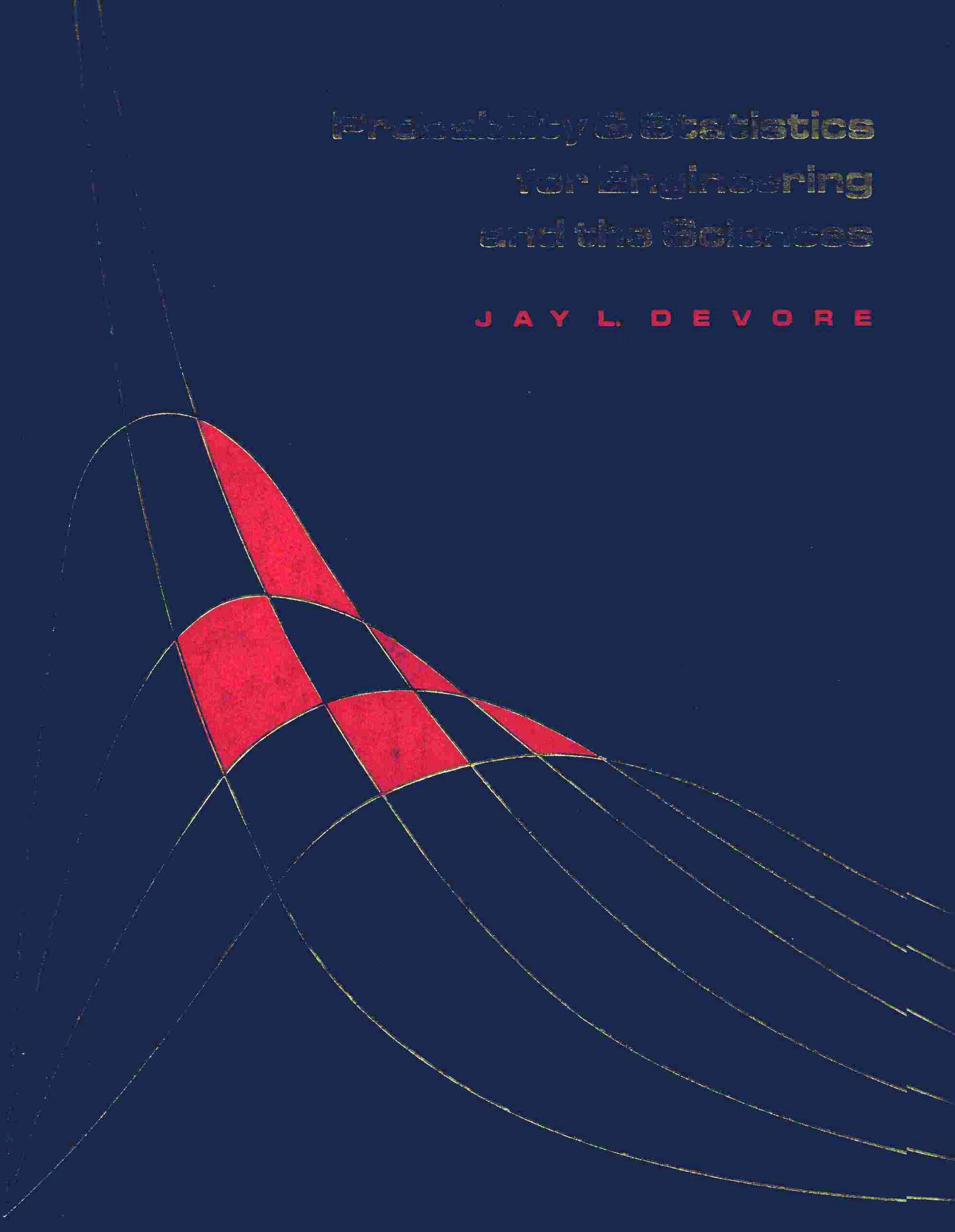
Probability & Statistics for Engineering and the Sciences

JAY L. DEVORE



Probability and Statistics for Engineering and the Sciences

JAY L. DEVORE

California Polytechnic State University
San Luis Obispo

Brooks/Cole Publishing Company Monterey, California

To Carol, Allie, and Teri

Brooks/Cole Publishing Company
A Division of Wadsworth, Inc.

Printed in the United States of America

10 9

Library of Congress Cataloging in Publication Data

Devore, Jay L.
Probability and statistics for engineering
and the sciences.

