EIGHTH EDITION

Fundamentals of Biostatistics

Bernard Rosner

Fundamentals of Biostatistics

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Fundamentals of Biostatistics

8TH EDITION

Bernard Rosner

Harvard University

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

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.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Fundamentals of Biostatistics, Eighth Edition
Bernard Rosner

Product Manager: Rita Lombard

Content Developer: Andrew Coppola

Associate Content Developer: Spencer Arritt

Product Assistant: Kathryn Schrumpf

Marketing Manager: Julie Schuster

Content Project Manager: Cheryll Linthicum

Art Director: Vernon Boes

Manufacturing Planner: Doug Bertke

Intellectual Property Analyst: Christina Ciaramella

Intellectual Property Project Manager: Farah Fard

Text and Cover Designer: C. Miller

Cover Image Credit: Abstract background: iStockPhoto.com/Pobytov; Office worker: Pressmaster/Shutterstock.com; financial diagram: iStockPhoto.com/Petrovich9; Test tube: iStockPhoto/HadelProductions; financial diagram: iStockPhoto.com/SergeyTimashov; Lab glass: iStocklPhoto.com/isak55.

Production Service and Compositor: Cenveo® Publisher Services

© 2016, 2011, 2006 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: 2015941787

ISBN: 978-1-305-26892-0

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

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This book is dedicated to my wife, Cynthia,
and my children, Sarah, David, and Laura

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CONTENTS

Preface / xiii

CHAPTER 1

General Overview / 1

CHAPTER 2

Descriptive Statistics / 5

2.1 Introduction / 5

2.2 Measures of Location / 6

2.3 Some Properties of the Arithmetic Mean / 14

2.4 Measures of Spread / 16

2.5 Some Properties of the Variance and Standard Deviation / 20

2.6 The Coefficient of Variation / 22

2.7 Grouped Data / 24

2.8 Graphic Methods / 27

2.9 Case Study 1: Effects of Lead Exposure on Neurological and Psychological Function in Children / 32

2.10 Case Study 2: Effects of Tobacco Use on Bone-Mineral Density in Middle-Aged Women / 32

2.11 Obtaining Descriptive Statistics on the Computer / 35

2.12 Summary / 35

Problems / 35

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CHAPTER 3

Probability / 42

3.1 Introduction / 42
3.2 Definition of Probability / 43
3.3 Some Useful Probabilistic Notation / 44
3.4 The Multiplication Law of Probability / 46
3.5 The Addition Law of Probability / 48
3.6 Conditional Probability / 50
3.7 Bayes' Rule and Screening Tests / 55
3.8 Bayesian Inference / 60
3.9 ROC Curves / 61
3.10 Prevalence and Incidence / 63
3.11 Summary / 64
Problems / 65

CHAPTER 4

Discrete Probability Distributions / 77

4.1 Introduction / 77
4.2 Random Variables / 78
4.3 The Probability-Mass Function for a Discrete Random Variable / 79
4.4 The Expected Value of a Discrete Random Variable / 81
4.5 The Variance of a Discrete Random Variable / 82
4.6 The Cumulative-Distribution Function of a Discrete Random Variable / 84
4.7 Permutations and Combinations / 85
4.8 The Binomial Distribution / 90
4.9 Expected Value and Variance of the Binomial Distribution / 96
4.10 The Poisson Distribution / 98
4.11 Computation of Poisson Probabilities / 101
4.12 Expected Value and Variance of the Poisson Distribution / 102
4.13 Poisson Approximation to the Binomial Distribution / 104
4.14 Summary / 106
Problems / 107

CHAPTER 5

Continuous Probability Distributions / 115

5.1 Introduction / 115
5.2 General Concepts / 115
5.3 The Normal Distribution / 118
5.4 Properties of the Standard Normal Distribution / 121
5.5 Conversion from an $N(\mu,\sigma^2)$ Distribution to an $N(0,1)$ Distribution / 127
5.6 Linear Combinations of Random Variables / 132
5.7 Normal Approximation to the Binomial Distribution / 133
5.8 Normal Approximation to the Poisson Distribution / 139
5.9 Summary / 141
Problems / 142

