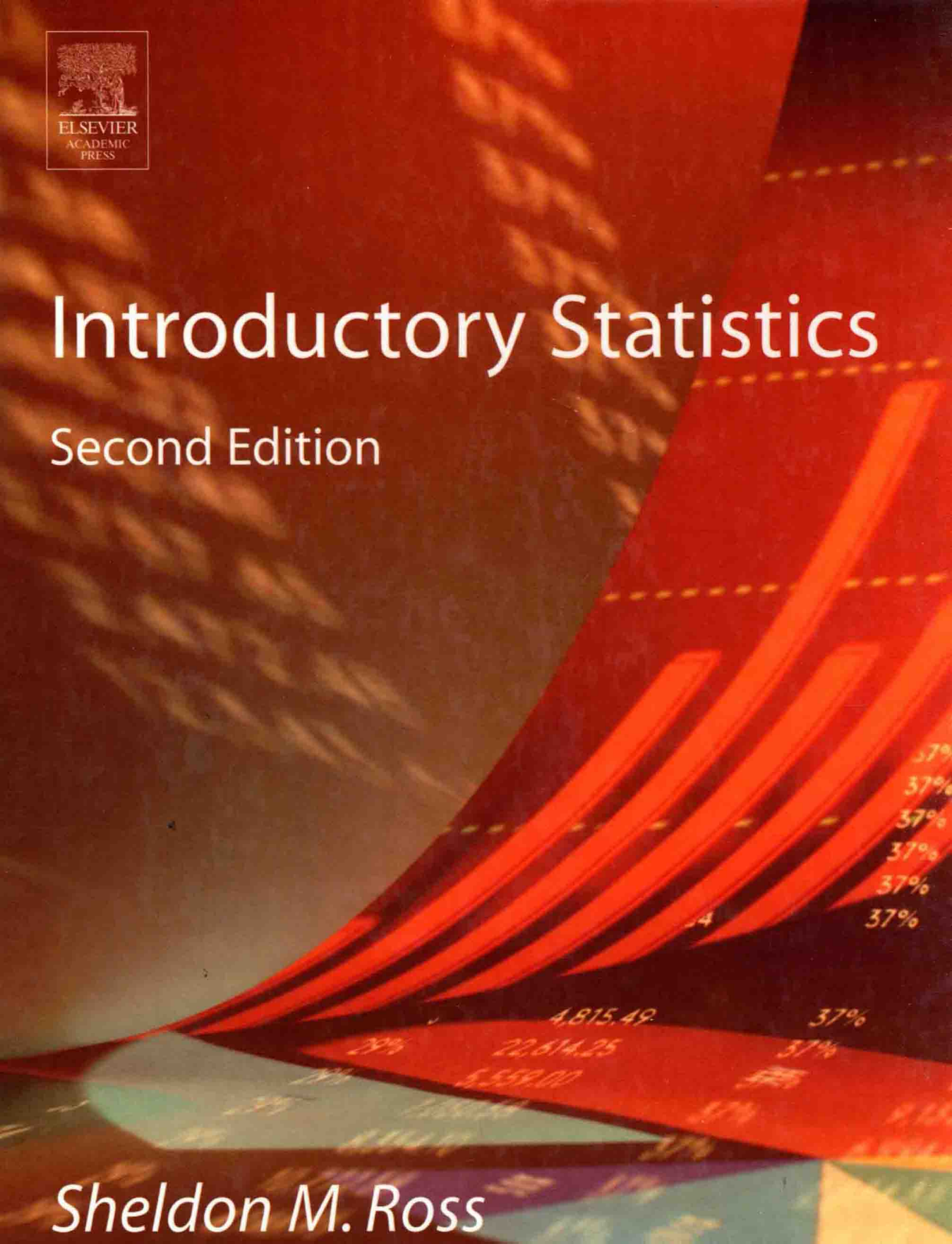
ELSEVIER
ACADEMIC
PRESS
Introductory Statistics
Second Edition
Sheldon M. Ross