Bibliography: p.
Includes index.
1. Probabilities. 2. Mathematical
statistics. I. Title.
QA273.D46 519.5 81-21744
ISBN 0-8185-0514-1 AACR2

Acquisition Editor: Craig Barth
Production Coordinator: Cece Munson
Book Design: Albert Burkhardt
Production: Greg Hubit Bookworks, Larkspur, California
Typesetting: Interactive Composition Corporation, Pleasant Hill, California

Probability and Statistics for Engineering and the Sciences

Preface

The use of probability models and statistical methods for analyzing data has become common practice in virtually all scientific disciplines. This book attempts to provide a comprehensive introduction to those models and methods most likely to be encountered and used by students in their careers in engineering and the natural sciences. Although the examples and exercises have been designed with scientists and engineers in mind, most of the methods covered are basic to statistical analyses in many other disciplines, so that students of business and the social sciences will also profit from reading the book.

Students in a statistics course designed to serve other majors may be initially skeptical of the value and relevance of the subject matter, but my experience is that students *can* be turned on to statistics by the use of good examples and exercises which blend their everyday experiences with their scientific interests. Consequently, I have worked hard to find examples of real, rather than artificial, data—data that someone thought was worth collecting and analyzing. Many of the methods presented, especially in the later chapters on statistical inference, are illustrated by analyzing data taken from a published source, and many of the exercises also involve working with such data. Sometimes the reader may be unfamiliar with the context of a particular problem (as indeed I often was), but I have found that students are more attracted by real problems with a somewhat strange context than by patently artificial problems in a familiar setting.

The exposition is relatively modest in terms of mathematical development. Substantial use of the calculus is made only in Chapter 4 and parts of Chapters 5 and 6. In particular, with the exception of an occasional remark or aside, calculus appears in the inference part of the book only in the second section of Chapter 6. A background in matrix algebra is a prerequisite only for Section 5 of the chapter on multiple regression, and no such background is required for the discussion of multiple regression models and computer analysis of such data given in Section 4 of that chapter. Thus almost all the exposition should be accessible to those whose mathematical background includes one semester or two quarters of differential and integral calculus.

Although the book's mathematical level should give most science and engineering students little difficulty, working toward an understanding of the concepts and gaining an appreciation for the logical development of the methodology may sometimes require substantial effort. To help students gain such an understanding and appreciation, I have provided numerous exercises ranging in difficulty from many that involve routine application of text material to some that ask the reader to extend concepts discussed in the text to somewhat new situations. Most of the more conceptually oriented exercises in Chapters 1–9, where the basic methodology is presented, appear in supplementary exercise sets at the end of each chapter (mixed in with more straightforward questions). I have not provided supplementary exercises for Chapters 10–15, where the more advanced methods appear, so each end-of-section exercise set contains a few more challenging problems as well as many relatively straightforward ones. There are many more exercises than most instructors would want to assign during any particular course, but I recommend that students be required to work a substantial number of them; in a problem-solving discipline, active involvement of this sort is the surest way to identify and close the gaps in understanding that inevitably arise.

There is enough material in the book for a full-year (30-week) course, so in courses of shorter duration a selection of topics will be necessary. Because goals, backgrounds, and abilities of students—and instructors' tastes—vary widely, I hesitate to make specific recommendations regarding coverage of topics. Experience in teaching a two-quarter sequence at Cal Poly (three lectures per week and no quiz or discussion sections) may provide helpful guidelines.

First-quarter coverage includes Chapter 1, most of Chapters 2 and 3, the first three sections of Chapter 4 (with little time spent on continuous families of distributions other than the normal), the last two sections of Chapter 5 (joint distributions and expected values, presented in the first two sections, are deemphasized), Section 1 of Chapter 6, and the first one or two sections of Chapters 7 and 9. In the second quarter we cover the remainder of Chapters 7 and 9, Chapter 8, selected material from Chapters 10 and 11 (typically Section 1 and parts of Sections 2 and 3 of Chapter 10, and brief mention of multifactor analyses), Chapter 12, selected portions from Sections 1 through 4 of Chapter 13, and material from the first three sections of Chapter 14. I have had success in introducing hypothesis testing right after a discussion of the binomial distribution (before any normal theory), so I have included such a section in Chapter 3, but this section can easily be skipped and the subject postponed until the course reaches Chapter 7. There always seems to be too little time in lectures to discuss all the topics that we statisticians think ought to be discussed. I hope that my presentation of material is readable enough so that students can be asked to read on their own selected portions of the text which have not or will not be covered in lecture.