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

CHAPTER 6

Estimation / 154

6.1 Introduction / 154

6.2 The Relationship Between Population and Sample / 155

6.3 Random-Number Tables / 157

6.4 Randomized Clinical Trials / 161

6.5 Estimation of the Mean of a Distribution / 165

6.6 Case Study: Effects of Tobacco Use on Bone-Mineral Density (BMD) in Middle-Aged Women / 180

6.7 Estimation of the Variance of a Distribution / 181

6.8 Estimation for the Binomial Distribution / 187

6.9 Estimation for the Poisson Distribution / 193

6.10 One-Sided Confidence Intervals / 197

6.11 The Bootstrap / 199

6.12 Summary / 202

Problems / 203

CHAPTER 7

Hypothesis Testing: One-Sample Inference / 211

7.1 Introduction / 211

7.2 General Concepts / 211

7.3 One-Sample Test for the Mean of a Normal Distribution: One-Sided Alternatives / 214

7.4 One-Sample Test for the Mean of a Normal Distribution: Two-Sided Alternatives / 222

7.5 The Relationship Between Hypothesis Testing and Confidence Intervals / 229

7.6 The Power of a Test / 232

7.7 Sample-Size Determination / 239

7.8 One-Sample χ^2 Test for the Variance of a Normal Distribution / 245

7.9 One-Sample Inference for the Binomial Distribution / 249

7.10 One-Sample Inference for the Poisson Distribution / 259

7.11 Case Study: Effects of Tobacco Use on Bone-Mineral Density in Middle-Aged Women / 265

7.12 Derivation of Selected Formulas / 265

7.13 Summary / 267

Problems / 269

CHAPTER 8

Hypothesis Testing: Two-Sample Inference / 279

8.1 Introduction / 279

8.2 The Paired *t* Test / 281

8.3 Interval Estimation for the Comparison of Means from Two Paired Samples / 285

8.4 Two-Sample *t* Test for Independent Samples with Equal Variances / 286

8.5 Interval Estimation for the Comparison of Means from Two Independent Samples (Equal Variance Case) / 290

8.6 Testing for the Equality of Two Variances / 292

8.7 Two-Sample *t* Test for Independent Samples with Unequal Variances / 298

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

8.8 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 305

8.9 Estimation of Sample Size and Power for Comparing Two Means / 307

8.10 The Treatment of Outliers / 312

8.11 Derivation of Equation 8.13 / 319

8.12 Summary / 320

Problems / 320

CHAPTER 9

Nonparametric Methods / 338

9.1 Introduction / 338

9.2 The Sign Test / 340

9.3 The Wilcoxon Signed-Rank Test / 345

9.4 The Wilcoxon Rank-Sum Test / 352

9.5 Case Study: Effects of Lead Exposure on Neurological and Psychological Function in Children / 358

9.6 Permutation Tests / 359

9.7 Summary / 364

Problems / 365

CHAPTER 10

Hypothesis Testing: Categorical Data / 372

10.1 Introduction / 372

10.2 Two-Sample Test for Binomial Proportions / 373

10.3 Fisher's Exact Test / 387

10.4 Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test) / 395

10.5 Estimation of Sample Size and Power for Comparing Two Binomial Proportions / 403

10.6 $R \times C$ Contingency Tables / 413

10.7 Chi-Square Goodness-of-Fit Test / 425

10.8 The Kappa Statistic / 431

10.9 Derivation of Selected Formulas / 436

10.10 Summary / 437

Problems / 439

CHAPTER 11

Regression and Correlation Methods / 457

11.1 Introduction / 457

11.2 General Concepts / 458

11.3 Fitting Regression Lines—The Method of Least Squares / 461

11.4 Inferences About Parameters from Regression Lines / 465

11.5 Interval Estimation for Linear Regression / 475

11.6 Assessing the Goodness of Fit of Regression Lines / 481

11.7 The Correlation Coefficient / 485

11.8 Statistical Inference for Correlation Coefficients / 490

11.9 Multiple Regression / 502

11.10 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 519

11.11 Partial and Multiple Correlation / 526

11.12 Rank Correlation / 529

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

11.13 Interval Estimation for Rank-Correlation Coefficients / 533
11.14 Derivation of Equation 11.26 / 537
11.15 Summary / 539
Problems / 540

CHAPTER 12

Multisample Inference / 551

12.1 Introduction to the One-Way Analysis of Variance / 551
12.2 One-Way ANOVA—Fixed-Effects Model / 552
12.3 Hypothesis Testing in One-Way ANOVA—Fixed-Effects Model / 553
12.4 Comparisons of Specific Groups in One-Way ANOVA / 559
12.5 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 579
12.6 Two-Way ANOVA / 589
12.7 The Kruskal-Wallis Test / 596
12.8 One-Way ANOVA—The Random-Effects Model / 604
12.9 The Intraclass Correlation Coefficient / 609
12.10 Mixed Models / 614
12.11 Derivation of Equation 12.30 / 619
12.12 Summary / 620
Problems / 621

CHAPTER 13

Design and Analysis Techniques for Epidemiologic Studies / 633

13.1 Introduction / 633
13.2 Study Design / 634
13.3 Measures of Effect for Categorical Data / 637
13.4 Attributable Risk / 647
13.5 Confounding and Standardization / 653
13.6 Methods of Inference for Stratified Categorical Data—The Mantel-Haenszel Test / 659
13.7 Multiple Logistic Regression / 673
13.8 Extensions to Logistic Regression / 694
13.9 Sample Size Estimation for Logistic Regression / 703
13.10 Meta-Analysis / 705
13.11 Equivalence Studies / 710
13.12 The Cross-Over Design / 713
13.13 Clustered Binary Data / 721
13.14 Longitudinal Data Analysis / 733
13.15 Measurement-Error Methods / 743
13.16 Missing Data / 753
13.17 Derivation of $100\% \times (1-\alpha)$ CI for the Risk Difference / 758
13.18 Summary / 761
Problems / 762

CHAPTER 14

Hypothesis Testing: Person-Time Data / 777

14.1 Measure of Effect for Person-Time Data / 777
14.2 One-Sample Inference for Incidence-Rate Data / 779
14.3 Two-Sample Inference for Incidence-Rate Data / 782
14.4 Power and Sample-Size Estimation for Person-Time Data / 790

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

14.5 Inference for Stratified Person-Time Data / 793
14.6 Power and Sample-Size Estimation for Stratified Person-Time Data / 800
14.7 Testing for Trend: Incidence-Rate Data / 805
14.8 Introduction to Survival Analysis / 808
14.9 Estimation of Survival Curves: The Kaplan-Meier Estimator / 811
14.10 The Log-Rank Test / 819
14.11 The Proportional-Hazards Model / 825
14.12 Power and Sample-Size Estimation under the Proportional-Hazards Model / 835
14.13 Parametric Survival Analysis / 839
14.14 Parametric Regression Models for Survival Data / 847
14.15 Derivation of Selected Formulas / 854
14.16 Summary / 856
Problems / 856

