



SEVENTH EDITION

An Introduction to
**Statistical Methods
& Data Analysis**

R. Lyman **Ott**
Michael **Longnecker**

An Introduction to

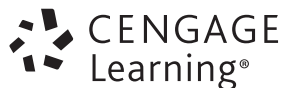
Statistical Methods & Data Analysis

An Introduction to

Statistical Methods & Data Analysis

SEVENTH EDITION

R. Lyman Ott
Michael Longnecker
Texas A&M University



This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

Important Notice: Media content referenced within the product description or the product text may not be available in the eBook version.

***An Introduction to Statistical Methods and
Data Analysis, Seventh Edition***

R. Lyman Ott, Michael Longnecker

Senior Product Team Manager:
Richard Stratton

Content Developer: Andrew Coppola

Associate Content Developer:
Spencer Arritt

Product Assistant: Kathryn Schrumpp

Marketing Manager: Julie Schuster

Content Project Manager: Cheryl Linthicum

Art Director: Vernon Boes

Manufacturing Planner: Sandee Milewski

Intellectual Property Analyst: Christina
Ciaramella

Intellectual Property Project Manager:
Farah Fard

Production Service and Compositor:
Cenveo Publishing Services

Photo and Text Researcher: Lumina
Datamatics, LTD

Copy Editor:

Illustrator: Macmillan Publishing Services/
Cenveo Publishing Services

Text and Cover Designer: C. Miller

Cover Image: polygraphus/Getty Images

© 2016, 2010 Cengage Learning

WCN: 02-200-203

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, Web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

For product information and technology assistance, contact us at
Cengage Learning Customer & Sales Support, 1-800-354-9706.

For permission to use material from this text or product,
submit all requests online at **www.cengage.com/permissions.**

Further permissions questions can be e-mailed to
permissionrequest@cengage.com

Library of Congress Control Number: 2015938496

ISBN: 978-1-305-26947-7

Cengage Learning

20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com**

Cengage Learning products are represented in Canada by
Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit
www.cengage.com

Purchase any of our products at your local college store or at our
preferred online store **www.cengagebrain.com**

Printed in the United States of America

Print Number: 01

Print Year: 2015

CONTENTS

Preface xi

PART 1 INTRODUCTION 1

CHAPTER 1

Statistics and the Scientific Method 2

- 1.1 Introduction 2
- 1.2 Why Study Statistics? 6
- 1.3 Some Current Applications of Statistics 9
- 1.4 A Note to the Student 13
- 1.5 Summary 13
- 1.6 Exercises 14

PART 2 COLLECTING DATA 17

CHAPTER 2

Using Surveys and Experimental Studies to Gather Data 18

- 2.1 Introduction and Abstract of Research Study 18
- 2.2 Observational Studies 20
- 2.3 Sampling Designs for Surveys 26
- 2.4 Experimental Studies 32
- 2.5 Designs for Experimental Studies 38
- 2.6 Research Study: Exit Polls Versus Election Results 48
- 2.7 Summary 50
- 2.8 Exercises 50

PART 3 SUMMARIZING DATA 59

CHAPTER 3

Data Description 60

- 3.1 Introduction and Abstract of Research Study 60
- 3.2 Calculators, Computers, and Software Systems 65
- 3.3 Describing Data on a Single Variable: Graphical Methods 66
- 3.4 Describing Data on a Single Variable: Measures of Central Tendency 82
- 3.5 Describing Data on a Single Variable: Measures of Variability 90
- 3.6 The Boxplot 104
- 3.7 Summarizing Data from More Than One Variable: Graphs and Correlation 109

- 3.8 Research Study: Controlling for Student Background in the Assessment of Teaching 119
- 3.9 R Instructions 124
- 3.10 Summary and Key Formulas 124
- 3.11 Exercises 125

CHAPTER 4**Probability and Probability Distributions 149**

- 4.1 Introduction and Abstract of Research Study 149
- 4.2 Finding the Probability of an Event 153
- 4.3 Basic Event Relations and Probability Laws 155
- 4.4 Conditional Probability and Independence 158
- 4.5 Bayes' Formula 161
- 4.6 Variables: Discrete and Continuous 164
- 4.7 Probability Distributions for Discrete Random Variables 166
- 4.8 Two Discrete Random Variables: The Binomial and the Poisson 167
- 4.9 Probability Distributions for Continuous Random Variables 177
- 4.10 A Continuous Probability Distribution: The Normal Distribution 180
- 4.11 Random Sampling 187
- 4.12 Sampling Distributions 190
- 4.13 Normal Approximation to the Binomial 200
- 4.14 Evaluating Whether or Not a Population Distribution Is Normal 203
- 4.15 Research Study: Inferences About Performance-Enhancing Drugs Among Athletes 208
- 4.16 R Instructions 211
- 4.17 Summary and Key Formulas 212
- 4.18 Exercises 214

PART 4 ANALYZING THE DATA, INTERPRETING THE ANALYSES, AND COMMUNICATING THE RESULTS 231**CHAPTER 5****Inferences About Population Central Values 232**