Introductory Statistics

Second Edition

Sheldon M. Ross
University of Southern California

ELSEVIER
ACADEMIC
PRESS

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Acquisition Editor: Barbara Holland
Project Manager: Sarah Hajduk
Associate Editor: Tom Singer
Marketing Manager: Linda Beattie
Cover Design: Paul Hodgson
Cover Image: Courtesy of Getty Images; Artist, Jean Louis Batt
Composition: Newgen Imaging Systems (P) Ltd.
Cover Printer: Phoenix Color
Interior Printer: RR Donnelley

Elsevier Academic Press
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data
APPLICATION SUBMITTED

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 0-12-597132-X

Printed in the United States of America
05 06 07 08 09 9 8 7 6 5 4 3 2 1

INTRODUCTORY STATISTICS

Sheldon M. Ross

1 Introduction to Statistics

Statistics: the art of learning from data
Descriptive statistics: describes and summarizes data
Inferential statistics: draws conclusions from data
Population: collection of elements of interest
Sample: the part of the population from which data is obtained

2 Describing Data Sets

Frequency and relative frequency tables and graphs
Histograms
Stem-and-leaf plots
Scatter plots for paired data

3 Using Statistics to Summarize Data Sets

Sample mean: $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$

Sample median: the middle value

Sample variance: $s^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2/\left(n - 1\right)$

Sample standard deviation: $s = \sqrt{s^2}$

Algebraic identity: $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Empirical rule for normal data sets:
- approximately 68% of the data lies within $\bar{x} \pm s$
- approximately 95% of the data lies within $\bar{x} \pm 2s$
- approximately 99.7% of the data lies within $\bar{x} \pm 3s$

Sample correlation coefficient:
$r = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/[(n-1)s_x s_y]$

4 Probability

$0 \le P(A) \le 1$
$P(S) = 1$, where S is the set of all possible values
$P(A \cup B) = P(A) + P(B)$, when A and B are disjoint
Probability of the complement: $P(A^c) = 1 - P(A)$
Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Conditional probability: $P(B|A) = P(A \cap B)/P(A)$
Multiplication rule: $P(A \cap B) = P(A)P(B|A)$
Independent events: $P(A \cap B) = P(A)P(B)$

5 Discrete Random Variables

Expected value (or mean): $E[X] = \sum_{i=1}^{n} x_i P\{X = x_i\}$

$E[X + Y] = E[X] + E[Y]$
Variance: $\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$
Standard deviation: $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$
$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent

Binomial random variable:

$$P\{X = i\} = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}, \; i = 0, \ldots, n$$

$E[X] = np \quad \mathrm{Var}(X) = np(1-p)$

6 Normal Random Variables

Normal random variable X: characterized by $\mu = E[X]$, $\sigma = \mathrm{SD}(X)$

Standard normal random variable Z: normal with $\mu = 0, \sigma = 1$

$P\{|Z| > x\} = 2P\{Z > x\}, x > 0$

$P\{Z < -x\} = P\{Z > x\}$

z_α is such that $P\{Z > z_\alpha\} = \alpha$

If X is normal then $Z = (X - \mu)/\sigma$ is standard normal.
Additive property: If X and Y are independent normals then $X + Y$ is normal with mean $\mu_x + \mu_y$, and variance $\sigma_x^2 + \sigma_y^2$

7 Distributions of Sampling Statistics

$X_1, \ldots, X_n$ is sample from population: $E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$
$E[\bar{X}] = \mu$
$\mathrm{Var}(\bar{X}) = \sigma^2/n$
Central limit theorem: $\sum_{i=1}^{n} X_i$ is, for large n, approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$; equivalently $\sqrt{n}(\bar{X} - \mu)/\sigma$ is approximately standard normal.

Normal approximation to binomial: If $np \ge 5, n(1-p) \ge 5$ then $[\mathrm{Bin}(n,p) - np]/\sqrt{np(1-p)}$ is approximately standard normal.

