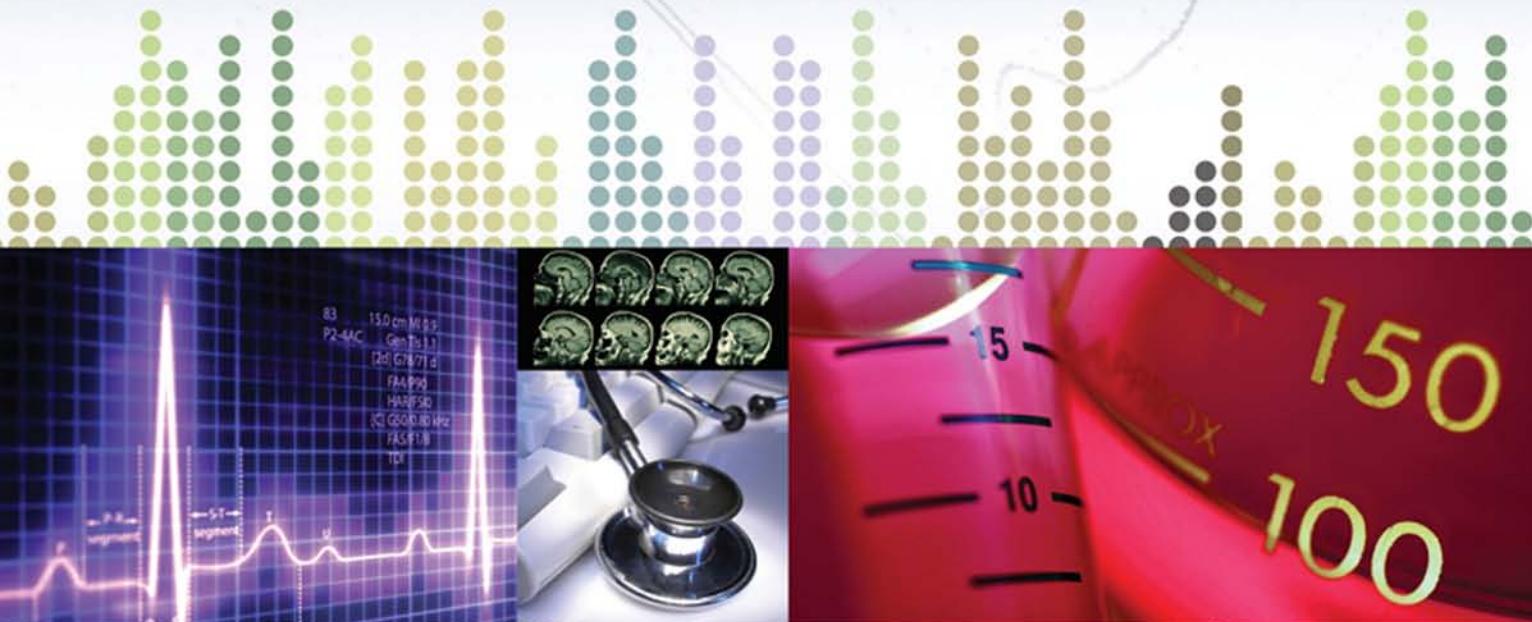


# FUNDAMENTALS OF BIOSTATISTICS

SEVENTH EDITION



BERNARD ROSNER

# **Fundamentals of Biostatistics**



# Fundamentals of Biostatistics

SEVENTH EDITION

**Bernard Rosner**

*Harvard University*



---

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

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](http://www.cengage.com/highered) to search by ISBN#, author, title, or keyword for materials in your areas of interest.

**Fundamentals of Biostatistics**  
**Seventh Edition**  
**Rosner**

Senior Sponsoring Editor: Molly Taylor  
Associate Editor: Daniel Seibert  
Editorial Assistant: Shaylin Walsh  
Marketing Manager: Ashley Pickering  
Marketing Coordinator: Erica O'Connell  
Marketing Communications Manager:  
Mary Anne Payumo  
Content Project Manager: Jessica Rasile  
Associate Media Editor: Andrew Coppola  
Art Director: Linda Helcher  
Senior Print Buyer: Diane Gibbons  
Senior Rights Specialist: Katie Huha  
Production Service/Composition: Cadmus  
Cover Design: Pier One Design  
Cover Images: ©Egorych/istockphoto,  
©enot-poloskun/istockphoto,  
©demio/istockphoto,  
©bcollet/istockphoto

© 2011, 2006 Brooks/Cole, Cengage Learning

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](http://www.cengage.com/permissions).  
Further permissions questions can be emailed to  
[permissionrequest@cengage.com](mailto:permissionrequest@cengage.com).

Library of Congress Control Number: 2010922638

ISBN-13: 978-0-538-73349-6

ISBN-10: 0-538-73349-7

**Brooks/Cole**  
20 Channel Center Street  
Boston, MA 02210  
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil and Japan. Locate your local office at [international.cengage.com/region](http://international.cengage.com/region)

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

For your course and learning solutions, visit [www.cengage.com](http://www.cengage.com).  
Purchase any of our products at your local college store or at our preferred online store [www.cengagebrain.com](http://www.cengagebrain.com).

Printed in Canada  
1 2 3 4 5 6 7 14 13 12 11 10

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



# 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 / 13
- 2.4 Measures of Spread / 15
- 2.5 Some Properties of the Variance and Standard Deviation / 18
- 2.6 The Coefficient of Variation / 20
- 2.7 Grouped Data / 22
- 2.8 Graphic Methods / 24

- 2.9 Case Study 1: Effects of Lead Exposure on Neurological and Psychological Function in Children / 29
- 2.10 Case Study 2: Effects of Tobacco Use on Bone-Mineral Density in Middle-Aged Women / 30
- 2.11 Obtaining Descriptive Statistics on the Computer / 31
- 2.12 Summary / 31

## PROBLEMS / 33

\*The new sections and the expanded sections for this edition are indicated by an asterisk.

**CHAPTER 3****Probability / 38**

- |  |   |
|--|---|
| 3.1 Introduction / 38<br>3.2 Definition of Probability / 39<br>3.3 Some Useful Probabilistic Notation / 40<br>3.4 The Multiplication Law of Probability / 42<br>3.5 The Addition Law of Probability / 44<br>3.6 Conditional Probability / 46 | 3.7 Bayes' Rule and Screening Tests / 51<br>3.8 Bayesian Inference / 56<br>3.9 ROC Curves / 57<br>3.10 Prevalence and Incidence / 59<br>3.11 Summary / 60 |
|--|---|

**PROBLEMS / 60****CHAPTER 4****Discrete Probability Distributions / 71**

- |  |   |
|--|---|
| 4.1 Introduction / 71<br>4.2 Random Variables / 72<br>4.3 The Probability-Mass Function for a Discrete Random Variable / 73<br>4.4 The Expected Value of a Discrete Random Variable / 75<br>4.5 The Variance of a Discrete Random Variable / 76<br>4.6 The Cumulative-Distribution Function of a Discrete Random Variable / 78<br>4.7 Permutations and Combinations / 79<br>4.8 The Binomial Distribution / 83 | 4.9 Expected Value and Variance of the Binomial Distribution / 88<br>4.10 The Poisson Distribution / 90<br>4.11 Computation of Poisson Probabilities / 93<br>4.12 Expected Value and Variance of the Poisson Distribution / 95<br>4.13 Poisson Approximation to the Binomial Distribution / 96<br>4.14 Summary / 99 |
|--|---|

**PROBLEMS / 99****CHAPTER 5****Continuous Probability Distributions / 108**

- |   |  |
|---|--|
| 5.1 Introduction / 108<br>5.2 General Concepts / 108<br>5.3 The Normal Distribution / 111<br>5.4 Properties of the Standard Normal Distribution / 114<br>5.5 Conversion from an $N(\mu, \sigma^2)$ Distribution to an $N(0, 1)$ Distribution / 120<br>5.6 Linear Combinations of Random Variables / 124 | 5.7 Normal Approximation to the Binomial Distribution / 129<br>5.8 Normal Approximation to the Poisson Distribution / 135<br>5.9 Summary / 137 |
|---|--|

**PROBLEMS / 138**

**CHAPTER 6****Estimation / 149**

- |   |   |
|---|---|
| 6.1 Introduction / 149<br>6.2 The Relationship Between Population and Sample / 150<br>6.3 Random-Number Tables / 152<br>6.4 Randomized Clinical Trials / 156<br>6.5 Estimation of the Mean of a Distribution / 160<br>6.6 Case Study: Effects of Tobacco Use on Bone-Mineral Density (BMD) in Middle-Aged Women / 175 | 6.7 Estimation of the Variance of a Distribution / 176<br>6.8 Estimation for the Binomial Distribution / 181<br>6.9 Estimation for the Poisson Distribution / 189<br>6.10 One-Sided CIs / 193<br>6.11 Summary / 195 |
|---|---|
- PROBLEMS / 196**

**CHAPTER 7****Hypothesis Testing: One-Sample Inference / 204**

- |  |   |
|--|---|
| 7.1 Introduction / 204<br>7.2 General Concepts / 204<br>7.3 One-Sample Test for the Mean of a Normal Distribution: One-Sided Alternatives / 207<br>7.4 One-Sample Test for the Mean of a Normal Distribution: Two-Sided Alternatives / 215<br>7.5 The Power of a Test / 221<br>7.6 Sample-Size Determination / 228<br>7.7 The Relationship Between Hypothesis Testing and Confidence Intervals / 235<br>7.8 Bayesian Inference / 237 | 7.9 One-Sample $\chi^2$ Test for the Variance of a Normal Distribution / 241<br>7.10 One-Sample Inference for the Binomial Distribution / 244<br>7.11 One-Sample Inference for the Poisson Distribution / 251<br>7.12 Case Study: Effects of Tobacco Use on Bone-Mineral Density in Middle-Aged Women / 256<br>7.13 Summary / 257 |
|--|---|
- PROBLEMS / 259**

**CHAPTER 8****Hypothesis Testing: Two-Sample Inference / 269**

- |  |   |
|--|---|
| 8.1 Introduction / 269<br>8.2 The Paired $t$ Test / 271<br>8.3 Interval Estimation for the Comparison of Means from Two Paired Samples / 275<br>8.4 Two-Sample $t$ Test for Independent Samples with Equal Variances / 276 | 8.5 Interval Estimation for the Comparison of Means from Two Independent Samples (Equal Variance Case) / 280<br>8.6 Testing for the Equality of Two Variances / 281<br>8.7 Two-Sample $t$ Test for Independent Samples with Unequal Variances / 287 |
|--|---|

- 8.8 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 293
  - 8.9 The Treatment of Outliers / 295
  - 8.10 Estimation of Sample Size and Power for Comparing Two Means / 301
  - 8.11 Sample-Size Estimation for Longitudinal Studies / 304
  - 8.12 Summary / 307
- PROBLEMS / 309**

## CHAPTER 9

### Nonparametric Methods / 327

- 9.1 Introduction / 327
- 9.2 The Sign Test / 329
- 9.3 The Wilcoxon Signed-Rank Test / 333
- 9.4 The Wilcoxon Rank-Sum Test / 339
- 9.5 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 344
- 9.6 Summary / 344

**PROBLEMS / 346**

## CHAPTER 10

### Hypothesis Testing: Categorical Data / 352

- 10.1 Introduction / 352
- 10.2 Two-Sample Test for Binomial Proportions / 353
- 10.3 Fisher's Exact Test / 367
- 10.4 Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test) / 373
- 10.5 Estimation of Sample Size and Power for Comparing Two Binomial Proportions / 381
- 10.6  $R \times C$  Contingency Tables / 390
- 10.7 Chi-Square Goodness-of-Fit Test / 401
- 10.8 The Kappa Statistic / 404
- 10.9 Summary / 408

**PROBLEMS / 409**

## CHAPTER 11

### Regression and Correlation Methods / 427

- 11.1 Introduction / 427
- 11.2 General Concepts / 428
- 11.3 Fitting Regression Lines—The Method of Least Squares / 431
- 11.4 Inferences About Parameters from Regression Lines / 435
- 11.5 Interval Estimation for Linear Regression / 443
- 11.6 Assessing the Goodness of Fit of Regression Lines / 448
- 11.7 The Correlation Coefficient / 452
- 11.8 Statistical Inference for Correlation Coefficients / 455
- 11.9 Multiple Regression / 468
- 11.10 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 484
- 11.11 Partial and Multiple Correlation / 491
- 11.12 Rank Correlation / 494
- \*11.13 Interval Estimation for Rank Correlation Coefficients / 499
- 11.14 Summary / 504

**PROBLEMS / 504**

**CHAPTER 12****Multisample Inference / 516**

- |   |  |
|---|--|
| 12.1 Introduction to the One-Way Analysis of Variance / 516<br>12.2 One-Way ANOVA—Fixed-Effects Model / 516<br>12.3 Hypothesis Testing in One-Way ANOVA—Fixed-Effects Model / 518<br>12.4 Comparisons of Specific Groups in One-Way ANOVA / 522<br>12.5 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children / 538 | 12.6 Two-Way ANOVA / 548<br>12.7 The Kruskal-Wallis Test / 555<br>12.8 One-Way ANOVA—The Random-Effects Model / 562<br>12.9 The Intraclass Correlation Coefficient / 568<br>*12.10 Mixed Models / 572<br>12.11 Summary / 576 |
|---|--|
- PROBLEMS / 577**

**CHAPTER 13****Design and Analysis Techniques for Epidemiologic Studies / 588**

- |   |   |
|---|---|
| 13.1 Introduction / 588<br>13.2 Study Design / 588<br>13.3 Measures of Effect for Categorical Data / 591<br>*13.4 Attributable Risk / 601<br>13.5 Confounding and Standardization / 607<br>13.6 Methods of Inference for Stratified Categorical Data—The Mantel-Haenszel Test / 612<br>13.7 Power and Sample-Size Estimation for Stratified Categorical Data / 625<br>13.8 Multiple Logistic Regression / 628 | *13.9 Extensions to Logistic Regression / 649<br>13.10 Meta-Analysis / 658<br>13.11 Equivalence Studies / 663<br>13.12 The Cross-Over Design / 666<br>*13.13 Clustered Binary Data / 674<br>*13.14 Longitudinal Data Analysis / 687<br>13.15 Measurement-Error Methods / 696<br>13.16 Missing Data / 706<br>13.17 Summary / 711 |
|---|---|
- PROBLEMS / 713**

**CHAPTER 14****Hypothesis Testing: Person-Time Data / 725**

- |   |   |
|---|---|
| 14.1 Measure of Effect for Person-Time Data / 725<br>14.2 One-Sample Inference for Incidence-Rate Data / 727<br>14.3 Two-Sample Inference for Incidence-Rate Data / 730<br>14.4 Power and Sample-Size Estimation for Person-Time Data / 738<br>14.5 Inference for Stratified Person-Time Data / 742 | 14.6 Power and Sample-Size Estimation for Stratified Person-Time Data / 750<br>14.7 Testing for Trend: Incidence-Rate Data / 755<br>14.8 Introduction to Survival Analysis / 758<br>14.9 Estimation of Survival Curves: The Kaplan-Meier Estimator / 760<br>14.10 The Log-Rank Test / 767<br>14.11 The Proportional-Hazards Model / 774 |
|---|---|

14.12 Power and Sample-Size Estimation under the Proportional-Hazards Model / 783	*14.14 Parametric Regression Models for Survival Data / 795
*14.13 Parametric Survival Analysis / 787	14.15 Summary / 802

## PROBLEMS / 802

**APPENDIX****Tables / 811**

- 1** Exact Binomial Probabilities  $Pr(X=k) = \binom{n}{k} p^k q^{n-k}$  / 811
- 2** Exact Poisson Probabilities  $Pr(X=k) = \frac{e^{-\mu} \mu^k}{k!}$  / 815
- 3** The Normal Distribution / 818
- 4** Table of 1000 Random Digits / 822
- 5** Percentage Points of the  $t$  Distribution ( $t_{d,u}$ ) / 823
- 6** Percentage Points of the Chi-Square Distribution ( $\chi^2_{d,u}$ ) / 824
- 7a** Exact Two-Sided 100%  $\times (1 - \alpha)$  Confidence Limits for Binomial Proportions ( $\alpha = .05$ ) / 825
- 7b** Exact Two-Sided 100%  $\times (1 - \alpha)$  Confidence Limits for Binomial Proportions ( $\alpha = .01$ ) / 826
- 8** Confidence Limits for the Expectation of a Poisson Variable ( $\mu$ ) / 827
- 9** Percentage Points of the  $F$  Distribution ( $F_{d_1, d_2, p}$ ) / 828
- 10** Critical Values for the ESD (Extreme Studentized Deviate) Outlier Statistic ( $ESD_{n, 1-\alpha}, \alpha = .05, .01$ ) / 830
- 11** Two-Tailed Critical Values for the Wilcoxon Signed-Rank Test / 830
- 12** Two-Tailed Critical Values for the Wilcoxon Rank-Sum Test / 831
- 13** Fisher's  $z$  Transformation / 833
- 14** Two-Tailed Upper Critical Values for the Spearman Rank-Correlation Coefficient ( $r_s$ ) / 834
- 15** Critical Values for the Kruskal-Wallis Test Statistic ( $H$ ) for Selected Sample Sizes for  $k = 3$  / 835
- 16** Critical Values for the Studentized Range Statistic  $q^*$ ,  $\alpha = .05$  / 836

**Answers to Selected Problems / 837****FLOWCHART: Methods of Statistical Inference / 841****Index of Data Sets / 847****Index / 849**

# 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 and Harvard Medical School 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 the 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 rapidly over the past 10 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."

## Changes in the Seventh Edition

For this edition, I have added seven new sections and added new content to one other section. 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](http://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 following topics:
  - Interval estimates for rank correlation coefficients (Section 11.13)
  - Mixed effect models (Section 12.10)
  - Attributable risk (Section 13.4)
  - Extensions to logistic regression (Section 13.9)
  - Regression models for clustered binary data (Section 13.13)
  - Longitudinal data analysis (Section 13.14)
  - Parametric survival analysis (Section 14.13)
  - Parametric regression models for survival data (Section 14.14)

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 1438 exercises; 244 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 in secure online format through Cengage's *Solution Builder* service. Adopting instructors can sign up for access at [www.cengage.com/solutionbuilder](http://www.cengage.com/solutionbuilder).

## Computation Method

The method of handling computations is similar to that used in the sixth 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*, Seventh 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

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.

**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. The fundamentals of Bayesian inference are explored.

**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.

**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. Interval estimates for rank correlation coefficients are covered for the first time. Methods for comparing correlation coefficients from dependent samples are also included.

**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. A section of mixed models is also included for the first time.

**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 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 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 covered for the first time.

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.

A flowchart of appropriate methods of statistical inference (see pages 841–846) 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 7–8 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.

## 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 Brian Claggett for updating solutions to problems for this edition, and to Daad Abraham for typing the Index of Applications. In addition, I wish to thank the manuscript reviewers, among them: Emilia Bagiella, Columbia University; Ron Brookmeyer, Johns Hopkins University; Mark van der Laan, University of California, Berkeley; and John Wilson, University of Pittsburgh. 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 instrumental 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 Molly Taylor, Daniel Seibert, Shaylin Walsh, and Laura Wheel, who were instrumental in providing editorial advice and in preparing the manuscript.

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 acknowledge John Hopper and Philip Landrigan for providing the data for our case studies.

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

Bernard Rosner

## 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 has three children, Sarah, David, and Laura, each of whom has contributed examples for this book.

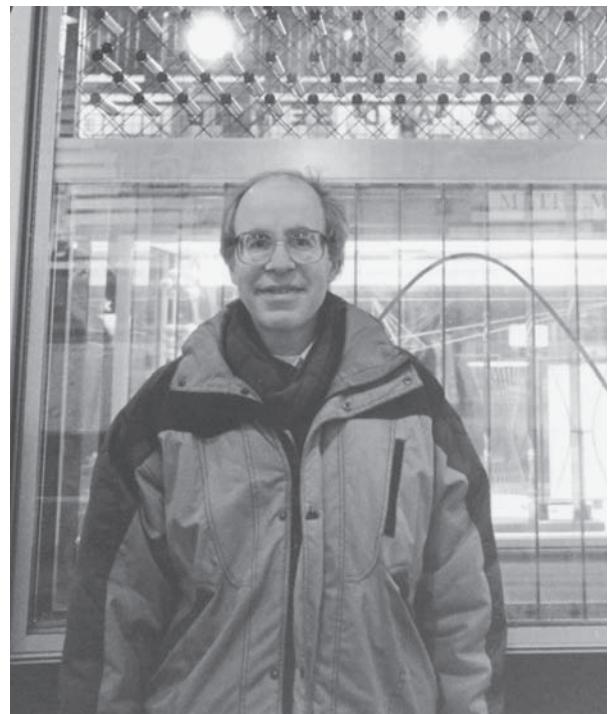


Photo courtesy of the Museum of Science, Boston



## 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 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 logically 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 during subsequent measurements. 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, sex, 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 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.

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 are 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 sex—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).

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

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

Location	Number of people	Systolic blood pressure (mm Hg)					
		Machine		Human		Difference	
		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: By permission of the American Heart Association, Inc.

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 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, 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., & Vandenburg, M. (1980). An evaluation of the Vita-Stat automatic blood pressure measuring device. *Hypertension*, 2(2), 221–227.

# 2

## 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 sex 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?

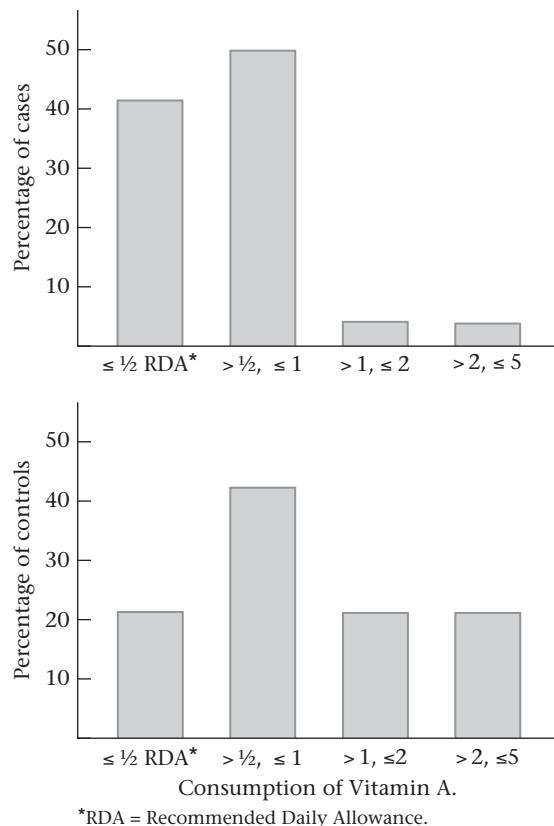
Before any formal attempt to answer this question can be made, the vitamin-A consumption among cases and controls must be described. Consider Figure 2.1. The **bar graphs** show that the controls consume more vitamin A than the cases do, particularly at consumption levels exceeding the Recommended Daily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

**Figure 2.1** Daily vitamin-A consumption among cancer cases and controls

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.

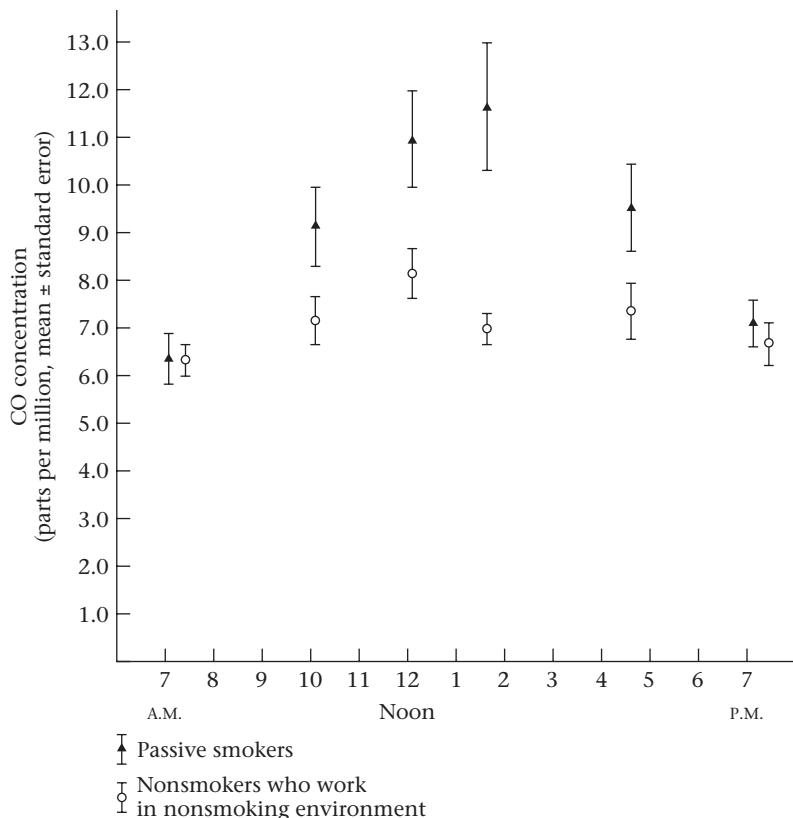
What makes a good graphic or numeric display? The main guideline is that the material should be as self-contained as possible and should be understandable without reading the text. These attributes require clear labeling. The captions, units, and axes on graphs should be clearly labeled, and the statistical 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  $x_1, \dots, x_n$ , where  $x_1$  corresponds to the first sample point and  $x_n$  corresponds to the

**Figure 2.2** Mean carbon-monoxide concentration ( $\pm$  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: Reproduced with permission of *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 birth-weights 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  $\bar{x}$ .

**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  $\Sigma$  (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  $(x_1 + x_2 + \dots + x_n)$ .

If  $a$  and  $b$  are integers, where  $a \leq b$ , then

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

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

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  $x_1 = 2$      $x_2 = 5$      $x_3 = -4$

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

**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.

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

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

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,

$$\begin{aligned} \text{Sample median} &= \text{average of the } 10\text{th and } 11\text{th largest observations} \\ &= (3245 + 3248)/2 = 3246.5 \text{ g} \end{aligned}$$

**Example 2.6**

**Infectious Disease** Consider the data set in Table 2.2, which consists of white-blood counts taken on admission of all patients entering a small hospital in Allentown, Pennsylvania, on a given day. Compute the median white-blood count.

**Table 2.2**

**Sample of admission white-blood counts ( $\times 1000$ ) for all patients entering a hospital in Allentown, PA, on a given day**

$i$	$x_i$	$i$	$x_i$
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

**Solution**

First, order the sample as follows: 3, 5, 7, 8, 8, 9, 10, 12, 35. Because  $n$  is odd, the sample median is given by the fifth largest point, which equals 8 or 8000 on the original scale.

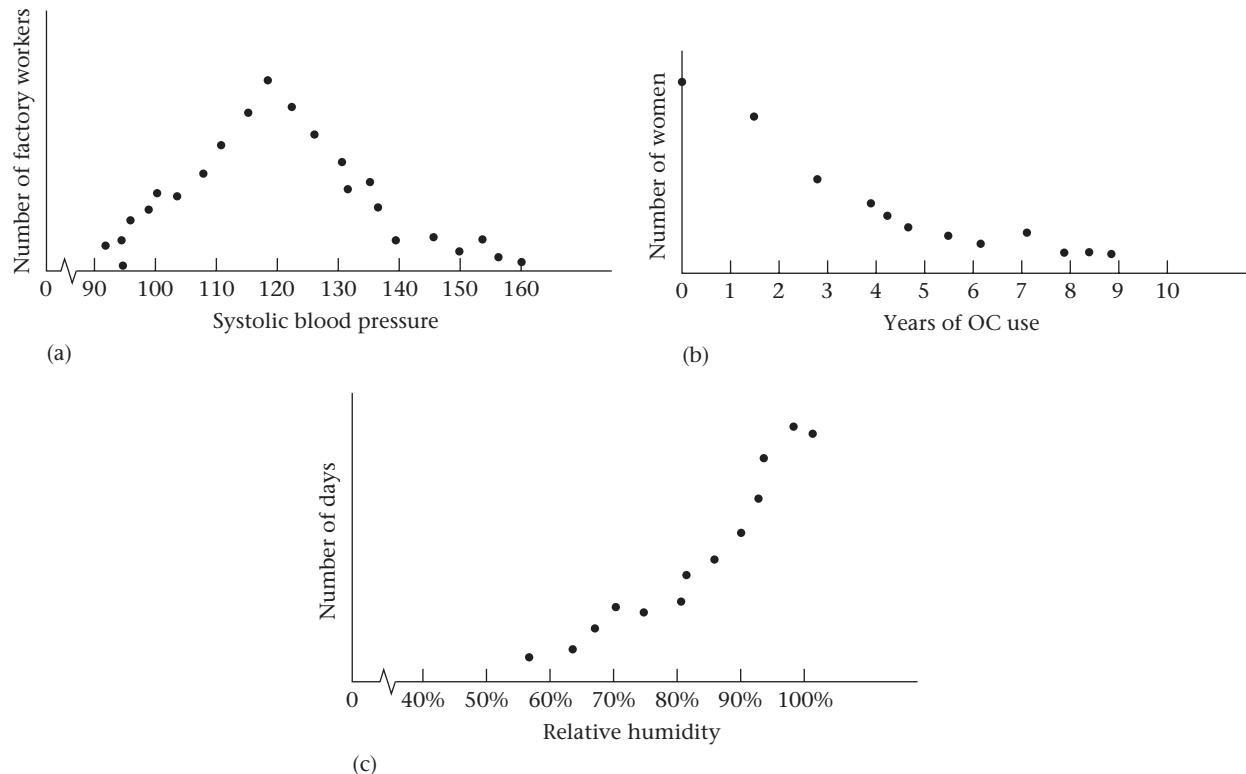
The main strength of the sample median is that it is insensitive to very large or very small values. In particular, if the second patient in Table 2.2 had a white count of 65,000 rather than 35,000, the sample median would remain unchanged, because the fifth largest value is still 8000. Conversely, the arithmetic mean would increase dramatically from 10,778 in the original sample to 14,111 in the new sample. The main weakness of the sample median is that it is determined mainly by the middle points in a sample and is less sensitive to the actual numeric values of the remaining data points.