APPENDIX

Tables / 867

1 Exact binomial probabilities $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$ / 867
2 Exact Poisson probabilities $Pr(X = k) = \frac{e^{-\mu}\mu^k}{k!}$ / 871
3 The normal distribution / 874
4 Table of 1000 random digits / 878
5 Percentage points of the *t* distribution $(t_{d,u})^a$ / 879
6 Percentage points of the chi-square distribution $(\chi^2_{d,u})^a$ / 880
7 Confidence limits for the expectation of a Poisson variable (μ) / 881
8 Percentage points of the *F* distribution $(F_{d_1,d_2,p})$ / 882
9 Critical values for the ESD (Extreme Studentized Deviate) outlier statistic ($ESD_{n,1-\alpha}$, $\alpha = .05, .01$) / 884
10 Two-tailed critical values for the Wilcoxon signed-rank test / 884
11 Two-tailed critical values for the Wilcoxon rank-sum test / 885
12 Fisher's *z* transformation / 887
13 Two-tailed upper critical values for the Spearman rank-correlation coefficient (r_s) / 888
14 Critical values for the Kruskal-Wallis test statistic (H) for selected sample sizes for $k = 3$ / 889
15 Critical values for the studentized range statistic q^*, $\alpha = .05$ / 890

Answers to Selected Problems / 891

Flowchart: Methods of Statistical Inference / 895

Index of Data Sets / 901

Index of Statistical Software / 903

Subject Index / 909

Index of Applications / 936

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

PREFACE

This introductory-level biostatistics text is designed for upper-level undergraduate or graduate students interested in medicine or other health-related areas. It requires no previous background in statistics, and its mathematical level assumes only a knowledge of algebra.

Fundamentals of Biostatistics evolved from notes that I have used in a biostatistics course taught to Harvard University undergraduates, Harvard Medical School, and Harvard School of Public Health students over the past 30 years. I wrote this book to help motivate students to master the statistical methods that are most often used in the medical literature. From the student's viewpoint, it is important that the example material used to develop these methods is representative of what actually exists in the literature. Therefore, most of the examples and exercises in this book are based either on actual articles from the medical literature or on actual medical research problems I have encountered during my consulting experience at the Harvard Medical School.

The Approach

Most introductory statistics texts either use a completely nonmathematical, cookbook approach or develop the material in a rigorous, sophisticated mathematical framework. In this book, however, I follow an intermediate course, minimizing the amount of mathematical formulation but giving complete explanations of all important concepts. Every new concept in this book is developed systematically through completely worked-out examples from current medical research problems. In addition, I introduce computer output where appropriate to illustrate these concepts.

I initially wrote this text for the introductory biostatistics course. However, the field has changed dramatically over the past 30 years; because of the increased power of newer statistical packages, we can now perform more sophisticated data analyses than ever before. Therefore, a second goal of this text is to present these new techniques *at an introductory level* so that students can become familiar with them without having to wade through specialized (and, usually, more advanced) statistical texts.

To differentiate these two goals more clearly, I included most of the content for the introductory course in the first 12 chapters. More advanced statistical techniques used in recent epidemiologic studies are covered in Chapter 13, "Design and Analysis Techniques for Epidemiologic Studies," and Chapter 14, "Hypothesis Testing: Person-Time Data."

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Changes in the Eighth Edition

For this edition, I have added three new sections and added new content to three other sections. Features new to this edition include the following:

- The data sets are now available on the book's Companion Website at www.cengage.com/statistics/rosner in an expanded set of formats, including Excel, Minitab®, SPSS, JMP, SAS, Stata, R, and ASCII formats.
- Data and medical research findings in Examples have been updated.
- New or expanded coverage of the followings topics has been added:
  - The Bootstrap (Section 6.11)
  - One-sample inference for the Binomial Distribution (Section 7.9)
  - Permutation Tests (Section 9.6)
  - Sample size estimation for logistic regression (Section 13.9)
  - Estimation of survival curves: The Kaplan-Meier Estimator (Section 14.9)
  - Derivation of selected formulas (Sections 7.12, 8.11, 10.9, 11.14, 12.11, 13.17, 14.15)

The new sections and the expanded sections for this edition have been indicated by an asterisk in the table of contents.

Exercises

This edition contains 1,490 exercises; 171 of these exercises are new. Data and medical research findings in the problems have been updated where appropriate. All problems based on the data sets are included. Problems marked by an asterisk (*) at the end of each chapter have corresponding brief solutions in the answer section at the back of the book. Based on requests from students for more completely solved problems, approximately 600 additional problems and complete solutions are presented in the *Study Guide* available on the Companion Website accompanying this text. In addition, approximately 100 of these problems are included in a Miscellaneous Problems section and are randomly ordered so that they are not tied to a specific chapter in the book. This gives the student additional practice in determining what method to use in what situation. Complete instructor solutions to all exercises are available at the instructor companion website at cengage.com/statistics/rosner.

Computation Method

The method of handling computations is similar to that used in the seventh edition. All intermediate results are carried to full precision (10+ significant digits), even though they are presented with fewer significant digits (usually 2 or 3) in the text. Thus, intermediate results may seem inconsistent with final results in some instances; this, however, is not the case.

Organization

Fundamentals of Biostatistics, Eighth Edition, is organized as follows.

Chapter 1 is an *introductory* chapter that contains an outline of the development of an actual medical study with which I was involved. It provides a unique sense of the role of biostatistics in medical research.

Chapter 2 concerns *descriptive statistics* and presents all the major numeric and graphic tools used for displaying medical data. This chapter is especially important

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

for both consumers and producers of medical literature because much information is actually communicated via descriptive material.

Chapters 3 through 5 discuss *probability*. The basic principles of probability are developed, and the most common probability distributions—such as the binomial and normal distributions—are introduced. These distributions are used extensively in later chapters of the book. The concepts of prior probability and posterior probability are also introduced.

Chapters 6 through 10 cover some of the basic methods of *statistical inference*.