- 5.1 Introduction and Abstract of Research Study 232
- 5.2 Estimation of μ 235
- 5.3 Choosing the Sample Size for Estimating μ 240
- 5.4 A Statistical Test for μ 242
- 5.5 Choosing the Sample Size for Testing μ 255
- 5.6 The Level of Significance of a Statistical Test 257
- 5.7 Inferences About μ for a Normal Population, σ Unknown 260
- 5.8 Inferences About μ When the Population Is Nonnormal and n Is Small: Bootstrap Methods 269
- 5.9 Inferences About the Median 275
- 5.10 Research Study: Percentage of Calories from Fat 280
- 5.11 Summary and Key Formulas 283
- 5.12 Exercises 285

CHAPTER 6**Inferences Comparing Two Population Central Values 300**

- 6.1 Introduction and Abstract of Research Study 300
- 6.2 Inferences About $\mu_1 - \mu_2$: Independent Samples 303

- 6.3 A Nonparametric Alternative:
The Wilcoxon Rank Sum Test 315
- 6.4 Inferences About $\mu_1 - \mu_2$: Paired Data 325
- 6.5 A Nonparametric Alternative:
The Wilcoxon Signed-Rank Test 329
- 6.6 Choosing Sample Sizes for Inferences About $\mu_1 - \mu_2$ 334
- 6.7 Research Study: Effects of an Oil Spill on Plant Growth 336
- 6.8 Summary and Key Formulas 341
- 6.9 Exercises 344

CHAPTER 7**Inferences About Population Variances 366**

- 7.1 Introduction and Abstract of Research Study 366
- 7.2 Estimation and Tests for a Population Variance 368
- 7.3 Estimation and Tests for Comparing Two Population Variances 376
- 7.4 Tests for Comparing $t > 2$ Population Variances 382
- 7.5 Research Study: Evaluation of Methods for Detecting *E. coli* 385
- 7.6 Summary and Key Formulas 390
- 7.7 Exercises 391

CHAPTER 8**Inferences About More Than Two Population Central Values 400**

- 8.1 Introduction and Abstract of Research Study 400
- 8.2 A Statistical Test About More Than Two Population Means:
An Analysis of Variance 403
- 8.3 The Model for Observations in a Completely Randomized Design 412
- 8.4 Checking on the AOV Conditions 414
- 8.5 An Alternative Analysis: Transformations of the Data 418
- 8.6 A Nonparametric Alternative: The Kruskal–Wallis Test 425
- 8.7 Research Study: Effect of Timing on the Treatment
of Port-Wine Stains with Lasers 428
- 8.8 Summary and Key Formulas 433
- 8.9 Exercises 435

CHAPTER 9**Multiple Comparisons 445**

- 9.1 Introduction and Abstract of Research Study 445
- 9.2 Linear Contrasts 447
- 9.3 Which Error Rate Is Controlled? 454
- 9.4 Scheffé's *S* Method 456
- 9.5 Tukey's *W* Procedure 458
- 9.6 Dunnett's Procedure: Comparison of Treatments to a Control 462
- 9.7 A Nonparametric Multiple-Comparison Procedure 464
- 9.8 Research Study: Are Interviewers' Decisions Affected by Different
Handicap Types? 467
- 9.9 Summary and Key Formulas 474
- 9.10 Exercises 475

CHAPTER 10

Categorical Data 482

- 10.1 Introduction and Abstract of Research Study 482
- 10.2 Inferences About a Population Proportion π 483
- 10.3 Inferences About the Difference Between Two Population Proportions, $\pi_1 - \pi_2$ 491
- 10.4 Inferences About Several Proportions: Chi-Square Goodness-of-Fit Test 501
- 10.5 Contingency Tables: Tests for Independence and Homogeneity 508
- 10.6 Measuring Strength of Relation 515
- 10.7 Odds and Odds Ratios 517
- 10.8 Combining Sets of 2×2 Contingency Tables 522
- 10.9 Research Study: Does Gender Bias Exist in the Selection of Students for Vocational Education? 525
- 10.10 Summary and Key Formulas 531
- 10.11 Exercises 533

CHAPTER 11

Linear Regression and Correlation 555

- 11.1 Introduction and Abstract of Research Study 555
- 11.2 Estimating Model Parameters 564
- 11.3 Inferences About Regression Parameters 574
- 11.4 Predicting New y -Values Using Regression 577
- 11.5 Examining Lack of Fit in Linear Regression 581
- 11.6 Correlation 587
- 11.7 Research Study: Two Methods for Detecting *E. coli* 598
- 11.8 Summary and Key Formulas 602
- 11.9 Exercises 604

CHAPTER 12

Multiple Regression and the General Linear Model 625

- 12.1 Introduction and Abstract of Research Study 625
- 12.2 The General Linear Model 635
- 12.3 Estimating Multiple Regression Coefficients 636
- 12.4 Inferences in Multiple Regression 644
- 12.5 Testing a Subset of Regression Coefficients 652
- 12.6 Forecasting Using Multiple Regression 656
- 12.7 Comparing the Slopes of Several Regression Lines 658
- 12.8 Logistic Regression 662
- 12.9 Some Multiple Regression Theory (Optional) 669
- 12.10 Research Study: Evaluation of the Performance of an Electric Drill 676
- 12.11 Summary and Key Formulas 683
- 12.12 Exercises 685

CHAPTER 13

Further Regression Topics 711

- 13.1 Introduction and Abstract of Research Study 711
- 13.2 Selecting the Variables (Step 1) 712
- 13.3 Formulating the Model (Step 2) 729
- 13.4 Checking Model Assumptions (Step 3) 745

- 13.5 Research Study: Construction Costs for Nuclear Power Plants 765
- 13.6 Summary and Key Formulas 772
- 13.7 Exercises 773