### Comparison of the Arithmetic Mean and the Median

If a distribution is **symmetric**, then the relative position of the points on each side of the sample median is the same. An example of a distribution that is expected to be roughly symmetric is the distribution of systolic blood-pressure measurements taken on all 30- to 39-year-old factory workers in a given workplace (Figure 2.3a).

If a distribution is **positively skewed** (skewed to the right), then points above the median tend to be farther from the median in absolute value than points below the median. An example of a positively skewed distribution is that of the number of years of oral contraceptive (OC) use among a group of women ages 20 to 29 years (Figure 2.3b). Similarly, if a distribution is **negatively skewed** (skewed to the left), then points below the median tend to be farther from the median in absolute value than points above the median. An example of a negatively skewed distribution is that of relative humidities observed in a humid climate at the same time of day over a number of days. In this case, most humidities are at or close to 100%, with a few very low humidities on dry days (Figure 2.3c).

In many samples, the relationship between the arithmetic mean and the sample median can be used to assess the symmetry of a distribution. In particular, for symmetric distributions the arithmetic mean is approximately the same as the median. For positively skewed distributions, the arithmetic mean tends to be larger than the median; for negatively skewed distributions, the arithmetic mean tends to be smaller than the median.

**Figure 2.3** Graphic displays of (a) symmetric, (b) positively skewed, and (c) negatively skewed distributions

## The Mode

Another widely used measure of location is the mode.

### **Definition 2.3**

The **mode** is the most frequently occurring value among all the observations in a sample.

### **Example 2.7**

**Gynecology** Consider the sample of time intervals between successive menstrual periods for a group of 500 college women age 18 to 21 years, shown in Table 2.3. The frequency column gives the number of women who reported each of the respective durations. The mode is 28 because it is the most frequently occurring value.

**Table 2.3** Sample of time intervals between successive menstrual periods (days) in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

**Example 2.8** Compute the mode of the distribution in Table 2.2.

**Solution** The mode is 8000 because it occurs more frequently than any other white-blood count.

Some distributions have more than one mode. In fact, one useful method of classifying distributions is by the number of modes present. A distribution with one mode is called **unimodal**; two modes, **bimodal**; three modes, **trimodal**; and so forth.

**Example 2.9** Compute the mode of the distribution in Table 2.1.

**Solution** There is no mode, because all the values occur exactly once.

Example 2.9 illustrates a common problem with the mode: It is not a useful measure of location if there is a large number of possible values, each of which occurs infrequently. In such cases the mode will be either far from the center of the sample or, in extreme cases, will not exist, as in Example 2.9. The mode is not used in this text because its mathematical properties are, in general, rather intractable, and in most common situations it is inferior to the arithmetic mean.

### The Geometric Mean

Many types of laboratory data, specifically data in the form of concentrations of one substance in another, as assessed by serial dilution techniques, can be expressed either as multiples of 2 or as a constant multiplied by a power of 2; that is, outcomes can only be of the form  $2^k c$ ,  $k = 0, 1, \dots$ , for some constant  $c$ . For example, the data in Table 2.4 represent the minimum inhibitory concentration (MIC) of penicillin G in the urine for *N. gonorrhoeae* in 74 patients [2]. The arithmetic mean is not appropriate as a measure of location in this situation because the distribution is very skewed.

However, the data do have a certain pattern because the only possible values are of the form  $2^k(0.03125)$  for  $k = 0, 1, 2, \dots$ . One solution is to work with the distribution of the logs of the concentrations. The log concentrations have the property that successive possible concentrations differ by a constant; that is,  $\log(2^{k+1}c) - \log(2^k c) = \log(2^{k+1}) + \log c - \log(2^k) - \log c = (k+1)\log 2 - k\log 2 = \log 2$ . Thus the log concentrations are equally spaced from each other, and the resulting distribution is now not as skewed as the concentrations themselves. The arithmetic mean can then be computed in the log scale; that is,

Text not available due to copyright restrictions

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

and used as a measure of location. However, it is usually preferable to work in the original scale by taking the antilogarithm of  $\overline{\log x}$  to form the geometric mean, which leads to the following definition:

**Definition 2.4** The **geometric mean** is the antilogarithm of  $\overline{\log x}$ , where

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

Any base can be used to compute logarithms for the geometric mean. The geometric mean is the same regardless of which base is used. The only requirement is that the logs and antilogs in Definition 2.4 should be in the same base. Bases often used in practice are base 10 and base  $e$ ; logs and antilogs using these bases can be computed using many pocket calculators.

**Example 2.10**

**Infectious Disease** Compute the geometric mean for the sample in Table 2.4.

**Solution**

(1) For convenience, use base 10 to compute the logs and antilogs in this example.

(2) Compute

$$\overline{\log x} = \left[ \frac{21\log(0.03125) + 6\log(0.0625) + 8\log(0.125)}{74} \right] / 74 = -0.846$$

(3) The geometric mean = the antilogarithm of  $-0.846 = 10^{-0.846} = 0.143$ .

## 2.3 Some Properties of the Arithmetic Mean

Consider a sample  $x_1, \dots, x_n$ , which will be referred to as the original sample. To create a **translated sample**  $x_1 + c, \dots, x_n + c$ , add a constant  $c$  to each data point. Let  $y_i = x_i + c$ ,  $i = 1, \dots, n$ . Suppose we want to compute the arithmetic mean of the translated sample. We can show that the following relationship holds:

**Equation 2.1**

If	$y_i = x_i + c, \quad i = 1, \dots, n$
then	$\bar{y} = \bar{x} + c$

Therefore, to find the arithmetic mean of the  $y$ 's, compute the arithmetic mean of the  $x$ 's and add the constant  $c$ .

This principle is useful because it is sometimes convenient to change the "origin" of the sample data—that is, to compute the arithmetic mean after the translation and then transform back to the original origin.

**Example 2.11**

To compute the arithmetic mean of the time interval between menstrual periods in Table 2.3, it is more convenient to work with numbers that are near zero than with numbers near 28. Thus a translated sample might first be created by subtracting 28 days from each outcome in Table 2.3. The arithmetic mean of the translated sample could then be found and 28 added to get the actual arithmetic mean. The calculations are shown in Table 2.5.

**Table 2.5** Translated sample for the duration between successive menstrual periods in college-age women

Value	Frequency	Value	Frequency	Value	Frequency
-4	5	1	96	6	7
-3	10	2	63	7	3
-2	28	3	24	8	2
-1	64	4	9	9	1
0	185	5	2	10	1

Note:  $\bar{y} = [(-4)(5) + (-3)(10) + \dots + (10)(1)] / 500 = 0.54$

$$\bar{x} = \bar{y} + 28 = 0.54 + 28 = 28.54 \text{ days}$$

Similarly, systolic blood-pressure scores are usually between 100 and 200. Therefore, to obtain the mean of the original sample it is easier to subtract 100 from each blood-pressure score, find the mean of the translated sample, and add 100.

What happens to the arithmetic mean if the units or scale being worked with changes? A **rescaled sample** can be created:

$$y_i = cx_i, \quad i = 1, \dots, n$$

The following result holds:

**Equation 2.2**

$$\text{If } y_i = cx_i, \quad i = 1, \dots, n$$

$$\text{then } \bar{y} = c\bar{x}$$

Therefore, to find the arithmetic mean of the  $y$ 's, compute the arithmetic mean of the  $x$ 's and multiply it by the constant  $c$ .

**Example 2.12**

Express the mean birthweight for the data in Table 2.1 in ounces rather than grams.

**Solution**

We know that 1 oz = 28.35 g and that  $\bar{x} = 3166.9$  g. Thus, if the data were expressed in terms of ounces,

$$c = \frac{1}{28.35} \quad \text{and} \quad \bar{y} = \frac{1}{28.35}(3166.9) = 111.71 \text{ oz}$$

Sometimes we want to change both the origin and the scale of the data at the same time. To do this, apply Equations 2.1 and 2.2 as follows:

**Equation 2.3**

Let  $x_1, \dots, x_n$  be the original sample of data and let  $y_i = c_1x_i + c_2, \quad i = 1, \dots, n$  represent a transformed sample obtained by multiplying each original sample point by a factor  $c_1$  and then shifting over by a constant  $c_2$ .

$$\text{If } y_i = c_1x_i + c_2, \quad i = 1, \dots, n$$

$$\text{then } \bar{y} = c_1\bar{x} + c_2$$

**Example 2.13** If we have a sample of temperatures in °C with an arithmetic mean of 11.75°C, then what is the arithmetic mean in °F?

**Solution** Let  $y_i$  denote the °F temperature that corresponds to a °C temperature of  $x_i$ . The required transformation to convert the data to °F would be

$$y_i = \frac{9}{5}x_i + 32, \quad i = 1, \dots, n$$

so the arithmetic mean would be

$$\bar{y} = \frac{9}{5}(11.75) + 32 = 53.15^{\circ}\text{F}$$

## 2.4 Measures of Spread

Consider Figure 2.4, which represents two samples of cholesterol measurements, each on the same person, but using different measurement techniques. The samples appear to have about the same center, and whatever measure of central location is used is probably about the same in the two samples. In fact, the arithmetic means are both 200 mg/dL. Visually, however, the two samples appear radically different. This difference lies in the greater **variability**, or **spread**, of the Autoanalyzer method relative to the Microenzymatic method. In this section, the notion of variability is quantified. Many samples can be well described by a combination of a measure of location and a measure of spread.

### The Range

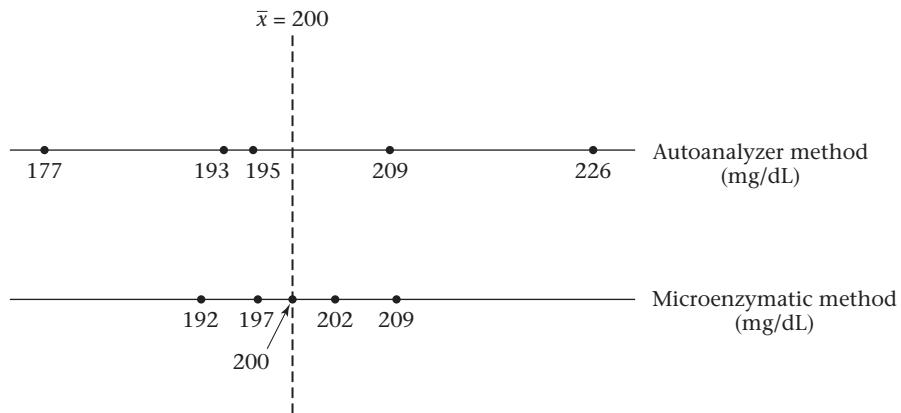
Several different measures can be used to describe the variability of a sample. Perhaps the simplest measure is the range.

---

**Definition 2.5** The **range** is the difference between the largest and smallest observations in a sample.

---

**Figure 2.4** Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



**Example 2.14** The range in the sample of birthweights in Table 2.1 is

$$4146 - 2069 = 2077 \text{ g}$$

**Example 2.15** Compute the ranges for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4, and compare the variability of the two methods.

**Solution** The range for the Autoanalyzer method =  $226 - 177 = 49 \text{ mg/dL}$ . The range for the Microenzymatic method =  $209 - 192 = 17 \text{ mg/dL}$ . The Autoanalyzer method clearly seems more variable.

One advantage of the range is that it is very easy to compute once the sample points are ordered. One striking disadvantage is that it is very sensitive to extreme observations. Hence, if the lightest infant in Table 2.1 weighed 500 g rather than 2069 g, then the range would increase dramatically to  $4146 - 500 = 3646 \text{ g}$ . Another disadvantage of the range is that it depends on the sample size ( $n$ ). That is, the larger  $n$  is, the larger the range tends to be. This complication makes it difficult to compare ranges from data sets of differing size.

## Quantiles

Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of **quantiles** or **percentiles**. Intuitively, the  $p$ th percentile is the value  $V_p$  such that  $p$  percent of the sample points are less than or equal to  $V_p$ . The median, being the 50th percentile, is a special case of a quantile. As was the case for the median, a different definition is needed for the  $p$ th percentile, depending on whether or not  $np/100$  is an integer.

**Definition 2.6** The  $p$ th percentile is defined by

- (1) The  $(k + 1)$ th largest sample point if  $np/100$  is not an integer (where  $k$  is the largest integer less than  $np/100$ )
- (2) The average of the  $(np/100)$ th and  $(np/100 + 1)$ th largest observations if  $np/100$  is an integer.

Percentiles are also sometimes called **quantiles**.

The spread of a distribution can be characterized by specifying several percentiles. For example, the 10th and 90th percentiles are often used to characterize spread. Percentiles have the advantage over the range of being less sensitive to outliers and of not being greatly affected by the sample size ( $n$ ).

**Example 2.16** Compute the 10th and 90th percentiles for the birthweight data in Table 2.1.

**Solution** Because  $20 \times .1 = 2$  and  $20 \times .9 = 18$  are integers, the 10th and 90th percentiles are defined by

$$\begin{aligned} \text{10th percentile: average of the second and third largest values} \\ = (2581 + 2759)/2 = 2670 \text{ g} \end{aligned}$$

$$\begin{aligned} \text{90th percentile: average of the 18th and 19th largest values} \\ = (3609 + 3649)/2 = 3629 \text{ g} \end{aligned}$$

We would estimate that 80% of birthweights will fall between 2670 g and 3629 g, which gives an overall impression of the spread of the distribution.

**Example 2.17** Compute the 20th percentile for the white-blood-count data in Table 2.2.

**Solution**

Because  $np/100 = 9 \times .2 = 1.8$  is not an integer, the 20th percentile is defined by the  $(1 + 1)$ th largest value = second largest value = 5000.

To compute percentiles, the sample points must be ordered. This can be difficult if  $n$  is even moderately large. An easy way to accomplish this is to use a stem-and-leaf plot (see Section 2.8) or a computer program.

There is no limit to the number of percentiles that can be computed. The most useful percentiles are often determined by the sample size and by subject-matter considerations. Frequently used percentiles are quartiles (25th, 50th, and 75th percentiles), quintiles (20th, 40th, 60th, and 80th percentiles), and deciles (10th, 20th, . . . , 90th percentiles). It is almost always instructive to look at some of the quantiles to get an overall impression of the spread and the general shape of a distribution.

## The Variance and Standard Deviation

The main difference between the Autoanalyzer- and Microenzymatic-method data in Figure 2.4 is that the Microenzymatic-method values are closer to the center of the sample than the Autoanalyzer-method values. If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean is needed; that is,

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$$

One simple measure that would seem to accomplish this goal is

$$d = \frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$$

Unfortunately, this measure will not work, because of the following principle:

**Equation 2.4**

The sum of the deviations of the individual observations of a sample about the sample mean is always zero.

**Example 2.18**

Compute the sum of the deviations about the mean for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

**Solution**

For the Autoanalyzer-method data,

$$\begin{aligned} d &= (177 - 200) + (193 - 200) + (195 - 200) + (209 - 200) + (226 - 200) \\ &= -23 - 7 - 5 + 9 + 26 = 0 \end{aligned}$$

For the Microenzymatic-method data,

$$\begin{aligned} d &= (192 - 200) + (197 - 200) + (200 - 200) + (202 - 200) + (209 - 200) \\ &= -8 - 3 + 0 + 2 + 9 = 0 \end{aligned}$$

Thus  $d$  does not help distinguish the difference in spreads between the two methods.

A second possible measure is

$$\sum_{i=1}^n |x_i - \bar{x}| / n$$

which is called the **mean deviation**. The mean deviation is a reasonable measure of spread but does not characterize the spread as well as the standard deviation (see Definition 2.8) if the underlying distribution is bell-shaped.

A third idea is to use the average of the squares of the deviations from the sample mean rather than the deviations themselves. The resulting measure of spread, denoted by  $s^2$ , is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

The more usual form for this measure is with  $n - 1$  in the denominator rather than  $n$ . The resulting measure is called the *sample variance* (or *variance*).

**Definition 2.7** The **sample variance**, or **variance**, is defined as follows:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

A rationale for using  $n - 1$  in the denominator rather than  $n$  is presented in the discussion of estimation in Chapter 6.

Another commonly used measure of spread is the sample standard deviation.

**Definition 2.8** The **sample standard deviation**, or **standard deviation**, is defined as follows:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

**Example 2.19** Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

**Solution**

**Autoanalyzer Method**

$$\begin{aligned}s^2 &= \left[ (177 - 200)^2 + (193 - 200)^2 + (195 - 200)^2 + (209 - 200)^2 + (226 - 200)^2 \right] / 4 \\ &= (529 + 49 + 25 + 81 + 676) / 4 = 1360 / 4 = 340 \\ s &= \sqrt{340} = 18.4\end{aligned}$$

**Microenzymatic Method**

$$\begin{aligned}s^2 &= \left[ (192 - 200)^2 + (197 - 200)^2 + (200 - 200)^2 + (202 - 200)^2 + (209 - 200)^2 \right] / 4 \\ &= (64 + 9 + 0 + 4 + 81) / 4 = 158 / 4 = 39.5 \\ s &= \sqrt{39.5} = 6.3\end{aligned}$$

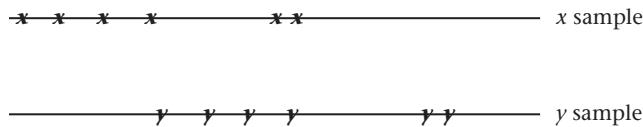
Thus the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

## 2.5 Some Properties of the Variance and Standard Deviation

The same question can be asked of the variance and standard deviation as of the arithmetic mean: namely, how are they affected by a change in origin or a change in the units being worked with? Suppose there is a sample  $x_1, \dots, x_n$  and all data points in the sample are shifted by a constant  $c$ ; that is, a new sample  $y_1, \dots, y_n$  is created such that  $y_i = x_i + c$ ,  $i = 1, \dots, n$ .

In Figure 2.5, we would clearly expect the variance and standard deviation to remain the same because the relationship of the points in the sample relative to one another remains the same. This property is stated as follows:

**Figure 2.5** Comparison of the variances of two samples, where one sample has an origin shifted relative to the other



**Equation 2.5**

Suppose there are two samples

$$x_1, \dots, x_n \quad \text{and} \quad y_1, \dots, y_n$$

$$\text{where } y_i = x_i + c, \quad i = 1, \dots, n$$

If the respective sample variances of the two samples are denoted by

$$s_x^2 \quad \text{and} \quad s_y^2$$

$$\text{then } s_y^2 = s_x^2$$

**Example 2.20**

Compare the variances and standard deviations for the menstrual-period data in Tables 2.3 and 2.5.

**Solution**

The variance and standard deviation of the two samples are the same because the second sample was obtained from the first by subtracting 28 days from each data value; that is,

$$y_i = x_i - 28$$

Suppose the units are now changed so that a new sample  $y_1, \dots, y_n$  is created such that  $y_i = cx_i$ ,  $i = 1, \dots, n$ . The following relationship holds between the variances of the two samples.

**Equation 2.6**

Suppose there are two samples

$$x_1, \dots, x_n \quad \text{and} \quad y_1, \dots, y_n$$

$$\text{where } y_i = cx_i, \quad i = 1, \dots, n, \quad c > 0$$

$$\text{Then } s_y^2 = c^2 s_x^2 \quad s_y = cs_x$$

This can be shown by noting that

$$\begin{aligned} s_y^2 &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (cx_i - c\bar{x})^2}{n-1} \\ &= \frac{\sum_{i=1}^n [c(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^n c^2(x_i - \bar{x})^2}{n-1} \\ &= \frac{c^2 \sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = c^2 s_x^2 \\ s_y &= \sqrt{c^2 s_x^2} = cs_x \end{aligned}$$

**Example 2.21**

Compute the variance and standard deviation of the birthweight data in Table 2.1 in both grams and ounces.

**Solution**

The original data are given in grams, so first compute the variance and standard deviation in these units.

$$\begin{aligned}s^2 &= \frac{(3265 - 3166.9)^2 + \dots + (2834 - 3166.9)^2}{19} \\&= 3,768,147.8/19 = 198,323.6 \text{ g}^2 \\s &= 445.3 \text{ g}\end{aligned}$$

To compute the variance and standard deviation in ounces, note that

$$1 \text{ oz} = 28.35 \text{ g} \quad \text{or} \quad y_i = \frac{1}{28.35} x_i$$

$$\text{Thus } s^2(\text{oz}) = \frac{1}{28.35^2} s^2(\text{g}) = 246.8 \text{ oz}^2$$

$$s(\text{oz}) = \frac{1}{28.35} s(\text{g}) = 15.7 \text{ oz}$$

Thus, if the sample points change in scale by a factor of  $c$ , the variance changes by a factor of  $c^2$  and the standard deviation changes by a factor of  $c$ . This relationship is the main reason why the standard deviation is more often used than the variance as a measure of spread: the standard deviation and the arithmetic mean are in the same units, whereas the variance and the arithmetic mean are not. Thus, as illustrated in Examples 2.12 and 2.21, both the mean and the standard deviation change by a factor of  $1/28.35$  in the birthweight data of Table 2.1 when the units are expressed in ounces rather than in grams.

The mean and standard deviation are the most widely used measures of location and spread in the literature. One of the main reasons for this is that the normal (or bell-shaped) distribution is defined explicitly in terms of these two parameters, and this distribution has wide applicability in many biological and medical settings. The normal distribution is discussed extensively in Chapter 5.

## 2.6 The Coefficient of Variation

It is useful to relate the arithmetic mean and the standard deviation to each other because, for example, a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 than if it is 1000. A special measure, the coefficient of variation, is often used for this purpose.

**Definition 2.9**

The coefficient of variation (CV) is defined by

$$100\% \times (s/\bar{x})$$

This measure remains the same regardless of what units are used because if the units change by a factor  $c$ , then both the mean and standard deviation change by the factor  $c$ ; the CV, which is the ratio between them, remains unchanged.

**Example 2.22** Compute the coefficient of variation for the data in Table 2.1 when the birthweights are expressed in either grams or ounces.

**Table 2.6** Reproducibility of cardiovascular risk factors in children, Bogalusa Heart Study, 1978–1979

	<i>n</i>	Mean	<i>sd</i>	CV(%)
Height (cm)	364	142.6	0.31	0.2
Weight (kg)	365	39.5	0.77	1.9
Triceps skin fold (mm)	362	15.2	0.51	3.4
Systolic blood pressure (mm Hg)	337	104.0	4.97	4.8
Diastolic blood pressure (mm Hg)	337	64.0	4.57	7.1
Total cholesterol (mg/dL)	395	160.4	3.44	2.1
HDL cholesterol (mg/dL)	349	56.9	5.89	10.4

**Solution**

$$CV = 100\% \times (s/\bar{x}) = 100\% \times (445.3\text{ g}/3166.9\text{ g}) = 14.1\%$$

If the data were expressed in ounces, then

$$CV = 100\% \times (15.7 \text{ oz}/111.71 \text{ oz}) = 14.1\%$$

The *CV* is most useful in comparing the variability of several different samples, each with different arithmetic means. This is because a higher variability is usually expected when the mean increases, and the *CV* is a measure that accounts for this variability. Thus, if we are conducting a study in which air pollution is measured at several sites and we wish to compare day-to-day variability at the different sites, we might expect a higher variability for the more highly polluted sites. A more accurate comparison could be made by comparing the *CVs* at different sites than by comparing the standard deviations.

The *CV* is also useful for comparing the reproducibility of different variables. Consider, for example, data from the Bogalusa Heart Study, a large study of cardiovascular risk factors in children [3] that began in the 1970s and continues up to the present time.

At approximately 3-year intervals, cardiovascular risk factors such as blood pressure, weight, and cholesterol levels were measured for each of the children in the study. In 1978, replicate measurements were obtained for a subset of the children a short time apart from regularly scheduled risk factor measurements. Table 2.6 presents reproducibility data on a selected subset of cardiovascular risk factors. We note that the *CV* ranges from 0.2% for height to 10.4% for HDL cholesterol. The standard deviations reported here are within-subject standard deviations based on repeated assessments on the same child. Details on how within- and between-subject variations are computed are covered in Chapter 12 in the discussion of the random-effects analysis-of-variance model.

### REVIEW QUESTIONS 2A

- When is it appropriate to use the arithmetic mean as opposed to the median?
- How does the geometric mean differ from the arithmetic mean? What type of data is the geometric mean used for?
- What is the difference between the standard deviation and the *CV*? When is it appropriate to use each measure?

**Table 2.7** Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

## 2.7 Grouped Data

Sometimes the sample size is too large to display all the raw data. Also, data are frequently collected in grouped form because the required degree of accuracy to measure a quantity exactly is often lacking due either to measurement error or to imprecise patient recall. For example, systolic blood-pressure measurements taken with a standard cuff are usually specified to the nearest 2 mm Hg because assessing them with any more precision is difficult using this instrument. Thus, a stated measurement of 120 mm Hg may actually imply that the reading is some number  $\geq 119$  mm Hg and  $< 121$  mm Hg. Similarly, because dietary recall is generally not very accurate, the most precise estimate of fish consumption might take the following form: 2–3 servings per day, 1 serving per day, 5–6 servings per week, 2–4 servings per week, 1 serving per week,  $< 1$  serving per week and  $\geq 1$  serving per month, never.

Consider the data set in Table 2.7, which represents the birthweights from 100 consecutive deliveries at a Boston hospital. Suppose we wish to display these data for publication purposes. How can we do this? The simplest way to display the data is to generate a frequency distribution using a statistical package.

### Definition 2.10

A **frequency distribution** is an ordered display of each value in a data set together with its **frequency**, that is, the number of times that value occurs in the data set. In addition, the percentage of sample points that take on a particular value is also typically given.

A frequency distribution of the sample of 100 birthweights in Table 2.7, generated using the MINITAB package, is displayed in Table 2.8.

The MINITAB Tally program provides the frequency (Count), relative frequency (Percent), cumulative frequency (CumCnt), and cumulative percent (CumPct) for each birthweight present in the sample. For any particular birthweight  $b$ , the cumulative frequency, or CumCnt, is the number of birthweights in the sample that are less than or equal to  $b$ . The Percent =  $100 \times \text{Count}/n$ , whereas CumPct =  $100 \times \text{CumCnt}/n$  = the percentage of birthweights less than or equal to  $b$ .

**Table 2.8** Frequency distribution of the birthweight data in Table 2.7 using the MINITAB Tally program

Birthweight	Count	CumCnt	Percent	CumPct
32	1	1	1.00	1.00
58	1	2	1.00	2.00
64	1	3	1.00	3.00
67	1	4	1.00	4.00
68	1	5	1.00	5.00
83	1	6	1.00	6.00
85	2	8	2.00	8.00
86	1	9	1.00	9.00
87	1	10	1.00	10.00
88	2	12	2.00	12.00
89	3	15	3.00	15.00
91	1	16	1.00	16.00
92	1	17	1.00	17.00
93	1	18	1.00	18.00
94	2	20	2.00	20.00
95	1	21	1.00	21.00
96	1	22	1.00	22.00
98	3	25	3.00	25.00
99	1	26	1.00	26.00
100	1	27	1.00	27.00
101	1	28	1.00	28.00
102	1	29	1.00	29.00
103	1	30	1.00	30.00
104	5	35	5.00	35.00
105	2	37	2.00	37.00
106	1	38	1.00	38.00
107	1	39	1.00	39.00
108	4	43	4.00	43.00
109	2	45	2.00	45.00
110	2	47	2.00	47.00
111	1	48	1.00	48.00
112	3	51	3.00	51.00
113	1	52	1.00	52.00
115	6	58	6.00	58.00
116	1	59	1.00	59.00
118	2	61	2.00	61.00
119	1	62	1.00	62.00
120	1	63	1.00	63.00
121	3	66	3.00	66.00
122	4	70	4.00	70.00
123	1	71	1.00	71.00
124	4	75	4.00	75.00
125	2	77	2.00	77.00
126	1	78	1.00	78.00
127	2	80	2.00	80.00
128	2	82	2.00	82.00
132	3	85	3.00	85.00
133	2	87	2.00	87.00
134	1	88	1.00	88.00
135	2	90	2.00	90.00
137	1	91	1.00	91.00
138	3	94	3.00	94.00
140	1	95	1.00	95.00
141	1	96	1.00	96.00
144	1	97	1.00	97.00
146	1	98	1.00	98.00
155	1	99	1.00	99.00
161	1	100	1.00	100.00
N =		100		

**Table 2.9 General layout of grouped data**

Group interval	Frequency
$y_1 \leq x < y_2$	$f_1$
$y_2 \leq x < y_3$	$f_2$
.	.
.	.
$y_i \leq x < y_{i+1}$	$f_i$
.	.
.	.
$y_k \leq x < y_{k+1}$	$f_k$

**Table 2.10 Grouped frequency distribution of birthweight (oz) from 100 consecutive deliveries**

Group interval	Frequency
$29.5 \leq x < 69.5$	5
$69.5 \leq x < 89.5$	10
$89.5 \leq x < 99.5$	11
$99.5 \leq x < 109.5$	19
$109.5 \leq x < 119.5$	17
$119.5 \leq x < 129.5$	20
$129.5 \leq x < 139.5$	12
$139.5 \leq x < 169.5$	6
	100

Note: If birthweight can only be measured to an accuracy of 0.1 oz, then a possible alternate representation of the group intervals in Table 2.10 could be 29.5–69.4, 69.5–89.4, to 139.5–169.5.

If the number of unique sample values is large, then a frequency distribution may still be too detailed a summary for publication purposes. Instead, the data could be grouped into broader categories. Here are some general instructions for categorizing the data:

- (1) Subdivide the data into  $k$  intervals, starting at some lower bound  $y_1$  and ending at some upper bound  $y_{k+1}$ .
- (2) The first interval is from  $y_1$  inclusive to  $y_2$  exclusive; the second interval is from  $y_2$  inclusive to  $y_3$  exclusive; . . . ; the  $k$ th and last interval is from  $y_k$  inclusive to  $y_{k+1}$  exclusive. The rationale for this representation is to make certain the group intervals include all possible values *and* do not overlap. In some published work, grouped data are presented, but the group boundaries are ambiguous (e.g., 0–5, 5–10, etc.).
- (3) The group intervals are generally chosen to be equal, although the appropriateness of equal group sizes should be dictated more by subject-matter considerations. Thus, equal intervals might be appropriate for the blood-pressure or birthweight data but not for the dietary-recall data, where the nature of the data dictates unequal group sizes corresponding to how most people remember what they eat.
- (4) A count is made of the number of units that fall in each interval, which is denoted by the frequency within that interval.
- (5) Finally, the group intervals and their frequencies,  $f_i$ , are then displayed concisely in a table such as Table 2.9.