8 Estimation

$\bar{X}$ is the estimator of the population mean μ.
$\hat{p}$, the proportion of the sample that has a certain property, estimates p, the population proportion having this property.
S^2 estimates σ^2, and S estimates σ.
$100(1-\alpha)$ confidence interval estimator for μ:
- data normal or n large, σ known: $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$
- data normal, σ unknown: $\bar{X} \pm t_{n-1,\alpha/2} S/\sqrt{n}$

$100(1-\alpha)$ confidence interval for p: $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

9 Testing Statistical Hypotheses

H_0 = null hypothesis: hypothesis that is to be tested

Significance level α: the (largest possible) probability of rejecting H_0 when it is true

p value: the smallest significance level at which H_0 would be rejected

Hypothesis Tests Concerning the Mean μ of a Population

Assumption: *Either the distribution is normal or sample size n is large.*

H_0	H_1	Test statistic TS	Significance level α test	p value if TS = v
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$†	Reject H_0 if $\lvert TS\rvert \geq z_{\alpha/2}$	$2P\{z \geq \lvert v\rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$†	Reject H_0 if $TS \geq z_\alpha$	$P\{Z \geq v\}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$	Reject H_0 if $\lvert TS\rvert \geq t_{n-1,\alpha/2}$	$2P\{T_{n-1} \geq \lvert v\rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$	Reject H_0 if $TS \geq t_{n-1,\alpha}$	$P\{T_{n-1} \geq v\}$

† *Assumption*: σ known.

Note: To test H_0: $\mu \geq \mu_0$, multiply data by -1 and use the above.

Hypothesis Tests Concerning p

(the proportion of a large population that has a certain characteristic)

X is the number of population members in a sample of size n that have the characteristic. B is a binomial random variable with parameters n and p_0.

H_0	H_1	Test statistic TS	p value if TS = x
$p \leq p_0$	$p > p_0$	X	$P\{B \geq x\}$
$p = p_0$	$p \neq p_0$	X	2 Min $\{P\{B \leq x\}, P\{B \geq x\}\}$

10 Hypotheses Tests Concerning Two Populations

Tests Concerning the Means of Two Populations When Samples Are Independent

The X sample of size n and the Y sample of size m are independent.

H_0	H_1	Test statistic TS	Assumptions	Significance level α test	p value if TS = v
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m large	Reject if $\lvert TS\rvert \geq z_{\alpha/2}$	$2P\{Z \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m large	Reject if $TS \geq z_\alpha$	$P\{Z \geq v\}$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Normal populations $\sigma_x = \sigma_y$	Reject if $TS \geq t_{n+m-2,\alpha/2}$	$2P\{T_{n+m-2} \geq \lvert v\rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Normal populations $\sigma_x = \sigma_y$	Reject if $TS \geq t_{n+m-2,\alpha}$	$P\{T_{n+m-2} \geq v\}$

$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2$ = pooled estimator of $\sigma_x^2 = \sigma_y^2$

Introductory Statistics

Second Edition

About the Author

Sheldon M. Ross. received his Ph.D. in Statistics at Stanford University in 1968 and then joined the Department of Industrial Engineering and Operations Research at the University of California at Berkeley. He remained at Berkeley until Fall 2004, when he became the Daniel J. Epstein Professor of Industrial and Systems Engineering in the Daniel J. Epstein Department of Industrial and Systems Engineering at the University of Southern California. He has published many technical articles and textbooks in the areas of statistics and applied probability. Among his texts are *A First Course in Probability* (sixth edition), *Introduction to Probability Models* (eighth edition), *Simulation* (third edition), and *Introduction to Probability and Statistics for Engineers and Scientists* (third edition).

Professor Ross is the founding and continuing editor of the journal *Probability in the Engineering and Informational Sciences*. He is a fellow of the Institute of Mathematical Statistics and a recipient of the Humboldt U.S. Senior Scientist Award.

For Rebecca and Elise

Preface

> Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.
>
> H. G. Wells (1866–1946)

In today's complicated world, very few issues are clear-cut and without controversy. In order to understand and form an opinion about an issue, one must usually gather information, or data. To learn from data, one must know something about statistics, which is the art of learning from data.