CHAPTER 14**Analysis of Variance for Completely Randomized Designs 798**

- 14.1 Introduction and Abstract of Research Study 798
- 14.2 Completely Randomized Design with a Single Factor 800
- 14.3 Factorial Treatment Structure 805
- 14.4 Factorial Treatment Structures with an Unequal Number of Replications 830
- 14.5 Estimation of Treatment Differences and Comparisons of Treatment Means 837
- 14.6 Determining the Number of Replications 841
- 14.7 Research Study: Development of a Low-Fat Processed Meat 846
- 14.8 Summary and Key Formulas 851
- 14.9 Exercises 852

CHAPTER 15**Analysis of Variance for Blocked Designs 865**

- 15.1 Introduction and Abstract of Research Study 865
- 15.2 Randomized Complete Block Design 866
- 15.3 Latin Square Design 878
- 15.4 Factorial Treatment Structure in a Randomized Complete Block Design 889
- 15.5 A Nonparametric Alternative—Friedman's Test 893
- 15.6 Research Study: Control of Leatherjackets 897
- 15.7 Summary and Key Formulas 902
- 15.8 Exercises 904

CHAPTER 16**The Analysis of Covariance 917**

- 16.1 Introduction and Abstract of Research Study 917
- 16.2 A Completely Randomized Design with One Covariate 920
- 16.3 The Extrapolation Problem 931
- 16.4 Multiple Covariates and More Complicated Designs 934
- 16.5 Research Study: Evaluation of Cool-Season Grasses for Putting Greens 936
- 16.6 Summary 942
- 16.7 Exercises 942

CHAPTER 17**Analysis of Variance for Some Fixed-, Random-, and Mixed-Effects Models 952**

- 17.1 Introduction and Abstract of Research Study 952
- 17.2 A One-Factor Experiment with Random Treatment Effects 955
- 17.3 Extensions of Random-Effects Models 959
- 17.4 Mixed-Effects Models 967
- 17.5 Rules for Obtaining Expected Mean Squares 971

- 17.6 Nested Factors 981
- 17.7 Research Study: Factors Affecting Pressure Drops
Across Expansion Joints 986
- 17.8 Summary 991
- 17.9 Exercises 992

CHAPTER 18

**Split-Plot, Repeated Measures,
and Crossover Designs 1004**

- 18.1 Introduction and Abstract of Research Study 1004
- 18.2 Split-Plot Designed Experiments 1008
- 18.3 Single-Factor Experiments with Repeated Measures 1014
- 18.4 Two-Factor Experiments with Repeated Measures on
One of the Factors 1018
- 18.5 Crossover Designs 1025
- 18.6 Research Study: Effects of an Oil Spill on Plant Growth 1033
- 18.7 Summary 1035
- 18.8 Exercises 1035

CHAPTER 19

**Analysis of Variance for Some Unbalanced
Designs 1050**

- 19.1 Introduction and Abstract of Research Study 1050
- 19.2 A Randomized Block Design with One or More
Missing Observations 1052
- 19.3 A Latin Square Design with Missing Data 1058
- 19.4 Balanced Incomplete Block (BIB) Designs 1063
- 19.5 Research Study: Evaluation of the Consistency
of Property Assessors 1070
- 19.6 Summary and Key Formulas 1074
- 19.7 Exercises 1075

Appendix: Statistical Tables 1085

Answers to Selected Exercises 1125

References 1151

Index 1157

PREFACE

INTENDED AUDIENCE

An Introduction to Statistical Methods and Data Analysis, Seventh Edition, provides a broad overview of statistical methods for advanced undergraduate and graduate students from a variety of disciplines. This book is intended to prepare students to solve problems encountered in research projects, to make decisions based on data in general settings both within and beyond the university setting, and finally to become critical readers of statistical analyses in research papers and in news reports. The book presumes that the students have a minimal mathematical background (high school algebra) and no prior course work in statistics. The first 11 chapters of the textbook present the material typically covered in an introductory statistics course. However, this book provides research studies and examples that connect the statistical concepts to data analysis problems that are often encountered in undergraduate capstone courses. The remaining chapters of the book cover regression modeling and design of experiments. We develop and illustrate the statistical techniques and thought processes needed to design a research study or experiment and then analyze the data collected using an intuitive and proven four-step approach. This should be especially helpful to graduate students conducting their MS thesis and PhD dissertation research.

MAJOR FEATURES OF TEXTBOOK

Learning from Data

In this text, we approach the study of statistics by considering a four-step process by which we can learn from data:

1. Defining the Problem
2. Collecting the Data
3. Summarizing the Data
4. Analyzing the Data, Interpreting the Analyses, and Communicating the Results

Case Studies

In order to demonstrate the relevance and critical nature of statistics in solving real-world problems, we introduce the major topic of each chapter using a case study. The case studies were selected from many sources to illustrate the broad applicability of statistical methodology. The four-step learning from data process is illustrated through the case studies. This approach will hopefully assist in overcoming

the natural initial perception held by many people that statistics is just another “math course.” The introduction of major topics through the use of case studies provides a focus on the central nature of applied statistics in a wide variety of research and business-related studies. These case studies will hopefully provide the reader with an enthusiasm for the broad applicability of statistics and the statistical thought process that the authors have found and used through their many years of teaching, consulting, and R & D management. The following research studies illustrate the types of studies we have used throughout the text.

