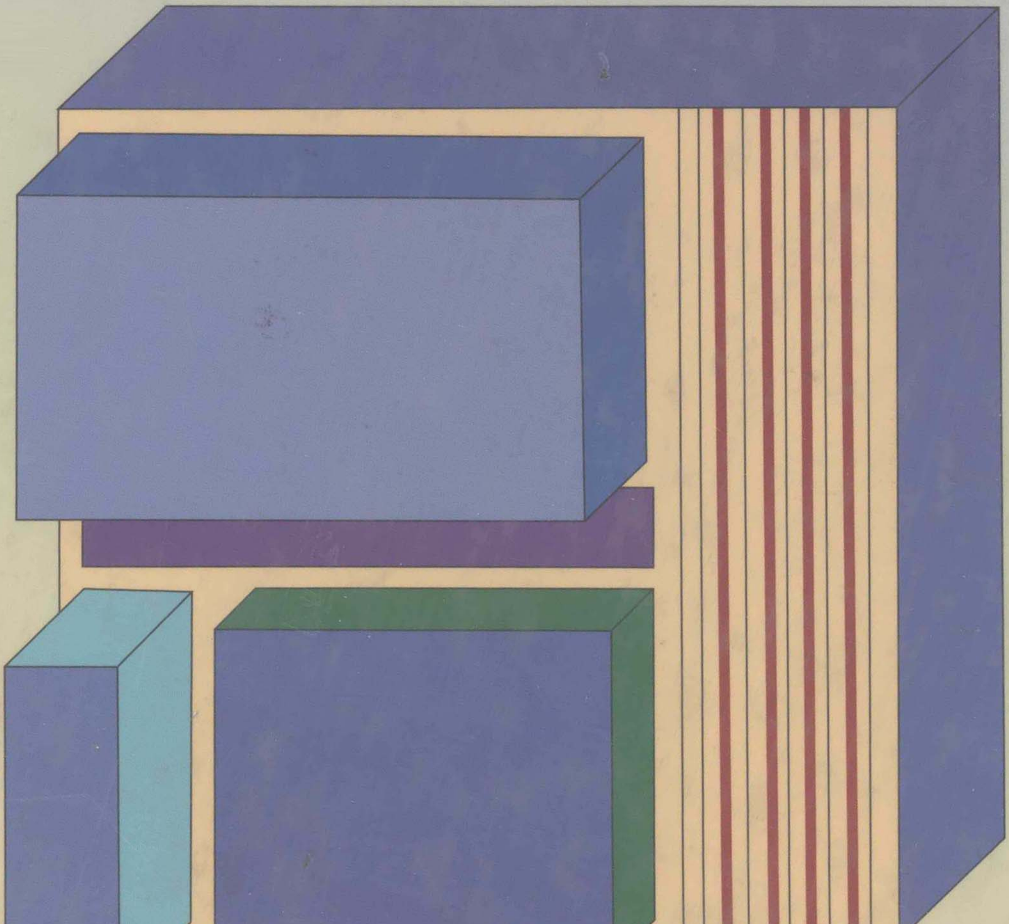
INTRODUCTORY STATISTICS



JAY DEVORE ROXY PECK

INTRODUCTORY STATISTICS

JAY DEVORE
California Polytechnic State University
San Luis Obispo

ROXY PECK
California Polytechnic State University
San Luis Obispo

WEST PUBLISHING COMPANY
St. Paul New York Los Angeles San Francisco

Composition: Syntax International
Copyediting: Susan Gerstein
Cover and interior design: Janet Bollow
Illustrations: Scientific Illustrators

50 W. Kellogg Boulevard
P.O. Box 64526
St. Paul, MN 55164-1003

Printed in the United States of America

97 96 95 94 93 92 91 90 8 7 6 5 4 3 2 1

Library of Congress Cataloging in Publication Data

Devore, Jay L.
Introductory statistics/Jay Devore, Roxy Peck.
p. cm.
ISBN 0-314-56884-0
1. Statistics. I. Peck, Roxy. II. Title.
QA276. 12.D48 1990
519.5—dc20 89-36670
CIP

