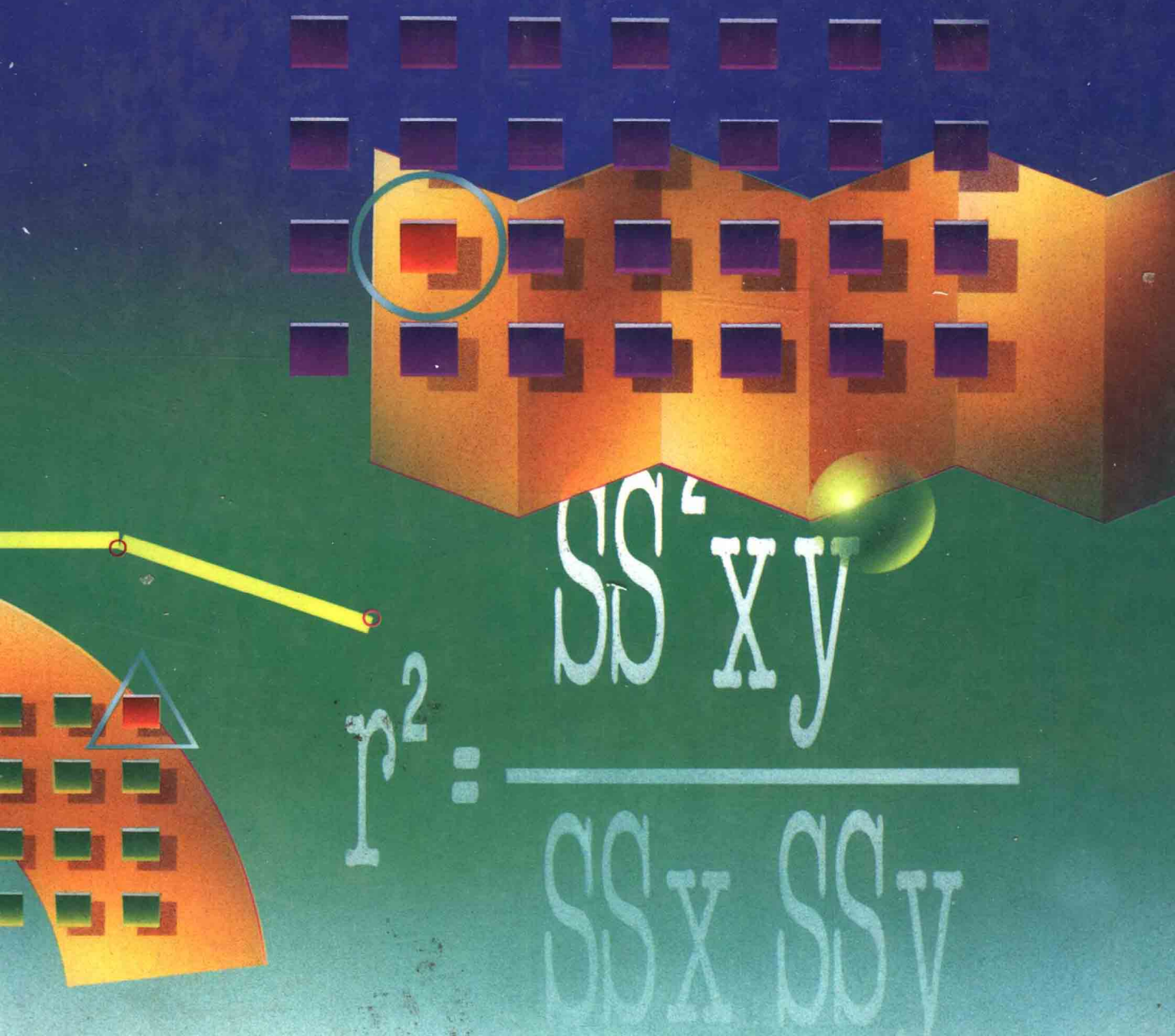
Warren Chase Fred Bown

# GENERAL STATISTICS

## Second Edition



# GENERAL STATISTICS

**Second Edition**

**Warren Chase**

**Fred Bown**

*Framingham State College*

**John Wiley & Sons, Inc.**

*New York* *Chichester* *Brisbane* *Toronto* *Singapore*

ACQUISITIONS EDITOR / Brad Wiley II
PRODUCTION MANAGER / Linda Muriello
DESIGNER / Kevin Murphy
PRODUCTION SUPERVISOR / Micheline Frederick
MANUFACTURING MANAGER / Lorraine Fumoso
COPY EDITOR / Elizabeth Swain
COVER AND INTERIOR OPENING ART / Roy Wiemann