- **Exit Polls Versus Election Results:** A study of why the exit polls from 9 of 11 states in the 2004 presidential election predicted John Kerry as the winner when in fact President Bush won 6 of the 11 states.
- **Evaluation of the Consistency of Property Assessors:** A study to determine if county property assessors differ systematically in their determination of property values.
- **Effect of Timing of the Treatment of Port-Wine Stains with Lasers:** A prospective study that investigated whether treatment at a younger age would yield better results than treatment at an older age.
- **Controlling for Student Background in the Assessment of Teaching:** An examination of data used to support possible improvements to the No Child Left Behind program while maintaining the important concepts of performance standards and accountability.

Each of the research studies includes a discussion of the whys and hows of the study. We illustrate the use of the four-step learning from data process with each case study. A discussion of sample size determination, graphical displays of the data, and a summary of the necessary ingredients for a complete report of the statistical findings of the study are provided with many of the case studies.

Examples and Exercises

We have further enhanced the practical nature of statistics by using examples and exercises from journal articles, newspapers, and the authors’ many consulting experiences. These will provide the students with further evidence of the practical usages of statistics in solving problems that are relevant to their everyday lives. Many new exercises and examples have been included in this edition of the book. The number and variety of exercises will be a great asset to both the instructor and students in their study of statistics.

Topics Covered

This book can be used for either a one-semester or a two-semester course. Chapters 1 through 11 would constitute a one-semester course. The topics covered would include

- Chapter 1—Statistics and the scientific method
- Chapter 2—Using surveys and experimental studies to gather data
- Chapters 3 & 4—Summarizing data and probability distributions
- Chapters 5–7—Analyzing data: inferences about central values and variances
- Chapters 8 & 9—One-way analysis of variance and multiple comparisons

Chapter 10—Analyzing data involving proportions

Chapter 11—Linear regression and correlation

The second semester of a two-semester course would then include model building and inferences in multiple regression analysis, logistic regression, design of experiments, and analysis of variance:

Chapters 11–13—Regression methods and model building: multiple regression and the general linear model, logistic regression, and building regression models with diagnostics

Chapters 14–19—Design of experiments and analysis of variance: design concepts, analysis of variance for standard designs, analysis of covariance, random and mixed effects models, split-plot designs, repeated measures designs, crossover designs, and unbalanced designs

Emphasis on Interpretation, not Computation

In the book are examples and exercises that allow the student to study how to calculate the value of statistical estimators and test statistics using the definitional form of the procedure. After the student becomes comfortable with the aspects of the data the statistical procedure is reflecting, we then emphasize the use of computer software in making computations in the analysis of larger data sets. We provide output from three major statistical packages: SAS, Minitab, and SPSS. We find that this approach provides the student with the experience of computing the value of the procedure using the definition; hence, the student learns the basics behind each procedure. In most situations beyond the statistics course, the student should be using computer software in making the computations for both expedience and quality of calculation. In many exercises and examples, the use of the computer allows for more time to emphasize the interpretation of the results of the computations without having to expend enormous amounts of time and effort in the actual computations.

In numerous examples and exercises, the importance of the following aspects of hypothesis testing are demonstrated:

1. The statement of the research hypothesis through the summarization of the researcher's goals into a statement about population parameters.
2. The selection of the most appropriate test statistic, including sample size computations for many procedures.
3. The necessity of considering both Type I and Type II error rates (α and β) when discussing the results of a statistical test of hypotheses.
4. The importance of considering both the statistical significance and the practical significance of a test result. Thus, we illustrate the importance of estimating effect sizes and the construction of confidence intervals for population parameters.
5. The statement of the results of the statistical test in nonstatistical jargon that goes beyond the statement “reject H_0 ” or “fail to reject H_0 .”

New to the Seventh Edition

- There are instructions on the use of R code. R is a free software package that can be downloaded from <http://lib.stat.cmu.edu/R/CRAN>.

Click your choice of platform (Linux, MacOS X, or Windows) for the precompiled binary distribution. Note the FAQs link to the left for additional information. Follow the instructions for installing the *base* system software (which is all you will need).

- New examples illustrate the breadth of applications of statistics to real-world problems.
- An alternative to the standard deviation, MAD, is provided as a measure of dispersion in a population/sample.
- The use of bootstrapping in obtaining confidence intervals and p-values is discussed.
- Instructions are included on how to use R code to obtain percentiles and probabilities from the following distributions: normal, binomial, Poisson, chi-squared, F , and t .
- A nonparametric alternative to the Pearson correlation coefficient: Spearman's rank correlation, is provided.
- The binomial test for small sample tests of proportions is presented.
- The McNemar test for paired count data has been added.
- The Akaike information criterion and Bayesian information criterion for variable selection are discussed.

Additional Features Retained from Previous Editions

- Many practical applications of statistical methods and data analysis from agriculture, business, economics, education, engineering, medicine, law, political science, psychology, environmental studies, and sociology have been included.
- The seventh edition contains over 1,000 exercises, with nearly 400 of the exercises new.
- Computer output from Minitab, SAS, and SPSS is provided in numerous examples. The use of computers greatly facilitates the use of more sophisticated graphical illustrations of statistical results.
- Attention is paid to the underlying assumptions. Graphical procedures and test procedures are provided to determine if assumptions have been violated. Furthermore, in many settings, we provide alternative procedures when the conditions are not met.
- The first chapter provides a discussion of “What Is Statistics?” We provide a discussion of why students should study statistics along with a discussion of several major studies that illustrate the use of statistics in the solution of real-life problems.

