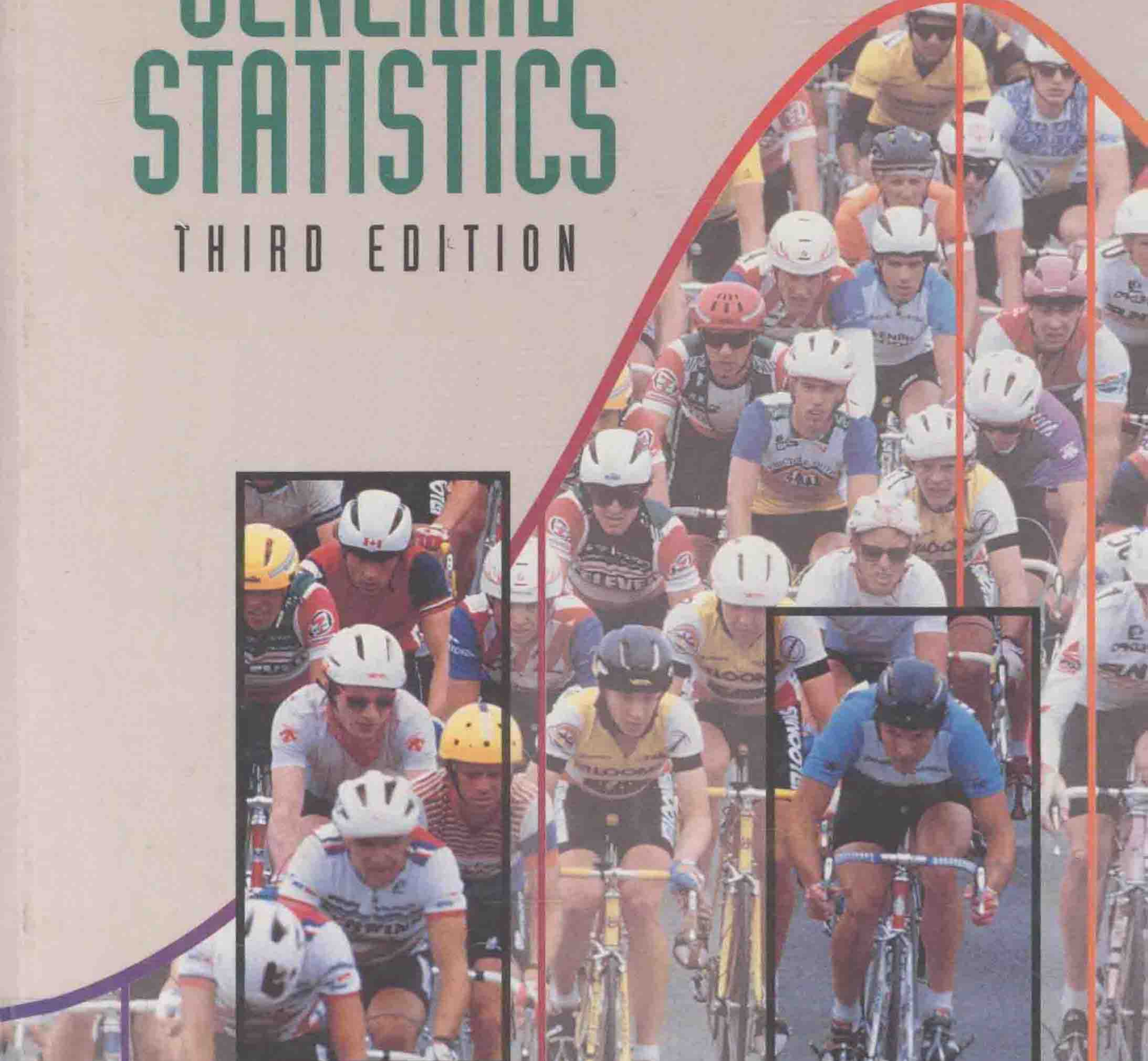
GENERAL STATISTICS

THIRD EDITION



WARREN CHASE FRED BOWN

GENERAL STATISTICS

THIRD EDITION

WARREN CHASE
FRED BOWN
Framingham State College

JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

ACQUISITIONS EDITOR Brad Wiley II
MARKETING MANAGER Jay Kirsch
PRODUCTION MANAGER Lucille Buonocore
SENIOR PRODUCTION EDITOR Tracey Kuehn
PRODUCTION COORDINATOR Susan L. Reiland
DESIGNER Ann Marie Renzi
MANUFACTURING MANAGER Dorothy Sinclair
ILLUSTRATION Jaime Perea
COVER DESIGN Kenny Beck
COVER PHOTO Charles Krebs/Tony Stone Images

This book was set in Palatino by Publication Services
and printed and bound by R. R. Donnelley/Willard. The cover was printed by Lehigh Press.