Chapter 6 introduces the concept of drawing random samples from populations. The difficult notion of a sampling distribution is developed and includes an introduction to the most common sampling distributions, such as the *t* and chi-square distributions. The basic methods of *estimation*, including an extensive discussion of confidence intervals, are also presented. *In addition, the bootstrap method for obtaining confidence limits is introduced for the first time.*

Chapters 7 and 8 contain the basic principles of *hypothesis testing*. The most elementary hypothesis tests for normally distributed data, such as the t test, are also fully discussed for one- and two-sample problems.

Chapter 9 covers the basic principles of *nonparametric statistics*. The assumptions of normality are relaxed, and distribution-free analogues are developed for the tests in Chapters 7 and 8. *The technique of permutation testing, which is widely used in genetic studies, is introduced for the first time.*

Chapter 10 contains the basic concepts of *hypothesis testing* as applied to categorical data, including some of the most widely used statistical procedures, such as the chi-square test and Fisher's exact test.

Chapter 11 develops the principles of *regression analysis*. The case of simple linear regression is thoroughly covered, and extensions are provided for the multiple-regression case. Important sections on goodness-of-fit of regression models are also included. Also, rank correlation is introduced, including methods for obtaining confidence intervals for rank correlation.

Chapter 12 introduces the basic principles of the *analysis of variance* (ANOVA). The one-way analysis of variance fixed- and random-effects models are discussed. In addition, two-way ANOVA, the analysis of covariance, and mixed effects models are covered. Finally, we discuss nonparametric approaches to one-way ANOVA. Multiple comparison methods including material on the false discovery rate are also provided.

Chapter 13 discusses methods of design and analysis for *epidemiologic studies*. The most important study designs, including the prospective study, the case-control study, the cross-sectional study, and the cross-over design are introduced. The concept of a confounding variable—that is, a variable related to both the disease and the exposure variable—is introduced, and methods for controlling for confounding, which include the Mantel-Haenszel test and multiple-logistic regression, are discussed in detail. Extensions to logistic regression models, including conditional logistic regression, polytomous logistic regression, and ordinal logistic regression, are discussed. *Methods of estimation of sample size for logistic regression models are provided for the first time.* This discussion is followed by the exploration of topics of current interest in epidemiologic data analysis, including meta-analysis (the combination of results from more than one study); correlated binary data techniques (techniques that can be applied when replicate measures, such as data from multiple teeth from the same person, are available for an individual); measurement error methods (useful when there is substantial measurement error in the exposure data collected); equivalence studies (whose objective it is to establish bioequivalence between two

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

treatment modalities rather than that one treatment is superior to the other); and missing-data methods for how to handle missing data in epidemiologic studies. Longitudinal data analysis and generalized estimating equation (GEE) methods are also briefly discussed.

Chapter 14 introduces methods of analysis for *person-time data*. The methods covered in this chapter include those for incidence-rate data, as well as several methods of survival analysis: the Kaplan-Meier survival curve estimator, the log-rank test, and the proportional-hazards model. Methods for testing the assumptions of the proportional-hazards model have also been included. Parametric survival analysis methods are also discussed.

Throughout the text—particularly in Chapter 13—I discuss the elements of study designs, including the concepts of matching; cohort studies; case-control studies; retrospective studies; prospective studies; and the sensitivity, specificity, and predictive value of screening tests. These designs are presented in the context of actual samples. In addition, Chapters 7, 8, 10, 11, 13, and 14 contain specific sections on sample-size estimation for different statistical situations.

There have been two important organizational changes in the presentation of material in the text. First, the derivation of more complex formulas have either been moved after the statement of an equation or to separate derivation sections at the end of the chapter, to enable students to access the main results in the equations more immediately. Second, there are numerous subsections entitled "Using the Computer to Perform a Specific Test" to more clearly highlight use of the computer to implement many of the methods in the text.

A flowchart of appropriate methods of statistical inference (see pages 895–900) is a handy reference guide to the methods developed in this book. Page references for each major method presented in the text are also provided. In Chapters 7 and 8 and Chapters 10–14, I refer students to this flowchart to give them some perspective on how the methods discussed in a given chapter fit with all the other statistical methods introduced in this book.

In addition, I have provided an index of applications, grouped by medical specialty, summarizing all the examples and problems this book covers.

Finally, we provide for the first time, an index of computer software to more clearly identify the computer commands in specific computer packages that are featured in the text.

Acknowledgments

I am indebted to Debra Sheldon, the late Marie Sheehan, and Harry Taplin for their invaluable help typing the manuscript, to Dale Rinkel for invaluable help in typing problem solutions, and to Marion McPhee for helping to prepare the data sets on the Companion Website. I am also indebted to Roland Matsouaka for updating solutions to problems for this edition, and to Virginia Piaseczny for typing the Index of Applications. In addition, I wish to thank the manuscript reviewers, among them: Shouhao Zhou, Daniela Szatmari-Voicu, Jianying Gu, Raid Amin, Claus Wilke, Glen Johnson, Kara Zografos, and Hui Zhao. I would also like to thank my colleagues Nancy Cook, who was instrumental in helping me develop the part of Section 12.4 on the false-discovery rate, and Robert Glynn, who was invaluable in developing Section 13.16 on missing data and Section 14.11 on testing the assumptions of the proportional-hazards model.

In addition, I wish to thank Spencer Arritt and Jay Campbell, whose input was critical in providing editorial advice and in preparing the manuscript.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

I am also indebted to my colleagues at the Channing Laboratory—most notably, the late Edward Kass, Frank Speizer, Charles Hennekens, the late Frank Polk, Ira Tager, Jerome Klein, James Taylor, Stephen Zinner, Scott Weiss, Frank Sacks, Walter Willett, Alvaro Munoz, Graham Colditz, and Susan Hankinson—and to my other colleagues at the Harvard Medical School, most notably, the late Frederick Mosteller, Eliot Berson, Robert Ackerman, Mark Abelson, Arthur Garvey, Leo Chylack, Eugene Braunwald, and Arthur Dempster, who inspired me to write this book. I also wish to express appreciation to John Hopper and Philip Landrigan for providing the data for our case studies.