For example, the raw data in Table 2.7 might be displayed in grouped form as shown in Table 2.10.

## 2.8 Graphic Methods

In Sections 2.1 through 2.7 we concentrated on methods for describing data in numeric and tabular form. In this section, these techniques are supplemented by presenting certain commonly used graphic methods for displaying data. The purpose of using graphic displays is to give a quick overall impression of data, which is sometimes difficult to obtain with numeric measures.

## Bar Graphs

One of the most widely used methods for displaying grouped data is the bar graph.

A **bar graph** can be constructed as follows:

- (1) The data are divided in a number of groups using the guidelines provided in Section 2.7.
- (2) For each group a rectangle is constructed with a base of a constant width and a height proportional to the frequency within that group.
- (3) The rectangles are generally not contiguous and are equally spaced from each other.

A bar graph of daily vitamin-A consumption among 200 cancer cases and 200 age-and sex-matched controls is shown in Figure 2.1.

## Stem-and-Leaf Plots

Two problems with bar graphs are that (1) they are somewhat difficult to construct and (2) the sense of what the actual sample points are within the respective groups is lost. One type of graphic display that overcomes these problems is the stem-and-leaf plot.

A **stem-and-leaf** plot can be constructed as follows:

- (1) Separate each data point into a stem component and a leaf component, respectively, where the stem component consists of the number formed by all but the rightmost digit of the number, and the leaf component consists of the rightmost digit. Thus the stem of the number 483 is 48, and the leaf is 3.
- (2) Write the smallest stem in the data set in the upper left-hand corner of the plot.
- (3) Write the second stem, which equals the first stem + 1, below the first stem.
- (4) Continue with step 3 until you reach the largest stem in the data set.
- (5) Draw a vertical bar to the right of the column of stems.
- (6) For each number in the data set, find the appropriate stem and write the leaf to the right of the vertical bar.

The collection of leaves thus formed takes on the general shape of the distribution of the sample points. Furthermore, the actual sample values are preserved and yet there is a grouped display for the data, which is a distinct advantage over a bar graph. Finally, it is also easy to compute the median and other quantiles from a stem-and-leaf plot. Figure 2.6 presents a stem-and-leaf plot using MINITAB for the birthweight data in Table 2.7. Thus the point 5|8 represents 58, 11|8 represents 118, and so forth. Notice how this plot gives an overall feel for the distribution without losing the individual values. Also, the cumulative frequency count from either the lowest or the highest value is given in the first column. For the 11 stem, the absolute count is given in parentheses (17) instead of the cumulative total because the highest or lowest value would exceed 50% (50).

In some variations of stem-and-leaf plots the leaf can consist of more than one digit. This might be appropriate for the birthweight data in Table 2.1 because the number of three-digit stems required would be very large relative to the number of

**Figure 2.6** Stem-and-leaf plot for the birthweight data (oz) in Table 2.7

Stem-and-leaf of birthwt N = 100		
Leaf Unit = 1.0		
1	3	2
1	4	
2	5	8
5	6	478
5	7	
15	8	3556788999
26	9	12344568889
45	10	012344445567888899
(17)	11	0012223555556889
38	12	0111222234445567788
18	13	222334557888
6	14	0146
2	15	5
1	16	1

data points. In this case, the leaf would consist of the rightmost two digits and the stem the leftmost two digits, and the pairs of digits to the right of the vertical bar would be underlined to distinguish between two different leaves. The stem-and-leaf display for the data in Table 2.1 is shown in Figure 2.7.

Another common variation on the ordinary stem-and-leaf plot if the number of leaves is large is to allow more than one line for each stem. Similarly, one can position the largest stem at the top of the plot and the smallest stem at the bottom of the plot. In Figure 2.8 some graphic displays using the SAS UNIVARIATE procedure illustrate this technique.

Notice that each stem is allowed two lines, with the leaves from 5 to 9 on the upper line and the leaves from 0 to 4 on the lower line. Furthermore, the leaves are ordered on each line, and a count of the number of leaves on each line is given under the # column to allow easy computation of the median and other quantiles. Thus the number 7 in the # column on the upper line for stem 12 indicates there are 7 birth weights from 125 to 129 oz in the sample, whereas the number 13 indicates there are 13 birth weights from 120 to 124 oz. Finally, a multiplication factor ( $m$ ) at the bottom of the display allows for representing decimal numbers in stem-and-leaf form. In particular, if no  $m$  is present, then it is assumed all numbers have actual value stem.leaf; whereas if  $m$  is present, then the actual value of the number is assumed to be stem.leaf  $\times 10^m$ . Thus, for example, because the multiplication factor is  $10^1$ , the value 6 | 4 on the stem-and-leaf plot represents the number  $6.4 \times 10^1 = 64$  oz.

## Box Plots

In the section beginning on page 6 we discussed the comparison of the arithmetic mean and the median as a method for looking at the skewness of a distribution. This goal can also be achieved by a graphic technique known as the **box plot**. A box plot uses the relationships among the median, upper quartile, and lower quartile to describe the skewness of a distribution.

The upper and lower quartiles can be thought of conceptually as the approximate 75th and 25th percentiles of the sample—that is, the points 3/4 and 1/4 along the way in the ordered sample.

**Figure 2.7** Stem-and-leaf plot for the birthweight data (g) in Table 2.1

20	69
21	
22	
23	
24	
25	81
26	
27	59
28	41 38 34
29	
30	31
31	01
32	65 60 45 00 48
33	23 14
34	84
35	41
36	49 09
37	
38	
39	
40	
41	46

How can the median, upper quartile, and lower quartile be used to judge the symmetry of a distribution?

- (1) If the distribution is symmetric, then the upper and lower quartiles should be approximately equally spaced from the median.
- (2) If the upper quartile is farther from the median than the lower quartile, then the distribution is positively skewed.
- (3) If the lower quartile is farther from the median than the upper quartile, then the distribution is negatively skewed.

These relationships are illustrated graphically in a box plot. In Figure 2.8 the top of the box corresponds to the upper quartile, whereas the bottom of the box corresponds to the lower quartile. A horizontal line is also drawn at the median value. Furthermore, in the SAS implementation of the box plot, the sample mean is indicated by a + sign distinct from the edges of the box.

### Example 2.23

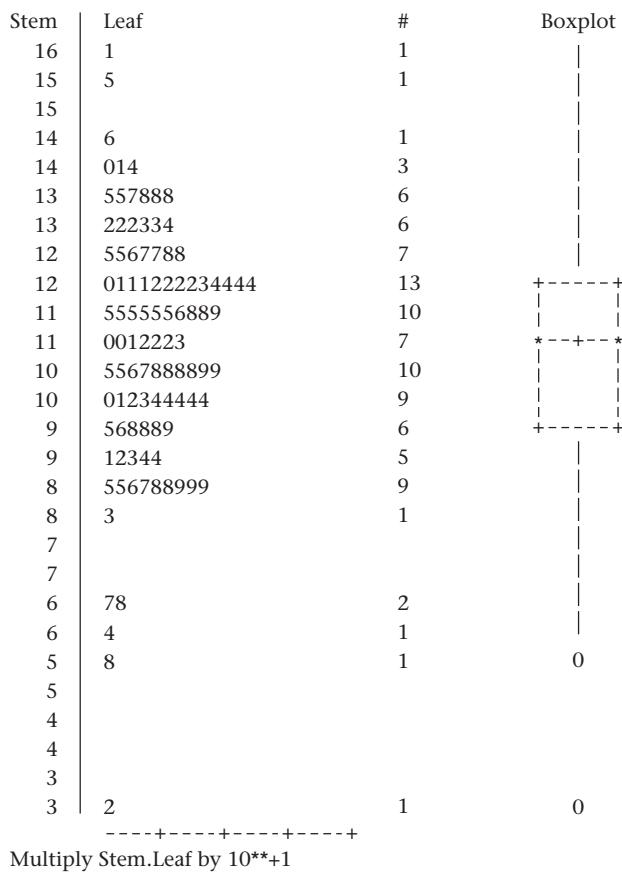
What can we learn about the symmetry properties of the distribution of birthweights from the box plot in Figure 2.8?

### Solution

In Figure 2.8, because the lower quartile is farther from the median than the upper quartile, the distribution is slightly negatively skewed. This pattern is true of many birthweight distributions.

In addition to displaying the symmetry properties of a sample, a box plot can also be used to visually describe the spread of a sample and can help identify possible outlying values—that is, values that seem inconsistent with the rest of the

**Figure 2.8** Stem-and-leaf and box plots for the birthweight data (oz) in Table 2.7 as generated by the SAS UNIVARIATE procedure



points in the sample. In the context of box plots, outlying values are defined as follows:

---

**Definition 2.11** An **outlying value** is a value  $x$  such that either

- (1)  $x > \text{upper quartile} + 1.5 \times (\text{upper quartile} - \text{lower quartile})$  or
  - (2)  $x < \text{lower quartile} - 1.5 \times (\text{upper quartile} - \text{lower quartile})$
- 

---

**Definition 2.12** An **extreme outlying value** is a value  $x$  such that either

- (1)  $x > \text{upper quartile} + 3.0 \times (\text{upper quartile} - \text{lower quartile})$  or
  - (2)  $x < \text{lower quartile} - 3.0 \times (\text{upper quartile} - \text{lower quartile})$
- 

The box plot is then completed by

- (1) Drawing a vertical bar from the upper quartile to the largest nonoutlying value in the sample
  - (2) Drawing a vertical bar from the lower quartile to the smallest nonoutlying value in the sample
  - (3) Individually identifying the outlying and extreme outlying values in the sample by zeroes (0) and asterisks (\*), respectively
-

**Example 2.24**

Using the box plot in Figure 2.8, comment on the spread of the sample in Table 2.7 and the presence of outlying values.

**Solution**

It can be shown from Definition 2.6 that the upper and lower quartiles are 124.5 and 98.5 oz, respectively. Hence, an outlying value  $x$  must satisfy the following relations:

$$\begin{aligned}x &> 124.5 + 1.5 \times (124.5 - 98.5) = 124.5 + 39.0 = 163.5 \\ \text{or } x &< 98.5 - 1.5 \times (124.5 - 98.5) = 98.5 - 39.0 = 59.5\end{aligned}$$

Similarly, an extreme outlying value  $x$  must satisfy the following relations:

$$\begin{aligned}x &> 124.5 + 3.0 \times (124.5 - 98.5) = 124.5 + 78.0 = 202.5 \\ \text{or } x &< 98.5 - 3.0 \times (124.5 - 98.5) = 98.5 - 78.0 = 20.5\end{aligned}$$

Thus the values 32 and 58 oz are outlying values but not extreme outlying values. These values are identified by 0's on the box plot. A vertical bar extends from 64 oz (the smallest nonoutlying value) to the lower quartile and from 161 oz (the largest nonoutlying value = the largest value in the sample) to the upper quartile. The accuracy of the two identified outlying values should probably be checked.

The methods used to identify outlying values in Definitions 2.11 and 2.12 are descriptive and unfortunately are sensitive to sample size, with more outliers detected for larger sample sizes. Alternative methods for identifying outliers based on a hypothesis-testing framework are given in Chapter 8.

Many more details on stem-and-leaf plots, box plots, and other exploratory data methods are given in Tukey [4].

**REVIEW QUESTIONS 2B**

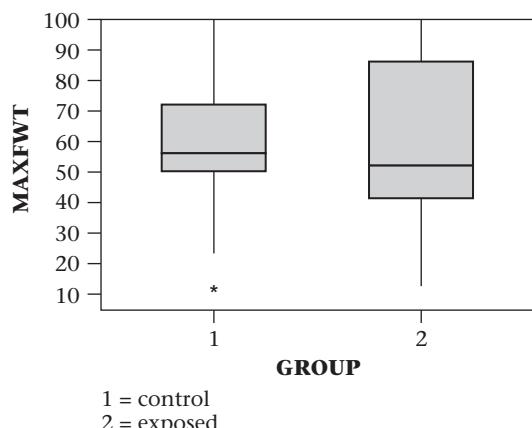
- 1** What is a stem-and-leaf plot? How does it differ from a bar graph?
- 2** Consider the bar graph in Figure 2.1. Is it possible to construct a stem-and-leaf plot from the data presented? If so, construct the plot.
- 3** Consider the stem-and-leaf plot in Figure 2.6. Is it possible to construct a bar graph from the data presented? If so, construct the plot.
- 4** What is a box plot? What additional information does this type of display give that is not available from either a bar graph or stem-and-leaf plot?

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

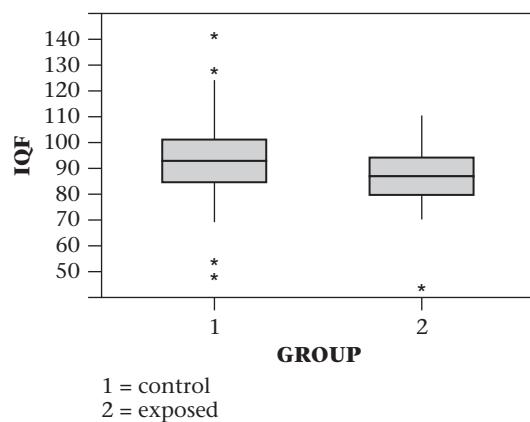
The effects of exposure to lead on the psychological and neurological well-being of children were studied [5]. Complete raw data for this study are in Data Set LEAD.DAT, and documentation for this file is in Data Set LEAD.DOC. Dr. Philip Landrigan, Mount Sinai Medical Center, New York City, provided this data set. All data sets are on the Companion Website.

In summary, blood levels of lead were measured in a group of children who lived near a lead smelter in El Paso, Texas. Forty-six children with blood-lead levels  $\geq 40 \mu\text{g}/\text{mL}$  were identified in 1972 (a few children were identified in 1973); this group is defined by the variable GROUP = 2. A control group of 78 children with blood-lead levels  $< 40 \mu\text{g}/\text{mL}$  were also identified in 1972 and 1973; this group is defined by the variable GROUP = 1. All children lived close to the lead smelter.

**Figure 2.9 Number of finger-wrist taps in the dominant hand for exposed and control groups, El Paso Lead Study**



**Figure 2.10 Wechsler full-scale IQ scores for exposed and control groups, El Paso Lead Study**



Two important outcome variables were studied: (1) the number of finger-wrist taps in the dominant hand (a measure of neurological function) and (2) the Wechsler full-scale IQ score. To explore the relationship of lead exposure to the outcome variables, we used MINITAB to obtain box plots for these two variables for children in the exposed and control groups. These box plots are shown in Figures 2.9 and 2.10, respectively. Because the dominant hand was not identified in the database, we used the larger of the finger-wrist tapping scores for the right and left hand as a proxy for the number of finger-wrist taps in the dominant hand.

We note that although there is considerable spread within each group, both finger-wrist tapping scores (MAXFWT) and full-scale IQ scores (IQF) seem slightly lower in the exposed group than in the control group. We analyze these data in more detail in later chapters, using *t* tests, analysis of variance, and regression methods.

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

A twin study was performed on the relationship between bone density and cigarette consumption [6]. Forty-one pairs of middle-aged female twins who were discordant for tobacco consumption (had different smoking histories) were enrolled in a study in Australia and invited to visit a hospital in Victoria, Australia, for a measurement of bone density. Additional information was also obtained from the participants via questionnaire, including details of tobacco use; alcohol, coffee, and tea consumption; intake of calcium from dairy products; menopausal, reproductive, and fracture history; use of oral contraceptives or estrogen replacement therapy; and assessment of physical activity. Dr. John Hopper, University of Melbourne, School of Population Health, Australia, provided the data set for this study, which is available on the Companion Website under the file name BONEDEN.DAT with documentation in BONEDEN.DOC.

Tobacco consumption was expressed in terms of *pack-years*. One pack-year is defined as 1 pack of cigarettes per day (usually about 20 cigarettes per pack) consumed for 1 year. One advantage of using twins in a study such as this is that genetic influences on bone density are inherently controlled for. To analyze the data,

the investigators first identified the heavier- and lighter-smoking twins in terms of pack-years. The lighter-smoking twin usually had 0 pack-years (indicating she had never smoked) or occasionally either smoked very few cigarettes per day and/or smoked for only a short time. The researchers then looked at the difference in bone-mineral density (BMD) (calculated by subtracting the BMD in the heavier-smoking twin from the BMD in the lighter-smoking twin, expressed as a percentage of the average bone density of the twins) as a function of the difference in tobacco use (calculated as pack-years for the heavier-smoking twin minus pack-years for the lighter-smoking twin). BMD was assessed separately at three sites: the lumbar spine (lower back), the femoral neck (hip), and the femoral shaft (hip). A *scatter plot* showing the relationship between the difference in BMD versus the difference in tobacco use is given in Figure 2.11.

Note that for the lumbar spine an inverse relationship appears between the difference in BMD and the difference in tobacco use (a downward trend). Virtually all the differences in BMD are below 0, especially for twins with a large difference in tobacco use ( $\geq 30$  pack-years), indicating that the heavier-smoking twin had a lower BMD than the lighter-smoking twin. A similar relationship holds for BMD in the femoral neck. Results are less clear for the femoral shaft.

This is a classic example of a *matched-pair study*, which we discuss in detail beginning in Chapter 8. For such a study, the exposed (heavier-smoking twin) and control (lighter-smoking twin) are matched on other characteristics related to the outcome (BMD). In this case, the matching is based on having similar genes. We analyze this data set in more detail in later chapters, using methods based on the binomial distribution, *t* tests, and regression analysis.

## 2.11 Obtaining Descriptive Statistics on the Computer

Numerous statistical packages can be used to obtain descriptive statistics as well as for other statistical functions used in probability, estimation, and hypothesis testing that are covered later in this book. The Companion Website for this book explains in detail how to use Microsoft Excel to perform these functions. Read the first chapter in the Companion Website for details on obtaining descriptive statistics using Excel. Functions available include Average (for the arithmetic mean), Median (for the median), Stdev (for the standard deviation), Var (for the variance), GeoMean (for the geometric mean), and Percentile (for obtaining arbitrary percentiles from a sample).

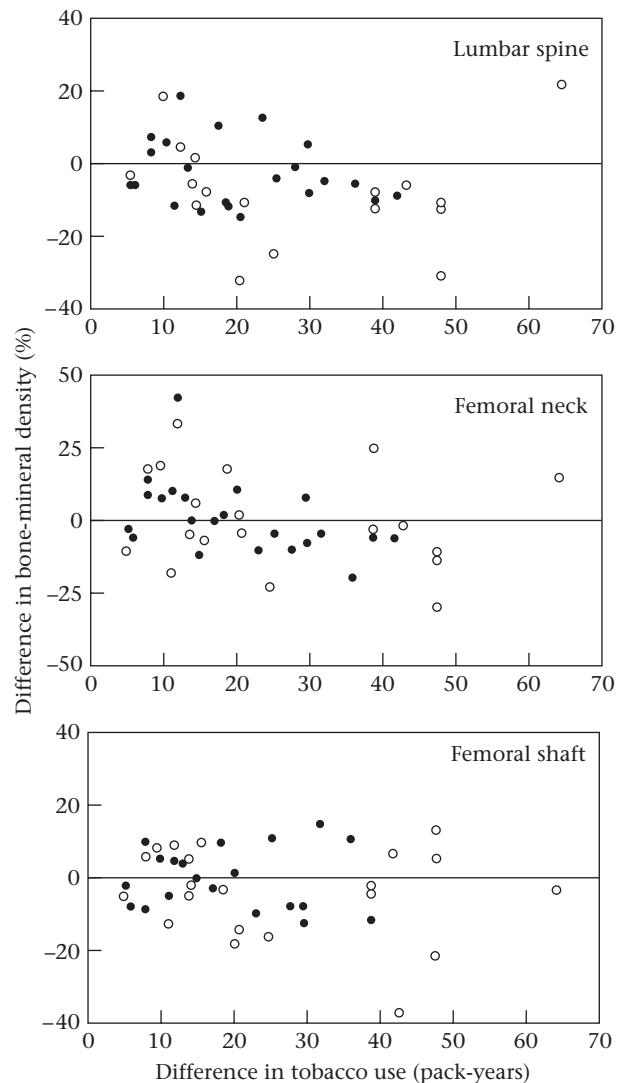
## 2.12 Summary

This chapter presented several **numeric and graphic methods for describing data**. These techniques are used to

- (1) quickly summarize a data set
- (2) and/or present results to others

In general, a data set can be described numerically in terms of a **measure of location** and a **measure of spread**. Several alternatives were introduced, including the **arithmetic mean**, **median**, **mode**, and **geometric mean**, as possible choices for measures of location, and the **standard deviation**, **quantiles**, and **range** as possible choices for measures of spread. Criteria were discussed for choosing the

**Figure 2.11** Within-pair differences in bone density at the lumbar spine, femoral neck, and femoral shaft as a function of within-pair differences in pack-years of tobacco use in 41 pairs of female twins. Monozygotic (identical) twins are represented by solid circles and dizygotic (fraternal) twins by open circles. The difference in bone density between members of a pair is expressed as the percentage of the mean bone density for the pair.



Source: From "The bone density of female twins discordant for tobacco use," by J. H. Hopper and E. Seeman, 1994, *The New England Journal of Medicine*, 330, 387–392. Copyright © 1994 Massachusetts Medical Society. All rights reserved.

appropriate measures in particular circumstances. Several graphic techniques for summarizing data, including traditional methods, such as the **bar graph**, and more modern methods characteristic of exploratory data analysis (EDA), such as the **stem-and-leaf plot** and **box plot**, were introduced.

How do the descriptive methods in this chapter fit in with the methods of statistical inference discussed later in this book? Specifically, if, based on some prespecified hypotheses, some interesting trends can be found using descriptive methods,

then we need some criteria to judge how “significant” these trends are. For this purpose, several commonly used **probability models** are introduced in Chapters 3 through 5 and approaches for testing the validity of these models using the methods of **statistical inference** are explored in Chapters 6 through 14.

## PROBLEMS

### Infectious Disease

The data in Table 2.11 are a sample from a larger data set collected on people discharged from a selected Pennsylvania hospital as part of a retrospective chart review of antibiotic usage in hospitals [7]. The data are also given in Data Set HOSPITAL.DAT with documentation in HOSPITAL.DOC on the Companion Website. Each data set on the Companion Website is available in six formats: ASCII, MINITAB-readable format, Excel-readable format, SAS-readable format, SPSS-readable format, and Stata-readable format.

**2.1** Compute the mean and median for the duration of hospitalization for the 25 patients.

**2.2** Compute the standard deviation and range for the duration of hospitalization for the 25 patients.

**2.3** It is of clinical interest to know if the duration of hospitalization is affected by whether a patient has received antibiotics. Answer this question descriptively using either numeric or graphic methods.

Suppose the scale for a data set is changed by multiplying each observation by a positive constant.

**Table 2.11 Hospital-stay data**

ID no.	Duration of hospital stay	Age	Sex 1 = M 2 = F	First temp. following admission	First WBC ( $\times 10^3$ ) following admission	Received antibiotic 1 = yes 2 = no	Received bacterial culture 1 = yes 2 = no	Service 1 = med. 2 = surg.
1	5	30	2	99.0	8	2	2	1
2	10	73	2	98.0	5	2	1	1
3	6	40	2	99.0	12	2	2	2
4	11	47	2	98.2	4	2	2	2
5	5	25	2	98.5	11	2	2	2
6	14	82	1	96.8	6	1	2	2
7	30	60	1	99.5	8	1	1	1
8	11	56	2	98.6	7	2	2	1
9	17	43	2	98.0	7	2	2	1
10	3	50	1	98.0	12	2	1	2
11	9	59	2	97.6	7	2	1	1
12	3	4	1	97.8	3	2	2	2
13	8	22	2	99.5	11	1	2	2
14	8	33	2	98.4	14	1	1	2
15	5	20	2	98.4	11	2	1	2
16	5	32	1	99.0	9	2	2	2
17	7	36	1	99.2	6	1	2	2
18	4	69	1	98.0	6	2	2	2
19	3	47	1	97.0	5	1	2	1
20	7	22	1	98.2	6	2	2	2
21	9	11	1	98.2	10	2	2	2
22	11	19	1	98.6	14	1	2	2
23	11	67	2	97.6	4	2	2	1
24	9	43	2	98.6	5	2	2	2
25	4	41	2	98.0	5	2	2	1

\***2.4** What is the effect on the median?

\***2.5** What is the effect on the mode?

**2.6** What is the effect on the geometric mean?

**2.7** What is the effect on the range?

\*Asterisk indicates that the answer to the problem is given in the Answer Section at the back of the book.

### Ophthalmology

Table 2.12 comes from a paper giving the distribution of astigmatism in 1033 young men ranging in age from 18 to 22 who were accepted for military service in Great Britain [8]. Assume that astigmatism is rounded to the nearest 10th of a diopter and each subject in a group has the average astigmatism within that group (e.g., for the group 0.2–0.3 diopters, the actual range is from 0.15 to 0.35 diopters), and assume that each man in the group has an astigmatism of  $(0.15 + 0.35)/2 = 0.25$  diopters.

**2.8** Compute the arithmetic mean.

**2.9** Compute the standard deviation.

**2.10** Construct a bar graph to display the distribution of astigmatism.

**2.11** Suppose we wish to construct an upper boundary for astigmatism so that no more than 5% of military recruits exceed this upper boundary. What is the smallest possible value for this upper boundary?

**Table 2.12 Distribution of astigmatism in 1033 young men age 18–22**

Degree of astigmatism (diopters)	Frequency
0.0 or less than 0.2	458
0.2–0.3	268
0.4–0.5	151
0.6–1.0	79
1.1–2.0	44
2.1–3.0	19
3.1–4.0	9
4.1–5.0	3
5.1–6.0	2
	1033

Source: Reprinted with permission of the Editor, the authors and the Journal from the *British Medical Journal*, May 7, 1394–1398, 1960.

### Cardiovascular Disease

The data in Table 2.13 are a sample of cholesterol levels taken from 24 hospital employees who were on a standard American diet and who agreed to adopt a vegetarian diet for 1 month. Serum-cholesterol measurements were made before adopting the diet and 1 month after.

**2.12** Compute the mean change in cholesterol.

**Table 2.13 Serum-cholesterol levels (mg/dL) before and after adopting a vegetarian diet**

Subject	Before	After	Difference*
1	195	146	49
2	145	155	-10
3	205	178	27
4	159	146	13
5	244	208	36
6	166	147	19
7	250	202	48
8	236	215	21
9	192	184	8
10	224	208	16
11	238	206	32
12	197	169	28
13	169	182	-13
14	158	127	31
15	151	149	2
16	197	178	19
17	180	161	19
18	222	187	35
19	168	176	-8
20	168	145	23
21	167	154	13
22	161	153	8
23	178	137	41
24	137	125	12

\*Before – after.

\***2.13** Compute the standard deviation of the change in cholesterol levels.

**2.14** Construct a stem-and-leaf plot of the cholesterol changes.

\***2.15** Compute the median change in cholesterol.

**2.16** Construct a box plot of the cholesterol changes to the right of the stem-and-leaf plot.

**2.17** Comment on the symmetry of the distribution of change scores based on your answers to Problems 2.12 through 2.16.

**2.18** Some investigators believe that the effects of diet on cholesterol are more evident in people with high rather than low cholesterol levels. If you split the data in Table 2.13 according to whether baseline cholesterol is above or below the median, can you comment descriptively on this issue?

### Hypertension

In an experiment that examined the effect of body position on blood pressure [9], 32 participants had their blood pressures

measured while lying down with their arms at their sides and again standing with their arms supported at heart level. The data are given in Table 2.14.

**2.19** Compute the arithmetic mean and median for the difference in systolic and diastolic blood pressure, respectively, taken in different positions (recumbent minus standing).

**2.20** Construct stem-and-leaf and box plots for the difference scores for each type of blood pressure.

**Table 2.14 Effect of position on blood pressure**

Participant	Blood pressure (mm Hg)			
	Recumbent, arm at side		Standing, arm at heart level	
B. R. A.	99 <sup>a</sup>	71 <sup>b</sup>	105 <sup>a</sup>	79 <sup>b</sup>
J. A. B.	126	74	124	76
F. L. B.	108	72	102	68
V. P. B.	122	68	114	72
M. F. B.	104	64	96	62
E. H. B.	108	60	96	56
G. C.	116	70	106	70
M. M. C.	106	74	106	76
T. J. F.	118	82	120	90
R. R. F.	92	58	88	60
C. R. F.	110	78	102	80
E. W. G.	138	80	124	76
T. F. H.	120	70	118	84
E. J. H.	142	88	136	90
H. B. H.	118	58	92	58
R. T. K.	134	76	126	68
W. E. L.	118	72	108	68
R. L. L.	126	78	114	76
H. S. M.	108	78	94	70
V. J. M.	136	86	144	88
R. H. P.	110	78	100	64
R. C. R.	120	74	106	70
J. A. R.	108	74	94	74
A. K. R.	132	92	128	88
T. H. S.	102	68	96	64
O. E. S.	118	70	102	68
R. E. S.	116	76	88	60
E. C. T.	118	80	100	84
J. H. T.	110	74	96	70
F. P. V.	122	72	118	78
P. F. W.	106	62	94	56
W. J. W.	146	90	138	94

<sup>a</sup>Systolic blood pressure

<sup>b</sup>Diastolic blood pressure

Source: C. E. Kossman (1946), "Relative importance of certain variables in the clinical determination of blood pressure," *American Journal of Medicine*, 1, 464–467. Reprinted with permission of the *American Journal of Medicine*.

**2.21** Based on your answers to Problems 2.19 and 2.20, comment on the effect of body position on the levels of systolic and diastolic blood pressure.

**2.22** Orthostatic hypertension is sometimes defined based on an unusual change in blood pressure after changing position. Suppose we define a normal range for change in systolic blood pressure (SBP) based on change in SBP from the recumbent to the standing position in Table 2.14 that is between the upper and lower decile. What should the normal range be?

