.8$ and $b>-a/2-0.45$.
14.3.3 $\hat{\lambda}=\overline{X}.\hat{p}=\sum Y_i/\sum X_i$.
14.4.5 $\hat{b}=\sum x_iy_i/\sum x_i^2, \hat{\sigma}^2=(1/n)\sum(y_i-\hat{b}x_i)^2$.
14.5.1 Reject H_0, $F=451.14>F_{0.05,4,6}=4.53$.
14.6.1 (i) [0.980, 1.061]; (ii) [1.014, 1.089]; (iii) [1.006, 1.076].
14.6.3 $\hat{b}x_0 \pm t_{\gamma/2,n-1}\sqrt{1+x_0^2/\sum x_i^2}\sqrt{\sum(y_i-\hat{b}x_i)^2/\sqrt{n-1}}$.
14.7.1 $13.646<x_0<17.883$.
14.8.1 $(\sum Y_i)/(\sum x_i)$.
14.11.3 (i)χ^2_{n-k}; (ii) Fn_k-1, n_1-1 distribution.
14.12.1 $F_W=2.94<F_{0.01,3,6}=10.72, F_D=15.16>F_{0.01,2,6}=9.76$. There is no effect of initial weight, but the type of diet affects final results.
14.13.1 $F_A=6.325>F_{0.05,2,18}=3.55, F_B=38.855>F_{0.05,2,18}=3.55$, $F_{AB}=4.970>F_{0.05,4,18}=2.93$. There is significant effect of the age, marijuana use, and their interaction on the level of emotional maturity.

CHAPTER 15

15.3.3 The joint density is $1/(4\sqrt{1-x^2})$ for $x^2+y^2 \leq 1$ and 0 otherwise/ Expected sample size is $8/\pi = 2.55$ under the "correct" scheme and $1+\pi/2 = 2.57$ under the "time saving" scheme.
15.3.5 Reject H_0 at 0.05 level.
15.3.7 $k > 1.85 + 1.36\sqrt{1.85+4m^2}$.
15.4.3 $Z = 1.647$, significant at 0.1 level.
15.4.5 $Z = 5.668$ for Problem 15.3.6 (i) and (ii), $Z = 4.808$ for Problem 15.3.7. Reject H_0 in all cases.
15.5.3 $R = 16$ or $R = 17$ depending on the value of d. $Z = 1.095$ and $Z = 1.596$, respectively.

CHAPTER 16

16.2.1 $Q^2 = 0.323 < \chi^2_{0.05,2} = 5.991$. Do not reject H_0.
16.2.3 $Q^2 = 1.307 < \chi^2_{0.1,4} = 7.779$. Do not reject H_0.
16.2.5 $Q^2 = 1.019 < \chi^2_{0.05,3} = 7.815$. Do not reject H_0.
16.3.1 $4/n^2$.
16.3.3 $Q^2 = 35.714$, p-value is below 0.005.
16.3.5 $Q^2 = 0.069 < \chi^2_{0.1,1} = 2.706$. Do not reject H_0.
16.4.3 $Q^2 = 18.714 > \chi^2_{0.05,4} = 9.488$. Reject H_0.
16.5.1 p-value= 0.5675.
16.5.3 $Z = 2.309$, the employment status in the county has changed.
16.6.1 For variables "gender-M/F" and "handedness-R/L" the values of $\hat{\gamma}$ are; $-1, -0.6875, 1$.
p-values for the tests of negative association in the first two tables are 0 and 0.058.
p-value for testing for positive association in the last table is 0.

INDEX

This Page Intentionally Left Blank