Finally, I would like to acknowledge Leslie Miller, Andrea Wagner, Ithamar Jotkowitz, Loren Fishman, and Frank Santopietro, without whose clinical help the current edition of this book would not have been possible.

Bernard Rosner

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

ABOUT THE AUTHOR

Bernard Rosner is Professor of Medicine (Biostatistics) at Harvard Medical School and Professor of Biostatistics in the Harvard School of Public Health. He received a B.A. in Mathematics from Columbia University in 1967, an M.S. in Statistics from Stanford University in 1968, and a Ph.D. in Statistics from Harvard University in 1971.

He has more than 30 years of biostatistical consulting experience with other investigators at the Harvard Medical School. Special areas of interest include cardiovascular disease, hypertension, breast cancer, and ophthalmology. Many of the examples and exercises used in the text reflect data collected from actual studies in conjunction with his consulting experience. In addition, he has developed new biostatistical methods, mainly in the areas of longitudinal data analysis, analysis of clustered data (such as data collected in families or from paired organ systems in the same person), measurement error methods, and outlier detection methods. You will see some of these methods introduced in this book at an elementary level. He was married in 1972 to his wife, Cynthia, and they have three children, Sarah, David, and Laura, each of whom has contributed examples to this book.

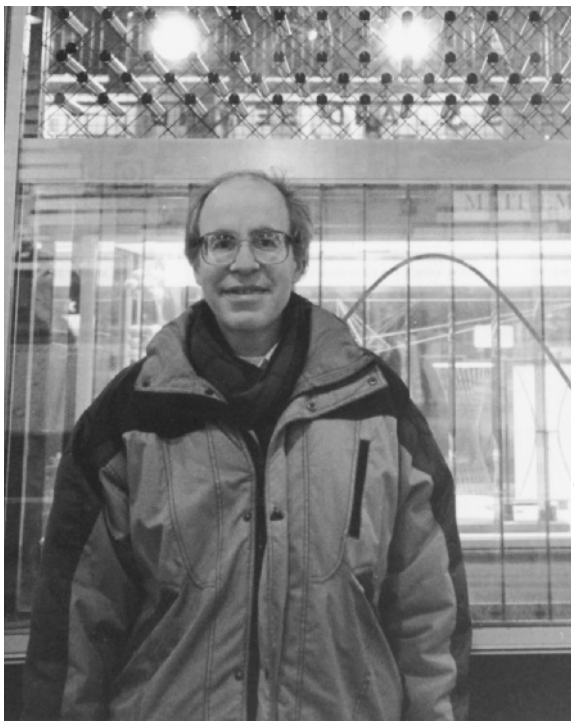


Photo courtesy of the Museum of Science, Boston

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

1 General Overview

Statistics is the science whereby inferences are made about specific random phenomena on the basis of relatively limited sample material. The field of statistics has two main areas: mathematical statistics and applied statistics. **Mathematical** statistics concerns the development of new methods of statistical inference and requires detailed knowledge of abstract mathematics for its implementation. Applied statistics involves applying the methods of mathematical statistics to specific subject areas, such as economics, psychology, and public health. **Biostatistics** is the branch of applied statistics that applies statistical methods to medical and biological problems. Of course, these areas of statistics overlap somewhat. For example, in some instances, given a certain biostatistical application, standard methods do not apply and must be modified. In this circumstance, biostatisticians are involved in developing new methods.

A good way to learn about biostatistics and its role in the research process is to follow the flow of a research study from its inception at the planning stage to its completion, which usually occurs when a manuscript reporting the results of the study is published. As an example, I will describe one such study in which I participated.

A friend called one morning and in the course of our conversation mentioned that he had recently used a new, automated blood-pressure measuring device of the type seen in many banks, hotels, and department stores. The machine had measured his average diastolic blood pressure on several occasions as 115 mm Hg; the highest reading was 130 mm Hg. I was very worried, because if these readings were accurate, my friend might be in imminent danger of having a stroke or developing some other serious cardiovascular disease. I referred him to a clinical colleague of mine who, using a standard blood-pressure cuff, measured my friend's diastolic blood pressure as 90 mm Hg. The contrast in readings aroused my interest, and I began to jot down readings from the digital display every time I passed the machine at my local bank. I got the distinct impression that a large percentage of the reported readings were in the hypertensive range. Although one would expect hypertensive individuals to be more likely to use such a machine, I still believed that blood-pressure readings from the machine might not be comparable with those obtained using standard methods

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

of blood-pressure measurement. I spoke with Dr. B. Frank Polk, a physician at Harvard Medical School with an interest in hypertension, about my suspicion and succeeded in interesting him in a small-scale evaluation of such machines. We decided to send a human observer, who was well trained in blood-pressure measurement techniques, to several of these machines. He would offer to pay participants 50¢ for the cost of using the machine if they would agree to fill out a short questionnaire and have their blood pressure measured by both a human observer and the machine.

At this stage we had to make several important decisions, each of which proved vital to the success of the study. These decisions were based on the following questions:

(1) How many machines should we test?

(2) How many participants should we test at each machine?

(3) In what order should we take the measurements? That is, should the human observer or the machine take the first measurement? Under ideal circumstances we would have taken both the human and machine readings simultaneously, but this was logistically impossible.

(4) What data should we collect on the questionnaire that might influence the comparison between methods?

(5) How should we record the data to facilitate computerization later?

(6) How should we check the accuracy of the computerized data?

We resolved these problems as follows:

(1) and (2) Because we were not sure whether all blood-pressure machines were comparable in quality, we decided to test four of them. However, we wanted to sample enough subjects from each machine so as to obtain an accurate comparison of the standard and automated methods for each machine. We tried to predict how large a discrepancy there might be between the two methods. Using the methods of sample-size determination discussed in this book, we calculated that we would need 100 participants at each site to make an accurate comparison.