### Pulmonary Disease

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. Data set FEV.DAT on the Companion Website contains determinations of FEV in 1980 on 654 children ages 3 through 19 who were seen in the Childhood Respiratory Disease (CRD) Study in East Boston, Massachusetts. These data are part of a longitudinal study to follow the change in pulmonary function over time in children [10].

The data in Table 2.15 are available for each child.

**Table 2.15 Format for FEV.DAT**

Column	Variable	Format or code
1–5	ID number	
7–8	Age (years)	
10–15	FEV (liters)	X.XXX
17–20	Height (inches)	XX.X
22	Sex	0 = female/1 = male
24	Smoking status	0 = noncurrent smoker/ 1 = current smoker

**2.23** For each variable (other than ID), obtain appropriate descriptive statistics (both numeric and graphic).

**2.24** Use both numeric and graphic measures to assess the relationship of FEV to age, height, and smoking status. (Do this separately for boys and girls.)

**2.25** Compare the pattern of growth of FEV by age for boys and girls. Are there any similarities? Any differences?

### Nutrition

The food-frequency questionnaire (FFQ) is an instrument often used in dietary epidemiology to assess consumption of specific foods. A person is asked to write down the number of servings per day typically eaten in the past year of over 100 individual food items. A food-composition table is then used to compute nutrient intakes (protein, fat, etc.) based on aggregating responses for individual foods. The FFQ is inexpensive to administer but is considered less accurate than the diet record (DR) (the gold standard of dietary epidemiology). For the DR, a participant writes down

the amount of each specific food eaten over the past week in a food diary and a nutritionist using a special computer program computes nutrient intakes from the food diaries. This is a much more expensive method of dietary recording. To validate the FFQ, 173 nurses participating in the Nurses' Health Study completed 4 weeks of diet recording about equally spaced over a 12-month period and an FFQ at the end of diet recording [11]. Data are presented in data set VALID.DAT on the Companion Website for saturated fat, total fat, total alcohol consumption, and total caloric intake for both the DR and FFQ. For the DR, average nutrient intakes were computed over the 4 weeks of diet recording. Table 2.16 shows the format of this file.

**Table 2.16 Format for VALID.DAT**

Column	Variable	Format or code
1–6	ID number	XXXXXX.XX
8–15	Saturated fat—DR (g)	XXXXXX.XX
17–24	Saturated fat—FFQ (g)	XXXXXX.XX
26–33	Total fat—DR (g)	XXXXXX.XX
35–42	Total fat—FFQ (g)	XXXXXX.XX
44–51	Alcohol consumption—DR (oz)	XXXXXX.XX
53–60	Alcohol consumption—FFQ (oz)	XXXXXX.XX
62–70	Total calories—DR	XXXXXXX.XX
72–80	Total calories—FFQ	XXXXXXX.XX

**2.26** Compute appropriate descriptive statistics for each nutrient for both DR and FFQ, using both numeric and graphic measures.

**2.27** Use descriptive statistics to relate nutrient intake for the DR and FFQ. Do you think the FFQ is a reasonably accurate approximation to the DR? Why or why not?

**2.28** A frequently used method for quantifying dietary intake is in the form of quintiles. Compute quintiles for each nutrient and each method of recording, and relate the nutrient composition for DR and FFQ using the quintile scale. (That is, how does the quintile category based on DR relate to the quintile category based on FFQ for the same individual?) Do you get the same impression about the concordance between DR and FFQ using quintiles as in Problem 2.27, in which raw (ungrouped) nutrient intake is considered?

In nutritional epidemiology, it is customary to assess nutrient intake in relation to total caloric intake. One measure used to accomplish this is *nutrient density*, which is defined as  $100\% \times (\text{caloric intake of a nutrient}/\text{total caloric intake})$ . For fat consumption, 1 g of fat is equivalent to 9 calories.

**2.29** Compute the nutrient density for total fat for the DR and FFQ, and obtain appropriate descriptive statistics for this variable. How do they compare?

**2.30** Relate the nutrient density for total fat for the DR versus the FFQ using the quintile approach in Problem 2.28. Is the concordance between total fat for DR and FFQ

stronger, weaker, or the same when total fat is expressed in terms of nutrient density as opposed to raw nutrient?

### Environmental Health, Pediatrics

In Section 2.9, we described Data Set LEAD.DAT (on the Companion Website) concerning the effect of lead exposure on neurological and psychological function in children.

**2.31** Compare the exposed and control groups regarding age and gender, using appropriate numeric and graphic descriptive measures.

**2.32** Compare the exposed and control groups regarding verbal and performance IQ, using appropriate numeric and graphic descriptive measures.

### Cardiovascular Disease

Activated-protein-C (APC) resistance is a serum marker that has been associated with thrombosis (the formation of blood clots often leading to heart attacks) among adults. A study assessed this risk factor among adolescents. To assess the reproducibility of the assay, a split-sample technique was used in which a blood sample was provided by 10 people; each sample was split into two aliquots (sub-samples), and each aliquot was assessed separately. Table 2.17 gives the results.

**Table 2.17 APC resistance split-samples data**

Sample number	A	B	A – B
1	2.22	1.88	0.34
2	3.42	3.59	-0.17
3	3.68	3.01	0.67
4	2.64	2.37	0.27
5	2.68	2.26	0.42
6	3.29	3.04	0.25
7	3.85	3.57	0.28
8	2.24	2.29	-0.05
9	3.25	3.39	-0.14
10	3.30	3.16	0.14

**2.33** Suppose the variation between split samples is thought to be a function of the mean level, where more variation is expected as the mean level increases. What measure of reproducibility can be used under these circumstances?

**2.34** Compute the measure in Problem 2.33 for each participant. Then obtain an average of this measure over the 10 participants. If this average is less than 10%, it indicates excellent reproducibility. Is this true for APC resistance? (Hint: Consider using Excel to answer this question.)

### Microbiology

A study was conducted to demonstrate that soy beans inoculated with nitrogen-fixing bacteria yield more and grow

adequately without expensive environmentally deleterious synthesized fertilizers. The trial was conducted under controlled conditions with uniform amounts of soil. The initial hypothesis was that inoculated plants would outperform their uninoculated counterparts. This assumption is based on the facts that plants need nitrogen to manufacture vital proteins and amino acids and that nitrogen-fixing bacteria would make more of this substance available to plants, increasing their size and yield. There were 8 inoculated plants (*I*) and 8 uninoculated plants (*U*). The plant yield as measured by pod weight for each plant is given in Table 2.18.

**2.35** Compute appropriate descriptive statistics for *I* and *U* plants.

**Table 2.18** Pod weight (g) from inoculated (*I*) and uninoculated (*U*) plants

Sample number	<i>I</i>	<i>U</i>
1	1.76	0.49
2	1.45	0.85
3	1.03	1.00
4	1.53	1.54
5	2.34	1.01
6	1.96	0.75
7	1.79	2.11
8	1.21	0.92

Note: The data for this problem were supplied by David Rosner.

**2.36** Use graphic methods to compare the two groups.

**2.37** What is your overall impression concerning the pod weight in the two groups?

### Endocrinology

In Section 2.10, we described Data Set BONEDEN.DAT (on the Companion Website) concerning the effect of tobacco use on BMD.

**2.38** For each pair of twins, compute the following for the lumbar spine:

$$A = \text{BMD for the heavier-smoking twin} - \text{BMD for the lighter-smoking twin} = x_1 - x_2$$

$$B = \text{mean BMD for the twinship} = (x_1 + x_2)/2$$

$$C = 100\% \times (A/B)$$

Derive appropriate descriptive statistics for *C* over the entire study population.

**2.39** Suppose we group the twin pairs according to the difference in tobacco use expressed in 10 pack-year groups (0–9.9 pack-years/10–19.9 pack-years/20–29.9 pack-years/30–39.9 pack-years/40+ pack-years). Compute appropriate descriptive statistics, and provide a scatter plot for *C* grouped by the difference in tobacco use in pack-years.

**2.40** What impression do you have of the relationship between BMD and tobacco use based on Problem 2.39?

**2.41–2.43** Answer Problems 2.38–2.40 for BMD for the femoral neck.

**2.44–2.46** Answer Problems 2.38–2.40 for BMD for the femoral shaft.

### REFERENCES

- [1] White, J. R., & Froeb, H. E. (1980). Small-airways dysfunction in nonsmokers chronically exposed to tobacco smoke. *New England Journal of Medicine*, 302(33), 720–723.
- [2] Pedersen, A., Wiesner, P., Holmes, K., Johnson, C., & Turck, M. (1972). Spectinomycin and penicillin G in the treatment of gonorrhea. *Journal of the American Medical Association*, 220(2), 205–208.
- [3] Foster, T. A., & Berenson, G. (1987). Measurement error and reliability in four pediatric cross-sectional surveys of cardiovascular disease risk factor variables—the Bogalusa Heart Study. *Journal of Chronic Diseases*, 40(1), 13–21.
- [4] Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- [5] Landrigan, P. J., Whitworth, R. H., Baloh, R. W., Staehling, N. W., Barthel, W. F., & Rosenblum, B. F. (1975, March 29). Neuropsychological dysfunction in children with chronic low-level lead absorption. *The Lancet*, 1, 708–715.
- [6] Hopper, J. H., & Seeman, E. (1994). The bone density of female twins discordant for tobacco use. *New England Journal of Medicine*, 330, 387–392.
- [7] Townsend, T. R., Shapiro, M., Rosner, B., & Kass, E. H. (1979). Use of antimicrobial drugs in general hospitals. I. Description of population and definition of methods. *Journal of Infectious Diseases*, 139(6), 688–697.
- [8] Sorsby, A., Sheridan, M., Leary, G. A., & Benjamin, B. (1960, May 7). Vision, visual acuity and ocular refraction of young men in a sample of 1033 subjects. *British Medical Journal*, 1(5183), 1394–1398.
- [9] Kossmann, C. E. (1946). Relative importance of certain variables in clinical determination of blood pressure. *American Journal of Medicine*, 1, 464–467.
- [10] Tager, I. B., Weiss, S. T., Rosner, B., & Speizer, F. E. (1979). Effect of parental cigarette smoking on pulmonary function in children. *American Journal of Epidemiology*, 110, 15–26.
- [11] Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H., & Speizer, F. E. (1985). Reproducibility and validity of a semi-quantitative food frequency questionnaire. *American Journal of Epidemiology*, 122, 51–65.

# 3

## Probability

### 3.1 Introduction

Chapter 2 outlined various techniques for concisely describing data. But we usually want to do more with data than just describe them. In particular, we might want to test certain specific inferences about the behavior of the data.

#### Example 3.1

**Cancer** One theory concerning the etiology of breast cancer states that women in a given age group who give birth to their first child relatively late in life (after age 30) are at greater risk for eventually developing breast cancer over some time period  $t$  than are women who give birth to their first child early in life (before age 20). Because women in upper social classes tend to have children later, this theory has been used to explain why these women have a higher risk of developing breast cancer than women in lower social classes. To test this hypothesis, we might identify 2000 women from a particular census tract who are currently ages 45–54 and have never had breast cancer, of whom 1000 had their first child before the age of 20 (call this group A) and 1000 after the age of 30 (group B). These 2000 women might be followed for 5 years to assess whether they developed breast cancer during this period. Suppose there are four new cases of breast cancer in group A and five new cases in group B.

Is this evidence enough to confirm a difference in risk between the two groups? Most people would feel uneasy about concluding that on the basis of such a limited amount of data.

Suppose we had a more ambitious plan and sampled 10,000 women each from groups A and B and at follow-up found 40 new cases in group A and 50 new cases in group B and asked the same question. Although we might be more comfortable with the conclusion because of the larger sample size, we would still have to admit that this apparent difference in the rates could be due to chance.

The problem is that we need a conceptual framework to make these decisions but have not explicitly stated what the framework is. This framework is provided by the underlying concept of **probability**. In this chapter, probability is defined and some rules for working with probabilities are introduced. Understanding probability is essential in calculating and interpreting  $p$ -values in the statistical tests of subsequent chapters. It also permits the discussion of sensitivity, specificity, and predictive values of screening tests in Section 3.7.

## 3.2 Definition of Probability

### Example 3.2

**Obstetrics** Suppose we are interested in the probability of a male live childbirth (or livebirth) among all livebirths in the United States. Conventional wisdom tells us this probability should be close to .5. We can explore this subject by looking at some vital-statistics data, as presented in Table 3.1 [1]. The probability of a male livebirth based on 1965 data is .51247; based on 1965–1969 data, .51248; and based on 1965–1974 data, .51268. These **empirical** probabilities are based on a finite amount of data. In principle, the sample size could be expanded indefinitely and an increasingly more precise estimate of the probability obtained.

**Table 3.1** Probability of a male livebirth during the period 1965–1974

Time period	Number of male livebirths (a)	Total number of livebirths (b)	Empirical probability of a male livebirth (a/b)
1965	1,927,054	3,760,358	.51247
1965–1969	9,219,202	17,989,361	.51248
1965–1974	17,857,857	34,832,051	.51268

This principle leads to the following definition of probability:

### Definition 3.1

The **sample space** is the set of all possible outcomes. In referring to probabilities of events, an **event** is any set of outcomes of interest. The **probability** of an event is the relative frequency (see p. 22) of this set of outcomes over an indefinitely large (or infinite) number of trials.

### Example 3.3

**Pulmonary Disease** The tuberculin skin test is a routine screening test used to detect tuberculosis. The results of this test can be categorized as either positive, negative, or uncertain. If the probability of a positive test is .1, it means that if a large number of such tests were performed, about 10% would be positive. The actual percentage of positive tests will be increasingly close to .1 as the number of tests performed increases.

### Example 3.4

**Cancer** The probability of developing a breast cancer over 30 years in 40-year-old women who have never had breast cancer is approximately 1/11. This probability means that over a large sample of 40-year-old women who have never had breast cancer, approximately 1 in 11 will develop the disease by age 70, with this proportion becoming increasingly close to 1 in 11 as the number of women sampled increases.

In real life, experiments cannot be performed an infinite number of times. Instead, probabilities of events are estimated from the empirical probabilities obtained from large samples (as in Examples 3.2–3.4). In other instances, theoretical-probability models are constructed from which probabilities of many different kinds of events can be computed. An important issue in statistical inference is to compare empirical probabilities with theoretical probabilities—that is, to assess the goodness-of-fit of probability models. This topic is covered in Section 10.7.

**Example 3.5**

**Cancer** The probability of developing stomach cancer over a 1-year period in 45- to 49-year-old women, based on SEER Tumor Registry data from 2002 to 2006, is 3.7 per 100,000 [2]. Suppose we have studied cancer rates in a small group of U.S. nurses over this period and want to compare how close the rates from this limited sample are to the tumor-registry figures. The value 3.7 per 100,000 would be the best estimate of the probability before collecting any data, and we would then see how closely our new sample data conformed with this probability.

From Definition 3.1 and from the preceding examples, we can deduce that probabilities have the following basic properties:

**Equation 3.1**

- (1) The probability of an event  $E$ , denoted by  $Pr(E)$ , always satisfies  $0 \leq Pr(E) \leq 1$ .
- (2) If outcomes  $A$  and  $B$  are two events that cannot both happen at the same time, then  $Pr(A \text{ or } B \text{ occurs}) = Pr(A) + Pr(B)$ .

**Example 3.6**

**Hypertension** Let  $A$  be the event that a person has normotensive diastolic blood-pressure (DBP) readings ( $DBP < 90$ ), and let  $B$  be the event that a person has borderline DBP readings ( $90 \leq DBP < 95$ ). Suppose that  $Pr(A) = .7$ , and  $Pr(B) = .1$ . Let  $Z$  be the event that a person has a  $DBP < 95$ . Then

$$Pr(Z) = Pr(A) + Pr(B) = .8$$

because the events  $A$  and  $B$  cannot occur at the same time.

**Definition 3.2**


---

Two events  $A$  and  $B$  are **mutually exclusive** if they cannot both happen at the same time.

---

Thus the events  $A$  and  $B$  in Example 3.6 are mutually exclusive.

**Example 3.7**

**Hypertension** Let  $X$  be DBP,  $C$  be the event  $X \geq 90$ , and  $D$  be the event  $75 \leq X \leq 100$ . Events  $C$  and  $D$  are *not* mutually exclusive, because they both occur when  $90 \leq X \leq 100$ .

### 3.3 Some Useful Probabilistic Notation

**Definition 3.3**


---

The symbol  $\{ \}$  is used as shorthand for the phrase “the event.”

---

**Definition 3.4**


---

$A \cup B$  is the event that either  $A$  or  $B$  occurs, or they both occur.

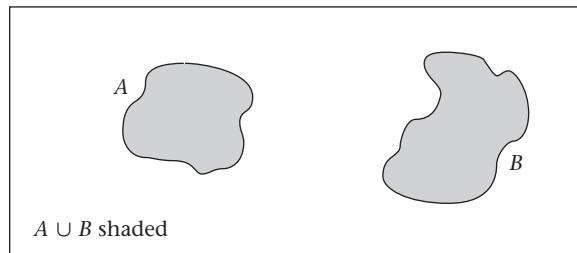
---

Figure 3.1 diagrammatically depicts  $A \cup B$  both for the case in which  $A$  and  $B$  are and are not mutually exclusive.

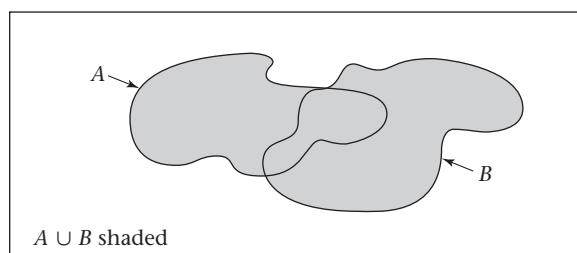
**Example 3.8**

**Hypertension** Let events  $A$  and  $B$  be defined as in Example 3.6:  $A = \{X < 90\}$ ,  $B = \{90 \leq X < 95\}$ , where  $X = \text{DBP}$ . Then  $A \cup B = \{X < 95\}$ .

**Figure 3.1** Diagrammatic representation of  $A \cup B$ : (a)  $A, B$  mutually exclusive; (b)  $A, B$  not mutually exclusive



(a)



(b)

**Example 3.9**

**Hypertension** Let events  $C$  and  $D$  be defined as in Example 3.7:

$$C = \{X \geq 90\} \quad D = \{75 \leq X \leq 100\}$$

$$\text{Then } C \cup D = \{X \geq 75\}$$

**Definition 3.5**

$A \cap B$  is the event that both  $A$  and  $B$  occur simultaneously.  $A \cap B$  is depicted diagrammatically in Figure 3.2.

**Example 3.10**

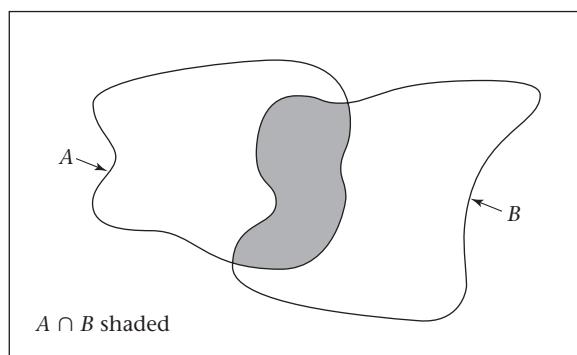
**Hypertension** Let events  $C$  and  $D$  be defined as in Example 3.7; that is,

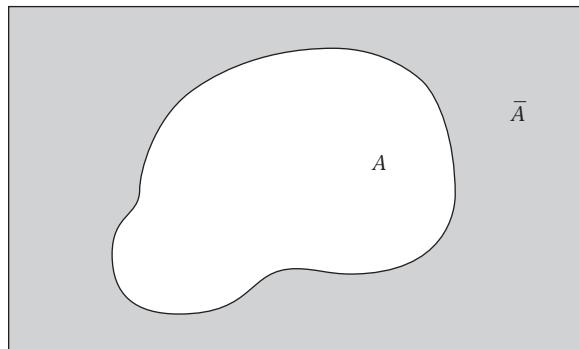
$$C = \{X \geq 90\} \quad D = \{75 \leq X \leq 100\}$$

$$\text{Then } C \cap D = \{90 \leq X \leq 100\}$$

**Figure 3.2**

Diagrammatic representation of  $A \cap B$



**Figure 3.3** Diagrammatic representation of  $\bar{A}$ 

Notice that  $A \cap B$  is not well defined for events  $A$  and  $B$  in Example 3.6 because both  $A$  and  $B$  cannot occur simultaneously. This is true for any mutually exclusive events.

**Definition 3.6**  $\bar{A}$  is the event that  $A$  does not occur. It is called the **complement** of  $A$ . Notice that  $Pr(\bar{A}) = 1 - Pr(A)$ , because  $\bar{A}$  occurs only when  $A$  does not occur. Event  $\bar{A}$  is diagrammed in Figure 3.3.

**Example 3.11** **Hypertension** Let events  $A$  and  $C$  be defined as in Examples 3.6 and 3.7; that is,

$$A = \{X < 90\} \quad C = \{X \geq 90\}$$

Then  $C = \bar{A}$ , because  $C$  can only occur when  $A$  does not occur. Notice that

$$Pr(C) = Pr(\bar{A}) = 1 - .7 = .3$$

Thus, if 70% of people have  $DBP < 90$ , then 30% of people must have  $DBP \geq 90$ .

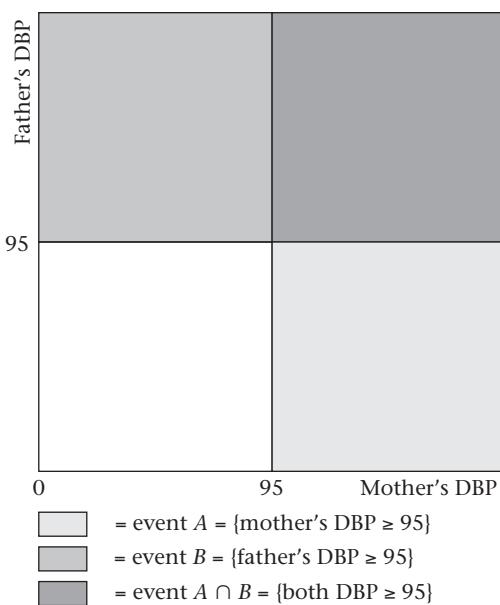
### 3.4 The Multiplication Law of Probability

In the preceding section, events in general were described. In this section, certain specific types of events are discussed.

**Example 3.12** **Hypertension, Genetics** Suppose we are conducting a hypertension-screening program in the home. Consider all possible pairs of DBP measurements of the mother and father within a given family, assuming that the mother and father are not genetically related. This sample space consists of all pairs of numbers of the form  $(X, Y)$  where  $X > 0$ ,  $Y > 0$ . Certain specific events might be of interest in this context. In particular, we might be interested in whether the mother or father is hypertensive, which is described, respectively, by events  $A = \{\text{mother's DBP} \geq 95\}$ ,  $B = \{\text{father's DBP} \geq 95\}$ . These events are diagrammed in Figure 3.4.

Suppose we know that  $Pr(A) = .1$ ,  $Pr(B) = .2$ . What can we say about  $Pr(A \cap B) = Pr(\text{mother's DBP} \geq 95 \text{ and father's DBP} \geq 95) = Pr(\text{both mother and father are hypertensive})$ ? We can say nothing unless we are willing to make certain assumptions.

**Figure 3.4** Possible diastolic blood-pressure measurements of the mother and father within a given family



**Definition 3.7** Two events  $A$  and  $B$  are called independent events if

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B)$$

**Example 3.13**

**Hypertension, Genetics** Compute the probability that both mother and father are hypertensive if the events in Example 3.12 are independent.

**Solution**

If  $A$  and  $B$  are independent events, then

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B) = .1(.2) = .02$$

One way to interpret this example is to assume that the hypertensive status of the mother does not depend at all on the hypertensive status of the father. Thus, if these events are independent, then in 10% of all households where the father is hypertensive the mother is also hypertensive, and in 10% of all households where the father is *not* hypertensive the mother is hypertensive. We would expect these two events to be independent if the primary determinants of elevated blood pressure were genetic. However, if the primary determinants of elevated blood pressure were, to some extent, environmental, then we would expect the mother would be more likely to have elevated blood pressure ( $A$  true) if the father had elevated blood pressure ( $B$  true) than if the father did not have elevated blood pressure ( $B$  not true). In this latter case the events would not be independent. The implications of this lack of independence are discussed later in this chapter.

If two events are not independent, then they are said to be dependent.

**Definition 3.8**

Two events  $A, B$  are **dependent** if

$$\Pr(A \cap B) \neq \Pr(A) \times \Pr(B)$$

Example 3.14 is a classic example of dependent events.

**Example 3.14**

**Hypertension, Genetics** Consider all possible DBP measurements from a mother and her first-born child. Let

$$A = \{\text{mother's DBP} \geq 95\} \quad B = \{\text{first-born child's DBP} \geq 80\}$$

$$\text{Suppose } Pr(A \cap B) = .05 \quad Pr(A) = .1 \quad Pr(B) = .2$$

$$\text{Then } Pr(A \cap B) = .05 > Pr(A) \times Pr(B) = .02$$

and the events  $A, B$  would be dependent.

This outcome would be expected because the mother and first-born child both share the same environment and are genetically related. In other words, the first-born child is more likely to have elevated blood pressure in households where the mother is hypertensive than in households where the mother is not hypertensive.

**Example 3.15**

**Sexually Transmitted Disease** Suppose two doctors, A and B, test all patients coming into a clinic for syphilis. Let events  $A^+ = \{\text{doctor A makes a positive diagnosis}\}$  and  $B^+ = \{\text{doctor B makes a positive diagnosis}\}$ . Suppose doctor A diagnoses 10% of all patients as positive, doctor B diagnoses 17% of all patients as positive, and both doctors diagnose 8% of all patients as positive. Are the events  $A^+, B^+$  independent?

**Solution**

We are given that

$$Pr(A^+) = .1 \quad Pr(B^+) = .17 \quad Pr(A^+ \cap B^+) = .08$$

$$\text{Thus } Pr(A^+ \cap B^+) = .08 > Pr(A^+) \times Pr(B^+) = .1(.17) = .017$$

and the events are dependent. This result would be expected because there should be a similarity between how two doctors diagnose patients for syphilis.

Definition 3.7 can be generalized to the case of  $k (> 2)$  independent events. This is often called the *multiplication law of probability*.

**Equation 3.2****Multiplication Law of Probability**

If  $A_1, \dots, A_k$  are mutually independent events,

$$\text{then } Pr(A_1 \cap A_2 \cap \dots \cap A_k) = Pr(A_1) \times Pr(A_2) \times \dots \times Pr(A_k)$$

### 3.5 The Addition Law of Probability

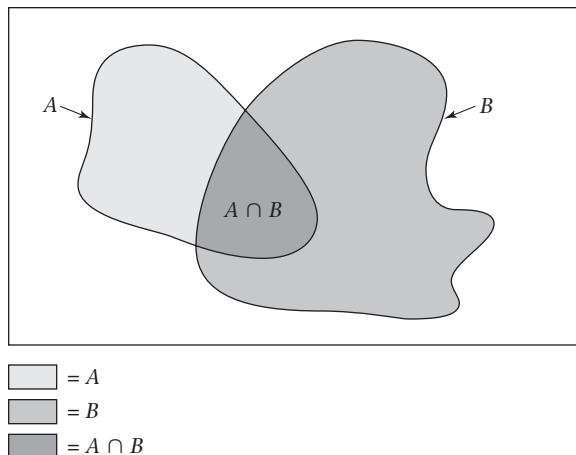
We have seen from the definition of probability that if  $A$  and  $B$  are mutually exclusive events, then  $Pr(A \cup B) = Pr(A) + Pr(B)$ . A more general formula for  $Pr(A \cup B)$  can be developed when events  $A$  and  $B$  are not necessarily mutually exclusive. This formula, the *addition law of probability*, is stated as follows:

**Equation 3.3****Addition Law of Probability**

If  $A$  and  $B$  are any events,

$$\text{then } Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

This principle is diagrammed in Figure 3.5. Thus, to compute  $Pr(A \cup B)$ , add the probabilities of  $A$  and  $B$  separately and then subtract the overlap, which is  $Pr(A \cap B)$ .

**Figure 3.5** Diagrammatic representation of the addition law of probability**Example 3.16**

**Sexually Transmitted Disease** Consider the data in Example 3.15. Suppose a patient is referred for further lab tests if either doctor A or B makes a positive diagnosis. What is the probability that a patient will be referred for further lab tests?

**Solution**

The event that either doctor makes a positive diagnosis can be represented by  $A^+ \cup B^+$ . We know that

$$Pr(A^+) = .1 \quad Pr(B^+) = .17 \quad Pr(A^+ \cap B^+) = .08$$

Therefore, from the addition law of probability,

$$Pr(A^+ \cup B^+) = Pr(A^+) + Pr(B^+) - Pr(A^+ \cap B^+) = .1 + .17 - .08 = .19$$

Thus 19% of all patients will be referred for further lab tests.

Special cases of the addition law are of interest. First, if events  $A$  and  $B$  are *mutually exclusive*, then  $Pr(A \cap B) = 0$  and the addition law reduces to  $Pr(A \cup B) = Pr(A) + Pr(B)$ . This property is given in Equation 3.1 for probabilities over any two mutually exclusive events. Second, if events  $A$  and  $B$  are *independent*, then by definition  $Pr(A \cap B) = Pr(A) \times Pr(B)$  and  $Pr(A \cup B)$  can be rewritten as  $Pr(A) + (B) - Pr(A) \times Pr(B)$ . This leads to the following important special case of the addition law.