*Library of Congress Cataloging in Publication Data:*
ISBN 0-471-61901-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2

**Printed and bound by R.R. Donnelley & Sons, Inc.**

# A Note from the Publisher

John Wiley & Sons, Inc., is committed to publishing textbooks of pedagogical integrity and technical accuracy. To these ends, the following measures were built into the revision program for *General Statistics*, 2e.

After both a survey of adopters and a review program, the authors rewrote many parts of the original text. After further review, the authors, publisher's copyeditors, and a professional statistics teacher each proofread the textual material, examples and solutions, student exercises, students' answers, and teacher's solutions manual for presentation consistency and mathematical and notational accuracy.

# Preface

The purpose of this book is to present a first course in statistics with an emphasis on statistical inference. It is appropriate for students in a wide variety of disciplines, the only prerequisite being a knowledge of intermediate high school algebra. Great care has been taken in writing the 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.

## New Features of the Second Edition

- The chapters on measures of central tendency and measures of dispersion have been combined and the discussion has been streamlined.
- The treatment of Minitab has been updated and expanded. With the exception of the introduction and the chapter on probability, all chapters contain optional sections at the end showing how to apply Minitab to the concepts developed in the chapter. These sections constitute a self-contained introduction to Minitab.
- The treatment of $P$ values has been expanded. A survey of our users of the first edition showed varying opinions on how much emphasis should be placed on $P$ values versus the traditional approach to hypothesis testing. Although some of our exercises on hypothesis testing ask for solutions using both the traditional approach and the $P$ value approach, most do not specify which approach to use. We give answers in terms of both approaches. The instructor can then decide which to use.
- Optional sections on sampling, box-and-whisker plots, counting techniques (combinatorics), and the design of experiments have been added.
- There has been a substantial increase in the number of exercises, and there is now a *Student Solutions Manual* containing the complete solutions to selected exercises.
- An *Instructor's Manual* contains the complete solutions to all exercises, a chapter on multiple regression that may be reproduced for class use, and a test bank. The test bank contains true–false, fill-in-the-blank, multiple-choice, and computational exercises. The test bank is also available on a floppy disk for either an IBM PC or a Macintosh hardware environment. The disk may be edited by the instructor.
- The *Minitab Supplement to Accompany General Statistics*, 2nd ed., by Anne Sevin, has been updated to the current release of Minitab. This supplement provides material for a lab course or homework assignments, giving further instruction in computation, analysis, and decision making with the com-

puter. A data disk that includes data on 1000 randomly selected subjects from the Framingham Heart Study is available for replication to any adopter of the supplement or the main text.

- A *Study Guide* to review the course has been prepared by Professor James Curl of Modesto Junior College. The study guide also includes a review of algebra and notation.

The book is organized to give the instructor maximum flexibility in choosing the order of topics in the course. For instance, some instructors like to introduce the students to regression and correlation early in the course. Although we cover these topics in Chapter 10, the first three sections of Chapter 10 can be used to provide an early introduction to these topics right after completing Chapter 3. In Chapter 7, where estimation and decision making (hypothesis testing) are introduced, the sections are written so that these topics may be treated in any order.

Another example of flexibility is our treatment of probability. We believe that students taking a first course in statistics need to develop both an understanding of the notions of event and the probability of an event, and the ability to calculate some simple probabilities. Instructors who wish to move quickly through probability may cover these topics in Chapter 4, sections 4.1–4.3. Compound events (including the Addition Rule and Multiplication Rule) are covered in sections 4.4–4.7, but may be omitted if desired. The text is written in such a way that only sections 4.1–4.3 are required. Even the treatment of contingency tables in Chapter 12, which in many books requires a prior understanding of independent events and the Multiplication Rule, is self-contained and requires no prior exposure to these topics.