(3) We then had to decide in what order to take the measurements for each person. According to some reports, one problem with obtaining repeated blood-pressure measurements is that people tense up during the initial measurement, yielding higher blood-pressure readings. Thus we would not always want to use either the automated or manual method first, because the effect of the method would get confused with the order-of-measurement effect. A conventional technique we used here was to **randomize** the order in which the measurements were taken, so that for any person it was equally likely that the machine or the human observer would take the first measurement. This random pattern could be implemented by flipping a coin or, more likely, by using a table of **random numbers** similar to Table 4 of the Appendix.

(4) We believed that the major extraneous factor that might influence the results would be body size (we might have more difficulty getting accurate readings from people with fatter arms than from those with leaner arms). We also wanted to get some idea of the type of people who use these machines. Thus we asked questions about age, gender, and previous hypertension history.

(5) To record the data, we developed a coding form that could be filled out on site and from which data could be easily entered into a computer for subsequent analysis. Each person in the study was assigned a unique identification (ID) number by which the computer could identify that person. The data on the coding forms were then keyed and verified. That is, the same form was entered twice and the two

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

records compared to make sure they were the same. If the records did not match, the form was re-entered.

(6) Checking each item on each form was impossible because of the large amount of data involved. Instead, after data entry we ran some editing programs to ensure that the data were accurate. These programs checked that the values for individual variables fell within specified ranges and printed out aberrant values for manual checking. For example, we checked that all blood-pressure readings were at least 50 mm Hg and no higher than 300 mm Hg, and we printed out all readings that fell outside this range. We also ran programs to detect outliers as discussed later in this book.

After completing the data-collection, data-entry, and data-editing phases, we were ready to look at the results of the study. The first step in this process is to get an impression of the data by summarizing the information in the form of several descriptive statistics. This descriptive material can be numeric or graphic. If numeric, it can be in the form of a few summary statistics, which can be presented in tabular form or, alternatively, in the form of a **frequency distribution**, which lists each value in the data and how frequently it occurs. If graphic, the data are summarized pictorially and can be presented in one or more figures. The appropriate type of descriptive material to use varies with the type of distribution considered. If the distribution is **continuous**—that is, if there is essentially an infinite number of possible values, as would be the case for blood pressure—then means and standard deviations may be the appropriate descriptive statistics. However, if the distribution is **discrete**—that is, if there are only a few possible values, as would be the case for gender—then percentages of people taking on each value are the appropriate descriptive measure. In some cases both types of descriptive statistics are used for continuous distributions by condensing the range of possible values into a few groups and giving the percentage of people that fall into each group (e.g., the percentages of people who have blood pressures between 120 and 129 mm Hg, between 130 and 139 mm Hg, and so on).

In this study we decided first to look at mean blood pressure for each method at each of the four sites. Table 1.1 summarizes this information [1].

You may notice from this table that we did not obtain meaningful data from all 100 people interviewed at each site. This was because we could not obtain valid readings from the machine for many of the people. This problem of missing data is very common in biostatistics and should be anticipated at the planning stage when deciding on sample size (which was not done in this study).

TABLE 1.1 Mean blood pressures and differences between machine and human readings at four locations

		Systolic blood pressure (mm Hg)					
		Machine		Human		Difference	
Location	Number of people	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
A	98	142.5	21.0	142.0	18.1	0.5	11.2
B	84	134.1	22.5	133.6	23.2	0.5	12.1
C	98	147.9	20.3	133.9	18.3	14.0	11.7
D	62	135.4	16.7	128.5	19.0	6.9	13.6

Source: Based on the American Heart Association, Inc.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Our next step in the study was to determine whether the apparent differences in blood pressure between machine and human measurements at two of the locations (C, D) were "real" in some sense or were "due to chance." This type of question falls into the area of **inferential statistics**. We realized that although there was a difference of 14 mm Hg in mean systolic blood pressure between the two methods for the 98 people we interviewed at location C, this difference might not hold up if we interviewed 98 other people at this location at a different time, and we wanted to have some idea as to the **error in the estimate** of 14 mm Hg. In statistical jargon, this group of 98 people represents a **sample** from the **population** of all people who might use that machine. We were interested in the population, and we wanted to use the sample to help us learn something about the population. In particular, we wanted to know how different the **estimated mean difference** of 14 mm Hg in our sample was likely to be from the **true mean difference** in the population of all people who might use this machine. More specifically, we wanted to know if it was still possible that there was no underlying difference between the two methods and that our results were due to chance. The 14-mm Hg difference in our group of 98 people is referred to as an **estimate** of the true mean difference (d) in the population. The problem of inferring characteristics of a population from a sample is the central concern of statistical inference and is a major topic in this text. To accomplish this aim, we needed to develop a **probability model**, which would tell us how likely it is that we would obtain a 14-mm Hg difference between the two methods in a sample of 98 people if there were no real difference between the two methods over the entire population of users of the machine. If this probability were small enough, then we would begin to believe a real difference existed between the two methods. In this particular case, using a probability model based on the t distribution, we concluded this probability was less than 1 in 1000 for each of the machines at locations C and D. This probability was sufficiently small for us to conclude there was a real difference between the automatic and manual methods of measuring blood pressure for two of the four machines tested.

We used a statistical package to perform the preceding data analyses. A package is a collection of statistical programs that describe data and perform various statistical tests on the data. Currently the most widely used statistical packages are SAS, SPSS, Stata, R, MINITAB, and Excel.

The final step in this study, after completing the data analysis, was to compile the results in a publishable manuscript. Inevitably, because of space considerations, we weeded out much of the material developed during the data-analysis phase and presented only the essential items for publication.