Recognizing the importance of preserving what has been written, it is a policy of John Wiley and Sons, Inc., to have books of enduring value published in the United States printed on acid-free paper, and we exert our best efforts to that end.

The paper in this book was manufactured by a mill whose forest management programs include sustained yield harvesting of its timberlands. Sustained yield harvesting principles ensure that the number of trees cut each year does not exceed the amount of new growth.

Library of Congress Cataloging-in-Publication Data:
Chase, Warren.
General statistics / Warren Chase, Fred Bown. – 3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-05584-0 (cloth : alk. paper)
1. Statistics. I. Bown, Fred. II. Title.
QA276.12.C456 1996
519.5–dc20 96-11200
CIP

ISBN 0-471-05584-0

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

The purpose of this book is to present a first course in statistics appropriate for students in a wide variety of disciplines, the only prerequisite being a knowledge of high school algebra. We feel that the objective of such a course should be to acquaint the student with the basic ideas of descriptive and inferential statistics. Great care has been taken in writing this book to make the subject matter understandable. All technical terms are defined in easy-to-grasp language; definitions, important formulas, and summaries of statistical tests are set off in boxes for quick reference. Concepts are introduced and reinforced with examples and exercises from a wide range of fields, from sports to medicine. All chapters begin with an introduction and end with a summary of the important ideas of the chapter. As in the second edition of the book, most exercises on hypothesis testing do not specify whether the traditional approach (finding the critical region), or the *P*-value approach should be used. We give answers in terms of both approaches. The instructor can then decide which to use.

New Features of the Third Edition

The third edition involves more use of **exploratory data analysis** (EDA). There is a growing consensus that all too often statistical formulas and techniques are blindly applied without looking carefully at the data to see whether these methods are appropriate. Techniques of exploratory data analysis, such as **stem-and-leaf plots** and **boxplots,** are used throughout the book to investigate these issues. For example, when investigating a population mean using a small sample, a t test would be inadvisable if the sample is strongly skewed or contains extreme outliers. A stem-and-leaf plot will detect these conditions.

Structural changes:

- Chapter 1 includes expanded coverage of sampling and design of experiments.
- Descriptive aspects of regression and correlation are treated earlier (Chapter 4) than in the second edition. Inference concerning regression and correlation is dealt with in Chapter 10, and has been expanded to include inference concerning slope of the regression line.
- The chapter on probability (Chapter 5) has been extensively reworked. All but the first four sections are optional. Traditionally, probability has been the topic that students find most difficult in a first course in statistics. We have tried to segregate the stumbling points into optional sections. For example, students do not seem to have trouble with the Addition Rule when the events are mutually exclusive, or the Multiplication Rule when the events are independent. But when the events are not mutually exclusive, or are dependent and conditional probability is involved, students encounter difficulty. We have placed these cases in Section 5, which is optional.

- The chi-square test for variance and the F test for comparing two variances have been placed in the Appendix. In practice, statisticians rarely use these tests due to their extreme sensitivity to departures from normality of the underlying populations.
- We have attempted to streamline various parts of the book. For example, all the introductory material on inference (both large- and small-sample inference for a mean μ and inference for a proportion p) are now in one chapter (Chapter 8).
- The exercise sets have been expanded and include more data sets.

Pedagogical changes:

- We have included an appendix containing data on 1000 randomly selected subjects from the Framingham Heart Study. These data are also available on a floppy disk. At the end of most chapters we have included optional sections entitled "Working with Data." These sections enable the student to apply the procedures developed in the chapter on data from the Framingham Heart Study.
- As in the second edition, most chapters have optional self-contained Minitab sections. In addition, we have included examples and exercises throughout the text that illustrate the interpretation of computer output from a variety of statistical packages. Also a number of exercises involve the use of output from Resampling Stats.
- A number of case studies have been added. We have chosen to integrate these within the exposition and exercises, rather than putting them in stand-alone boxes that can be ignored. Thus the Hazelwood race discrimination case is the principal vehicle for explaining hypothesis testing concerning a population proportion, rather than merely an appendage to the textual material. The Game Show Paradox of *Parade Magazine*'s Marilyn Vos Savant is used to illustrate most of the probability techniques involving compound events.