Acknowledgments

I gratefully acknowledge the numerous suggestions and constructive criticisms provided by the following reviewers: Lyle Broemeling, Oklahoma State University; Maurice C. Bryson, Los Alamos National Laboratory; Louis J. Cote, Purdue University, Jonathan Cryer, University of Iowa; Arthur Dayton, Kansas State University; Janice DuBien, Western Michigan University; John Groves, California State Polytechnic University; William Lesso, University of Texas, Austin; Frederick Morgan, Clemson University; Larry Ringer, Texas A&M University; Paul Shaman, University of Pennsylvania; Martyn Smith, Michigan Technical University; and Richard Van Nostrand, Eastman Kodak Company. Typing chores were admirably handled by Lynda Alamo, Carol Devore, Fran Fairbrother, and Meri Kay Gurnee. Finally, I very much appreciate the editorial and production services rendered by Craig Barth, Greg Hubit, Joan Marsh, Cece Munson, Mike Needham, and the staff at Brooks/Cole.

Jay L. Devore

Probability and Statistics for Engineering and the Sciences

Contents

CHAPTER 1

Introduction and Descriptive Statistics

1.1 An Overview of Probability and Statistics

In our own work, through conversations with others, and through contact with media of various sorts (books, television, newspapers, and the like), we are continually being confronted with collections of facts or **data.** Statistics is the branch of scientific inquiry which provides methods for organizing and summarizing data, and for using information in the data to draw various conclusions.

Frequently we wish to acquire information or draw some conclusion about an entire **population** consisting of all individuals or objects of a particular type. The population of interest might consist of all radial tires manufactured by a particular company during the previous calendar year, or it might be the collection of all individuals who had been inoculated with a particular flu vaccine, or it might consist of all U.S. colleges and universities. In this last case, we might be interested specifically in the number of students enrolled at each school. If so, rather than think of the schools as the population members, we may speak of the **numerical population** in which each population member is an enrollment figure such as 1536 or 21,311. In the inoculation example, we might wish to focus on whether or not individuals had subsequently exhibited a certain condition (a rash, dizziness, and so on). We might then visualize the population as consisting of Y's (for yes, the condition was present) and N's (no, the condition was not present). This is an example of a **dichotomous** (two-valued) population. In general, we will define the population to reflect our particular interests at the time of the investigation.

The data at our disposal frequently consists of a portion or subset of the population; any such subset is called a **sample.** If the population is all U.S. colleges and universities, one sample would be {Stanford University, University of Washington, Oberlin College, California Institute of Technology, Iowa State University}. If the population comprises all college and university enrollment figures, a sample might consist of {13,043, 35,234, 2756, 21,831}.

The objectives of organization and summarization of data have been pursued for hundreds of years. The part of statistics that deals with methods for performing these operations is called **descriptive statistics.** Descriptive methods can be used either when we have a list of all population members (a **census**), or when the data consists of a sample.

When the data is a sample and the objective is to go beyond the sample to draw conclusions about the population based on sample information, methods from **inferential statistics** are used. With a few isolated exceptions, the development of inferential statistics has occurred only since the early 1900s, making it of much more recent vintage than descriptive statistics. Yet much of the interest and activity in statistics today, particularly as it relates to scientific activity and experimentation, concerns inferences rather than just description. The psychologist who tries her behavior modification technique on a sample of obese individuals would like to infer something about what the technique would do if applied to all such individuals. The engineer who accumulates data on a sample of computer systems will ultimately wish to draw conclusions about all such systems. The medical team that develops a new vaccine for a disease threatening a particular population is interested in what would happen if the vaccine were administered to all people in the population. The marketing expert may test a product in a few "representative" areas; from the resulting information he will draw conclusions about what would happen if the product were made available to all potential purchasers.

The main focus of this book is on presenting and illustrating methods of inferential statistics which are useful in scientific work. The three important types of inferential procedures—point estimation, hypothesis testing, and estimation by confidence intervals—are introduced in Chapters 6–9, and then used in more complicated settings in Chapters 10–15 (a brief introduction to hypothesis testing appears earlier, in Section 3.5). The remainder of this chapter presents methods from descriptive statistics which are most used in the development of inference.