Ancillaries

- Student Solutions Manual (ISBN-10: 1-305-26948-9; ISBN-13: 978-1-305-26948-4), containing select worked solutions for problems in the textbook.
- A Companion Website at www.cengage.com/statistics/ott, containing downloadable data sets for Excel, Minitab, SAS, SPSS, and others, plus additional resources for students and faculty.

Acknowledgments

There are many people who have made valuable, constructive suggestions for the development of the original manuscript and during the preparation of the subsequent editions. We are very appreciative of the insightful and constructive comments from the following reviewers:

Naveen Bansal, Marquette University

Kameryn Denaro, San Diego State University

Mary Gray, American University

Craig Leth-Steensen, Carleton University

Jing Qian, University of Massachusetts

Mark Riggs, Abilene Christian University

Elaine Spiller, Marquette University

We also appreciate of the preparation assistance received from Molly Taylor and Jay Campbell; the scheduling of the revisions by Mary Tindle, the Senior Project Manager at Cengage Publisher Services, who made sure that the book was completed in a timely manner. The authors of the solutions manual, Soma Roy, California Polytechnic State University, and John Draper, The Ohio State University, provided me with excellent input which resulted in an improved set of exercises for the seventh edition. The person who assisted me the greatest degree in the preparation of the seventh edition, was Sherry Goldbecker, the copy editor. Sherry not only corrected my many grammatical errors but also provided rephrasing of many sentences which made for a more straight forward explanation of statistical concepts. The students, who use this book in their statistics classes, will be most appreciative of Sherry's many contributions.

Introduction

CHAPTER 1 **Statistics and the Scientific Method**

CHAPTER 1

Statistics and the Scientific Method

- 1.1 Introduction
- 1.2 Why Study Statistics?
- 1.3 Some Current Applications of Statistics
- 1.4 A Note to the Student
- 1.5 Summary
- 1.6 Exercises

1.1 Introduction

Statistics is the science of designing studies or experiments, collecting data, and modeling/analyzing data for the purpose of decision making and scientific discovery when the available information is both limited and variable. That is, statistics is the science of *Learning from Data*.

Almost everyone, including social scientists, medical researchers, superintendents of public schools, corporate executives, market researchers, engineers, government employees, and consumers, deals with data. These data could be in the form of quarterly sales figures, percent increase in juvenile crime, contamination levels in water samples, survival rates for patients undergoing medical therapy, census figures, or information that helps determine which brand of car to purchase. In this text, we approach the study of statistics by considering the four-step process in *Learning from Data*: (1) defining the problem, (2) collecting the data, (3) summarizing the data, and (4) analyzing the data, interpreting the analyses, and communicating the results. Through the use of these four steps in *Learning from Data*, our study of statistics closely parallels the Scientific Method, which is a set of principles and procedures used by successful scientists in their pursuit of knowledge. The method involves the formulation of research goals, the design of observational studies and/or experiments, the collection of data, the modeling/analysis of the data in the context of research goals, and the testing of hypotheses. The conclusion of these steps is often the formulation of new research goals for another study. These steps are illustrated in the schematic given in Figure 1.1.

This book is divided into sections corresponding to the four-step process in *Learning from Data*. The relationship among these steps and the chapters of the book is shown in Table 1.1. As you can see from this table, much time is spent discussing how to analyze data using the basic methods presented in Chapters 5–19.

FIGURE 1.1
Scientific Method
Schematic

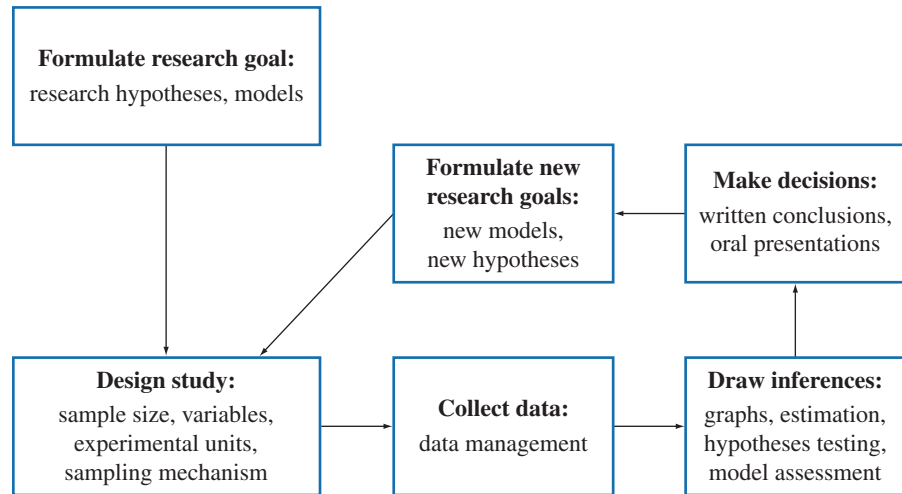


TABLE 1.1
Organization of the text

The Four-Step Process	Chapters
1 Defining the Problem	1 Statistics and the Scientific Method
2 Collecting the Data	2 Using Surveys and Experimental Studies to Gather Data
3 Summarizing the Data	3 Data Description
	4 Probability and Probability Distributions
4 Analyzing the Data, Interpreting the Analyses, and Communicating the Results	5 Inferences about Population Central Values
	6 Inferences Comparing Two Population Central Values
	7 Inferences about Population Variances
	8 Inferences about More Than Two Population Central Values
	9 Multiple Comparisons
	10 Categorical Data
	11 Linear Regression and Correlation
	12 Multiple Regression and the General Linear Model
	13 Further Regression Topics
	14 Analysis of Variance for Completely Randomized Designs
	15 Analysis of Variance for Blocked Designs
	16 The Analysis of Covariance
	17 Analysis of Variance for Some Fixed-, Random-, and Mixed-Effects Models
	18 Split-Plot, Repeated Measures, and Crossover Designs
	19 Analysis of Variance for Some Unbalanced Designs