To the memory of my beloved brother Paul

J. D.

To Lygia and Kyle

R. P.

PREFACE

Introductory Statistics provides traditional coverage of beginning probability and statistics with an emphasis on applications and data analysis. This book has been written with a one semester or two quarter course in mind. With judicious topic selection, it could also be used for a course as short as one quarter. Although the book does not presuppose the use of a statistical computer package, the role of the computer in data analysis is illustrated with some examples that show output from the more widely used computer packages, such as MINITAB, SPSS, BMDP, and SAS.

Throughout the text we have made an effort to avoid the use of contrived examples and exercises, and have spent a great deal of time seeking real applications taken from journals and other published sources in a wide variety of disciplines. We feel that this effort has been worthwhile in that it lends an air of credibility to the topics covered, and shows students that the techniques presented are, in fact, widely used.

There are a great many worked examples. An exercise set appears at the end of each section, and a supplementary exercise set appears at the end of each chapter. A summary of key concepts and formulas appears in each chapter just prior to the supplementary exercises. In addition, a student solutions manual containing worked solutions to the odd numbered problems, an instructor's manual (which includes worked solutions to all problems), transparency masters, and a test bank are all available from the publisher.

ACKNOWLEDGMENTS

Many people have made valuable contributions to the preparation of this book. We especially appreciated the help of our colleagues John Groves, who prepared the solutions manual and instructor's manual, and Susan Russell, who helped with typing and reviewed examples and exercise solutions for accuracy.

We would also like to thank the following reviewers, who provided helpful comments and suggestions:

Steve Bajgier
Drexel University
Warren Burch
Brevard Community College
Dan Brunk
Oregon State University
Veronica Czitrom
University of Texas—San Antonio
James Hilton
Grossmont College
K. G. Janardan
Eastern Michigan University
Don Jeffries
Orange Coast Community College
Robert Lacher
South Dakota State University
Marvin Lentner
Virginia Polytechnic Institute and State University
Maita Levine
University of Cincinnati
John Martin
Tarrant County Jr. College
Derek Mpinga
Tarrant County Jr. College
Larry Ringer
Texas A & M University
Gerald Rogers
New Mexico State University
Helen Roberts
Montclair State College
Mary Russo
University of Iowa
Rich J. VanAmerongon
Portland Community College
Timothy Wittig
Mississippi State University

The staff at West Publishing Company were a pleasure to work with, and we especially thank our editor Peter Marshall, our developmental editor Maralene Bates, and our production editor Laura Nelson. Susan Gerstein did an excellent job of copyediting for us. Finally, the support of our families and friends throughout the ordeal of completing this book was appreciated more than they will ever know!

Jay Devore
Roxy Peck

A NOTE TO THE STUDENT

In all likelihood, you have started reading this book because it is the text for an introductory statistics course required of all students in your major. You may well be thinking to yourself that if it weren't for this requirement, you wouldn't be enrolled in a statistics course and could then spend your time in more interesting and productive ways. Perhaps you are even somewhat apprehensive about your ability to do well in the course, since you've probably heard through the grapevine that mastering statistics requires some facility for mathematical reasoning and manipulation. If you are indeed ambivalent about studying statistics and a bit fearful of what lies ahead, please realize that these feelings are shared by many other students. We hope to lay these fears to rest in short order by convincing you that statistics is important for gaining a better understanding of the world around you, relevant to your particular interests and field of study, and accessible even if you have a very modest mathematical background. To this end, the book emphasizes concepts and an intuitive presentation of the core methodology used in a wide variety of applications. Statistics does rest on a mathematical foundation, but we have tried to keep the notation and mathematical development simple. We hope the result is a friendly and informal survey that will help you in various ways long after the course is finished.