Chapters 2–5 present material from the discipline of probability. This material ultimately forms a bridge between the descriptive and inferential techniques and leads to a better understanding of how inferential procedures are developed and used, how statistical conclusions can be translated into everyday language and interpreted, and when and where pitfalls may occur in applying the methods. Probability and statistics both deal with questions involving populations and samples, but do so in an "inverse manner" to one another.

In a probability problem, properties of the population under study are assumed known (in a numerical population, for example, some specified distribution of the population values may be assumed), and questions regarding a sample taken from the population are posed and answered. In a statistics problem, characteristics of a sample are available to the experimenter, and this information enables the experimenter to draw conclusions about the population. The relationship between the two disciplines can be summarized by saying that probability reasons from the population to the sample (deductive reasoning), while statistics reasons from the sample to the population (inductive reasoning). This is illustrated in Figure 1.1.

Before we can understand what a particular sample can tell us about the population, we should first understand the uncertainty associated with taking a

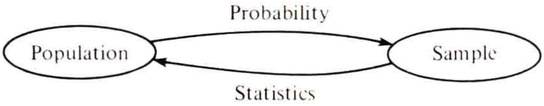


Figure 1.1 The relationship between probability and inferential statistics

sample from a given population. This is why we study probability before statistics. The following two examples pursue the distinction between the two disciplines and the types of questions asked within each.

Example 1.1 A new type of gasoline pump nozzle has been developed in order to minimize the emission of pollutants into the atmosphere during pumping. One potential design defect is that "topping off" the tanks may cause gasoline to be sucked back into the underground storage tank. (This phenomenon was described in an article which appeared in a May 1981 issue of the *Los Angeles Times.*) Suppose that 10,000 nozzles manufactured by a particular company are currently in use at Southern California gas stations. In probability, we might assume that 10% of them have the defect described above (an assumption about the population of nozzles) and ask, "How likely is it that a sample of 25 nozzles will include at least five which are defective?" or, "How many defective nozzles can we expect in a sample of 25?" On the other hand, in statistics we might find that there were five defectives in a sample of 25 nozzles (sample information) and then ask, "Does this strongly indicate that at least 10% of all nozzles currently in use are defective?" The last question concerns the population as a whole.

Example 1.2 The following measurements (reported in *Science* 167, pp. 277–279) of the ratio of the mass of the earth to that of the moon were obtained during several different spacecraft flights: 81.3001 (*Mariner 2*), 81.3015 (*Mariner 4*), 81.3006 (*Mariner 5*), 81.3011 (*Mariner 6*), 81.2997 (*Mariner 7*), 81.3005 (*Pioneer 6*), and 81.3021 (*Pioneer 7*). These numbers differ from one another (and presumably from the true ratio) because of measurement error. In probability, we might assume that the distribution of all possible measurements is bell shaped and centered at the true value of 81.3035, and then ask, "How likely is it that all seven measurements actually made result in observed ratios which are less than the true value?" In statistics, having been given the above seven measurements, we might ask, "Does this sample information conclusively demonstrate that the ratio is something other than 81.3035?" or, "How confident can we be that the true ratio is in the interval (81.2998, 81.3018)?"

In Example 1.1, the population is a well defined concrete or existing one—all nozzles presently in use. In Example 1.2, however, while the sample consists of seven observations, the population does not actually exist; instead, the population here is one consisting of all possible measurements that might be made under similar experimental conditions. Such a population is referred to as a **conceptual** or **hypothetical population.** There are a number of problems in which we fit questions into the framework of inferential statistics by conceptualizing a population. As another example, imagine that a sample of five catalytic converters with a new design is experimentally manufactured and tested. Then the inferences made will refer to the conceptual population consisting of all converters of this type which could be manufactured.

Many newcomers to the study of statistics are unaware of the broad potential application of probability and statistical methods. To remedy this, a number of statisticians contributed short, nontechnical essays to a book of readings edited by Judith Tanur, entitled *Statistics: A Guide to the Unknown* (Holden-Day, 1978). These essays are about areas of application, rather than particular methods of analysis, and serve as an excellent supplement to the problem-oriented textbook.

Exercises / Section 1.1

1. List a sample of size four from each of the following populations.

- **a.** The population of all daily newspapers published in the United States.
- **b.** The population of all U.S. corporations.
- **c.** The population of all companies which manufacture hand-held calculators.
- **d.** The population of all radio stations.

2. For each of the following hypothetical populations, list a plausible sample of size four.

- **a.** The population of all distances which might result when you throw a football.
- **b.** The population of page lengths of books which you might read during the next year.
- **c.** The population of all possible earthquake strength measurements (Richter scale) which might be recorded in California during the next year.
- **d.** The population consisting of all pH measurements made on soil samples from a particular region.