Chapters 1–8 would be suitable for the core of a three-semester-hour introductory course, and Chapters 1–10 would form the core of a 4-semester-hour course. In each case, additional material could be covered. After Chapter 8, the remaining chapters can be covered in any order. (See prerequisite chart in Figure 1.)

Supplements

- An *Instructor's Manual*, which includes complete solutions to all exercises, chapters on the Poisson distribution, the hypergeometric distribution, and multiple regression, and a list of answers to the even exercises. Any portion may be reproduced for class use.
- A *Student's Solutions Manual* with complete solutions to all odd-numbered exercises.
- A *Study Guide*, which includes a review of algebra and notation.
- A *Test Bank* with true–false, fill-in-the-blank, multiple choice, and computational exercises. This is available in hard copy and on a floppy disk that may be edited to include the instructor's test items.
- A Data Disk in ASCII file format is available from the publisher for each adopter of this text. The floppy disk contains data on 1000 randomly selected subjects from the Framingham Heart Study (which also appears in the Appendix).

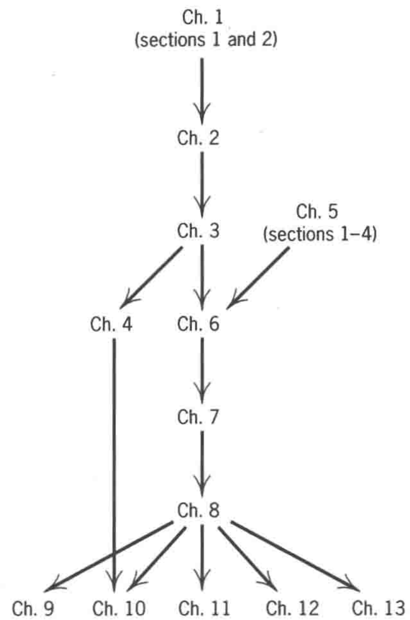


Figure 1
Prerequisite Chart. Chapter at Tail of Arrow Is a Prerequisite for Chapter at Point of Arrow.

Acknowledgments

We wish to thank Dr. William Castelli of the Framingham Heart Study for providing us with data and reprints of papers from the heart study. We also wish to thank Robert Garrison, Michael Hartman, and Paul Sorlie of the National Institutes of Health for their role in compiling the heart study data and for providing us with computer tapes containing the data.

We want to thank the following reviewers of the third edition for their many valuable suggestions:

Susan Barton
Brigham Young University-Hawaii

Beverly R. Broomell
Suffolk Community College

James Curl
Modesto Junior College

Stanley Fraser
State University of New York at Canton

R. K. Goodrich
University of Colorado-Boulder

Lawrence J. Kratz
Idaho State University

Susan Mills
Riverside Community College

Mike Scroggins
Lewis and Clark Community College

Thomas Short
Villanova University

Special thanks are in order for Mack Hill of Worcester State College, who worked through all the examples and exercises and checked the manuscript of the third edition for accuracy. We also want to express our gratitude to our editor, Brad Wiley; his hard work, enthusiasm, and encouragement contributed in large part to making the third edition a reality.

W.C.
F.B.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 THE NATURE OF STATISTICS

To get an idea of what statistics is all about, we consider a few concrete illustrations of the kinds of problems that involve statistics.

- A pollster interested in the outcome of an upcoming election interviews a certain number of voters and, based on the results obtained, makes a prediction as to who will probably win the election.
- The Environmental Protection Agency conducts tests on a certain number of cars of the same make and model to estimate the average gas mileage for all cars of that make and model.
- Each year the Federal Bureau of Investigation publishes the *Uniform Crime Reports.* Among other things, this document reports the violent crime rate (number per 100,000 people) for each of the Metropolitan Statistical Areas in the United States.
- To get an idea of the economic status of a town, an economist obtains the salaries of all wage earners in the town and then computes the average.

- To estimate the average age of all adult Americans, a sociologist selects 1000 adults, computes their average age, and then uses this figure as an estimate.