The key to success in your statistics course, as in so many endeavors, is to start with a positive attitude and resolve to invest a reasonable amount of time and effort. It won't always be easy and may occasionally be frustrating. (We ourselves sometimes get quite frustrated when attempting to learn new material.) But with the right attitude and commitment of your resources, we think that understanding, enjoyment, and a sense of accomplishment will quickly follow.

CONTENTS

CHAPTER 1

INTRODUCTION

Statistical methods for summary and analysis provide investigators with powerful tools for making sense out of data. Statistical techniques are being employed with increasing frequency in business, medicine, agriculture, social sciences, natural sciences, and the applied sciences (such as engineering). The pervasiveness of statistical analyses in such diverse fields has led to increased recognition that statistical literacy—a familiarity with the goals and methods of statistics—is a basic component of a well-rounded educational program. In this chapter we begin our march toward such literacy by first introducing two basic components of most statistical problems, *population* and *sample*, and then discussing various types of data that arise in statistical analyses.

1.1 POPULATIONS, SAMPLES, AND STATISTICS

For hundreds of years, individuals have been using statistical tools to organize and summarize data. Many of these tools—bar charts, tabular displays, various plots of economic data, averages and percentages—appear regularly in newspapers, magazines, and technical journals. Methods that organize and summarize data aid in effective presentation and increased understanding; such methods constitute a branch of the discipline called **descriptive statistics**.

Often the individuals or objects studied by an investigator come from a much larger collection, and the researcher's interest goes beyond just data summarization. It is frequently the larger collection about which the investigator wishes to draw conclusions. The entire collection of individuals or objects about which information is desired is called the **population** of interest. A **sample** is a subset of the population selected in some prescribed manner for study. The second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected. This type of generalization involves some risk, since a conclusion about the population will be reached on the basis of available, but incomplete, information. It may happen that the sample is, in some sense, unrepresentative of the population from which it came. An important aspect in the development of inference techniques involves quantifying the associated risks.

Considering some examples will help you to develop a preliminary appreciation for the scope and power of statistical methodology. We describe here three problems that can be handled using techniques to be presented in this text. First, suppose that a university has just implemented a new phone registration system. Students interact with the computer by entering information from a Touch-Tone® phone to select classes for the term. In order to assess student opinion regarding the effectiveness of the system, a survey of students is to be undertaken. Each student in a sample of 400 will be asked a variety of questions (such as the number of units received, the number of attempts required to get a phone connection, etc.) The result of such a survey will be a rather large and unwieldy data set. In order to make sense out of the raw data and to describe student responses, it is desirable to summarize the data. This would also make the results more accessible to others. Descriptive techniques to be presented in Chapters 2 and 3 could be used to accomplish this task. In addition, inferential methods from Chapters 8 and 9 could be employed in order to use the sample information to draw various conclusions about the experiences of all students at the university who used the registration system.

As a second example, consider a business application. Suppose that a publisher of college textbooks has two different machines that are used to bind the printed pages. One characteristic that affects the overall quality of the finished book is the strength of the binding. The publisher would like to determine if there is a significant difference between the two machines with respect to average binding strength. Strength could be measured for one sample of books bound by the first machine and for a second sample of books bound by the second machine. Hypothesis testing techniques (to be introduced in Chapters 9 and 10) could then be used to analyze the resulting data and provide the publisher with an answer to the question posed.

. A final example comes from the discipline of forestry. When a fire occurs in a forested area, decisions must be made as to the best way to combat the fire. One possibility is to try to contain the fire by building a fire line. If building a fire line requires four hours, the decision as to where the line should be built involves making a prediction of how far the fire will spread during this period. Many factors must be taken into account, including wind speed, temperature, humidity, and time elapsed since the last rainfall. Regression techniques (Chapter 12) will enable us to develop a model for the prediction of fire spread using information available from past fires.