3. List three different examples of concrete populations and three different examples of hypothetical populations.

4. For one each of your concrete and your conceptual examples (exercise 3), give an example of a probability question and an example of an inferential statistics question.

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Descriptive statistics can be divided into two general subject areas. In this section we will discuss the first of these areas—representing a data set using visual techniques. In Sections 1.3 and 1.4, we will develop some numerical summary measures for data sets. Many visual techniques may already be familiar to you: frequency tables, tally sheets, histograms, pie charts, bar graphs, scatter diagrams, and the like. Here we focus on a selected few of these techniques which are most useful and relevant to probability and inferential statistics.

Notation

Some general notation will make it easier to apply our methods and formulas to a wide variety of practical problems. The number of observations in a single data set will often be denoted by n, so that $n = 4$ for the sample of universities {Stanford, Iowa State, Wyoming, Rochester} and also for the sample of pH measurements {6.3, 6.2, 5.9, 6.5}. If two data sets are simultaneously under consideration, either m and n or n_1 and n_2 may be used to denote the numbers of observations. Thus if {29.7, 31.6, 30.9} and {28.7, 29.5, 29.4, 30.3} are thermal efficiency measurements for two different types of diesel engines, then $m = 3$ and $n = 4$.

Given a data set consisting of n observations, the observations themselves will be represented by subscripting a selected letter. We will frequently represent the observations by $x_1, x_2, x_3, \ldots, x_n$ (though any other letter could be used in place of x). The subscript bears no relation to the magnitude of a particular observation, so that x_1 will not in general be the smallest observation in the set, nor will x_n typically be the largest. In many applications x_1 will be the first observation gathered by the experimenter, x_2 the second, and so on. The ith observation in the data set will be denoted by x_i.

Stem and Leaf Displays

Suppose we have a data set $x_1, x_2, \ldots, x_n$ for which each x_i consists of at least two digits. A quick way to obtain an informative visual representation of the data set is to construct a stem and leaf display. To do this, split each x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, which consists of the remaining digits. Thus if the data set consists of exam scores between 0 and 100, then we would split the score 83 into the stem 8 and the leaf 3. If the data set consists of automobile gas mileages, each recorded to a tenth of a mile per gallon, lying between 7.1 and 47.8, then a reasonable choice of stems would be 0, 1, 2, 3, and 4; 32.6 would then have stem 3 and leaf 2.6, and 7.1 would have stem 0 and leaf 7.1. In general, the stems should be chosen so that there are relatively few stems compared with the number of observations—between five and 20 stems is usually desirable. For the gas mileages, choosing stems 7, 8, . . ., 47 and each leaf as the digit to the right of the decimal point (so 32.6 has stem 32 and leaf .6) results in too many stems unless the data set is huge.

Once the set of stems has been established, the stem values are listed out along the left-hand margin of the page, and beside each stem all leaves corresponding to data values are listed out in the order in which they are encountered as we proceed through the set.

Example 1.3 The following data on motor octane ratings for various gasoline blends is taken from an article in *Technometrics* (vol. 19, p. 425), a journal devoted to applications of statistics in the physical sciences and engineering:

88.5, 87.7, 83.4, 86.7, 87.5, 91.5, 88.6, 100.3, 95.6, 93.3, 94.7, 91.1, 91.0,
94.2, 87.8, 89.9, 88.3, 87.6, 84.3, 86.7, 88.2, 90.8, 88.3, 98.8, 94.2, 92.7,
93.2, 91.0, 90.3, 93.4, 88.5, 90.1, 89.2, 88.3, 85.3, 87.9, 88.6, 90.9, 89.0,
96.1, 93.3, 91.8, 92.3, 90.4, 90.1, 93.0, 88.7, 89.9, 89.8, 89.6, 87.4, 88.4,
88.9, 91.2, 89.3, 94.4, 92.7, 91.8, 91.6, 90.4, 91.1, 92.6, 89.8, 90.6, 91.1,
90.4, 89.3, 89.7, 90.3, 91.6, 90.5, 93.7, 92.7, 92.2, 92.2, 91.2, 91.0, 92.2,
90.0, 90.7.

Since the smallest observation is 83.4 and the largest is 100.3, we choose as stem values the numbers 83, 84, . . ., 100. The resulting stem and leaf display is given