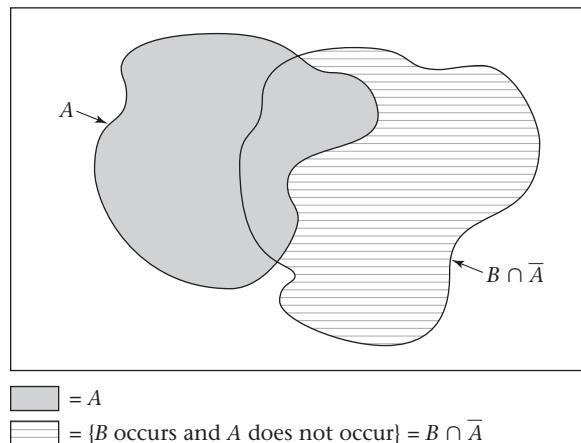
**Equation 3.4****Addition Law of Probability for Independent Events**

If two events  $A$  and  $B$  are independent, then

$$Pr(A \cup B) = Pr(A) + Pr(B) \times [1 - Pr(A)]$$

This special case of the addition law can be interpreted as follows: The event  $A \cup B$  can be separated into two mutually exclusive events: { $A$  occurs} and { $B$  occurs and  $A$  does not occur}. Furthermore, because of the independence of  $A$  and  $B$ , the probability of the latter event can be written as  $Pr(B) \times [1 - Pr(A)]$ . This probability is diagrammed in Figure 3.6.

**Figure 3.6** Diagrammatic representation of the addition law of probability for independent events



### Example 3.17

**Hypertension** Look at Example 3.12, where

$$A = \{\text{mother's DBP} \geq 95\} \quad \text{and} \quad B = \{\text{father's DBP} \geq 95\}$$

$Pr(A) = .1$ ,  $Pr(B) = .2$ , and assume  $A$  and  $B$  are independent events. Suppose a “hypertensive household” is defined as one in which either the mother or the father is hypertensive, with hypertension defined for the mother and father, respectively, in terms of events  $A$  and  $B$ . What is the probability of a hypertensive household?

### Solution

$Pr(\text{hypertensive household})$  is

$$Pr(A \cup B) = Pr(A) + Pr(B) \times [1 - Pr(A)] = .1 + .2(.9) = .28$$

Thus 28% of all households will be hypertensive.

It is possible to extend the addition law to more than two events. In particular, if there are three events  $A$ ,  $B$ , and  $C$ , then

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(A \cap C) - Pr(B \cap C) + Pr(A \cap B \cap C)$$

This result can be generalized to an arbitrary number of events, although that is beyond the scope of this text (see [3]).

## 3.6 Conditional Probability

Suppose we want to compute the probability of several events occurring simultaneously. If the events are independent, then we can use the multiplication law of probability to do so. If some of the events are dependent, then a quantitative measure of dependence is needed to extend the multiplication law to the case of dependent events. Consider the following example:

**Example 3.18**

**Cancer** Physicians recommend that all women over age 50 be screened for breast cancer. The definitive test for identifying breast tumors is a breast biopsy. However, this procedure is too expensive and invasive to recommend for *all* women over the age of 50. Instead, women in this age group are encouraged to have a mammogram every 1 to 2 years. Women with positive mammograms are then tested further with a biopsy. Ideally, the probability of breast cancer among women who are mammogram positive would be 1 and the probability of breast cancer among women who are mammogram negative would be 0. The two events {mammogram positive} and {breast cancer} would then be completely dependent; the results of the screening test would automatically determine the disease state. The opposite extreme is achieved when the events {mammogram positive} and {breast cancer} are completely independent. In this case, the probability of breast cancer would be the same regardless of whether the mammogram is positive or negative, and the mammogram would not be useful in screening for breast cancer and should not be used.

These concepts can be quantified in the following way. Let  $A = \{\text{mammogram}^+\}$ ,  $B = \{\text{breast cancer}\}$ , and suppose we are interested in the probability of breast cancer ( $B$ ) given that the mammogram is positive ( $A$ ). This probability can be written  $\Pr(A \cap B)/\Pr(A)$ .

**Definition 3.9**

The quantity  $\Pr(A \cap B)/\Pr(A)$  is defined as the **conditional probability of  $B$  given  $A$** , which is written  $\Pr(B|A)$ .

However, from Section 3.4 we know that, by definition of the multiplication law of probability, if two events are independent, then  $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$ . If both sides are divided by  $\Pr(A)$ , then  $\Pr(B) = \Pr(A \cap B)/\Pr(A) = \Pr(B|A)$ . Similarly, we can show that if  $A$  and  $B$  are independent events, then  $\Pr(B|\bar{A}) = \Pr(B|A) = \Pr(B)$ . This relationship leads to the following alternative interpretation of independence in terms of conditional probabilities.

**Equation 3.5**

- (1) If  $A$  and  $B$  are independent events, then  $\Pr(B|A) = \Pr(B) = \Pr(B|\bar{A})$ .
- (2) If two events  $A, B$  are dependent, then  $\Pr(B|A) \neq \Pr(B) \neq \Pr(B|\bar{A})$  and  $\Pr(A \cap B) \neq \Pr(A) \times \Pr(B)$ .

**Definition 3.10**

The **relative risk (RR)** of  $B$  given  $A$  is

$$\Pr(B|A)/\Pr(B|\bar{A})$$

Notice that if two events  $A, B$  are independent, then the *RR* is 1. If two events  $A, B$  are dependent, then the *RR* is different from 1. Heuristically, the more the dependence between events increases, the further the *RR* will be from 1.

**Example 3.19**

**Cancer** Suppose that among 100,000 women with negative mammograms 20 will be diagnosed with breast cancer within 2 years, or  $\Pr(B|\bar{A}) = 20/10^5 = .0002$ , whereas 1 woman in 10 with positive mammograms will be diagnosed with breast cancer within 2 years, or  $\Pr(B|A) = .1$ . The two events  $A$  and  $B$  would be highly dependent, because

$$RR = \Pr(B|A)/\Pr(B|\bar{A}) = .1/.0002 = 500$$

In other words, women with positive mammograms are 500 times more likely to develop breast cancer over the next 2 years than are women with negative mammograms. This is the rationale for using the mammogram as a screening test for breast cancer. If events  $A$  and  $B$  were independent, then the  $RR$  would be 1; women with positive or negative mammograms would be equally likely to have breast cancer, and the mammogram would not be useful as a screening test for breast cancer.

**Example 3.20**

**Sexually Transmitted Disease** Using the data in Example 3.15, find the conditional probability that doctor B makes a positive diagnosis of syphilis given that doctor A makes a positive diagnosis. What is the conditional probability that doctor B makes a positive diagnosis of syphilis given that doctor A makes a negative diagnosis? What is the  $RR$  of  $B^+$  given  $A^+$ ?

**Solution**

$$Pr(B^+|A^+) = Pr(B^+ \cap A^+)/Pr(A^+) = .08/.1 = .8$$

Thus doctor B will confirm doctor A's positive diagnoses 80% of the time. Similarly,

$$Pr(B^+|A^-) = Pr(B^+ \cap A^-)/Pr(A^-) = Pr(B^+ \cap A^-)/.9$$

We must compute  $Pr(B^+ \cap A^-)$ . We know that if doctor B diagnoses a patient as positive, then doctor A either does or does not confirm the diagnosis. Thus

$$Pr(B^+) = Pr(B^+ \cap A^+) + Pr(B^+ \cap A^-)$$

because the events  $B^+ \cap A^+$  and  $B^+ \cap A^-$  are mutually exclusive. If we subtract  $Pr(B^+ \cap A^+)$  from both sides of the equation, then

$$Pr(B^+ \cap A^-) = Pr(B^+) - Pr(B^+ \cap A^+) = .17 - .08 = .09$$

$$\text{Therefore, } Pr(B^+|A^-) = .09/.9 = .1$$

Thus, when doctor A diagnoses a patient as negative, doctor B will contradict the diagnosis 10% of the time. The  $RR$  of the event  $B^+$  given  $A^+$  is

$$Pr(B^+|A^+)/Pr(B^+|A^-) = .8/.1 = 8$$

This indicates that doctor B is 8 times as likely to diagnose a patient as positive when doctor A diagnoses the patient as positive than when doctor A diagnoses the patient as negative. These results quantify the dependence between the two doctors' diagnoses.

**REVIEW QUESTIONS 3A**

- 1 What is the frequency definition of probability?
- 2 What is the difference between independent and dependent events?
- 3 What are mutually exclusive events?
- 4 What is the addition law of probability?
- 5 What is conditional probability? How does it differ from unconditional probability?
- 6 What is relative risk? How do you interpret it?

**Total-Probability Rule**

The conditional ( $Pr(B|A), Pr(B|\bar{A})$ ) and unconditional ( $Pr(B)$ ) probabilities mentioned previously are related in the following way:

**Equation 3.6**

For any events  $A$  and  $B$ ,

$$Pr(B) = Pr(B|A) \times Pr(A) + Pr(B|\bar{A}) \times Pr(\bar{A})$$

This formula tells us that the unconditional probability of  $B$  is the sum of the conditional probability of  $B$  given  $A$  *times* the unconditional probability of  $A$  *plus* the conditional probability of  $B$  given  $A$  *not* occurring *times* the unconditional probability of  $A$  *not* occurring.

To derive this, we note that if the event  $B$  occurs, it must occur either with  $A$  or without  $A$ . Therefore,

$$Pr(B) = Pr(B \cap A) + Pr(B \cap \bar{A})$$

From the definition of conditional probability, we see that

$$Pr(B \cap A) = Pr(A) \times Pr(B|A)$$

and

$$Pr(B \cap \bar{A}) = Pr(\bar{A}) \times Pr(B|\bar{A})$$

By substitution, it follows that

$$Pr(B) = Pr(B|A)Pr(A) + Pr(B|\bar{A})Pr(\bar{A})$$

**Example 3.21**

**Cancer** Let  $A$  and  $B$  be defined as in Example 3.19, and suppose that 7% of the general population of women will have a positive mammogram. What is the probability of developing breast cancer over the next 2 years among women in the general population?

**Solution**

$$Pr(B) = Pr(\text{breast cancer})$$

$$\begin{aligned} &= Pr(\text{breast cancer} | \text{mammogram}^+) \times Pr(\text{mammogram}^+) \\ &\quad + Pr(\text{breast cancer} | \text{mammogram}^-) \times Pr(\text{mammogram}^-) \\ &= .1(.07) + .0002(.93) = .00719 = 719 / 10^5 \end{aligned}$$

Thus the unconditional probability of developing breast cancer over the next 2 years in the general population ( $719/10^5$ ) is a weighted average of the conditional probability of developing breast cancer over the next 2 years among women with a positive mammogram (.1) and the conditional probability of developing breast cancer over the next 2 years among women with a negative mammogram ( $20/10^5$ ).

In Equation 3.6 the probability of event  $B$  is expressed in terms of two mutually exclusive events  $A$  and  $\bar{A}$ . In many instances the probability of an event  $B$  will need to be expressed in terms of more than two mutually exclusive events, denoted by  $A_1, A_2, \dots, A_k$ .

**Definition 3.11**

A set of events  $A_1, \dots, A_k$  is exhaustive if at least one of the events must occur.

Assume that events  $A_1, \dots, A_k$  are mutually exclusive and exhaustive; that is, at least one of the events  $A_1, \dots, A_k$  must occur and no two events can occur simultaneously. Thus, exactly one of the events  $A_1, \dots, A_k$  must occur.

**Equation 3.7****Total-Probability Rule**

Let  $A_1, \dots, A_k$  be mutually exclusive and exhaustive events. The unconditional probability of  $B$  ( $Pr(B)$ ) can then be written as a weighted average of the conditional probabilities of  $B$  given  $A_i$  ( $Pr(B|A_i)$ ) as follows:

$$Pr(B) = \sum_{i=1}^k Pr(B|A_i) \times Pr(A_i)$$

To show this, we note that if  $B$  occurs, then it must occur together with one and only one of the events,  $A_1, \dots, A_k$ . Therefore,

$$Pr(B) = \sum_{i=1}^k Pr(B \cap A_i)$$

Also, from the definition of conditional probability,

$$Pr(B \cap A_i) = Pr(A_i) \times Pr(B|A_i)$$

By substitution, we obtain Equation 3.7.

An application of the total-probability rule is given in the following example:

**Example 3.22**

**Ophthalmology** We are planning a 5-year study of cataract in a population of 5000 people 60 years of age and older. We know from census data that 45% of this population is 60–64 years of age, 28% are 65–69 years of age, 20% are 70–74 years of age, and 7% are 75 or older. We also know from the Framingham Eye Study that 2.4%, 4.6%, 8.8%, and 15.3% of the people in these respective age groups will develop cataract over the next 5 years [4]. What percentage of the population in our study will develop cataract over the next 5 years, and how many people with cataract does this percentage represent?

**Solution**

Let  $A_1 = \{\text{ages 60–64}\}$ ,  $A_2 = \{\text{ages 65–69}\}$ ,  $A_3 = \{\text{ages 70–74}\}$ ,  $A_4 = \{\text{ages 75+}\}$ . These events are mutually exclusive and exhaustive because each person in our population must be in one and only one age group. Furthermore, from the conditions of the problem we know that  $Pr(A_1) = .45$ ,  $Pr(A_2) = .28$ ,  $Pr(A_3) = .20$ ,  $Pr(A_4) = .07$ ,  $Pr(B|A_1) = .024$ ,  $Pr(B|A_2) = .046$ ,  $Pr(B|A_3) = .088$ , and  $Pr(B|A_4) = .153$ , where  $B = \{\text{develop cataract in the next 5 years}\}$ . Finally, using the total-probability rule,

$$\begin{aligned} Pr(B) &= Pr(B|A_1) \times Pr(A_1) + Pr(B|A_2) \times Pr(A_2) \\ &\quad + Pr(B|A_3) \times Pr(A_3) + Pr(B|A_4) \times Pr(A_4) \\ &= .024(.45) + .046(.28) + .088(.20) + .153(.07) = .052 \end{aligned}$$

Thus 5.2% of this population will develop cataract over the next 5 years, which represents a total of  $5000 \times .052 = 260$  people with cataract.

The definition of conditional probability allows the multiplication law of probability to be extended to the case of dependent events.

**Equation 3.8****Generalized Multiplication Law of Probability**

If  $A_1, \dots, A_k$  are an arbitrary set of events, then

$$\begin{aligned} Pr(A_1 \cap A_2 \cap \dots \cap A_k) \\ = Pr(A_1) \times Pr(A_2|A_1) \times Pr(A_3|A_2 \cap A_1) \times \dots \times Pr(A_k|A_{k-1} \cap \dots \cap A_2 \cap A_1) \end{aligned}$$

If the events are independent, then the conditional probabilities on the right-hand side of Equation 3.8 reduce to unconditional probabilities and the generalized multiplication law reduces to the multiplication law for independent events given in Equation 3.2. Equation 3.8 also generalizes the relationship  $Pr(A \cap B) = Pr(A) \times Pr(B|A)$  given in Definition 3.9 for two events to the case of more than two events.

### REVIEW QUESTIONS 3B

- 1 What is the total-probability rule?
- 2 Suppose the rate of type II diabetes mellitus (DM) in 40- to 59-year-olds is 7% among Caucasians, 10% among African-Americans, 12% among Hispanics, and 5% among Asian-Americans. Suppose the ethnic distribution in Houston, Texas, among 40- to 59-year-olds is 30% Caucasian, 25% African-American, 40% Hispanic, and 5% Asian-American. What is the overall probability of type II DM among 40- to 59-year-olds in Houston?

## 3.7 Bayes' Rule and Screening Tests

The mammography test data given in Example 3.18 illustrate the general concept of the predictive value of a screening test, which can be defined as follows:

### Definition 3.12

The **predictive value positive ( $PV^+$ )** of a screening test is the probability that a person has a disease given that the test is positive.

$$Pr(\text{disease} | \text{test}^+)$$

The **predictive value negative ( $PV^-$ )** of a screening test is the probability that a person does *not* have a disease given that the test is negative.

$$Pr(\text{no disease} | \text{test}^-)$$

### Example 3.23

**Cancer** Find  $PV^+$  and  $PV^-$  for mammography given the data in Example 3.19.

### Solution

We see that  $PV^+ = Pr(\text{breast cancer} | \text{mammogram}^+) = .1$

whereas  $PV^- = Pr(\text{breast cancer} | \text{mammogram}^-)$

$$= 1 - Pr(\text{breast cancer} | \text{mammogram}^-) = 1 - .0002 = .9998$$

Thus, if the mammogram is negative, the woman is virtually certain *not* to develop breast cancer over the next 2 years ( $PV^- \approx 1$ ); whereas if the mammogram is positive, the woman has a 10% chance of developing breast cancer ( $PV^+ = .10$ ).

A symptom or a set of symptoms can also be regarded as a screening test for disease. The higher the  $PV$  of the screening test or symptoms, the more valuable the test will be. Ideally, we would like to find a set of symptoms such that both  $PV^+$  and  $PV^-$  are 1. Then we could accurately diagnose disease for each patient. However, this is usually impossible.

Clinicians often cannot directly measure the  $PV$  of a set of symptoms. However, they can measure how often specific symptoms occur in diseased and normal people. These measures are defined as follows:

### Definition 3.13

The **sensitivity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is present given that the person has a disease.

**Definition 3.14**

The **specificity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is *not* present given that the person does *not* have a disease.

**Definition 3.15**

A **false negative** is defined as a negative test result when the disease or condition being tested for is actually present. A **false positive** is defined as a positive test result when the disease or condition being tested for is not actually present.

For a symptom to be effective in predicting disease, it is important that both the sensitivity and specificity be high.

**Example 3.24**

**Cancer** Suppose the disease is lung cancer and the symptom is cigarette smoking. If we assume that 90% of people with lung cancer and 30% of people without lung cancer (essentially the entire general population) are smokers, then the sensitivity and specificity of smoking as a screening test for lung cancer are .9 and .7, respectively. Obviously, cigarette smoking cannot be used by itself as a screening criterion for predicting lung cancer because there will be too many false positives (people without cancer who are smokers).

**Example 3.25**

**Cancer** Suppose the disease is breast cancer in women and the symptom is having a family history of breast cancer (either a mother or a sister with breast cancer). If we assume 5% of women with breast cancer have a family history of breast cancer but only 2% of women without breast cancer have such a history, then the sensitivity of a family history of breast cancer as a predictor of breast cancer is .05 and the specificity is  $.98 = (1 - .02)$ . A family history of breast cancer cannot be used by itself to diagnose breast cancer because there will be too many false negatives (women with breast cancer who do not have a family history).

**REVIEW QUESTIONS 3C**

- 1 What is the sensitivity and specificity of a screening test?
- 2 What are the  $PV^+$  and  $PV^-$  of a screening test? How does  $PV$  differ from sensitivity and specificity?
- 3 The level of prostate-specific antigen (PSA) in the blood is frequently used as a screening test for prostate cancer. Punglia et al. [5] reported the following data regarding the relationship between a positive PSA test ( $\geq 4.1$  ng/dL) and prostate cancer.

**Table 3.2** Association between PSA and prostate cancer

PSA test result	Prostate cancer	Frequency
+	+	92
+	-	27
-	+	46
-	-	72

- (a) What are the sensitivity and specificity of the test?
- (b) What are the  $PV^+$  and  $PV^-$  of the test?

## Bayes' Rule

Review Question 3C.3 assumes that each PSA<sup>+</sup> and PSA<sup>-</sup> participant (or at least a representative sample of PSA<sup>+</sup> and PSA<sup>-</sup> participants) is evaluated for the presence of prostate cancer. Thus one can directly evaluate  $PV^+$  and  $PV^-$  from the data provided. Instead, in many screening studies, a random sample of cases and controls is obtained. One can estimate sensitivity and specificity from such a design. However, because cases are usually oversampled relative to the general population (e.g., if there are an equal number of cases and controls), one cannot directly estimate  $PV^+$  and  $PV^-$  from the frequency counts available in a typical screening study. Instead, an indirect method known as Bayes' rule is used for this purpose.

The general question then becomes how can the sensitivity and specificity of a symptom (or set of symptoms or diagnostic test), which are quantities a physician can estimate, be used to compute  $PVs$ , which are quantities a physician needs to make appropriate diagnoses?

Let  $A$  = symptom and  $B$  = disease. From Definitions 3.12, 3.13, and 3.14, we have

$$\text{Predictive value positive} = PV^+ = Pr(B|A)$$

$$\text{Predictive value negative} = PV^- = Pr(\bar{B}|\bar{A})$$

$$\text{Sensitivity} = Pr(A|B)$$

$$\text{Specificity} = Pr(\bar{A}|\bar{B})$$

Let  $Pr(B)$  = probability of disease in the reference population. We wish to compute  $Pr(B|A)$  and  $Pr(\bar{B}|\bar{A})$  in terms of the other quantities. This relationship is known as Bayes' rule.

### Equation 3.9

#### Bayes' Rule

Let  $A$  = symptom and  $B$  = disease.

$$PV^+ = Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\bar{B}) \times Pr(\bar{B})}$$

In words, this can be written as

$$PV^+ = \frac{\text{Sensitivity} \times x}{\text{Sensitivity} \times x + (1 - \text{Specificity}) \times (1 - x)}$$

where  $x = Pr(B)$  = prevalence of disease in the reference population. Similarly,

$$PV^- = \frac{\text{Specificity} \times (1 - x)}{\text{Specificity} \times (1 - x) + (1 - \text{Sensitivity}) \times x}$$

To derive this, we have, from the definition of conditional probability,

$$PV^+ = Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)}$$

Also, from the definition of conditional probability,

$$Pr(B \cap A) = Pr(A|B) \times Pr(B)$$

Finally, from the total-probability rule,

$$Pr(A) = Pr(A|B) \times Pr(B) + Pr(A|\bar{B}) \times Pr(\bar{B})$$

If the expressions for  $Pr(B \cap A)$  and  $Pr(A)$  are substituted into the equation for  $PV^+$ , we obtain

$$PV^+ = Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\bar{B}) \times Pr(\bar{B})}$$

That is,  $PV^+$  can be expressed as a function of sensitivity, specificity, and the probability of disease in the reference population. A similar derivation can be used to obtain  $PV^-$ .

### Example 3.26

**Hypertension** Suppose 84% of hypertensives and 23% of normotensives are classified as hypertensive by an automated blood-pressure machine. What are the  $PV^+$  and  $PV^-$  of the machine, assuming 20% of the adult population is hypertensive?

### Solution

The sensitivity = .84 and specificity =  $1 - .23 = .77$ . Thus, from Bayes' rule it follows that

$$\begin{aligned} PV^+ &= (.84)(.2) / [( .84)(.2) + (.23)(.8)] \\ &= .168 / .352 = .48 \end{aligned}$$

$$\begin{aligned} \text{Similarly, } PV^- &= (.77)(.8) / [(.77)(.8) + (.16)(.2)] \\ &= .616 / .648 = .95 \end{aligned}$$

Thus a negative result from the machine is reasonably predictive because we are 95% sure a person with a negative result from the machine is normotensive. However, a positive result is not very predictive because we are only 48% sure a person with a positive result from the machine is hypertensive.

Example 3.26 considered only two possible disease states: hypertensive and normotensive. In clinical medicine there are often more than two possible disease states. We would like to be able to predict the most likely disease state given a specific symptom (or set of symptoms). Let's assume that the probability of having these symptoms among people in a given disease state is known from clinical experience, as is the probability of each disease state in the reference population. This leads us to the generalized Bayes' rule:

### Equation 3.10

#### Generalized Bayes' Rule

Let  $B_1, B_2, \dots, B_k$  be a set of mutually exclusive and exhaustive disease states; that is, at least one disease state must occur and no two disease states can occur at the same time. Let  $A$  represent the presence of a symptom or set of symptoms. Then

$$Pr(B_i|A) = Pr(A|B_i) \times Pr(B_i) / \left[ \sum_{j=1}^k Pr(A|B_j) \times Pr(B_j) \right]$$

This result is obtained similarly to the result of Bayes' rule for two disease states in Equation 3.9. Specifically, from the definition of conditional probability, note that

$$Pr(B_i|A) = \frac{Pr(B_i \cap A)}{Pr(A)}$$

Also, from the definition of conditional probability,

$$Pr(B_i \cap A) = Pr(A|B_i) \times Pr(B_i)$$

From the total-probability rule,

$$Pr(A) = Pr(A|B_1) \times Pr(B_1) + \dots + Pr(A|B_k) \times Pr(B_k)$$

If the expressions for  $Pr(B_i \cap A)$  and  $Pr(A)$  are substituted, we obtain

$$Pr(B_i|A) = \frac{Pr(A|B_i) \times Pr(B_i)}{\sum_{j=1}^k Pr(A|B_j) \times Pr(B_j)}$$

### Example 3.27

**Pulmonary Disease** Suppose a 60-year-old man who has never smoked cigarettes presents to a physician with symptoms of a chronic cough and occasional breathlessness. The physician becomes concerned and orders the patient admitted to the hospital for a lung biopsy. Suppose the results of the lung biopsy are consistent either with lung cancer or with sarcoidosis, a fairly common, nonfatal lung disease. In this case

$$A = \{\text{chronic cough, results of lung biopsy}\}$$

Disease state  $\begin{cases} B_1 = \text{normal} \\ B_2 = \text{lung cancer} \\ B_3 = \text{sarcoidosis} \end{cases}$

Suppose that  $Pr(A|B_1) = .001$   $Pr(A|B_2) = .9$   $Pr(A|B_3) = .9$

and that in 60-year-old, never-smoking men

$$Pr(B_1) = .99 \quad Pr(B_2) = .001 \quad Pr(B_3) = .009$$

The first set of probabilities  $Pr(A|B_i)$  could be obtained from clinical experience with the previous diseases, whereas the latter set of probabilities  $Pr(B_i)$  would have to be obtained from age-, sex-, and smoking-specific prevalence rates for the diseases in question. The interesting question now becomes what are the probabilities  $Pr(B_i|A)$  of the three disease states given the previous symptoms?

### Solution

Bayes' rule can be used to answer this question. Specifically,

$$\begin{aligned} Pr(B_1|A) &= Pr(A|B_1) \times Pr(B_1) / \left[ \sum_{j=1}^3 Pr(A|B_j) \times Pr(B_j) \right] \\ &= .001(.99) / [ .001(.99) + .9(.001) + .9(.009) ] \\ &= .00099 / .00999 = .099 \\ Pr(B_2|A) &= .9(.001) / [ .001(.99) + .9(.001) + .9(.009) ] \\ &= .00090 / .00999 = .090 \\ Pr(B_3|A) &= .9(.009) / [ .001(.99) + .9(.001) + .9(.009) ] \\ &= .00810 / .00999 = .811 \end{aligned}$$

Thus, although the unconditional probability of sarcoidosis is very low (.009), the conditional probability of the disease given these symptoms and this age-sex-smoking group is .811. Also, although the symptoms and diagnostic tests are consistent with both lung cancer and sarcoidosis, the latter is much more likely among patients in this age-sex-smoking group.

**Example 3.28**

**Pulmonary Disease** Now suppose the patient in Example 3.27 smoked two packs of cigarettes per day for 40 years. Then assume  $Pr(B_1) = .98$ ,  $Pr(B_2) = .015$ , and  $Pr(B_3) = .005$  in this type of person. What are the probabilities of the three disease states for this type of patient, given these symptoms?

**Solution**

$$\begin{aligned} Pr(B_1|A) &= .001(.98)/[.001(.98)+.9(.015)+.9(.005)] \\ &= .00098/.01898=.052 \end{aligned}$$

$$Pr(B_2|A) = .9(.015)/.01898 = .01350/.01898 = .711$$

$$Pr(B_3|A) = .9(.005)/.01898 = .237$$

Thus in this type of patient lung cancer is the most likely diagnosis.

**REVIEW QUESTIONS 3D**

- 1 What is Bayes' rule? How is it used?
- 2 What is the generalized Bayes' rule?
- 3 Refer to Review Question 3B.2. Suppose a 40- to 59-year-old person in Houston has type II DM. What is the probability that this person is African-American? Hispanic? Caucasian? Asian-American? (*Hint:* Use the generalized Bayes' rule.)
- 4 Answer Review Question 3D.3 for a nondiabetic 40- to 59-year-old person in Houston.

**3.8 Bayesian Inference**

The definition of probability given in Definition 3.1 is sometimes called the **frequency definition of probability**. This definition forms the basis for the frequentist method of inference, which is the main approach to statistical inference featured in this book and used in statistical practice. However, Bayesian inference is an alternative definition of probability and inference, espoused by a vocal minority of statisticians. The Bayesian school of inference rejects the idea of the frequency definition of probability, considering that it is a theoretical concept that can never be realized in practice. Instead, Bayesians conceive of two types of probability: a prior probability and a posterior probability.

**Definition 3.16**

The **prior probability** of an event is the best guess by the observer of an event's likelihood in the absence of data. This prior probability may be a single number, or it may be a range of likely values for the probability, perhaps with weights attached to each possible value.

**Example 3.29**

**Hypertension** What is the prior probability of hypertension in Example 3.26?

**Solution**

The prior probability of hypertension in the absence of additional data is .20 because 20% of the adult population is hypertensive.

**Definition 3.17**

The **posterior probability** of an event is the likelihood that an event will occur after collecting some empirical data. It is obtained by integrating information from the prior probability with additional data related to the event in question.

**Example 3.30**

**Hypertension** What is the posterior probability of hypertension given that an automated blood-pressure machine has classified a person as hypertensive?

**Solution**

If we refer to Example 3.26 and let the event {true hypertensive} be denoted by  $B$  and the event {classified as hypertensive by an automated blood-pressure machine} be denoted by  $A$ , we see that the posterior probability is given by  $PV^+ = Pr(B|A) = .48$ .

**Example 3.31**

**Hypertension** What is the posterior probability of hypertension given that an automated blood-pressure machine has classified a person as normotensive?

**Solution**

The posterior probability  $= Pr(B|\bar{A}) = 1 - Pr(\bar{B}|\bar{A}) = 1 - PV^- = .05$ . Thus the initial prior probability of 20% has been integrated with the automated blood-pressure machine data to yield posterior probabilities of .48 and .05, for people who are classified as hypertensive and normotensive by the automated blood-pressure machine, respectively.

The main problem with Bayesian inference lies in specifying the prior probability. Two different people may provide different prior probabilities for an event and may reach different conclusions (obtain different posterior probabilities), even with the same data. However, in some cases the prior probability is well defined. Also, having sufficient data diminishes the impact of the prior probability on the posterior inference. Although most of the methodologies used in this book are based on the frequentist approach to inference, we use Bayesian inference in several additional examples in later chapters to provide further insight into data analysis.