It seems to us that one of the problems encountered in teaching (and studying) statistics is that it's easy to get bogged down in details of the early topics in the course, such as descriptive statistics and probability. We favor moving rapidly through these topics so that more time can be spent on inferential statistics (studying populations using samples). At Framingham State College, topics from the first 10 chapters constitute an introductory statistics course for liberal arts majors. We move very rapidly through the first 4 chapters. About all the student needs to know from these chapters is measures of central tendency (mean, median, mode), variance, standard deviation, percentiles, some graphical techniques, and what is meant by probability of an event (the first 3 sections of Chapter 4). We spend a fair amount of time on Chapter 6 (Normal Distributions) and then cover selected topics from Chapters 7 through 10. Depending on how rapidly one moves through Chapters 1 through 4, additional material might be covered such as Analysis of Variance (Chapter 11) or Analysis of Categorical Data (Chapter 12). After Chapter 9, the remaining chapters are independent—none is a prerequisite of any of the others. See Figure 1 for a chart indicating prerequisites.

In selecting topics for a course, one thing to keep in mind is that some sections are labeled as "optional" because most instructors would agree that they indeed are optional. That doesn't mean that other sections may not be

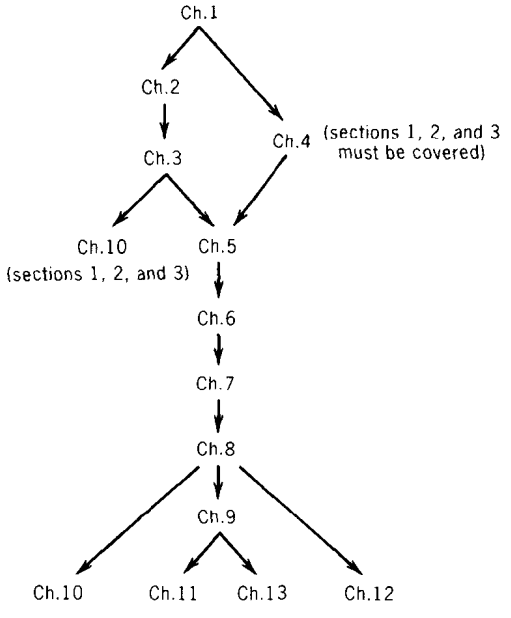



**FIGURE 1**
Prerequisite Chart. Chapter at Tail of Arrow Is a Prerequisite for Chapter at Point of Arrow.

legitimately viewed as optional by some instructors. Chapter 4 is a case in point. We view the last four sections as optional, although only one is labeled as such.

The theme underlying this book is that statistics is a very general and powerful discipline, and its methods are applicable in a variety of areas. We feel that a first course in statistics should stress this idea. Students who will make substantial use of statistics in their field may wish to follow up a general statistics course with a statistics course that is oriented toward their own discipline.

## 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 second edition for their many valuable suggestions:

Geraldine C. Banda
Glassboro State College

Carole Bernett
Harper College

J. Stanley Laughlin
Idaho State University

Kenneth Schoen
Worcester State College

Louis F. Bush
San Diego City College

James M. Edmondson
Santa Barbara City College

Janet Evert
Erie Community College

Shu-Ping Hodgson
Central Michigan University

Nancy Schoeps
University of North Carolina

Robert A. Schley
Mt. San Jacinto Community College

Lynn Smith
Gloucester County College

John D. Spurrier
University of South Carolina

Special thanks are in order for James Curl of Modesto Community College whose exhaustive comparative review of our first edition and a number of other books, provided a wealth of information and ideas; and also for Mack Hill of Worcester State College who worked through all the examples and exercises and checked the manuscript of the second edition for accuracy. We also want to express our gratitude to our editor, Brad Wiley; his hard work, ethusiasm, and encouragement contributed in large part to making the second edition a reality.

W.C.
F.B.

# Contents

## 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 in order 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 (the number per 100,000 people) for each of the Metropolitan Statistical Areas in the United States.
- In order 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 the above illustrations make use of the raw material of statistics, namely, **data** (or **data values**). Here 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 it might be nonnumerical information, 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 so as 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.** Since the major emphasis in this book will be on inferential statistics, let us examine this concept carefully. 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 motivates 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 is expressible as a number.)

For another example of these terms consider the population of all voters (or votes) 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.