This introductory statistics text is written for college-level students in any field of study. It can be used in a quarter, semester, or full-year course. Its only prerequisite is high school algebra. Our goal in writing it is to present statistical concepts and techniques in a manner that will teach students not only how and when to utilize the statistical procedures developed, but also to understand why these procedures should be used. As a result we have made a great effort to explain the ideas behind the statistical concepts and techniques presented. Concepts are motivated, illustrated, and explained in a way that attempts to increase one's intuition. It is only when a student develops a feel or intuition for statistics that she or he is really on the path toward making sense of data.

To illustrate the diverse applications of statistics and to offer students different perspectives about the use of statistics, we have provided a wide variety of text examples and problems to be worked by students. Most refer to real-world issues, such as gun control, stock price models, health issues, driving age limits, school admission ages, public policy issues, gender issues, use of helmets, sports, disputed authorship, scientific fraud, and Vitamin C, among many others. Many of them use data that not only are real but are themselves of interest. The examples have been posed in a clear and concise manner and include many thought-provoking problems that emphasize thinking and problem-solving skills. In addition, some of the problems are designed to be open ended and can be used as starting points for term projects.

Some Special Features of the Text

Introduction The first numbered section of each chapter is an introduction that poses a realistic statistical situation to help students gain perspective on what they will encounter in the chapter.

Statistics in Perspective Statistics in Perspective highlights are placed throughout the book to illustrate real-world application of statistical techniques and concepts. These perspectives are designed to help students analyze and interpret data as well as to utilize proper statistical techniques and methodology.

Real Data Throughout the text discussions, examples, perspective highlights, and problems, real data sets are used to enhance the students' understanding of the material. These data sets provide information for the study of current issues in a variety of disciplines, such as health, medicine, sports, business, and education.

Historical Perspectives These enrichment sections profile prominent statisticians and historical events, giving students an understanding of how the discipline of statistics has evolved.

Problems/Review Problems This text includes hundreds of exercises placed at the end of each section within a chapter, as well as more comprehensive review problems at the end of each chapter. Many of these problems utilize real data and are designed to assess the students' conceptual as well as computational understanding of the material. Selected problems are open-ended and offer excellent opportunity for extended discussion, group activities, or student projects.

Summary/Key Terms An end-of-chapter summary provides a detailed review of important concepts and formulas covered in the chapter. Key terms and their definitions are listed that serve as a working glossary within each chapter.

Formula Summary Important tables and formulas that students often refer to and utilize are included on the inside front and back covers of the book. These can serve as a quick reference when doing homework or studying for an exam.

Program CD-ROM A CD-ROM is provided with each volume that includes programs that can be used to solve basic statistical computation problems. Please refer to Appendix E for a listing of these programs.

The Text

In Chap. 1 we introduce the subject matter of statistics and present its two branches. The first of these, called descriptive statistics, is concerned with the collection, description, and summarization of data. The second branch, called inferential statistics, deals with the drawing of conclusions from data.

Chapters 2 and 3 are concerned with descriptive statistics. In Chap. 2 we discuss tabular and graphical methods of presenting a set of data. We see that an effective presentation of a data set can often reveal certain of its essential features. Chap. 3 shows how to summarize certain features of a data set.

In order to be able to draw conclusions from data it is necessary to have some understanding of what they represent. For instance, it is often assumed that the

data constitute a "random sample from some population." In order to understand exactly what this and similar phrases signify, it is necessary to have some understanding of probability, and that is the subject of Chap. 4. The study of probability is often a troublesome issue in an introductory statistics class because many students find it a difficult subject. As a result, certain textbooks have chosen to downplay its importance and present it in a rather cursory style. We have chosen a different approach and attempted to concentrate on its essential features and to present them in a clear and easily understood manner. Thus, we have briefly but carefully dealt with the concepts of the events of an experiment, the properties of the probabilities that are assigned to the events, and the idea of conditional probability and independence. Our study of probability is continued in Chap. 5, where discrete random variables are introduced, and in Chap. 6, which deals with the normal and other continuous random variables.