## 3.9 ROC Curves

In some instances, a test provides several categories of response rather than simply providing positive or negative results. In some instances, the results of the test are reported as a continuous variable. In either case, designation of a cutoff point for distinguishing a test result as positive versus negative is arbitrary.

**Example 3.32**

**Radiology** The following data, provided by Hanley and McNeil [6], are ratings of computed tomography (CT) images by a single radiologist in a sample of 109 subjects with possible neurological problems. The true disease status is also known for

**Table 3.3 Ratings of 109 CT images by a single radiologist vs. true disease status**

True disease status	CT rating						Total
	Definitely normal (1)	Probably normal (2)	Questionable (3)	Probably abnormal (4)	Definitely abnormal (5)		
Normal	33	6	6	11	2		58
Abnormal	3	2	2	11	33		51
Total	36	8	8	22	35		109

each of these subjects. The data are presented in Table 3.3. How can we quantify the diagnostic accuracy of the test?

Unlike previous examples, this one has no obvious cutoff point to use for designating a subject as positive for disease based on the CT scan. For example, if we designate a subject as test-positive if he or she is either probably abnormal or definitely abnormal (a rating of 4 or 5, or 4+), then the sensitivity of the test is  $(11 + 33)/51 = 44/51 = .86$ , whereas the specificity is  $(33 + 6 + 6)/58 = 45/58 = .78$ . In Table 3.4, we compute the sensitivity and specificity of the radiologist's ratings according to different criteria for test-positive.

**Table 3.4** Sensitivity and specificity of the radiologist's ratings according to different test-positive criteria based on the data in Table 3.3

Test-positive criteria	Sensitivity	Specificity
1 +	1.0	0
2 +	.94	.57
3 +	.90	.67
4 +	.86	.78
5 +	.65	.97
6 +	0	1.0

To display these data, we construct a receiver operating characteristic (ROC) curve.

**Definition 3.18** A **receiver operating characteristic (ROC) curve** is a plot of the sensitivity versus  $(1 - \text{specificity})$  of a screening test, where the different points on the curve correspond to different cutoff points used to designate test-positive.

**Example 3.33**

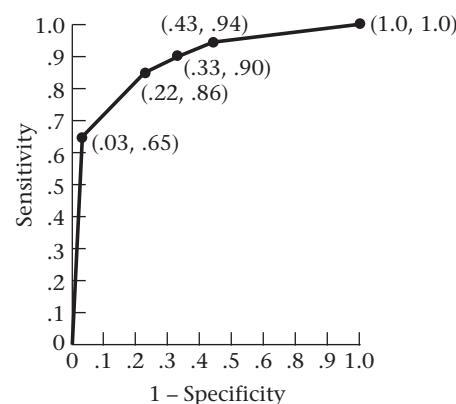
**Radiology** Construct an ROC curve based on the data in Table 3.4.

**Solution**

We plot sensitivity on the  $y$ -axis versus  $(1 - \text{specificity})$  on the  $x$ -axis using the data in Table 3.4. The plot is shown in Figure 3.7.

The area under the ROC curve is a reasonable summary of the overall diagnostic accuracy of the test. It can be shown [6] that this area, when calculated by the

**Figure 3.7** ROC curve for the data in Table 3.4\*



\*Each point represents  $(1 - \text{specificity}, \text{sensitivity})$  for different test-positive criteria.

trapezoidal rule, corresponds to the probability that for a randomly selected pair of normal and abnormal subjects, the radiologist will correctly identify the normal subject given the CT ratings. It is assumed that for untied ratings the radiologist designates the subject with the lower test score as normal and the subject with the higher test score as abnormal. For tied ratings, it is assumed that the radiologist randomly chooses one patient as normal and the other as abnormal.

**Example 3.34**

**Radiology** Calculate the area under the ROC curve in Figure 3.7, and interpret what it means.

**Solution**

The area under the ROC curve, when evaluated by the trapezoidal rule, is given by

$$\begin{aligned} &.5(.94+1.0)(.57)+.5(.90+.94)(.10)+.5(.86+.90)(.11)+.5(.65+.86)(.19) \\ &+.5(0+.65)(.03)=.89 \end{aligned}$$

This means the radiologist has an 89% probability of correctly distinguishing a normal from an abnormal subject based on the relative ordering of their CT ratings. For normal and abnormal subjects with the same ratings, it is assumed the radiologist selects one of the two subjects at random.

In general, of two screening tests for the same disease, the test with the higher area under its ROC curve is considered the better test, unless some particular level of sensitivity or specificity is especially important in comparing the two tests.

## 3.10 Prevalence and Incidence

In clinical medicine, the terms *prevalence* and *incidence* denote probabilities in a special context and are used frequently in this text.

**Definition 3.19**

The **prevalence** of a disease is the probability of currently having the disease regardless of the duration of time one has had the disease. Prevalence is obtained by dividing the number of people who currently have the disease by the number of people in the study population.

**Example 3.35**

**Hypertension** The prevalence of hypertension among adults (age 17 and older) was reported to be 20.3%, as assessed by the NHANES study conducted in 1999–2000 [7]. It was computed by dividing the number of people who had reported taking a prescription for hypertension and were 17 years of age and older (1225) by the total number of people 17 years of age and older in the study population (6044).

**Definition 3.20**

The **cumulative incidence** of a disease is the probability that a person with no prior disease will develop a new case of the disease over some specified time period.

In Chapter 14 we distinguish between *cumulative incidence*, which is defined over a long period of time, and *incidence density*, which is defined over a very short (or instantaneous) period of time. For simplicity, before Chapter 14 we use the abbreviated term *incidence* to denote *cumulative incidence*.

**Example 3.36**

**Cancer** The cumulative-incidence rate of breast cancer in 40- to 44-year-old U.S. women over the time period 2002–2006 was approximately 118.4 per 100,000 [2]. This means that on January 1, 2002, about 118 in 100,000 women 40 to 44 years of age who had never had breast cancer would develop breast cancer by December 31, 2002.

**REVIEW QUESTIONS 3E**

- 1** Suppose that of 25 students in a class, 5 are currently suffering from hay fever. Is the proportion 5 of 25 (20%) a measure of prevalence, incidence, or neither?
- 2** Suppose 50 HIV-positive men are identified, 5 of whom develop AIDS over the next 2 years. Is the proportion 5 of 50 (10%) a measure of prevalence, incidence, or neither?

**3.11 Summary**

In this chapter, probabilities and how to work with them using the addition and multiplication laws were discussed. An important distinction was made between independent events, which are unrelated to each other, and dependent events, which are related to each other. The general concepts of conditional probability and *RR* were introduced to quantify the dependence between two events. These ideas were then applied to the special area of screening populations for disease. In particular, the notions of sensitivity, specificity, and *PV*, which are used to define the accuracy of screening tests, were developed as applications of conditional probability. We also used an ROC curve to extend the concepts of sensitivity and specificity when the designation of the cutoff point for test-positive versus test-negative is arbitrary.

On some occasions, only sensitivities and specificities are available and we wish to compute the *PV* of screening tests. This task can be accomplished using Bayes' rule. The use of Bayes' rule in the context of screening tests is a special case of Bayesian inference. In Bayesian inference, we specify a prior probability for an event, which, after data are collected, is then modified to a posterior probability. Finally, prevalence and incidence, which are probabilistic parameters that are often used to describe the magnitude of disease in a population, were defined.

In the next two chapters, these general principles of probability are applied to derive some of the important probabilistic models often used in biomedical research, including the binomial, Poisson, and normal models. These models will eventually be used to test hypotheses about data.

**PROBLEMS**

Consider a family with a mother, father, and two children. Let  $A_1 = \{\text{mother has influenza}\}$ ,  $A_2 = \{\text{father has influenza}\}$ ,  $A_3 = \{\text{first child has influenza}\}$ ,  $A_4 = \{\text{second child has influenza}\}$ ,  $B = \{\text{at least one child has influenza}\}$ ,  $C = \{\text{at least one parent has influenza}\}$ , and  $D = \{\text{at least one person in the family has influenza}\}$ .

- \***3.1** What does  $A_1 \cup A_2$  mean?
- \***3.2** What does  $A_1 \cap A_2$  mean?
- \***3.3** Are  $A_3$  and  $A_4$  mutually exclusive?
- \***3.4** What does  $A_3 \cup B$  mean?
- \***3.5** What does  $A_3 \cap B$  mean?
- \***3.6** Express  $C$  in terms of  $A_1$ ,  $A_2$ ,  $A_3$ , and  $A_4$ .
- \***3.7** Express  $D$  in terms of  $B$  and  $C$ .
- \***3.8** What does  $\bar{A}_1$  mean?
- \***3.9** What does  $\bar{A}_2$  mean?

\***3.10** Represent  $\bar{C}$  in terms of  $A_1$ ,  $A_2$ ,  $A_3$ , and  $A_4$ .

\***3.11** Represent  $\bar{D}$  in terms of  $B$  and  $C$ .

Suppose an influenza epidemic strikes a city. In 10% of families the mother has influenza; in 10% of families the father has influenza; and in 2% of families both the mother and father have influenza.

**3.12** Are the events  $A_1 = \{\text{mother has influenza}\}$  and  $A_2 = \{\text{father has influenza}\}$  independent?

Suppose there is a 20% chance each child will get influenza, whereas in 10% of two-child families both children get the disease.

**3.13** What is the probability that at least one child will get influenza?

**3.14** Based on Problem 3.12, what is the conditional probability that the father has influenza given that the mother has influenza?

**Table 3.5 Prevalence of Alzheimer's disease (cases per 100 population)**

Age group	Males	Females
65–69	1.6	0.0
70–74	0.0	2.2
75–79	4.9	2.3
80–84	8.6	7.8
85 +	35.0	27.9

**3.15** Based on Problem 3.12, what is the conditional probability that the father has influenza given that the mother does not have influenza?

### Mental Health

Estimates of the prevalence of Alzheimer's disease have recently been provided by Pfeffer et al. [8]. The estimates are given in Table 3.5.

Suppose an unrelated 77-year-old man, 76-year-old woman, and 82-year-old woman are selected from a community.

**3.16** What is the probability that all three of these individuals have Alzheimer's disease?

**3.17** What is the probability that at least one of the women has Alzheimer's disease?

**3.18** What is the probability that at least one of the three people has Alzheimer's disease?

**3.19** What is the probability that exactly one of the three people has Alzheimer's disease?

**3.20** Suppose we know one of the three people has Alzheimer's disease, but we don't know which one. What is the conditional probability that the affected person is a woman?

**3.21** Suppose we know two of the three people have Alzheimer's disease. What is the conditional probability that they are both women?

**3.22** Suppose we know two of the three people have Alzheimer's disease. What is the conditional probability that they are both younger than 80 years of age?

Suppose the probability that both members of a married couple, each of whom is 75–79 years of age, will have Alzheimer's disease is .0015.

**3.23** What is the conditional probability that the man will be affected given that the woman is affected? How does this value compare with the prevalence in Table 3.5? Why should it be the same (or different)?

**3.24** What is the conditional probability that the woman will be affected given that the man is affected? How does this value compare with the prevalence in Table 3.5? Why should it be the same (or different)?

**3.25** What is the probability that at least one member of the couple is affected?

**Table 3.6 Age–sex distribution of retirement community**

Age group	Male (%) <sup>a</sup>	Female (%) <sup>a</sup>
65–69	5	10
70–74	9	17
75–79	11	18
80–84	8	12
85 +	4	6

<sup>a</sup>Percentage of total population.

Suppose a study of Alzheimer's disease is proposed in a retirement community with people 65+ years of age, where the age–sex distribution is as shown in Table 3.6.

**3.26** What is the expected overall prevalence of Alzheimer's disease in the community if the prevalence estimates in Table 3.5 for specific age–sex groups hold?

**3.27** If 1000 people 65+ years of age are in the community, then what is the expected number of cases of Alzheimer's disease in the community?

### Occupational Health

A study is conducted among men 50–69 years of age working in a chemical plant. We are interested in comparing the mortality experience of the workers in the plant with national mortality rates. Suppose that of the 500 plant workers in this age group, 35% are 50–54 years of age, 30% are 55–59, 20% are 60–64, and 15% are 65–69.

**\*3.28** If the annual national mortality rates are 0.9% in 50- to 54-year-old men, 1.4% in 55- to 59-year-old men, 2.2% in 60- to 64-year-old men, and 3.3% in 65- to 69-year-old men, then what is the projected annual mortality rate in the plant as a whole?

The standardized mortality ratio (SMR) is often used in occupational studies as a measure of risk. It is defined as 100% times the observed number of events in the exposed group divided by the expected number of events in the exposed group (based on some reference population).

**\*3.29** If 15 deaths are observed over 1 year among the 500 workers, what is the SMR?

### Genetics

Suppose that a disease is inherited via a **dominant mode of inheritance** and that only one of the two parents is affected with the disease. The implications of this mode of inheritance are that the probability is 1 in 2 that any particular offspring will get the disease.

**3.30** What is the probability that in a family with two children, both siblings are affected?

**3.31** What is the probability that exactly one sibling is affected?

**3.32** What is the probability that neither sibling is affected?

**3.33** Suppose the older child is affected. What is the probability that the younger child is affected?

**3.34** If  $A, B$  are two events such that  $A = \{\text{older child is affected}\}$ ,  $B = \{\text{younger child is affected}\}$ , then are the events  $A, B$  independent?

Suppose that a disease is inherited via an **autosomal recessive mode of inheritance**. The implications of this mode of inheritance are that the children in a family each have a probability of 1 in 4 of inheriting the disease.

**3.35** What is the probability that in a family with two children, both siblings are affected?

**3.36** What is the probability that exactly one sibling is affected?

**3.37** What is the probability that neither sibling is affected?

Suppose that a disease is inherited via a **sex-linked mode of inheritance**. The implications of this mode of inheritance are that each male offspring has a 50% chance of inheriting the disease, whereas the female offspring have no chance of getting the disease.

**3.38** In a family with one male and one female sibling, what is the probability that both siblings are affected?

**3.39** What is the probability that exactly one sibling is affected?

**3.40** What is the probability that neither sibling is affected?

**3.41** Answer Problem 3.38 for families with two male siblings.

**3.42** Answer Problem 3.39 for families with two male siblings.

**3.43** Answer Problem 3.40 for families with two male siblings.

Suppose that in a family with two male siblings, both siblings are affected with a genetically inherited disease. Suppose also that, although the genetic history of the family is unknown, only a dominant, recessive, or sex-linked mode of inheritance is possible.

**3.44** Assume that the dominant, recessive, and sex-linked modes of inheritance follow the probability laws given in Problems 3.30, 3.35, and 3.38 and that, without prior knowledge about the family in question, each mode of inheritance is equally likely. What is the posterior probability of each mode of inheritance in this family?

**3.45** Answer Problem 3.44 for a family with two male siblings in which only one sibling is affected.

**3.46** Answer Problem 3.44 for a family with one male and one female sibling in which both siblings are affected.

**3.47** Answer Problem 3.46 where only the male sibling is affected.

## Obstetrics

The following data are derived from the Monthly Vital Statistics Report (October 1999) issued by the National Center for Health Statistics [9]. These data are pertinent to livebirths only.

Suppose that infants are classified as low birthweight if they have a birthweight  $<2500$  g and as normal birthweight if they have a birthweight  $\geq 2500$  g. Suppose that infants are also classified by length of gestation in the following five categories: <28 weeks, 28–31 weeks, 32–35 weeks, 36 weeks, and  $\geq 37$  weeks. Assume the probabilities of the different periods of gestation are as given in Table 3.7.

**Table 3.7 Distribution of length of gestation**

Length of gestation	Probability
<28 weeks	.007
28–31 weeks	.012
32–35 weeks	.050
36 weeks	.037
$\geq 37$ weeks	.893

Also assume that the probability of low birthweight is .949 given a gestation of <28 weeks, .702 given a gestation of 28–31 weeks, .434 given a gestation of 32–35 weeks, .201 given a gestation of 36 weeks, and .029 given a gestation of  $\geq 37$  weeks.

**\*3.48** What is the probability of having a low birthweight infant?

**3.49** Show that the events {length of gestation  $\leq 31$  weeks} and {low birthweight} are not independent.

**\*3.50** What is the probability of having a length of gestation  $\leq 36$  weeks given that an infant is low birthweight?

## Pulmonary Disease

The familial aggregation of respiratory disease is a well-established clinical phenomenon. However, whether this aggregation is due to genetic or environmental factors or both is somewhat controversial. An investigator wishes to study a particular environmental factor, namely the relationship of cigarette-smoking habits in the parents to the presence or absence of asthma in their oldest child age 5 to 9 years living in the household (referred to below as their offspring). Suppose the investigator finds that (1) if both the mother and father are current smokers, then the probability of their offspring having asthma is .15; (2) if the mother is a current smoker and the father is not, then the probability of their offspring having asthma is .13; (3) if the father is a current smoker and the mother is not, then the probability of their offspring having asthma is .05; and (4) if neither parent is a current smoker, then the probability of their offspring having asthma is .04.

**\*3.51** Suppose the smoking habits of the parents are independent and the probability that the mother is a current smoker is .4, whereas the probability that the father is a current smoker is .5. What is the probability that both the father and mother are current smokers?

**\*3.52** Consider the subgroup of families in which the mother is not a current smoker. What is the probability that the father is a current smoker among such families? How does this probability differ from that calculated in Problem 3.51?

Suppose, alternatively, that if the father is a current smoker, then the probability that the mother is a current smoker is .6; whereas if the father is not a current smoker, then the probability that the mother is a current smoker is .2. Also assume that statements 1, 2, 3, and 4 above hold.

**\*3.53** If the probability that the father is a current smoker is .5, what is the probability that the father is a current smoker and that the mother is not a current smoker?

**\*3.54** Are the current smoking habits of the father and the mother independent? Why or why not?

**\*3.55** Under the assumptions made in Problems 3.53 and 3.54, find the unconditional probability that the offspring will have asthma.

**\*3.56** Suppose a child has asthma. What is the posterior probability that the father is a current smoker?

**\*3.57** What is the posterior probability that the mother is a current smoker if the child has asthma?

**\*3.58** Answer Problem 3.56 if the child does not have asthma.

**\*3.59** Answer Problem 3.57 if the child does not have asthma.

**\*3.60** Are the child's asthma status and the father's smoking status independent? Why or why not?

**\*3.61** Are the child's asthma status and the mother's smoking status independent? Why or why not?

### Pulmonary Disease

Smoking cessation is an important dimension in public-health programs aimed at preventing cancer and heart and lung diseases. For this purpose, data were accumulated starting in 1962 on a group of currently smoking men as part of the Normative Aging Study, a longitudinal study of the Veterans Administration in Boston. No interventions were attempted on this group of men, but the data in Table 3.8 were obtained as to annual quitting rates among initially healthy men who remained healthy during the entire period [10].

Note that the quitting rates increased from 1967 to 1970, which was around the time of the first U.S. Surgeon General's report on cigarette smoking.

**3.62** Suppose a man was a light smoker on January 1, 1962. What is the probability that he quit smoking by the end of 1975? (Assume he remained a light smoker until just prior to quitting and remained a quitter until 1975.)

**Table 3.8 Annual quitting rates of men who smoked, from the Normative Aging Study, 1962–1975**

Time period	Average annual quitting rate per 100 light smokers (< one pack per day)	Average annual quitting rate per 100 heavy smokers (> one pack per day)
1962–1966	3.1	2.0
1967–1970	7.1	5.0
1971–1975	4.7	4.1

**3.63** Answer Problem 3.62 for a heavy smoker on January 1, 1962. (Assume he remained a heavy smoker until just prior to quitting and remained a quitter until 1975.)

### Pulmonary Disease

Research into cigarette-smoking habits, smoking prevention, and cessation programs necessitates accurate measurement of smoking behavior. However, decreasing social acceptability of smoking appears to cause significant under-reporting. Chemical markers for cigarette use can provide objective indicators of smoking behavior. One widely used noninvasive marker is the level of saliva thiocyanate (SCN). In a Minneapolis school district, 1332 students in eighth grade (ages 12–14) participated in a study [11] whereby they

- (1) Viewed a film illustrating how recent cigarette use could be readily detected from small samples of saliva
- (2) Provided a personal sample of SCN
- (3) Provided a self-report of the number of cigarettes smoked per week

The results are given in Table 3.9.

**Table 3.9 Relationship between SCN levels and self-reported cigarettes smoked per week**

Self-reported cigarettes smoked in past week	Number of students	Percent with SCN $\geq 100 \mu\text{g/mL}$
None	1163	3.3
1–4	70	4.3
5–14	30	6.7
15–24	27	29.6
25–44	19	36.8
45 +	23	65.2

Source: Reprinted with permission from the *American Journal of Public Health*, 71(12), 1320, 1981.

Suppose the self-reports are completely accurate and are representative of the number of eighth-grade students who smoke in the general community. We are considering using

an SCN level  $\geq 100 \mu\text{g/mL}$  as a test criterion for identifying cigarette smokers. Regard a student as positive if he or she smokes one or more cigarettes per week.

\***3.64** What is the sensitivity of the test for light-smoking students (students who smoke  $\leq 14$  cigarettes per week)?

\***3.65** What is the sensitivity of the test for moderate-smoking students (students who smoke 15–44 cigarettes per week)?

\***3.66** What is the sensitivity of the test for heavy-smoking students (students who smoke  $\geq 45$  cigarettes per week)?

\***3.67** What is the specificity of the test?

\***3.68** What is the  $PV^+$  of the test?

\***3.69** What is the  $PV^-$  of the test?

Suppose we regard the self-reports of all students who report some cigarette consumption as valid but estimate that 20% of students who report no cigarette consumption actually smoke 1–4 cigarettes per week and an additional 10% smoke 5–14 cigarettes per week.

\***3.70** If we assume the percentage of students with SCN  $\geq 100 \mu\text{g/mL}$  in these two subgroups is the same as in those who truly report 1–4 and 5–14 cigarettes per week, then compute the specificity under these assumptions.

\***3.71** Compute the  $PV^-$  under these altered assumptions. How does the true  $PV^-$  using a screening criterion of SCN  $\geq 100 \mu\text{g/mL}$  for identifying smokers compare with the  $PV^-$  based on self-reports obtained in Problem 3.69?

## Hypertension

Laboratory measures of cardiovascular reactivity are receiving increasing attention. Much of the expanded interest is based on the belief that these measures, obtained under challenge from physical and psychological stressors, may yield a more biologically meaningful index of cardiovascular function than more traditional static measures. Typically, measurement of cardiovascular reactivity involves the use of an automated blood-pressure monitor to examine the changes in blood pressure before and after a stimulating experience (such as playing a video game). For this purpose, blood-pressure measurements were made with the Vita-Stat blood-pressure machine both before and after playing a video game. Similar measurements were obtained using manual methods for measuring blood pressure. A person was classified as a “reactor” if his or her DBP increased by 10 mm Hg or more after playing the game and as a non-reactor otherwise. The results are given in Table 3.10.

**3.72** If the manual measurements are regarded as the “true” measure of reactivity, then what is the sensitivity of automated DBP measurements?

**3.73** What is the specificity of automated DBP measurements?

**3.74** If the population tested is representative of the general population, then what are the  $PV^+$  and  $PV^-$  using this test?

**Table 3.10 Classification of cardiovascular reactivity using an automated and a manual sphygmomanometer**

	$\Delta\text{DBP}$ , manual	
$\Delta\text{DBP}$ , automated	<10	$\geq 10$
<10	51	7
$\geq 10$	15	6

## Otolaryngology

The data set in Table 3.11 is based on 214 children with acute otitis media (otitis media with effusion, or OME) who participated in a randomized clinical trial [12]. Each child had OME at the beginning of the study in either one (unilateral cases) or both (bilateral cases) ears and was randomly assigned to receive a 14-day course of one of two antibiotics, either cefaclor (CEF) or amoxicillin (AMO). The data here concern the 203 children whose middle-ear status was determined during a 14-day follow-up visit. The data in Table 3.11 are presented in data set EAR.DAT (on the Companion Website).

**3.75** Does there seem to be any difference in the effect of the antibiotics on clearance of otitis media? Express your results in terms of *relative risk (RR)*. Consider separate analyses for unilateral and bilateral cases. Also consider an analysis combining the two types of cases.

**3.76** The investigators recorded the ages of the children because they felt this might be an important factor in determining outcome. Were they right? Try to express your results in terms of *RR*.

**3.77** While controlling for age, propose an analysis comparing the effectiveness of the two antibiotics. Express your results in terms of *RR*.

**3.78** Another issue in this trial is the possible dependence between ears for the bilateral cases. Comment on this issue based on the data collected.

The concept of a **randomized clinical trial** is discussed more completely in Chapter 6. The analysis of **contingency-table data** is studied in Chapters 10 and 13, in which many of the formal methods for analyzing this type of data are discussed.

**Table 3.11 Format for EAR.DAT**

Column	Variable	Format or code
1–3	ID	
5	Clearance by 14 days	1 = yes/0 = no
7	Antibiotic	1 = CEF/2 = AMO
9	Age	1 = < 2 yrs/2 = 2–5 yrs 3 = 6+ yrs
11	Ear	1 = 1st ear/2 = 2nd ear

## Gynecology

A drug company is developing a new pregnancy-test kit for use on an outpatient basis. The company uses the pregnancy test on 100 women who are known to be pregnant, for whom 95 test results are positive. The company uses the pregnancy test on 100 other women who are known to *not* be pregnant, of whom 99 test negative.

**\*3.79** What is the sensitivity of the test?

**\*3.80** What is the specificity of the test?

The company anticipates that of the women who will use the pregnancy-test kit, 10% will actually be pregnant.

**\*3.81** What is the  $PV^+$  of the test?

**\*3.82** Suppose the “cost” of a false negative ( $2c$ ) is twice that of a false positive ( $c$ ) (because for a false negative prenatal care would be delayed during the first trimester of pregnancy). If the standard home pregnancy-test kit (made by another drug company) has a sensitivity of .98 and a specificity of .98, then which test (the new or standard) will cost the least per woman using it in the general population and by how much?

## Mental Health

The Chinese Mini-Mental Status Test (CMMS) consists of 114 items intended to identify people with Alzheimer’s disease and senile dementia among people in China [13]. An extensive clinical evaluation of this instrument was performed, whereby participants were interviewed by psychiatrists and nurses and a definitive diagnosis of dementia was made. Table 3.12 shows the results obtained for the subgroup of people with at least some formal education.

Suppose a cutoff value of  $\leq 20$  on the test is used to identify people with dementia.

**3.83** What is the sensitivity of the test?

**3.84** What is the specificity of the test?

**Table 3.12 Relationship of clinical dementia to outcome on the Chinese Mini-Mental Status Test**

CMMS score	Nondemented	Demented
0–5	0	2
6–10	0	1
11–15	3	4
16–20	9	5
21–25	16	3
26–30	18	1
Total	46	16

**3.85** The cutoff value of 20 on the CMMS used to identify people with dementia is arbitrary. Suppose we consider changing the cutoff. What are the sensitivity and specificity if cutoffs of 5, 10, 15, 20, 25, or 30 are used? Make a table of your results.

**3.86** Construct an ROC curve based on the table constructed in Problem 3.85.

**3.87** Suppose we want both the sensitivity and specificity to be at least 70%. Use the ROC curve to identify the possible value(s) to use as the cutoff for identifying people with dementia, based on these criteria.

**3.88** Calculate the area under the ROC curve. Interpret what this area means in words in the context of this problem.

## Demography

A study based on data collected from the Medical Birth Registry of Norway looked at fertility rates according to survival outcomes of previous births [14]. The data are presented in Table 3.13.

**3.89** What is the probability of having a livebirth (L) at a second birth given that the outcome of the first pregnancy was a stillbirth (D), that is, death?

**3.90** Answer Problem 3.89 if the outcome of the first pregnancy was a livebirth.

**3.91** What is the probability of 0, 1, and 2+ additional pregnancies if the first birth was a stillbirth?

**3.92** Answer Problem 3.91 if the first birth was a live birth.

## Mental Health

The  $\epsilon 4$  allele of the gene encoding apolipoprotein E (APOE) is strongly associated with Alzheimer’s disease, but its value in making the diagnosis remains uncertain. A study was conducted among 2188 patients who were evaluated at autopsy for Alzheimer’s disease by previously established pathological criteria [15]. Patients were also evaluated clinically for the presence of Alzheimer’s disease. The data in Table 3.14 were presented.

Suppose the pathological diagnosis is considered the gold standard for Alzheimer’s disease.

**3.93** If the clinical diagnosis is considered a screening test for Alzheimer’s disease, then what is the sensitivity of this test?

**3.94** What is the specificity of this test?

To possibly improve on the diagnostic accuracy of the clinical diagnosis for Alzheimer’s disease, information on both the APOE genotype as well as the clinical diagnosis were considered. The data are presented in Table 3.15.

Suppose we consider the combination of both a clinical diagnosis for Alzheimer’s disease and the presence of  $\geq 1 \epsilon 4$  allele as a screening test for Alzheimer’s disease.

**3.95** What is the sensitivity of this test?

**3.96** What is the specificity of this test?

**Table 3.13 Relationship of fertility rates to survival outcome of previous births in Norway**

	First birth	Continuing to second birth	Second birth outcome	Continuing to third birth	Third birth outcome
	n	n	n	n	n
Perinatal outcome					
D	7022	5924	D 368 L 5556	277 3916	D 39 L 238 D 115 L 3801
L	350,693	265,701	D 3188 L 262,513	2444 79,450	D 140 L 2304 D 1005 L 78,445

Note: D = dead, L = alive at birth and for at least one week.

**Table 3.14 Relationship between clinical and pathological diagnoses of Alzheimer's disease**

Pathological diagnosis		
Clinical diagnosis	Alzheimer's disease	Other causes of dementia
Alzheimer's disease	1643	190
Other causes of dementia	127	228

### Cardiovascular Disease

A fascinating subject of recent interest is the “Hispanic paradox”: Census data “show” that coronary heart disease (CHD) has a lower prevalence in Hispanic people than in non-Hispanic whites (NHW) based on health interviews of representative samples of people from different ethnic groups from the U.S. population, although the risk-factor profile of Hispanics is generally worse (more hypertension, diabetes, and obesity in this group than in NHW). To study this further, researchers looked at a group of 1000 Hispanic men ages 50–64 from several counties in Texas who were free of CHD in 1990 and followed them for 5 years. They found that 100 of the men had developed CHD

(either fatal cases or nonfatal cases in which the men survived a heart attack).