This review of our blood-pressure study should give you some idea of what medical research is about and the role of biostatistics in this process. The material in this text parallels the description of the data-analysis phase of the study. Chapter 2 summarizes different types of descriptive statistics. Chapters 3 through 5 present some basic principles of probability and various probability models for use in later discussions of inferential statistics. Chapters 6 through 14 discuss the major topics of inferential statistics as used in biomedical practice. Issues of study design or data collection are brought up only as they relate to other topics discussed in the text.

REFERENCE

[1] Polk, B. F., Rosner, B., Feudo, R., & Vandenburgh, M. (1980). An evaluation of the Vita-Stat automatic blood pressure measuring device. *Hypertension, 2*(2), 221–227.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

Descriptive Statistics

2.1 INTRODUCTION

The first step in looking at data is to describe the data at hand in some concise way. In smaller studies this step can be accomplished by listing each data point. In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

EXAMPLE 2.1 **Cancer, Nutrition** Some investigators have proposed that consumption of vitamin A prevents cancer. To test this theory, a dietary questionnaire might be used to collect data on vitamin-A consumption among 200 hospitalized cancer patients (cases) and 200 controls. The controls would be matched with regard to age and gender with the cancer cases and would be in the hospital at the same time for an unrelated disease. What should be done with these data after they are collected?

Before any formal attempt to answer this question can be made, the vitamin-A consumption among cases and controls must be described. Consider Figure 2.1. The **bar graphs** show that the controls consume more vitamin A than the cases do, particularly at consumption levels exceeding the Recommended Daily Allowance (RDA).

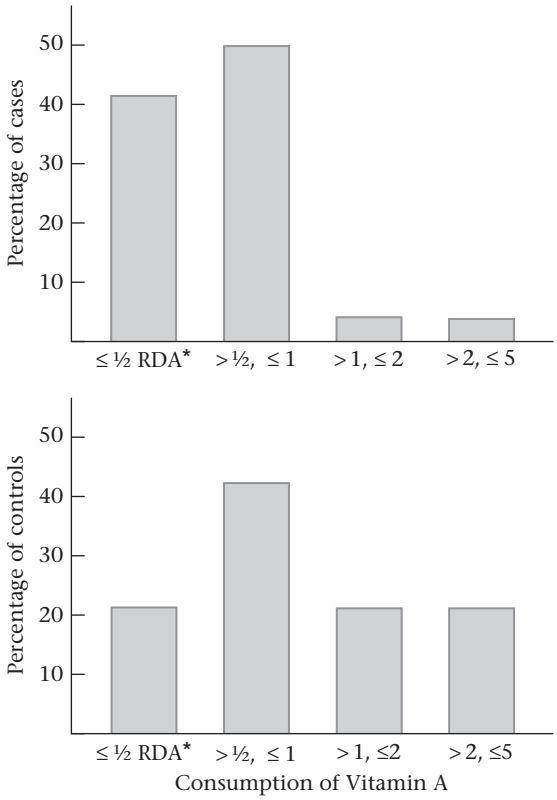
EXAMPLE 2.2 **Pulmonary Disease** Medical researchers have often suspected that passive smokers—people who themselves do not smoke but who live or work in an environment in which others smoke—might have impaired pulmonary function as a result. In 1980 a research group in San Diego published results indicating that passive smokers did indeed have significantly lower pulmonary function than comparable nonsmokers who did not work in smoky environments [1]. As supporting evidence, the authors measured the carbon-monoxide (CO) concentrations in the working environments of passive smokers and of nonsmokers whose companies did not permit smoking in the workplace to see if the relative CO concentration changed over the course of the day. These results are displayed as a **scatter plot** in Figure 2.2.

Figure 2.2 clearly shows that the CO concentrations in the two working environments are about the same early in the day but diverge widely in the middle of the day and then converge again after the workday is over at 7 P.M.

Graphic displays illustrate the important role of descriptive statistics, which is to quickly display data to give the researcher a clue as to the principal trends in the data and suggest hints as to where a more detailed look at the data, using the

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

FIGURE 2.1 Daily vitamin-A consumption among cancer cases and controls



*RDA = Recommended Daily Allowance.

methods of inferential statistics, might be worthwhile. Descriptive statistics are also crucially important in conveying the final results of studies in written publications. Unless it is one of their primary interests, most readers will not have time to critically evaluate the work of others but will be influenced mainly by the descriptive statistics presented.

What makes a good graphic or numeric display? The main guideline is that the material should be as self-contained as possible and should be understandable without reading the text. These attributes require clear labeling. The captions, units, and axes on graphs should be clearly labeled, and the statistical terms used in tables and figures should be well defined. The quantity of material presented is equally important. If bar graphs are constructed, then care must be taken to display neither too many nor too few groups. The same is true of tabular material.