Some individuals regard conclusions based on statistical analyses with a great deal of suspicion. Extreme skeptics, usually speaking out of ignorance, characterize the discipline as a subcategory of lying—something used for deception rather than for positive ends. However, we believe that statistical methods, used intelligently, constitute a set of powerful tools for gaining insight into the world around us. We hope that this text will help you to understand the logic behind statistical reasoning, prepare you to apply statistical methods appropriately, and enable you to recognize when others are not doing so.

EXERCISES 1.1–1.7 SECTION 1.1

1.1 Give a brief definition of the terms *descriptive statistics* and *inferential statistics.*

1.2 Give a brief definition of the terms *population* and *sample.*

1.3 The student senate at a university with 15,000 students is interested in the proportion of students who favor a change in the grading system to allow for + and − grades (i.e., B−, B, B+, rather that just B). Two hundred students are interviewed to determine their attitude toward this proposed change. What is the population of interest? What group of students constitutes the sample in this problem?

1.4 The supervisors of a rural county are interested in the proportion of property owners who support the construction of a sewer system. Because it is too costly to contact all 7000 property owners, a survey of 500 (selected at random) is undertaken. Describe the population and sample for this problem.

1.5 Representatives of the insurance industry wished to investigate the monetary loss due to damage to single-family dwellings in Pasadena, California, resulting from an earthquake that occurred on December 3, 1988. One hundred homes were selected for inspection from the set of all single-family homes in Pasadena. Describe the population and sample for this problem.

1.6 A consumer group conducts crash tests of new model cars. To determine the severity of damage to 1989 Mazda 626s resulting from a 10-mph crash into a concrete wall, six cars of this type are tested and the amount of damage is assessed. Describe the population and sample for this problem.

1.7 A building contractor has a chance to buy an odd lot of 5000 used bricks at an auction. She is interested in determining the proportion of bricks in the lot that are cracked and therefore unusable for her current project, but she does not have enough time to inspect all 5000 bricks. Instead, she checks 100 bricks to determine whether each is cracked. Describe the population and sample for this problem.

1.2 TYPES OF DATA

The individuals or objects in any particular group typically possess many attributes that might be studied. Consider as an example a group of students currently enrolled in a statistics course. One attribute is the brand of calculator owned by each student (Sharp, Hewlett–Packard, Casio, and so on). Another attribute of potential interest is the number of courses for which each student is registered (1, 2, 3, . . .), and yet another is the distance from the university to each student's permanent residence. In this example, *calculator brand* is a categorical attribute, since each student's response to the query ''What brand of calculator do you own?'' is a category. The collection of responses from all these students forms a **categorical data set**. The other two attributes, *number of units* and *distance*, are both numerical in nature. Determining the value of such a numerical attribute (by counting or measuring) for each student results in a **numerical data set**.

EXAMPLE 1

A sample of 15 people who belong to a certain tennis club is selected. Each one is then asked which brand of racket he or she uses. The resulting set of responses is

{Head, Prince, Prince, Wilson, Yonex, Head, Yamaha, Head, Head, Prince, Yamaha, Kennex, Prince, Wilson, Yonex}.*

This a categorical data set.

EXAMPLE 2

A sample of 20 automobiles of a certain type is selected and the fuel efficiency (miles per gallon or mpg) is determined for each one. The resulting numerical data set is

{29.8, 27.6, 28.3, 28.7, 27.9, 29.9, 30.1, 28.0, 28.7, 27.9, 28.5, 29.5, 27.2, 26.9, 28.4, 27.9, 28.0, 30.0, 29.6, 29.1}.

In both of the preceding examples, the data sets consisted of observations (categorical responses or numbers) on a single attribute. Such data sets are called *univariate*.

* We will often use braces to enclose the members of a set.

DEFINITION

A data set consisting of observations on a single attribute is a **univariate data set**. A univariate data set is **categorical** (or **qualitative**) if the individual observations are categorical responses; it is **numerical** (or **quantitative**) if the observations are numbers.