**3.97** Is the proportion 100 out of 1000 a prevalence rate, an incidence rate, or neither?

Given other surveys over the same time period among NHW in these counties, the researchers expected that the comparable rate of CHD for NHW would be 8%.

Another important parameter in the epidemiology of CHD is the *case-fatality rate* (the proportion of people who die among those who have a heart attack). Among the 100 CHD cases ascertained among Hispanics, 50 were fatal.

**3.98** What is the expected proportion of Hispanic men who will be identified by health surveys as having a previous heart attack in the past 5 years (who are by definition survivors) if we assume that the proportion of men with more than one nonfatal heart attack is negligible? What is the comparable proportion for NHW men if the expected case-fatality rate is 20% among NHW men with CHD?

**3.99** Are these proportions prevalence rates, incidence rates, or neither? Do the results in this problem give insight into why the Hispanic paradox occurs (do Hispanic men truly have lower risk of CHD as government surveys would indicate)? Why or why not?

**Table 3.15 Influence of the APOE genotype in diagnosing Alzheimer's disease (AD)**

APOE genotype	Both clinical and pathological criteria for AD	Only clinical criteria for AD	Only pathological criteria for AD	Neither clinical nor pathological criteria for AD
≥1 ε4 allele	1076	66	66	67
No ε4 allele	567	124	61	161
Total	1643	190	127	228

## Genetics

A dominantly inherited genetic disease is identified over several generations of a large family. However, about half the families have dominant disease with *complete penetrance*, whereby if a parent is affected there is a 50% probability that any one offspring will be affected. Similarly, about half the families have dominant disease with *reduced penetrance*, whereby if a parent is affected there is a 25% probability that any one offspring will be affected.

Suppose in a particular family one parent and two of the two offspring are affected.

**3.100** What is the probability that exactly two of the two offspring will be affected in a family with dominant disease with complete penetrance?

**3.101** What is the probability that exactly two of the two offspring will be affected in a family with dominant disease with reduced penetrance?

**3.102** What is the probability that the mode of transmission for this particular family is dominant with complete penetrance? Is this a prior probability or a posterior probability?

**3.103** Suppose you are a genetic counselor and are asked by the parents what the probability is that if they have another (a third) child he or she will be affected by the disease. What is the answer?

## SIMULATION—CLASS PROJECT

### Infectious Disease

Suppose a standard antibiotic kills a particular type of bacteria 80% of the time. A new antibiotic is reputed to have better efficacy than the standard antibiotic. Researchers propose to try the new antibiotic on 100 patients infected with the bacteria. Using principles of hypothesis testing (covered in Chapter 7), researchers will deem the new antibiotic “significantly better” than the standard one if it kills the bacteria in at least 88 out of the 100 infected patients.

**3.104** Suppose there is a true probability (true efficacy) of 85% that the new antibiotic will work *for an individual patient*. Perform a “simulation study” on the computer, based on random number generation (using, for example, MINITAB or Excel) for a group of 100 randomly simulated patients. Repeat this exercise 20 times with separate columns for each simulated sample of 100 patients. For what percentage of the 20 samples is the new antibiotic considered “significantly better” than the standard antibiotic? (This percentage is referred to as the *statistical power* of the experiment.) Compare results for different students in the class.

**3.105** Repeat the procedure in Problem 3.104 for each simulated patient, assuming the true efficacy of the new antibiotic is (a), 80%, (b) 90%, and (c) 95%, and compute the statistical power for each of (a), (b), and (c).

**3.106** Plot the statistical power versus the true efficacy. Do you think 100 patients is a sufficiently large sample to discover whether the new drug is “significantly better” if the true efficacy of the drug is 90%? Why or why not?

## Infectious Disease, Cardiovascular Disease

A validation study is to be performed in a local hospital to check the accuracy of assessment of hospital-acquired infection (INF) following coronary bypass surgery (coronary-artery bypass graft, or CABG). In a given year the hospital performs 1100 CABG procedures. A Centers for Disease Control and Prevention (CDC) algorithm is currently used to categorize subjects as having INF. To validate this algorithm, all CDC<sup>+</sup> subjects ( $N = 100$ ) and a random sample of CDC<sup>-</sup> subjects ( $N = 1000$ ) will be ascertained by an infectious-disease (ID) fellow and a detailed investigation will be performed, including a chart review and documentation of antibiotic use. Assume the ID-fellow’s determination is correct.

Suppose 100 CDC<sup>+</sup> subjects are ascertained, of whom the ID fellow confirms 80. Because there are a large number of CDC<sup>-</sup> subjects (1000), only a sample of 100 is studied, of whom the ID fellow confirms 90.

**3.107** What is the  $PV^+$  of the CDC algorithm?

**3.108** What is the  $PV^-$  of the CDC algorithm?

**3.109** What is the sensitivity of the CDC algorithm?

**3.110** What is the specificity of the CDC algorithm?

### Genetics

Suppose a birth defect has a recessive form of inheritance. In a study population, the recessive gene (a) initially has a prevalence of 25%. A subject has the birth defect if both maternal and paternal genes are of type a.

**3.111** In the general population, what is the probability that an individual will have the birth defect, assuming that maternal and paternal genes are inherited independently?

A further study finds that after 10 generations ( $\approx 200$  years) a lot of inbreeding has taken place in the population. Two subpopulations (populations A and B), consisting of 30% and 70% of the general population, respectively, have formed. Within population A, prevalence of the recessive gene is 40%, whereas in population B it is 10%.

**3.112** Suppose that in 25% of marriages both people are from population A, in 65% both are from population B, and in 10% there is one partner from population A and one from population B. What is the probability of a birth defect in the next generation?

**3.113** Suppose that a baby is born with a birth defect, but the baby’s ancestry is unknown. What is the posterior probability that the baby will have both parents from population A, both parents from population B, or mixed ancestry, respectively? (*Hint:* Use Bayes’ rule.)

## Orthopedics

Piriformis syndrome is a pelvic condition that involves malfunction of the piriformis muscle (a deep buttock muscle), which often causes back and buttock pain with sciatica (pain radiating down the leg). An electrophysiologic test to detect piriformis syndrome involves measuring nerve-conduction velocity (NCV) at two nerves in the leg (the tibial and peroneal nerves) with the leg flexed in a specific position. Increases in NCV in these nerves are often associated with piriformis syndrome. The resulting test, called the flexion abduction and internal rotation (FAIR) test, is positive if the average NCV in these nerves is delayed by 2+ seconds relative to normal.

A small study compared the FAIR test results with patient self-reports of how they feel on a visual analog scale (VAS) of 0–10, with 0 indicating no pain and 10 very severe pain. The results were as shown in Table 3.16.

**Table 3.16 FAIR test results on piriformis syndrome patients**

Clinical response	VAS	FAIR $\geq 2$	FAIR $< 2$	Total
Best	$\leq 2$	5	14	19
	3–4	3	12	15
	5–6	7	6	13
Worst	$\geq 7$	7	6	13
Total		22	38	60

Suppose physicians consider the FAIR test the gold standard, with a FAIR test result of  $\geq 2$  defined as a true positive and a FAIR test result of  $< 2$  defined as a true negative. Suppose a VAS of  $\leq 4$  is considered a good clinical response based on self-report (a test-negative) and a VAS of  $\geq 5$  is considered a bad clinical response (a test-positive).

**3.114** What is the sensitivity of the VAS?

**3.115** What is the specificity of the VAS?

**3.116** The cutoff points of  $\geq 5$  for a VAS test-positive and  $\leq 4$  for a VAS test-negative are arbitrary. Compute and graph the ROC curve for the VAS test by varying the cutoff point for a test-positive. (Use the cutoff points VAS  $\geq 0$ , VAS  $\geq 3$ , VAS  $\geq 5$ , VAS  $\geq 7$ , and VAS  $\geq 11$  as possible criteria for test-positive.)

**3.117** The area under the ROC curve is 65%. What does it mean?

## Cancer

Breast cancer is considered largely a hormonal disease. An important hormone in breast-cancer research is estradiol. The data in Table 3.17 on serum estradiol levels were obtained from 213 breast-cancer cases and 432 age-matched controls. All women were age 50–59 years.

**Table 3.17 Serum-estradiol data**

Serum estradiol (pg/mL)	Cases (N = 213)	Controls (N = 432)
1–4	28	72
5–9	96	233
10–14	53	86
15–19	17	26
20–24	10	6
25–29	3	5
30+	6	4

Suppose a serum-estradiol level of 20+ pg/mL is proposed as a screening criterion for identifying breast-cancer cases.

**3.118** What is the sensitivity of this test?

**3.119** What is the specificity of this test?

The preceding sample was selected to oversample cases. In the general population, the prevalence of breast cancer is about 2% among women 50–59 years of age.

**3.120** What is the probability of breast cancer among 50- to 59-year-old women in the general population who have a serum-estradiol level of  $\geq 20$  pg/mL? What is another name for this quantity?

## Cardiovascular Disease

Mayo Clinic investigators have tracked coronary-heart-disease (CHD) mortality in Olmstead County, Minnesota, for the past 20 years[16]. Mayo Clinic physicians provided virtually all medical care to Olmstead County residents. Deaths from CHD were subdivided into those that occurred in hospital and those that occurred out of hospital. In-hospital death rates are thought to be influenced mainly by advances in medical care. Out-of-hospital death rates are thought to be influenced mainly by changes in risk-factor levels over time. For men, out-of-hospital CHD death rates were 280 cases per 100,000 men per year and in-hospital CHD death rates were 120 cases per 100,000 men per year in 1998. For women, out-of-hospital CHD death rates were 100 cases per 100,000 women per year; in-hospital CHD death rates were 40 cases per 100,000 women per year in 1998.

**3.121** If 50% of the Olmstead County population is male and 50% is female, what was the overall CHD mortality rate in Olmstead County in 1998?

The investigators reported that for both men and women, in-hospital CHD death rates were declining at a rate of 5.3% per year, whereas out-of-hospital CHD death rates were declining by 1.8% per year.

**3.122** What is the expected overall CHD mortality rate in Olmstead County in 2011 if these trends continue?

**3.123** In 2011, what proportion of the CHD deaths will occur in women?

## Genetics

Suppose a disease is caused by a single major gene with two alleles (*a*) and (*A*) with frequencies .95 and .05, respectively.

**3.124** What is the probability that an individual will have genotype (*aa*), (*aA*), and (*AA*), if we assume that each allele is inherited independently?

**3.125** Suppose the *A* allele is the deleterious allele but that the gene is only partially penetrant, meaning that the probability of developing the disease is .9 if one has two *A* alleles, .5 if one has one *A* allele, and .1 if one has no *A* alleles (sporadic cases). What is the overall probability of developing the disease in the population?

**3.126** Suppose an individual has the disease. What is the probability that he or she will have no, one, or two *A* alleles?

## Cardiovascular Disease

The ankle-arm blood pressure index (AAI) is defined as the ratio of ankle systolic blood pressure/arm systolic blood pressure and is used for the diagnosis of lower extremity arterial disease. A study was conducted to investigate whether the AAI can be used as a screening test for atherosclerotic diseases in general [17]. The subjects were 446 male workers in a copper smelter in Japan. Each subject had an AAI determination as well as an electrocardiogram (ECG). From the ECG, an S-T segment depression was defined as an S-T segment  $\geq 0.1$  mV below the baseline in at least 1 of 12 leads in a resting ECG. S-T segment depression is often used as one characterization of an abnormal ECG. The data in Table 3.18 were presented relating AAI to S-T segment depression.

**Table 3.18 Association between ankle-arm blood pressure index (AAI) and S-T segment depression**

	S-T segment depression	
	+	-
AAI < 1.0	20	95
AAI $\geq 1.0$	13	318

**3.127** If an abnormal ECG as determined by S-T segment depression is regarded as the gold standard for the presence of heart disease and an AAI of  $< 1.0$  is regarded as a possible test criterion for heart disease, then what is the sensitivity of the test?

**3.128** What is the specificity of the test?

**3.129** What is the  $PV^+$ ? (*Hint:* Assume that the subjects in this study are a random sample from the general population of Japan.)

**3.130** What is the  $PV^-$ ?

**3.131** Suppose the reproducibility of the AAI test were improved using better technology. Would the sensitivity of the test increase, decrease, or remain the same?

## Obstetrics, Health Promotion

A study was performed to assess the accuracy of self-reported exposure to cigarette smoking in-utero. A comparison was made between daughters' reports of smoking by their mothers during pregnancy with the mother's self-reports of their own smoking while pregnant with their daughters. The results were as shown in Table 3.19.

**Table 3.19 Relationship between mothers' self-reports of smoking while pregnant and daughters' reports of fetal smoke exposure**

Daughter's report of fetal smoke exposure	Mother's report of smoking during pregnancy	<i>N</i>
yes	yes	6685
yes	no	1126
no	yes	1222
no	no	23,227

**3.132** If a mother's self-report is considered completely accurate, then what is the  $PV^+$  of the daughter's report, in which positive indicates smoking and negative indicates not smoking?

**3.133** If a mother's self-report is considered completely accurate, then what is the  $PV^-$  of the daughter's report?

Additional data on self-reported smoking indicate that the mother is *not always completely accurate*. Saliva cotinine is a biochemical marker that, if elevated, is a 100% accurate indication of recent smoking.

Suppose if the mother states she is a nonsmoker during pregnancy that saliva cotinine is elevated 5% of the time, whereas if the mother states she is a smoker during pregnancy that saliva cotinine is elevated 97% of the time. Assume also that a daughter report adds no further information regarding the probability of an elevated cotinine level once the mother's self-report is known.

**3.134** What is the probability that the saliva cotinine level in a mother is elevated during pregnancy if the daughter reports that the mother smoked in-utero?

**3.135** What is the probability that the saliva cotinine level in the mother is not elevated during pregnancy if the daughter reports that the mother did not smoke in-utero?

## REFERENCES

- [1] National Center for Health Statistics. (1976, February 13). *Monthly vital statistics report, advance report, final natality statistics (1974)*, 24(11) (Suppl. 2).
- [2] Horner, M. J., Ries L. A. G., Krapcho M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S. F., Feuer, E. J., Huang, L., Mariotto, A., Miller, B. A., Lewis, D. R., Eisner, M. P., Stinchcomb, D. G., & Edwards, B. K., (eds). *SEER Cancer Statistics Review, 1975–2006*. Bethesda, MD: National Cancer Institute; [http://seer.cancer.gov/csr/1975\\_2006/](http://seer.cancer.gov/csr/1975_2006/), based on November 2008 SEER data submission, posted to SEER website, 2009.
- [3] Feller, W. (1960). *An introduction to probability theory and its applications* (Vol. 1). New York: Wiley.
- [4] Podgor, M. J., Leske, M. C., & Ederer, F. (1983). Incidence estimates for lens changes, macular changes, open-angle glaucoma, and diabetic retinopathy. *American Journal of Epidemiology*, 118(2), 206–212.
- [5] Punglia, R. S., D'Amico, A. V., Catalona, W. J., Roehl, K. A., & Kuntz, K. M. (2003). Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. *New England Journal of Medicine*, 349(4), 335–342.
- [6] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Diagnostic Radiology*, 143, 29–36.
- [7] Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, CDC, 2000.
- [8] Pfeffer, R. I., Afifi, A. A., & Chance, J. M. (1987). Prevalence of Alzheimer's disease in a retirement community. *American Journal of Epidemiology*, 125(3), 420–436.
- [9] National Center for Health Statistics. (1999, October). *Monthly vital statistics report, final natality statistics*.
- [10] Garvey, A. J., Bossé, R., Glynn, R. J., & Rosner, B. (1983). Smoking cessation in a prospective study of healthy adult males: Effects of age, time period, and amount smoked. *American Journal of Public Health*, 73(4), 446–450.
- [11] Luepker, R. V., Pechacek, T. F., Murray, D. M., Johnson, C. A., Hund, F., & Jacobs, D. R. (1981). Saliva thiocyanate: A chemical indicator of cigarette smoking in adolescents. *American Journal of Public Health*, 71(12), 1320.
- [12] Mandel, E., Bluestone, C. D., Rockette, H. E., Blatter, M. M., Reisinger, K. S., Wucher, E. P., & Harper, J. (1982). Duration of effusion after antibiotic treatment for acute otitis media: Comparison of cefaclor and amoxicillin. *Pediatric Infectious Diseases*, 1, 310–316.
- [13] Katzman, R., Zhang, M. Y., Ouang-Ya-Qu, Wang, Z. Y., Liu, W. T., Yu, E., Wong, S. C., Salmon, D. P., & Grant, I. (1988). A Chinese version of the Mini-Mental State Examination: impact of illiteracy in a Shanghai dementia survey. *Journal of Clinical Epidemiology*, 41(10), 971–978.
- [14] Skjaerven, R., Wilcox, A. J., Lie, R. T., & Irgens, L. M. (1988). Selective fertility and the distortion of perinatal mortality. *American Journal of Epidemiology*, 128(6), 1352–1363.
- [15] Mayeux, R., Saunders, A. M., Shea, S., Mirra, S., Evans, D., Roses, A. D., Hyman, B. T., Crain, B., Tang, M. X., & Phelps, C. H. (1998). Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. *New England Journal of Medicine*, 338(8), 506–511.
- [16] Goraya, T. Y., Jacobsen, S. J., Kottke, T. E., Frye, R. L., Weston, S. A., & Roger, V. L. (2003). Coronary heart disease death and sudden cardiac death: A 20-year population-based study. *American Journal of Epidemiology*, 157, 763–770.
- [17] Shinozaki, T., Hasegawa, T., & Yanoa, E. (1998). Ankle-arm index as an indicator of atherosclerosis: its application as a screening method. *Journal of Clinical Epidemiology*, 51(12), 1263–1269.

# 4

## Discrete Probability Distributions

### 4.1 Introduction

Chapter 3 defined probability and introduced some basic tools used in working with probabilities. We now look at problems that can be put into a probabilistic framework. That is, by assessing the probabilities of certain events from actual past data, we can consider specific probability models that fit our problems.

#### Example 4.1

**Ophthalmology** Retinitis pigmentosa is a progressive ocular disease that in some cases eventually results in blindness. The three main genetic types of the disease are dominant, recessive, and sex-linked. Each genetic type has a different rate of progression, with the dominant mode being the slowest to progress and the sex-linked mode the fastest. Suppose the prior history of disease in a family is unknown, but one of the two male children is affected and the one female child is not affected. Can this information help identify the genetic type?

The **binomial distribution** can be applied to calculate the probability of this event occurring (one of two males affected, none or one female affected) under each of the genetic types mentioned, and these results can then be used to infer the most likely genetic type. In fact, this distribution can be used to make an inference for any family for which we know  $k_1$  of  $n_1$  male children are affected and  $k_2$  of  $n_2$  female children are affected.

#### Example 4.2

**Cancer** A second example of a commonly used probability model concerns a cancer scare in Woburn, Massachusetts. A news story reported an “excessive” number of cancer deaths in young children in this town and speculated about whether this high rate was due to the dumping of industrial wastes in the northeastern part of town [1]. Suppose 12 cases of leukemia were reported in a town where 6 would normally be expected. Is this enough evidence to conclude that the town has an excessive number of leukemia cases?

The **Poisson distribution** can be used to calculate the probability of 12 or more cases if this town had typical national rates for leukemia. If this probability were small enough, we would conclude that the number was excessive; otherwise, we would decide that longer surveillance of the town was needed before arriving at a conclusion.

This chapter introduces the general concept of a discrete random variable and describes the binomial and Poisson distributions in depth. This forms the basis for the discussion (in Chapters 7 and 10) of hypothesis testing based on the binomial and Poisson distributions.

## 4.2 Random Variables

In Chapter 3 we dealt with very specific events, such as the outcome of a tuberculin skin test or blood-pressure measurements taken on different members of a family. We now want to introduce ideas that let us refer, in general terms, to different types of events having the *same probabilistic structure*. For this purpose let's consider the concept of a random variable.

### Definition 4.1

A **random variable** is a function that assigns numeric values to different events in a sample space.

Two types of random variables are discussed in this text: discrete and continuous.

### Definition 4.2

A random variable for which there exists a discrete set of numeric values is a **discrete random variable**.

### Example 4.3

**Otolaryngology** Otitis media, a disease of the middle ear, is one of the most common reasons for visiting a doctor in the first 2 years of life other than a routine well-baby visit. Let  $X$  be the random variable that represents the number of episodes of otitis media in the first 2 years of life. Then  $X$  is a discrete random variable, which takes on the values 0, 1, 2, and so on.

### Example 4.4

**Hypertension** Many new drugs have been introduced in the past several decades to bring hypertension under control—that is, to reduce high blood pressure to normotensive levels. Suppose a physician agrees to use a new antihypertensive drug on a trial basis on the first four untreated hypertensives she encounters in her practice, before deciding whether to adopt the drug for routine use. Let  $X$  = the number of patients of four who are brought under control. Then  $X$  is a discrete random variable, which takes on the values 0, 1, 2, 3, 4.

### Definition 4.3

A random variable whose possible values cannot be enumerated is a **continuous random variable**.

### Example 4.5

**Environmental Health** Possible health effects on workers of exposure to low levels of radiation over long periods of time are of public health interest. One problem in assessing this issue is how to measure the cumulative exposure of a worker. A study was performed at the Portsmouth Naval Shipyard, where each exposed worker wore a badge, or dosimeter, which measured annual radiation exposure in rem [2]. The

cumulative exposure over a worker's lifetime could then be obtained by summing the yearly exposures. Cumulative lifetime exposure to radiation is a good example of a continuous random variable because it varied in this study from 0.000 to 91.414 rem; this would be regarded as taking on an essentially infinite number of values, which cannot be enumerated.

### 4.3 The Probability-Mass Function for a Discrete Random Variable

The values taken by a discrete random variable and its associated probabilities can be expressed by a rule or relationship called a *probability-mass function* (pmf).

#### Definition 4.4

A **probability-mass function** is a mathematical relationship, or rule, that assigns to any possible value  $r$  of a discrete random variable  $X$  the probability  $Pr(X = r)$ . This assignment is made for all values  $r$  that have positive probability. The probability-mass function is sometimes also called a **probability distribution**.

The probability-mass function can be displayed in a table giving the values and their associated probabilities, or it can be expressed as a mathematical formula giving the probabilities of all possible values.

#### Example 4.6

**Hypertension** Consider Example 4.4. Suppose from previous experience with the drug, the drug company expects that for any clinical practice the probability that 0 patients of 4 will be brought under control is .008, 1 patient of 4 is .076, 2 patients of 4 is .265, 3 patients of 4 is .411, and all 4 patients is .240. This probability-mass function, or probability distribution, is displayed in Table 4.1.

**Table 4.1** Probability-mass function for the hypertension-control example

$Pr(X = r)$	.008	.076	.265	.411	.240
$r$	0	1	2	3	4

Notice that for any probability-mass function, the probability of any particular value must be between 0 and 1 and the sum of the probabilities of all values must exactly equal 1. Thus  $0 < Pr(X = r) \leq 1$ ,  $\sum Pr(X = r) = 1$ , where the summation is taken over all possible values that have positive probability.

#### Example 4.7

**Hypertension** In Table 4.1, for any clinical practice, the probability that between 0 and 4 hypertensives are brought under control is 1; that is,

$$.008 + .076 + .265 + .411 + .240 = 1$$

### Relationship of Probability Distributions to Frequency Distributions

In Chapters 1 and 2 we discussed the concept of a **frequency distribution** in the context of a sample. It was described as a list of each value in the data set and a corresponding count of how frequently the value occurs. If each count is divided by the

total number of points in the sample, then the frequency distribution can be considered as a sample analog to a probability distribution. In particular, a probability distribution can be thought of as a model based on an infinitely large sample, giving the fraction of data points in a sample that *should* be allocated to each specific value. Because the frequency distribution gives the actual proportion of points in a sample that correspond to specific values, the appropriateness of the model can be assessed by comparing the observed sample-frequency distribution with the probability distribution. The formal statistical procedure for making this comparison, called a **goodness-of-fit test**, is discussed in Chapter 10.

**Example 4.8**

**Hypertension** How can the probability-mass function in Table 4.1 be used to judge whether the drug behaves with the same efficacy in actual practice as predicted by the drug company? The drug company might provide the drug to 100 physicians and ask each of them to treat their first four untreated hypertensives with it. Each physician would then report his or her results to the drug company, and the combined results could be compared with the expected results in Table 4.1. For example, suppose that of 100 physicians who agree to participate, 19 bring all their first four untreated hypertensives under control with the drug, 48 bring three of four hypertensives under control, 24 bring two of four under control, and the remaining 9 bring only one of four under control. The sample-frequency distribution can be compared with the probability distribution given in Table 4.1, as shown in Table 4.2 and Figure 4.1.

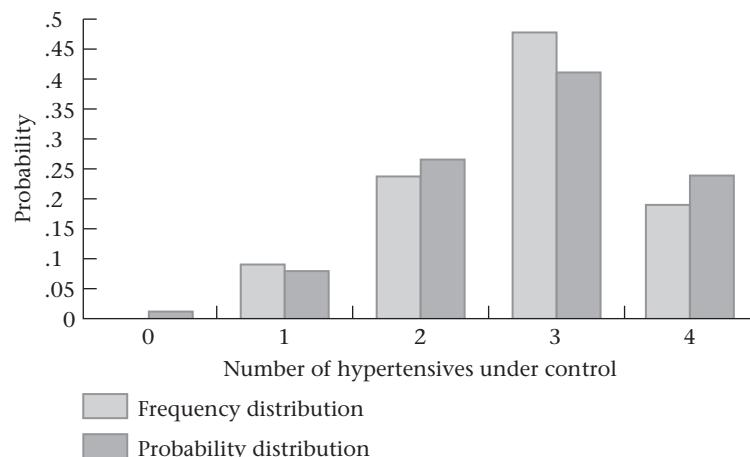
**Table 4.2**

**Comparison of the sample-frequency distribution and the theoretical-probability distribution for the hypertension-control example**

Number of hypertensives under control = $r$	Probability distribution $Pr(X = r)$	Frequency distribution
0	.008	.000 = 0/100
1	.076	.090 = 9/100
2	.265	.240 = 24/100
3	.411	.480 = 48/100
4	.240	.190 = 19/100

**Figure 4.1**

**Comparison of the frequency and probability distribution for the hypertension-control example**



The distributions look reasonably similar. The role of statistical inference is to compare the two distributions to judge whether differences between the two can be attributed to chance or whether real differences exist between the drug's performance in actual clinical practice and expectations from previous drug-company experience.

Students often ask where a probability-mass function comes from. In some instances previous data can be obtained on the same type of random variable being studied, and the probability-mass function can be computed from these data. In other instances previous data may not be available, but the probability-mass function from some well-known distribution can be used to see how well it fits actual sample data. This approach was used in Table 4.2, where the probability-mass function was derived from the binomial distribution and then compared with the frequency distribution from the sample of 100 physician practices.

## 4.4 The Expected Value of a Discrete Random Variable

If a random variable has a large number of values with positive probability, then the probability-mass function is not a useful summary measure. Indeed, we face the same problem as in trying to summarize a sample by enumerating each data value.

Measures of location and spread can be developed for a random variable in much the same way as they were developed for samples. The analog to the arithmetic mean  $\bar{x}$  is called the expected value of a random variable, or population mean, and is denoted by  $E(X)$  or  $\mu$ . The expected value represents the "average" value of the random variable. It is obtained by multiplying each possible value by its respective probability and summing these products over all the values that have positive (that is, nonzero) probability.

**Definition 4.5** The expected value of a discrete random variable is defined as

$$E(X) \equiv \mu = \sum_{i=1}^R x_i Pr(X = x_i)$$

where the  $x_i$ 's are the values the random variable assumes with positive probability.

Note that the sum in the definition of  $\mu$  is over  $R$  possible values.  $R$  may be either finite or infinite. In either case, the individual values must be distinct from each other.

**Example 4.9** **Hypertension** Find the expected value for the random variable shown in Table 4.1.

**Solution**  $E(X) = 0(.008) + 1(.076) + 2(.265) + 3(.411) + 4(.240) = 2.80$

Thus on average about 2.8 hypertensives would be expected to be brought under control for every 4 who are treated.

**Example 4.10** **Otolaryngology** Consider the random variable mentioned in Example 4.3 representing the number of episodes of otitis media in the first 2 years of life. Suppose this random variable has a probability-mass function as given in Table 4.3.

**Table 4.3** Probability-mass function for the number of episodes of otitis media in the first 2 years of life

$r$	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

**Solution**

What is the expected number of episodes of otitis media in the first 2 years of life?

$$E(X) = 0(.129) + 1(.264) + 2(.271) + 3(.185) + 4(.095) + 5(.039) + 6(.017) = 2.038$$

Thus on average a child would be expected to have about two episodes of otitis media in the first 2 years of life.

In Example 4.8 the probability-mass function for the random variable representing the number of previously untreated hypertensives brought under control was compared with the actual number of hypertensives brought under control in 100 clinical practices. In much the same way, the expected value of a random variable can be compared with the actual sample mean in a data set ( $\bar{x}$ ).

**Example 4.11**

**Hypertension** Compare the average number of hypertensives brought under control in the 100 clinical practices ( $\bar{x}$ ) with the expected number of hypertensives brought under control ( $\mu$ ) per 4-patient practice.

**Solution**

From Table 4.2 we have

$$\bar{x} = [0(0) + 1(9) + 2(24) + 3(48) + 4(19)]/100 = 2.77$$

hypertensives controlled per 4-patient clinical practice, while  $\mu = 2.80$ . This agreement is rather good. The specific methods for comparing the observed average value and expected value of a random variable ( $\bar{x}$  and  $\mu$ ) are covered in the material on statistical inference in Chapter 7. Notice that  $\bar{x}$  could be written in the form

$$\bar{x} = 0(0/100) + 1(9/100) + 2(24/100) + 3(48/100) + 4(19/100)$$

that is, a weighted average of the number of hypertensives brought under control, where the weights are the observed probabilities. The expected value, in comparison, can be written as a similar weighted average, where the weights are the theoretical probabilities:

$$\mu = 0(.008) + 1(.076) + 2(.265) + 3(.411) + 4(.240)$$

Thus the two quantities are actually obtained in the same way, one with weights given by the “observed” probabilities and the other with weights given by the “theoretical” probabilities.