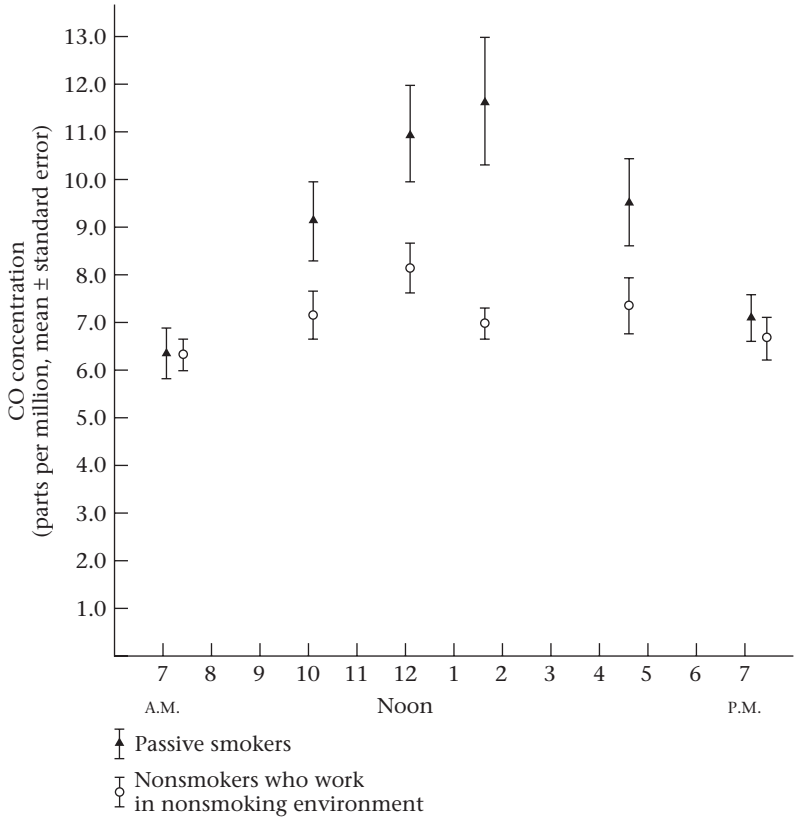
Many methods are available for summarizing data in both numeric and graphic form. In this chapter these methods are summarized and their strengths and weaknesses noted.

2.2 MEASURES OF LOCATION

The basic problem of statistics can be stated as follows: Consider a sample of data $x_1, \ldots, x_n$, where x_1 corresponds to the first sample point and x_n corresponds to the

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

FIGURE 2.2 **Mean carbon-monoxide concentration (± standard error) by time of day as measured in the working environment of passive smokers and in nonsmokers who work in a nonsmoking environment**



Source: Based on *The New England Journal of Medicine, 302,* 720–723, 1980.

*n*th sample point. Presuming that the sample is drawn from some population *P*, what inferences or conclusions can be made about *P* from the sample?

Before this question can be answered, the data must be summarized as succinctly as possible; this is because the number of sample points is often large, and it is easy to lose track of the overall picture when looking at individual sample points. One type of measure useful for summarizing data defines the center, or middle, of the sample. This type of measure is a **measure of location**.

The Arithmetic Mean

How to define the middle of a sample may seem obvious, but the more you think about it, the less obvious it becomes. Suppose the sample consists of the birthweights of all live-born infants born at a private hospital in San Diego, California, during a 1-week period. This sample is shown in Table 2.1.

One measure of location for this sample is the arithmetic mean (colloquially called the *average*). The arithmetic mean (or mean or sample mean) is usually denoted by $\bar{x}$.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

TABLE 2.1 **Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period**

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

DEFINITION 2.1 The **arithmetic mean** is the sum of all the observations divided by the number of observations. It is written in statistical terms as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sign Σ (sigma) in Definition 2.1 is a summation sign. The expression

$$\sum_{i=1}^{n} x_i$$

is simply a short way of writing the quantity $(x_1 + x_2 + \cdots + x_n)$.

If a and b are integers, where $a < b$, then

$$\sum_{i=a}^{b} x_i$$

means $x_a + x_{a+1} + \cdots + x_b$.

If $a = b$, then $\sum_{i=a}^{b} x_i = x_a$. One property of summation signs is that if each term in the summation is a multiple of the same constant c, then c can be factored out from the summation; that is,

$$\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$$

EXAMPLE 2.3

If $x_1 = 2 \quad x_2 = 5 \quad x_3 = -4$

find $\sum_{i=1}^{3} x_i \quad \sum_{i=2}^{3} x_i \quad \sum_{i=1}^{3} x_i^2 \quad \sum_{i=1}^{3} 2x_i$

Solution:

$$\sum_{i=1}^{3} x_i = 2 + 5 - 4 = 3 \quad \sum_{i=2}^{3} x_i = 5 - 4 = 1$$

$$\sum_{i=1}^{3} x_i^2 = 4 + 25 + 16 = 45 \quad \sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 6$$

It is important to become familiar with summation signs because they are used extensively throughout the remainder of the text.

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

EXAMPLE 2.4 What is the arithmetic mean for the sample of birthweights in Table 2.1?

$$\bar{x} = (3265 + 3260 + \cdots + 2834)/20 = 3166.9 \text{ g}$$

The arithmetic mean is, in general, a very natural measure of location. One of its main limitations, however, is that it is oversensitive to extreme values. In this instance, it may not be representative of the location of the great majority of sample points. For example, if the first infant in Table 2.1 happened to be a premature infant weighing 500 g rather than 3265 g, then the arithmetic mean of the sample would fall to 3028.7 g. In this instance, 7 of the birthweights would be lower than the arithmetic mean, and 13 would be higher than the arithmetic mean. It is possible in extreme cases for all but one of the sample points to be on one side of the arithmetic mean. In these types of samples, the arithmetic mean is a poor measure of central location because it does not reflect the center of the sample. Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

The Median

An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the **median** or, more precisely, the **sample median**.

Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

DEFINITION 2.2 The **sample median** is

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

The rationale for these definitions is to ensure an equal number of sample points on both sides of the sample median. The median is defined differently when n is even and odd because it is impossible to achieve this goal with one uniform definition. Samples with an odd sample size have a unique central point; for example, for samples of size 7, the fourth largest point is the central point in the sense that 3 points are smaller than it and 3 points are larger. Samples with an even sample size have no unique central point, and the middle two values must be averaged. Thus, for samples of size 8 the fourth and fifth largest points would be averaged to obtain the median, because neither is the central point.

EXAMPLE 2.5 Compute the sample median for the sample in Table 2.1.

Solution: First, arrange the sample in ascending order:

2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Because n is even,

Sample median = average of the 10th and 11th largest observations
= (3245 + 3248)/2 = 3246.5 g

Copyright 2016 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s).
Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.