All of these illustrations make use of the raw material of statistics, namely, **data** (or **data values**), also called **observations.** We use these terms in a very broad sense. That is, a data value is simply a piece of information that might be numerical, such as the annual snowfall in Boston, or a person's weight or age. Or the information might be nonnumerical, such as the color of a car, a person's ethnic status, or whom you favor in the next presidential election. We give the following definition of **statistics:**

> **Definition** *Statistics* is the science of collecting, simplifying, and describing data, as well as making inferences (drawing conclusions) based on the analysis of data.

As the definition suggests, there are two branches of statistics. The branch that deals with collecting, simplifying, and giving properties of data is called **descriptive statistics.** An important objective of descriptive statistics is to organize, summarize, or describe the data to make it more comprehensible. For example, suppose that we obtained a list of the salaries of all the wage earners in Boston. This list would be so long that it would be incomprehensible. But if we were to find the average of all these salaries, then we would understand something about the economic status of the residents of Boston.

The other branch of statistics, which involves drawing conclusions based on the analysis of data, is called **inferential statistics.** The pollster who predicts the outcome of an election based on a knowledge of only *some* of the votes, and the sociologist who estimates the average age of *all* adult Americans based on a knowledge of the average of the ages of *some* of these adults, are both using inferential statistics. Apparently, it was impractical for these researchers to obtain all the data they were interested in (i.e., all the votes or all the ages); therefore, in both cases a judgment was made about the larger body of data that was being studied by means of information obtained from only some of these data values. This leads to the following definition:

> **Definition** The entire collection of all the elements we are interested in is called a *population.* (These elements might be people, automobiles, data values, etc.) A collection of some of the elements obtained from the population is called a *sample* from the population.

In an investigation such as a voter preference study, we may think of the population as consisting of all the voters or all the votes they will cast. The votes are the **data values** of interest, and it is these values that we are really investigating. In general, we are ultimately interested in data values in statistics. For this reason, in this book, we will often think of populations as consisting of data values.

Now that we have introduced the terms *population* and *sample,* we can give a more precise definition of inferential statistics.

> **Definition** *Inferential statistics* is concerned with making judgments (or inferences) about a population based on the properties of some sample obtained from the population.

We said that trying to estimate the average age of all adult Americans by using the average of the ages of some of these adults is a typical problem in inferential statistics. The average age for the population of all adult Americans is an example of what is called a **parameter.** The average of the ages for a sample of adult Americans is an example of a **statistic.**

> **Definition** A numerical property of a population is called a *parameter.* A numerical property of a sample is called a *statistic.* (By numerical property we mean a property that can be expressed as a number.)

Consider the population of all voters in the 1936 presidential election. The percentage of the voters that were for Roosevelt is a parameter. Viewing the voters in Peoria, Illinois, as a sample from this population, we see that the percentage in this sample that were for Roosevelt is a statistic.

Many problems in inferential statistics involve estimating the value of, or making some decision concerning, a parameter based on the value of a statistic.

The methods used in making inferences in statistics are probabilistic. For example, suppose that a pollster interviews 100 voters selected by chance and finds that 96 of them favor candidate A in an upcoming election. Then the pollster would say that the evidence points to candidate A winning the election, because it would be highly unlikely or improbable that so many voters (in the sample of 100 voters) would be in favor of candidate A if candidate A were not going to win the election. **Probability, which deals with the laws of chance, plays an important role in statistics.** Therefore, we will study this topic in some detail in this book.

EXERCISES

In Exercises 1.1 and 1.2, discuss similarities and differences between the following terms.

1.1 Population and sample

1.2 Parameter and statistic

In Exercises 1.3–1.5, which are true?

1.3 Often, the value of a parameter is unknown.

1.4 The value of a parameter always remains unchanged from sample to sample.

1.5 The value of a statistic always remains unchanged from sample to sample.

In Exercises 1.6–1.10, determine whether the results given are examples of descriptive or inferential statistics.

1.6 In the 1992 presidential election, voters in Massachusetts cast 1,315,016 votes for Bill Clinton, 803,974 for George Bush, and 630,440 for H. Ross Perot (*Source:* *World Almanac,* 1993, p. 73).