Chapter 7 is concerned with the probability distributions of sampling statistics. In this chapter we learn why the normal distribution is of such importance in statistics.

Chapter 8 deals with the problem of using data to estimate certain parameters of interest. For instance, we might want to estimate the proportion of people who are presently in favor of congressional term limits. Two types of estimators are studied. The first of these estimates the quantity of interest with a single number (for instance, it might estimate that 52 percent of the voting population favors term limits). The second type provides an estimator in the form of an interval of values (for instance, it might estimate that between 49 and 55 percent of the voting population favors term limits).

Chapter 9 introduces the important topic of statistical hypothesis testing, which is concerned with using data to test the plausibility of a specified hypothesis. For instance, such a test might reject the hypothesis that over 60 percent of the voting population favors term limits. The concept of p value, which measures the degree of plausibility of the hypothesis after the data have been observed, is introduced.

Whereas the tests in Chap. 9 deal with a single population, the ones in Chap. 10 relate to two separate populations. For instance, we might be interested in testing whether the proportions of men and of women that favor term limits are the same.

Probably the most widely used statistical inference technique is that of the analysis of variance; this is introduced in Chap. 11. This technique allows us to test inferences about parameters that are affected by many different factors. Both one- and two-factor analysis of variance problems are considered in this chapter.

In Chap. 12 we learn about linear regression and how it can be used to relate the value of one variable (say, the height of a man) to that of another (the height of his father). The concept of regression to the mean is discussed, and the regression fallacy is introduced and carefully explained. We also learn about the relation between regression and correlation. Also, in an optional section, we use regression to the mean along with the central limit theorem to present a simple, original argument to explain why biological data sets often appear to be normally distributed.

In Chap. 13 we present goodness-of-fit tests, which can be used to test whether a proposed model is consistent with data. This chapter also considers populations

classified according to two characteristics and shows how to test whether the characteristics of a randomly chosen member of the population are independent.

Chapter 14 deals with nonparametric hypothesis test, which are tests that can be used in situations where the ones of earlier chapters are inappropriate. Chapter 15 introduces the subject matter of quality control, a key statistical technique in manufacturing and production processes.

New to this Edition

This second edition has many new and updated examples and exercises. The following are among the new sections.

- Section 4.7, an optional section on counting principles
- Section 5.7, an optional section introducing Poisson random variables
- Section 12.10, on assessing the linear regression model by analyzing residuals
- New sections in Chapter 15 on Quality Control, introducing exponentially weighted moving-average and cumulative-sum control charts.

Acknowledgments

We would like to thank the following reviewers of the second edition:

James Wright, Bucknell University
Rodney Wong, University of California at Berkeley
William Owen, Case Western University
Jaechoul Lee, Boise State University
Steven Garren, James Madison University
Pierre A. Grillet, Tulane University
Vincent Lariccia, University of Delaware
John J. Deely, Purdue University
Cen-Tsong Lin, Central Washington Unviersity
Emily Silverman, University of Michigan

In addition we wish to thank Margaret Lin, Erol Pekoz, and the following reviewers of the first edition for their many helpful comments: William H. Beyer, University of Akron; Patricia Buchanan, Pennsylvania State University; Michael Eurgubian, Santa Rosa Junior College; Larry Griffey, Florida Community College, Jacksonville; James E. Holstein, University of Missouri; James Householder, Humboldt State University; Robert Lacher, South Dakota State University; Jacinta Mann, Seton Hill College; C. J. Park, San Diego State University; Ronald Pierce, Eastern Kentucky University; Lawrence Riddle, Agnes Scott College; Gaspard T. Rizzuto, University of Southwestern Louisiana; Jim Robison-Cox, Montana State University; Walter Rosenkrantz, University of Massachusetts, Amherst; Bruce Sisko, Belleville Area College; Glen Swindle, University of California, Santa Barbara; Paul Vetrano, Santa Rose Junior College; Joseph J. Walker, Georgia State University; Deborah White, College of the Redwoods; and Cathleen Zucco, LeMoyne College.

Sheldon M. Ross

Introductory Statistics

Contents