However, you must remember that for each data set requiring analysis, someone has defined the problem to be examined (Step 1), developed a plan for collecting data to address the problem (Step 2), and summarized the data and prepared the data for analysis (Step 3). Then following the analysis of the data, the results of the analysis must be interpreted and communicated either verbally or in written form to the intended audience (Step 4).

All four steps are important in Learning from Data; in fact, unless the problem to be addressed is clearly defined and the data collection carried out properly, the interpretation of the results of the analyses may convey misleading information because the analyses were based on a data set that did not address the problem or that was incomplete and contained improper information. Throughout the text,

we will try to keep you focused on the bigger picture of Learning from Data through the four-step process. Most chapters will end with a summary section that emphasizes how the material of the chapter fits into the study of statistics—Learning from Data.

To illustrate some of the above concepts, we will consider four situations in which the four steps in Learning from Data could assist in solving a real-world problem.

1. Problem: Inspection of ground beef in a large beef-processing facility.

A beef-processing plant produces approximately half a million packages of ground beef per week. The government inspects packages for possible improper labeling of the packages with respect to the percent fat in the meat. The inspectors must open the ground beef package in order to determine the fat content of the ground beef. The inspection of every package would be prohibitively costly and time consuming. An alternative approach is to select 250 packages for inspection from the daily production of 100,000 packages. The fraction of packages with improper labeling in the sample of 250 packages would then be used to estimate the fraction of packages improperly labeled in the complete day's production. If this fraction exceeds a set specification, action is then taken against the meat processor. In later chapters, a procedure will be formulated to determine how well the sample fraction of improperly labeled packages approximates the fraction of improperly labeled packages for the whole day's output.

2. Problem: Is there a relationship between quitting smoking and gaining weight?

To investigate the claim that people who quit smoking often experience a subsequent weight gain, researchers selected a random sample of 400 participants who had successfully participated in programs to quit smoking. The individuals were weighed at the beginning of the program and again 1 year later. The average change in weight of the participants was an increase of 5 pounds. The investigators concluded that there was evidence that the claim was valid. We will develop techniques in later chapters to assess when changes are truly significant changes and not changes due to random chance.

3. Problem: What effect does nitrogen fertilizer have on wheat production?

For a study of the effects of nitrogen fertilizer on wheat production, a total of 15 fields was available to the researcher. She randomly assigned three fields to each of the five nitrogen rates under investigation. The same variety of wheat was planted in all 15 fields. The fields were cultivated in the same manner until harvest, and the number of pounds of wheat per acre was then recorded for each of the 15 fields. The experimenter wanted to determine the optimal level of nitrogen to apply to *any* wheat field, but, of course, she was limited to running experiments on a limited number of fields. After determining the amount of nitrogen that yielded the largest production of wheat in the study fields, the experimenter then concluded that similar results would hold for wheat fields possessing characteristics somewhat the same as the study fields. Is the experimenter justified in reaching this conclusion?

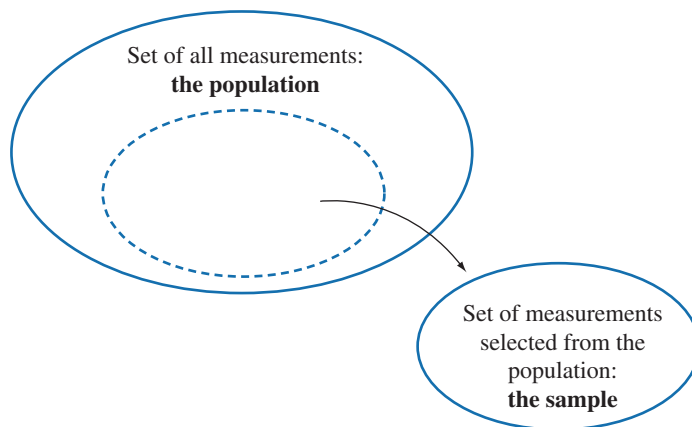
4. Problem: Determining public opinion toward a question, issue, product, or candidate. Similar applications of statistics are brought to mind by the frequent use of the *New York Times/CBS News*, *Washington Post/ABC News*, *Wall Street Journal/NBC News*, *Harris, Gallup/Newsweek*, and *CNN/Time* polls. How can these pollsters determine the opinions of more than 195 million Americans who are of voting age? They certainly do not contact every potential voter in the United States. Rather, they sample the opinions of a small number of potential voters, perhaps as few as 1,500, to estimate the reaction of every person of voting age in the country. The amazing result of this process is that if the selection of the voters is done in an unbiased way and voters are asked unambiguous, nonleading questions, the fraction of those persons contacted who hold a particular opinion will closely match the fraction in the total population holding that opinion at a particular time. We will supply convincing supportive evidence of this assertion in subsequent chapters.