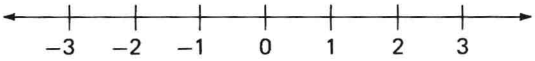
In some studies, attention focuses simultaneously on two different attributes. For example, both the height (in.) and weight (lb) might be recorded for each individual in a group. The resulting data set consists of pairs of numbers, such as (68, 146). This is called a **bivariate data set. Multivariate data** results from obtaining a category or value for each of three or more attributes: for example, height, weight, pulse rate, and systolic blood pressure for each individual. Much of this book will focus on methods for analyzing univariate data. In the last several chapters we consider briefly the analysis of some bivariate and multivariate data.

TWO TYPES OF NUMERICAL DATA

With numerical data, it is useful to make a further distinction. Visualize a number line (Figure 1) for locating values of the numerical attribute being studied. To every possible number (2, 3.125, −8.12976, etc.) there corresponds exactly one point on the number line. Now suppose that the attribute of interest is the number of cylinders of an automobile engine. The possible values of 4, 6, and 8 are identified in Figure 2(a) by the dots at the points marked 4, 6, and 8. These possible values are isolated from one another on the line; around any possible value we can place an interval that is small enough so that no other possible value is included in the interval. On the other hand, the line segment in Figure 2(b) identifies a plausible set of possible values for quarter-mile time. Here the possible values comprise an entire interval on the number line, and no possible value is isolated from the other possible values.

FIGURE 1

A number line.

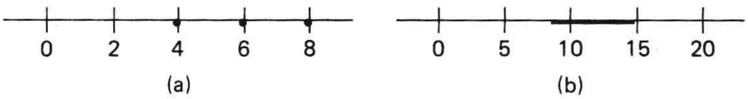


DEFINITION

Numerical data is **discrete** if the possible values are isolated points on the number line. Numerical data is **continuous** if the set of possible values forms an entire interval on the number line.

FIGURE 2

Possible values of a variable.
(a) Number of cylinders
(b) Quarter-mile time



Discrete data usually arises when each observation is determined by counting (the number of classes for which a student is registered, the number of petals on a certain type of flower, etc.).

EXAMPLE 3

The number of telephone calls to a drug hotline is recorded for each different 24-hour period. The resulting data set is

$$\{3, \quad 0, \quad 4, \quad 3, \quad 1, \quad 0, \quad 6, \quad 2, \quad 0, \quad 0, \quad 1, \quad 2\}.$$

Possible values for the *number of calls* are 0, 1, 2, 3, . . . ; these are isolated points on the number line, so we have a sample consisting of discrete numerical data.

The sample of fuel efficiencies in Example 2 is an example of continuous data. A car's fuel efficiency could be 27.0, 27.13, 27.12796, or any other value in an entire interval. Other examples of continuous data arise when task completion times are observed, body temperatures are recorded, or packages are weighed. In general, data is continuous when observations involve making measurements, as opposed to counting.

In practice, measuring instruments do not have infinite accuracy. Thus possible measured values do not form a continuum on the number line. The distinction between discrete and continuous data will nevertheless be important in our discussion of probability models.

EXERCISES 1.8–1.11 SECTION 1.2

1.8 Classify each of the following attributes as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

a. Number of students in a class of 35 who turn in a term paper before the due date
b. Sex of the next baby born at a particular hospital
c. Amount of fluid (oz) dispensed by a machine used to fill bottles with soda pop
d. Thickness of the gelatin coating of a vitamin E capsule
e. Birth classification (only child, first born, middle child, last born) of a math major

1.9 Classify each of the following attributes as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

a. Brand of personal computer purchased by a customer
b. State of birth for someone born in the United States
c. Price of a textbook
d. Concentration of a contaminant (micrograms/cm^3 or μg/cm^3) in a water sample
e. Zip code (Think carefully about this one.)
f. Actual weight of coffee in a 1-lb can

1.10 For the following numerical attributes, state whether each is discrete or continuous.

a. The number of checks received by a grocery store during a given month that bounce