1.7 As of January 1, 1992, the Nielsen Company estimated various percentages for United States homes (*Source:* *World Almanac,* 1993, p. 305). Percentage with

Color TV sets	98%	Black and white only	2%
Two or more sets	65%	One set	35%

1.8 The mid-1993 population of the United States was estimated to be 258.3 million, and the 1990 median age was estimated to be 32.9 years (*Source:* *The 1994 Information Please Almanac,* 1994, p. 3). *Note:* Estimating a median age of 32.9 years is saying that 50% of the population is estimated to be younger than 32.9 years of age.

1.9 In the 1992 presidential election, Bill Clinton, George Bush, and H. Ross Perot received 370, 168, and 0 electoral votes, respectively (*Source:* *World Almanac,* 1993, p. 73).

1.10 The population of Worcester, Massachusetts, in 1990 was 169,759, an increase of 4.92% from the 1980 population of 161,799 (*Source:* *World Almanac,* 1993, p. 412).

1.11 The Bureau of the Census estimates that 5% of the black population was missed in the 1980 population census (*Source:* Hansen, M. H., and B. A. Bailar, "How to Count Better," in *Statistics: A Guide to the Unknown,* 3rd ed., p. 209). Is 5% the value of a parameter or a statistic?

1.12 A researcher is looking into aspects of life at a large eastern university. The researcher wishes to estimate the proportion of students at the university who are commuters. The researcher does not know that 4/10 of the students are commuters. Is 4/10 the value of a parameter or a statistic?

1.13 A linguist was interested in the population of words in James Joyce's *Ulysses.* The word *the* occurs 14,887 times (*Source:* Simon, H. A., "The Sizes of Things," in *Statistics: A Guide to the Unknown,* 3rd ed., p. 143). Is 14,887 the value of a parameter or a statistic?

1.14 A sample of 50 federal employees in a large midwestern city showed an average age of 38.2 years. Is 38.2 the value of a parameter or a statistic?

1.15 Consider the problem of estimating the average grade point average (GPA) of the 750 seniors at a college. (The average, which is unknown, is the sum of 750 GPAs divided by 750.)

(a) What is the population? How many data values are in the population?

(b) What is the parameter of interest?

(c) Suppose that a sample of 10 seniors is selected, and their GPAs are 2.72, 2.81, 2.65, 2.69, 3.17, 2.74, 2.57, 2.17, 3.48, 3.10. Calculate a statistic that you would use to estimate the parameter.

(d) Suppose that another sample of 10 seniors was selected. Would it be likely that the value of the statistic is the same as in part (c)? Why or why not? Would the value of the parameter remain the same?

1.16 Unaware that 35% of the 10,000 voters in his district still support him, a politician decides to estimate his political strength. A sample of 200 voters shows that 40% support him.

(a) What is the population?

(b) What is the value of the parameter of interest?

(c) What is the value of the statistic of interest?

(d) Compare your answers in (b) and (c). Is it surprising they are different? If the politician were to sample another 200 voters, which of the two numbers would most likely change? Explain.

1.17 A sociologist was interested in estimating some aspects of family life in a town. Information about the entire town (unknown to the sociologist) and the results of a sample obtained by the sociologist follow.

Number of Children per Family	Number of Families in Town	Number of Families in Sample
0	120	6
1	180	10
2	270	12
3	300	8
4	80	6
5	50	8

(a) Identify the elements of the population and give the population size.

(b) Identify the elements of the sample and give the sample size.

(c) Suppose that the sociologist were interested in the *number* of families with more than two children.
  (i) Calculate a statistic to estimate the parameter of interest. (*Hint:* Assume that the sociologist knows that the population is 20 times the size of the sample.)
  (ii) What is the value of the parameter?

(d) Suppose that the sociologist were interested in the *proportion* of families with less than four children.
  (i) What is the value of the statistic?
  (ii) What is the value of the parameter?

1.18 In the Massachusetts State Lottery Megabucks game, six numbers are selected from the set of 42 numbers $\{1, 2, 3, 4, \ldots, 39, 40, 41, 42\}$. A player thought that too high a percentage of single-digit numbers was being selected. He obtained a partial list of past drawings and observed that 26.4% of the numbers were single digit.

(a) If the numbers were generated randomly, what would the value of the parameter of interest be?

(b) What is the value of the statistic?

1.2 SAMPLING

We said that in inferential statistics we use samples to make judgments about populations. We want the samples we obtain to be **representative** of the population, that is, to resemble the population. There are many ways of obtaining samples; we mention only a few here.

Random Samples

Random sampling is one of the most important types of sampling in statistics.