These problems illustrate the four-step process in Learning from Data. First, there was a problem or question to be addressed. Next, for each problem a study or experiment was proposed to collect meaningful data to solve the problem. The government meat inspection agency had to decide both how many packages to inspect per day and how to select the sample of packages from the total daily output in order to obtain a valid prediction. The polling groups had to decide how many voters to sample and how to select these individuals in order to obtain information that is representative of the population of all voters. Similarly, it was necessary to carefully plan how many participants in the weight-gain study were needed and how they were to be selected from the list of all such participants. Furthermore, what variables did the researchers have to measure on each participant? Was it necessary to know each participant's age, sex, physical fitness, and other health-related variables, or was weight the only important variable? The results of the study may not be relevant to the general population if many of the participants in the study had a particular health condition. In the wheat experiment, it was important to measure both the soil characteristics of the fields and the environmental conditions, such as temperature and rainfall, to obtain results that could be generalized to fields not included in the study. The design of a study or experiment is crucial to obtaining results that can be generalized beyond the study.

Finally, having collected, summarized, and analyzed the data, it is important to report the results in unambiguous terms to interested people. For the meat inspection example, the government inspection agency and the personnel in the beef-processing plant would need to know the distribution of fat content in the daily production of ground beef. Based on this distribution, the agency could then impose fines or take other remedial actions against the production facility. Also, knowledge of this distribution would enable company production personnel to make adjustments to the process in order to obtain acceptable fat content in their ground beef packages. Therefore, the results of the statistical analyses cannot be presented in ambiguous terms; decisions must be made from a well-defined knowledge base. The results of the weight-gain study would be of vital interest to physicians who have patients participating in the smoking-cessation program. If a significant increase in weight was recorded for those individuals who had quit smoking, physicians would have to recommend diets so that the former smokers

FIGURE 1.2

Population and sample



would not go from one health problem (smoking) to another (elevated blood pressure due to being overweight). It is crucial that a careful description of the participants—that is, age, sex, and other health-related information—be included in the report. In the wheat study, the experiment would provide farmers with information that would allow them to economically select the optimum amount of nitrogen required for their fields. Therefore, the report must contain information concerning the amount of moisture and types of soils present on the study fields. Otherwise, the conclusions about optimal wheat production may not pertain to farmers growing wheat under considerably different conditions.

population**sample**

To infer validly that the results of a study are applicable to a larger group than just the participants in the study, we must carefully define the **population** (see Definition 1.1) to which inferences are sought and design a study in which the **sample** (see Definition 1.2) has been appropriately selected from the designated population. We will discuss these issues in Chapter 2.

DEFINITION 1.1

A **population** is the set of all measurements of interest to the sample collector. (See Figure 1.2.)

DEFINITION 1.2

A **sample** is any subset of measurements selected from the population. (See Figure 1.2.)

1.2 Why Study Statistics?

We can think of many reasons for taking an introductory course in statistics. One reason is that you need to know how to evaluate published numerical facts. Every person is exposed to manufacturers' claims for products; to the results of sociological, consumer, and political polls; and to the published results of scientific research. Many of these results are inferences based on sampling. Some inferences are valid; others are invalid. Some are based on samples of adequate size; others are not. Yet all these published results bear the ring of truth. Some people (particularly statisticians) say that statistics can be made to support almost

anything. Others say it is easy to lie with statistics. Both statements are true. It is easy, purposely or unwittingly, to distort the truth by using statistics when presenting the results of sampling to the uninformed. It is thus crucial that you become an informed and critical reader of data-based reports and articles.

A second reason for studying statistics is that your profession or employment may require you to interpret the results of sampling (surveys or experimentation) or to employ statistical methods of analysis to make inferences in your work. For example, practicing physicians receive large amounts of advertising describing the benefits of new drugs. These advertisements frequently display the numerical results of experiments that compare a new drug with an older one. Do such data really imply that the new drug is more effective, or is the observed difference in results due simply to random variation in the experimental measurements?

Recent trends in the conduct of court trials indicate an increasing use of probability and statistical inference in evaluating the quality of evidence. The use of statistics in the social, biological, and physical sciences is essential because all these sciences make use of observations of natural phenomena, through sample surveys or experimentation, to develop and test new theories. Statistical methods are employed in business when sample data are used to forecast sales and profit. In addition, they are used in engineering and manufacturing to monitor product quality. The sampling of accounts is a useful tool to assist accountants in conducting audits. Thus, statistics plays an important role in almost all areas of science, business, and industry; persons employed in these areas need to know the basic concepts, strengths, and limitations of statistics.

The article ***“What Educated Citizens Should Know About Statistics and Probability,”*** by J. Utts (2003), contains a number of statistical ideas that need to be understood by users of statistical methodology in order to avoid confusion in the use of their research findings. Misunderstandings of statistical results can lead to major errors by government policymakers, medical workers, and consumers of this information. The article selected a number of topics for discussion. We will summarize some of the findings in the article. A complete discussion of all these topics will be given throughout the book.

1. One of the most frequent misinterpretations of statistical findings is when a statistically significant relationship is established between two variables and it is then concluded that a change in the explanatory variable *causes* a change in the response variable. As will be discussed in the book, this conclusion can be reached only under very restrictive constraints on the experimental setting. Utts examined a recent *Newsweek* article discussing the relationship between the strength of religious beliefs and physical healing. Utts’ article discussed the problems in reaching the conclusion that the stronger a patient’s religious beliefs, the more likely the patient would be cured of his or her ailment. Utts showed that there are numerous other factors involved in a patient’s health and the conclusion that religious beliefs cause a cure cannot be validly reached.
2. A common confusion in many studies is the difference between *(statistically) significant* findings in a study and *(practically) significant* findings. This problem often occurs when large data sets are involved in a study or experiment. This type of problem will be discussed in detail throughout the book. We will use a number of examples that will illustrate how this type of confusion can be avoided by researchers when reporting the findings of their experimental results.

Utts' article illustrated this problem with a discussion of a study that found a statistically significant difference in the average heights of military recruits born in the spring and in the fall. There were 507,125 recruits in the study and the difference in average height was about 1/4 inch. So, even though there may be a difference in the actual average heights of recruits in the spring and the fall, the difference is so small (1/4 inch) that it is of no practical importance.

3. The size of the sample also may be a determining factor in studies in which statistical significance is *not* found. A study may not have selected a sample size large enough to discover a difference between the several populations under study. In many government-sponsored studies, the researchers do not receive funding unless they are able to demonstrate that the sample sizes selected for their study are of an appropriate size to detect specified differences in populations if in fact they exist. Methods to determine appropriate sample sizes will be provided in the chapters on hypotheses testing and experimental design.
4. Surveys are ubiquitous, especially during the years in which national elections are held. In fact, market surveys are nearly as widespread as political polls. There are many sources of bias that can creep into the most reliable of surveys. The manner in which people are selected for inclusion in the survey, the way in which questions are phrased, and even the manner in which questions are posed to the subject may affect the conclusions obtained from the survey. We will discuss these issues in Chapter 2.
5. Many students find the topic of probability to be very confusing. One of these confusions involves conditional probability where the probability of an event occurring is computed under the condition that a second event has occurred with certainty. For example, a new diagnostic test for the pathogen *Escherichia coli* in meat is proposed to the U.S. Department of Agriculture (USDA). The USDA evaluates the test and determines that the test has both a low *false positive* rate and a low *false negative* rate. That is, it is very unlikely that the test will declare the meat contains *E. coli* when in fact it does not contain *E. coli*. Also, it is very unlikely that the test will declare the meat does not contain *E. coli* when in fact it does contain *E. coli*. Although the diagnostic test has a very low false positive rate and a very low false negative rate, the probability that *E. coli* is in fact present in the meat when the test yields a positive test result is *very* low for those situations in which a particular strain of *E. coli* occurs very infrequently. In Chapter 4, we will demonstrate how this probability can be computed in order to provide a true assessment of the performance of a diagnostic test.
6. Another concept that is often misunderstood is the role of the degree of variability in interpreting what is a “normal” occurrence of some naturally occurring event. Utts' article provided the following example. A company was having an odor problem with its wastewater treatment plant. It attributed the problem to “abnormal” rainfall during the period in which the odor problem was occurring. A company official stated that the facility experienced 170% to 180% of its “normal” rainfall during this period, which resulted in the water in

the holding ponds taking longer to exit for irrigation. Thus, there was more time for the pond to develop an odor. The company official did not point out that yearly rainfall in this region is extremely variable. In fact, the historical range for rainfall is between 6.1 and 37.4 inches with a median rainfall of 16.7 inches. The rainfall for the year of the odor problem was 29.7 inches, which was well within the “normal” range for rainfall. There was a confusion between the terms “average” and “normal” rainfall. The concept of natural variability is crucial to correct interpretation of statistical results. In this example, the company official should have evaluated the percentile for an annual rainfall of 29.7 inches in order to demonstrate the abnormality of such a rainfall. We will discuss the ideas of data summaries and percentiles in Chapter 3.

The types of problems expressed above and in Utts’ article represent common and important misunderstandings that can occur when researchers use statistics in interpreting the results of their studies. We will attempt throughout the book to discuss possible misinterpretations of statistical results and how to avoid them in your data analyses. More importantly, we want the reader of this book to become a discriminating reader of statistical findings, the results of surveys, and project reports.

1.3 Some Current Applications of Statistics

Defining the Problem: Obtaining Information from Massive Data Sets

Data mining is defined to be a process by which useful information is obtained from large sets of data. Data mining uses statistical techniques to discover patterns and trends that are present in a large data set. In most data sets, important patterns would not be discovered by using traditional data exploration techniques because the types of relationships between the many variables in the data set are either too complex or because the data sets are so large that they mask the relationships.

The patterns and trends discovered in the analysis of the data are defined as data mining models. These models can be applied to many different situations, such as:

- Forecasting: Estimating future sales, predicting demands on a power grid, or estimating server downtime
- Assessing risk: Choosing the rates for insurance premiums, selecting best customers for a new sales campaign, determining which medical therapy is most appropriate given the physiological characteristics of the patient
- Identifying sequences: Determining customer preferences in online purchases, predicting weather events
- Grouping: Placing customers or events into cluster of related items, analyzing and predicting relationships between demographic characteristics and purchasing patterns, identifying fraud in credit card purchases

A new medical procedure referred to as gene editing has the potential to assist thousands of people suffering many different diseases. An article in the *Houston Chronicle* (2013), describes how data mining techniques are used to