## 4.5 The Variance of a Discrete Random Variable

The analog to the sample variance ( $s^2$ ) for a random variable is called the **variance of the random variable**, or **population variance**, and is denoted by  $Var(X)$  or  $\sigma^2$ . The variance represents the spread, relative to the expected value, of all values that have positive probability. In particular, the variance is obtained by multiplying the squared distance of each possible value from the expected value by its respective probability and summing over all the values that have positive probability.

**Definition 4.6**

The **variance of a discrete random variable**, denoted by  $Var(X)$ , is defined by

$$Var(X) = \sigma^2 = \sum_{i=1}^R (x_i - \mu)^2 Pr(X = x_i)$$

where the  $x_i$ 's are the values for which the random variable takes on positive probability. The **standard deviation of a random variable**  $X$ , denoted by  $sd(X)$  or  $\sigma$ , is defined by the square root of its variance.

The population variance can also be expressed in a different (“short”) form as follows:

**Equation 4.1**

A short form for the population variance is given by

$$\sigma^2 = E(X - \mu)^2 = \sum_{i=1}^R x_i^2 Pr(X = x_i) - \mu^2$$

**Example 4.12**

**Otolaryngology** Compute the variance and standard deviation for the random variable depicted in Table 4.3.

**Solution**

We know from Example 4.10 that  $\mu = 2.038$ . Furthermore,

$$\begin{aligned}\sum_{i=1}^R x_i^2 Pr(X = x_i) &= 0^2 (.129) + 1^2 (.264) + 2^2 (.271) + 3^2 (.185) \\ &\quad + 4^2 (.095) + 5^2 (.039) + 6^2 (.017) \\ &= 0(.129) + 1(.264) + 4(.271) + 9(.185) \\ &\quad + 16(.095) + 25(.039) + 36(.017) \\ &= 6.12\end{aligned}$$

Thus  $Var(X) = \sigma^2 = 6.12 - (2.038)^2 = 1.967$ . The standard deviation of  $X$  is  $\sigma = \sqrt{1.967} = 1.402$ .

How can we interpret the standard deviation of a random variable? The following often-used principle is true for many, but not all, random variables:

**Equation 4.2**

Approximately 95% of the probability mass falls within two standard deviations ( $2\sigma$ ) of the mean of a random variable.

If  $1.96\sigma$  is substituted for  $2\sigma$  in Equation 4.2, this statement holds exactly for normally distributed random variables and approximately for certain other random variables. Normally distributed random variables are discussed in detail in Chapter 5.

**Example 4.13**

**Otolaryngology** Find  $a, b$  such that approximately 95% of infants will have between  $a$  and  $b$  episodes of otitis media in the first 2 years of life.

**Solution**

The random variable depicted in Table 4.3 has mean ( $\mu$ ) = 2.038 and standard deviation ( $\sigma$ ) = 1.402. The interval  $\mu \pm 2\sigma$  is given by

$$2.038 \pm 2(1.402) = 2.038 \pm 2.805$$

or from  $-0.77$  to  $4.84$ . Because only positive-integer values are possible for this random variable, the valid range is from  $a = 0$  to  $b = 4$  episodes. Table 4.3 gives the probability of having  $\leq 4$  episodes as

$$.129 + .264 + .271 + .185 + .095 = .944$$

The rule lets us quickly summarize the range of values that have most of the probability mass for a random variable without specifying each individual value. Chapter 6 discusses the type of random variable for which Equation 4.2 applies.

## 4.6 The Cumulative-Distribution Function of a Discrete Random Variable

Many random variables are displayed in tables or figures in terms of a cumulative-distribution function rather than a distribution of probabilities of individual values as in Table 4.1. The basic idea is to assign to each individual value the sum of probabilities of all values that are no larger than the value being considered. This function is defined as follows:

### Definition 4.7

The cumulative-distribution function (**cdf**) of a random variable  $X$  is denoted by  $F(X)$  and, for a specific value  $x$  of  $X$ , is defined by  $Pr(X \leq x)$  and denoted by  $F(x)$ .

### Example 4.14

**Otolaryngology** Compute the cdf for the otitis-media random variable in Table 4.3 and display it graphically.

#### Solution

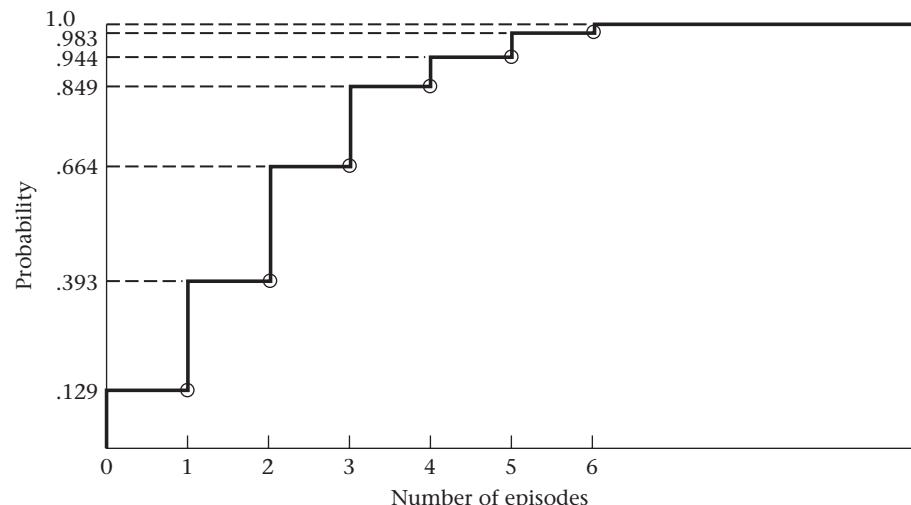
The cdf is given by

$$\begin{aligned} F(x) &= 0 && \text{if } x < 0 \\ F(x) &= .129 && \text{if } 0 \leq x < 1 \\ F(x) &= .393 && \text{if } 1 \leq x < 2 \\ F(x) &= .664 && \text{if } 2 \leq x < 3 \\ F(x) &= .849 && \text{if } 3 \leq x < 4 \\ F(x) &= .944 && \text{if } 4 \leq x < 5 \\ F(x) &= .983 && \text{if } 5 \leq x < 6 \\ F(x) &= 1.0 && \text{if } x \geq 6 \end{aligned}$$

The function is displayed in Figure 4.2.

Another way to distinguish between a discrete and continuous random variable is by each variable's cdf. For a discrete random variable, the cdf looks like a series of steps and is sometimes called a *step function*. For a continuous random variable, the cdf is a smooth curve. As the number of values increases, the cdf for a discrete random variable approaches that of a smooth curve. In Chapter 5, we discuss in more detail the cdf for continuous random variables.

**Figure 4.2** Cumulative-distribution function for the number of episodes of otitis media in the first 2 years of life



## REVIEW QUESTIONS 4A

- 1 What is the difference between a frequency distribution and a probability distribution?
- 2 What is the difference between a probability-mass function (pmf) and a cumulative-distribution function (cdf)?
- 3 In Table 4.4 the random variable  $X$  represents the number of boys in families with 4 children.

**Table 4.4** Number of boys in families with 4 children

$X$	$Pr(X = x)$
0	1/16
1	1/4
2	3/8
3	1/4
4	1/16

- (a) What is the expected value of  $X$ ? What does it mean?
- (b) What is the standard deviation of  $X$ ?
- (c) What is the cdf of  $X$ ?

## 4.7 Permutations and Combinations

Sections 4.2 through 4.6 introduced the concept of a discrete random variable in very general terms. The remainder of this chapter focuses on some specific discrete random variables that occur frequently in medical and biological work. Consider the following example.

### Example 4.15

**Infectious Disease** One of the most common laboratory tests performed on any routine medical examination is a blood count. The two main aspects of a blood count are (1) counting the number of white blood cells (the “white count”) and (2) differentiating the white blood cells that do exist into five categories—namely, neutrophils, lymphocytes, monocytes, eosinophils, and basophils (called the “differential”). Both the white count and the differential are used extensively in making clinical diagnoses. We concentrate here on the differential, particularly on the distribution of the number of neutrophils  $k$  out of 100 white blood cells (which is the typical number counted). We will see that the number of neutrophils follows a binomial distribution.

To study the binomial distribution, **permutations** and **combinations**—important topics in probability—must first be understood.

### Example 4.16

**Mental Health** Suppose we identify 5 men ages 50–59 with schizophrenia in a community, and we wish to match these subjects with normal controls of the same sex and age living in the same community. Suppose we want to employ a **matched-pair design**, where each case is matched with a normal control of the same sex and age. Five psychologists are employed by the study, each of whom interviews a single case and his matched control. If there are 10 eligible 50- to 59-year-old male controls in the community (labeled  $A, B, \dots, J$ ), then how many ways are there of choosing controls for the study if a control can never be used more than once?

**Solution**

The first control can be any of  $A, \dots, J$  and thus can be chosen in 10 ways. Once the first control is chosen, he can no longer be selected as the second control; therefore, the second control can be chosen in 9 ways. Thus the first two controls can be chosen in any one of  $10 \times 9 = 90$  ways. Similarly, the third control can be chosen in any one of 8 ways, the fourth control in 7 ways, and the fifth control in 6 ways, and so on. In total, there are  $10 \times 9 \times 8 \times 7 \times 6 = 30,240$  ways of choosing the 5 controls. For example, one possible selection is  $ACDFE$ . This means control  $A$  is matched to the first case, control  $C$  to the second case, and so on. The selection order of the controls is important because different psychologists may be assigned to interview each matched pair. Thus the selection  $ABCDE$  differs from  $CBAED$ , even though the same group of controls is selected.

We can now ask the general question: How many ways can  $k$  objects be selected out of  $n$  where the order of selection matters? Note that the first object can be selected in any one of  $n = (n + 1) - 1$  ways. Given that the first object has been selected, the second object can be selected in any one of  $n - 1 = (n + 1) - 2$  ways,  $\dots$ ; the  $k$ th object can be selected in any one of  $n - k + 1 = (n + 1) - k$  ways.

**Definition 4.8**

The number of **permutations** of  $n$  things taken  $k$  at a time is

$${}_nP_k = n(n-1) \times \dots \times (n-k+1)$$

It represents the number of ways of selecting  $k$  items of  $n$ , where the order of selection is important.

**Example 4.17**

**Mental Health** Suppose 3 schizophrenic women ages 50–59 and 6 eligible controls live in the same community. How many ways are there of selecting 3 controls?

**Solution**

To answer this question, consider the number of permutations of 6 things taken 3 at a time.

$${}_6P_3 = 6 \times 5 \times 4 = 120$$

Thus there are 120 ways of choosing the controls. For example, one way is to match control  $A$  to case 1, control  $B$  to case 2, and control  $C$  to case 3 ( $ABC$ ). Another way would be to match control  $F$  to case 1, control  $C$  to case 2, and control  $D$  to case 3 ( $FCD$ ). The order of selection is important because, for example, the selection  $ABC$  differs from the selection  $BCA$ .

In some instances we are interested in a special type of permutation: selecting  $n$  objects out of  $n$ , where order of selection matters (ordering  $n$  objects). By the preceding principle,

$${}_nP_n = n(n-1) \times \dots \times [(n-n+1)] = n(n-1) \times \dots \times 2 \times 1$$

The special symbol generally used for this quantity is  $n!$ , which is called  $n$  factorial and is defined as follows:

**Definition 4.9**

$n! = n$  factorial is defined as  $n(n-1) \times \dots \times 2 \times 1$

**Example 4.18**

Evaluate 5 factorial.

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

The quantity  $0!$  has no intuitive meaning, but for consistency it will be defined as 1. Another way of writing  $_n P_k$  is in terms of factorials. Specifically, from Definition 4.8 we can re-express  $_n P_k$  in the form

$$\begin{aligned}_n P_k &= n(n-1) \times \cdots \times (n-k+1) \\&= \frac{n(n-1) \times \cdots \times (n-k+1) \times (n-k) \times (n-k-1) \times \cdots \times 1}{(n-k) \times (n-k-1) \times \cdots \times 1} \\&= n!/(n-k)!\end{aligned}$$

**Equation 4.3****Alternative Form for Permutations**

An alternative formula expressing permutations in terms of factorials is given by

$$_n P_k = n!/(n-k)!$$

**Example 4.19**

**Mental Health** Suppose 4 schizophrenic women and 7 eligible controls live in the same community. How many ways are there of selecting 4 controls?

**Solution**

The number of ways =  ${}_7 P_4 = 7(6)(5)(4) = 840$ .

Alternatively,  ${}_7 P_4 = 7!/3! = 5040/6 = 840$ .

**Example 4.20**

**Mental Health** Consider a somewhat different design for the study described in Example 4.16. Suppose an **unmatched study design**, in which *all* cases and controls are interviewed by the same psychologist, is used. If there are 10 eligible controls, then how many ways are there of choosing 5 controls for the study?

**Solution**

In this case, because the same psychologist interviews all patients, what is important is which controls are selected, not the order of selection. Thus the question becomes how many ways can 5 of 10 eligible controls be selected, where order is not important? Note that for each set of 5 controls (say *A, B, C, D, E*), there are  $5 \times 4 \times 3 \times 2 \times 1 = 5!$  ways of ordering the controls among themselves (e.g., *ACBED* and *DBCAE* are two possible orders). Thus the number of ways of selecting 5 of 10 controls for the study without respect to order = (number of ways of selecting 5 controls of 10 where order is important)/ $5! = {}_{10} P_5 / 5! = (10 \times 9 \times 8 \times 7 \times 6) / 120 = 30,240 / 120 = 252$  ways. Thus, *ABCDE* and *CDFIJ* are two possible selections. Also, *ABCDE* and *BCADE* are *not* counted twice.

The number of ways of selecting 5 objects of 10 without respect to order is referred to as the number of **combinations** of 10 things taken 5 at a time and is denoted by  ${}_{10} C_5$  or  $\binom{10}{5} = 252$ .

This discussion can be generalized to evaluate the number of combinations of  $n$  things taken  $k$  at a time. Note that for every selection of  $k$  distinct items of  $n$ , there are  $k(k-1) \times \cdots \times 2 \times 1 = k!$  ways of ordering the items among themselves. Thus, we have the following definition:

**Definition 4.10**

The number of combinations of  $n$  things taken  $k$  at a time is

$${}_n C_k = \binom{n}{k} = \frac{n(n-1) \times \cdots \times (n-k+1)}{k!}$$

Alternatively, if we express permutations in terms of factorials, as in Equation 4.3, we obtain

$$\begin{aligned} {}_n C_k &= \binom{n}{k} = {}_n P_k / k! \\ &= n! / [(n - k)! k!] \end{aligned}$$

Thus we have the following alternative definition of combinations:

---

**Definition 4.11** The number of combinations of  $n$  things taken  $k$  at a time is

$${}_n C_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

It represents the number of ways of selecting  $k$  objects out of  $n$  where the order of selection does not matter.

---

**Example 4.21** Evaluate  ${}_7 C_3$ .

$${}_7 C_3 = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 7 \times 5 = 35$$

Henceforth, for consistency we will always use the more common notation  $\binom{n}{k}$  for combinations. In words, this is expressed as “ $n$  choose  $k$ .”

A special situation arises upon evaluating  $\binom{n}{0}$ . By definition,  $\binom{n}{0} = n! / (0!n!)$ , and  $0!$  was defined as 1. Hence,  $\binom{n}{0} = 1$  for any  $n$ .

Frequently,  $\binom{n}{k}$  will need to be computed for  $k = 0, 1, \dots, n$ . The combinatorials have the following symmetry property, which makes this calculation easier than it appears at first.

**Equation 4.4**

For any non-negative integers  $n, k$ , where  $n \geq k$ ,

$$\binom{n}{k} = \binom{n}{n-k}$$

To see this, note from Definition 4.11 that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

If  $n - k$  is substituted for  $k$  in this expression, then we obtain

$$\binom{n}{n-k} = \frac{n!}{(n-k)![n-(n-k)]!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

Intuitively, this result makes sense because  $\binom{n}{k}$  represents the number of ways of selecting  $k$  objects of  $n$  without regard to order. However, for every selection of  $k$  objects, we have also, in a sense, identified the other  $n - k$  objects that were not selected. Thus the number of ways of selecting  $k$  objects of  $n$  without regard to order should be the same as the number of ways of selecting  $n - k$  objects of  $n$  without regard to order.

Hence we need only evaluate combinatorials  $\binom{n}{k}$  for the integers  $k \leq n/2$ . If  $k > n/2$ , then the relationship  $\binom{n}{n-k} = \binom{n}{k}$  can be used.

**Example 4.22** Evaluate

$$\binom{7}{0}, \binom{7}{1}, \dots, \binom{7}{7}$$

**Solution**  $\binom{7}{0} = 1$     $\binom{7}{1} = 7$     $\binom{7}{2} = \frac{7 \times 6}{2 \times 1} = 21$     $\binom{7}{3} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$   
 $\binom{7}{4} = \binom{7}{3} = 35$     $\binom{7}{5} = \binom{7}{2} = 21$     $\binom{7}{6} = \binom{7}{1} = 7$     $\binom{7}{7} = \binom{7}{0} = 1$

**REVIEW QUESTIONS 4B**

- Suppose we select 3 students randomly out of a class of 10 students to read a paper from the literature and summarize the results for the class. How many ways can the students be selected? Is this a permutation, a combination, or neither?
- Suppose we select 2 students randomly from a class of 20 students. The first student selected will analyze a data set on the computer and prepare summary tables, and the second student will present the results to the class. How many ways can the students be selected for these tasks? Is this a permutation, a combination, or neither?

## 4.8 The Binomial Distribution

All examples involving the binomial distribution have a common structure: a sample of  $n$  independent trials, each of which can have only two possible outcomes, which are denoted as “success” and “failure.” Furthermore, the probability of a success at each trial is assumed to be some constant  $p$ , and hence the probability of a failure at each trial is  $1 - p = q$ . The term “success” is used in a general way, without any specific contextual meaning.

For Example 4.15,  $n = 100$  and a “success” occurs when a cell is a neutrophil.

**Example 4.23**

**Infectious Disease** Reconsider Example 4.15 with 5 cells rather than 100, and ask the more limited question: What is the probability that the second and fifth cells considered will be neutrophils and the remaining cells non-neutrophils, given a probability of .6 that any one cell is a neutrophil?

**Solution**

If a neutrophil is denoted by an  $x$  and a non-neutrophil by an  $o$ , then the question being asked is: What is the probability of the outcome  $oxoox = Pr(oxoox)$ ? Because the probabilities of success and failure are given, respectively, by .6 and .4, and the outcomes for different cells are presumed to be independent, then the probability is

$$q \times p \times q \times q \times p = p^2 q^3 = (.6)^2 (.4)^3$$

**Example 4.24**

**Infectious Disease** Now consider the more general question: What is the probability that any 2 cells out of 5 will be neutrophils?

**Solution**

The arrangement *oxoax* is only one of 10 possible orderings that result in 2 neutrophils. Table 4.5 gives the 10 possible orderings.

**Table 4.5****Possible orderings for 2 neutrophils of 5 cells**

<i>xxooo</i>	<i>oxxoo</i>	<i>oooxo</i>
<i>xoxoo</i>	<i>oxoxo</i>	<i>oooxx</i>
<i>xooxo</i>	<i>oxoox</i>	
<i>xoooo</i>	<i>ooxxx</i>	

In terms of combinations, the number of orderings = the number of ways of selecting 2 cells to be neutrophils out of 5 cells =  $\binom{5}{2} = (5 \times 4) / (2 \times 1) = 10$ .

The probability of any of the orderings in Table 4.5 is the same as that for the ordering *oxoax*, namely,  $(.6)^2(.4)^3$ . Thus the probability of obtaining 2 neutrophils in 5 cells is  $\binom{5}{2}(.6)^2(.4)^3 = 10(.6)^2(.4)^3 = .230$ .

Suppose the neutrophil problem is now considered more generally, with  $n$  trials rather than 5 trials, and the question is asked: What is the probability of  $k$  successes (rather than 2 successes) in these  $n$  trials? The probability that the  $k$  successes will occur at  $k$  specific trials within the  $n$  trials and that the remaining trials will be failures is given by  $p^k(1-p)^{n-k}$ . To compute the probability of  $k$  successes in any of the  $n$  trials, this probability must be multiplied by the number of ways in which  $k$  trials for the successes and  $n - k$  trials for the failures can be selected =  $\binom{n}{k}$ , as was done in Table 4.5. Thus the probability of  $k$  successes in  $n$  trials, or  $k$  neutrophils in  $n$  cells, is

$$\binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} p^k q^{n-k}$$

**Equation 4.5**

The distribution of the number of successes in  $n$  statistically independent trials, where the probability of success on each trial is  $p$ , is known as the **binomial distribution** and has a probability-mass function given by

$$Pr(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n$$

**Example 4.25**

What is the probability of obtaining 2 boys out of 5 children if the probability of a boy is .51 at each birth and the sexes of successive children are considered independent random variables?

**Solution**

Use a binomial distribution with  $n = 5$ ,  $p = .51$ ,  $k = 2$ . Compute

$$\begin{aligned} Pr(X = 2) &= \binom{5}{2} (.51)^2 (.49)^3 = \frac{5 \times 4}{2 \times 1} (.51)^2 (.49)^3 \\ &= 10 (.51)^2 (.49)^3 = .306 \end{aligned}$$

**Using Binomial Tables**

Often a number of binomial probabilities need to be evaluated for the same  $n$  and  $p$ , which would be tedious if each probability had to be calculated from Equation 4.5.

Instead, for small  $n$  ( $n \leq 20$ ) and selected values of  $p$ , refer to Table 1 in the Appendix, where individual binomial probabilities are calculated. In this table, the number of trials ( $n$ ) is provided in the first column, the number of successes ( $k$ ) out of the  $n$  trials is given in the second column, and the probability of success for an individual trial ( $p$ ) is given in the first row. Binomial probabilities are provided for  $n = 2, 3, \dots, 20; p = .05, .10, \dots, .50$ .

**Example 4.26**

**Infectious Disease** Evaluate the probability of 2 lymphocytes out of 10 white blood cells if the probability that any one cell is a lymphocyte is .2.

**Solution**

Refer to Table 1 with  $n = 10, k = 2, p = .20$ . The appropriate probability, given in the  $k = 2$  row and  $p = .20$  column under  $n = 10$ , is .3020.

**Example 4.27**

**Pulmonary Disease** An investigator notices that children develop chronic bronchitis in the first year of life in 3 of 20 households in which both parents have chronic bronchitis, as compared with the national incidence of chronic bronchitis, which is 5% in the first year of life. Is this difference “real,” or can it be attributed to chance? Specifically, how likely are infants in at least 3 of 20 households to develop chronic bronchitis if the probability of developing disease in any one household is .05?

**Solution**

Suppose the underlying rate of disease in the offspring is .05. Under this assumption, the number of households in which the infants develop chronic bronchitis will follow a binomial distribution with parameters  $n = 20, p = .05$ . Thus among 20 households the probability of observing  $k$  with bronchitic children is given by

$$\binom{20}{k} (.05)^k (.95)^{20-k}, \quad k = 0, 1, \dots, 20$$

The question is: What is the probability of observing at least 3 households with a bronchitic child? The answer is

$$Pr(X \geq 3) = \sum_{k=3}^{20} \binom{20}{k} (.05)^k (.95)^{20-k} = 1 - \sum_{k=0}^2 \binom{20}{k} (.05)^k (.95)^{20-k}$$

These three probabilities in the sum can be evaluated using the binomial table (Table 1). Refer to  $n = 20, p = .05$ , and note that  $Pr(X = 0) = .3585, Pr(X = 1) = .3774, Pr(X = 2) = .1887$ . Thus

$$Pr(X \geq 3) = 1 - (.3585 + .3774 + .1887) = .0754$$

Thus  $X \geq 3$  is an unusual event, but not very unusual. Usually .05 or less is the range of probabilities used to identify unusual events. This criterion is discussed in more detail in our work on  $p$ -values in Chapter 7. If 3 infants of 20 were to develop the disease, it would be difficult to judge whether the familial aggregation was real until a larger sample was available.

One question sometimes asked is why a criterion of  $Pr(X \geq 3)$  cases, rather than  $Pr(X = 3)$  cases, was used to define unusualness in Example 4.27? The latter is what we actually observed. An intuitive answer is that if the number of households studied in which both parents had chronic bronchitis were very large (for example,  $n = 1500$ ), then the probability of any specific occurrence would be small. For example, suppose 75 cases occurred among 1500 households in which both parents had

chronic bronchitis. If the incidence of chronic bronchitis were .05 in such families, then the probability of 75 cases among 1500 households would be

$$\binom{1500}{75}(.05)^{75}(.95)^{1425} = .047$$

This result is exactly consistent with the national incidence rate (5% of households with cases in the first year of life) and yet yields a small probability. This doesn't make intuitive sense. The alternative approach is to calculate the probability of obtaining a result at least as extreme as the one obtained (a probability of at least 75 cases out of 1500 households) if the incidence rate of .05 were applicable to families in which both parents had chronic bronchitis. This would yield a probability of approximately .50 in the preceding example and would indicate that nothing very unusual is occurring in such families, which is clearly the correct conclusion. If this probability were small enough, then it would cast doubt on the assumption that the true incidence rate was .05 for such families. This approach was used in Example 4.27 and is developed in more detail in our work on hypothesis testing in Chapter 7. Alternative approaches to the analysis of these data also exist, based on *Bayesian inference*, and are discussed in Chapters 6 and 7.

One question that arises is how to use the binomial tables if the probability of success on an individual trial ( $p$ ) is greater than .5. Recall that

$$\binom{n}{k} = \binom{n}{n-k}$$

Let  $X$  be a binomial random variable with parameters  $n$  and  $p$ , and let  $Y$  be a binomial random variable with parameters  $n$  and  $q = 1 - p$ . Then Equation 4.5 can be rewritten as

$$\text{Equation 4.6} \quad \Pr(X = k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{n-k} q^{n-k} p^k = \Pr(Y = n - k)$$

In words, the probability of obtaining  $k$  successes for a binomial random variable  $X$  with parameters  $n$  and  $p$  is the same as the probability of obtaining  $n - k$  successes for a binomial random variable  $Y$  with parameters  $n$  and  $q$ . Clearly, if  $p > .5$ , then  $q = 1 - p < .5$ , and Table 1 can be used with sample size  $n$ , referring to the  $n - k$  row and the  $q$  column to obtain the appropriate probability.

### Example 4.28

**Infectious Disease** Evaluate the probabilities of obtaining  $k$  neutrophils out of 5 cells for  $k = 0, 1, 2, 3, 4, 5$ , where the probability that any one cell is a neutrophil is .6.

#### Solution

Because  $p > .5$ , refer to the random variable  $Y$  with parameters  $n = 5$ ,  $p = 1 - .6 = .4$ .

$$\Pr(X = 0) = \binom{5}{0} (.6)^0 (.4)^5 = \binom{5}{5} (.4)^5 (.6)^0 = \Pr(Y = 5) = .0102$$

on referring to the  $k = 5$  row and  $p = .40$  column under  $n = 5$ . Similarly,

$$\Pr(X = 1) = \Pr(Y = 4) = .0768 \text{ on referring to the } 4 \text{ row and } .40 \text{ column under } n = 5$$

$$\Pr(X = 2) = \Pr(Y = 3) = .2304 \text{ on referring to the } 3 \text{ row and } .40 \text{ column under } n = 5$$

$$\Pr(X = 3) = \Pr(Y = 2) = .3456 \text{ on referring to the } 2 \text{ row and } .40 \text{ column under } n = 5$$

$$\Pr(X = 4) = \Pr(Y = 1) = .2592 \text{ on referring to the } 1 \text{ row and } .40 \text{ column under } n = 5$$

$$\Pr(X = 5) = \Pr(Y = 0) = .0778 \text{ on referring to the } 0 \text{ row and } .40 \text{ column under } n = 5$$

## Using “Electronic” Tables

In many instances we want to evaluate binomial probabilities for  $n > 20$  and/or for values of  $p$  not given in Table 1 of the Appendix. For sufficiently large  $n$ , the normal distribution can be used to approximate the binomial distribution, and tables of the normal distribution can be used to evaluate binomial probabilities. This procedure is usually less tedious than evaluating binomial probabilities directly using Equation 4.5 and is studied in detail in Chapter 5. Alternatively, if the sample size is not large enough to use the normal approximation and if the value of  $n$  or  $p$  is not in Table 1, then an electronic table can be used to evaluate binomial probabilities.

One example of an electronic table is provided by Microsoft Excel. A menu of statistical functions is available to the user, including calculation of probabilities for many probability distributions, including but not limited to those discussed in this text. For example, one function in this menu is the binomial-distribution function, which is called BINOMDIST and is discussed in detail on the Companion Website. Using this function, we can calculate the probability-mass function and cdf for virtually any binomial distribution.

### Example 4.29

**Pulmonary Disease** Compute the probability of obtaining exactly 75 cases of chronic bronchitis and the probability of obtaining at least 75 cases of chronic bronchitis in the first year of life among 1500 families in which both parents have chronic bronchitis, if the underlying incidence rate of chronic bronchitis in the first year of life is .05.

### Solution

We use the BINOMDIST function of Excel 2007 to solve this problem. Table 4.6 gives the results. First we compute  $Pr(X = 75)$ , which is .047, which is unusual. We then use the cdf option to compute  $Pr(X \leq 74)$ , which equals .483. Finally, we compute the probability of obtaining at least 75 cases by

$$Pr(X \geq 75) = 1 - Pr(X \leq 74) = .517$$

Hence, obtaining 75 cases out of 1500 children is clearly not unusual.

**Table 4.6**

Calculation of binomial probabilities using Excel 2007

<b>n</b>	<b>1500</b>
<b>k</b>	<b>75</b>
<b>p</b>	<b>0.05</b>
<b>Pr(X = 75)</b>	<b>0.047210 = BINOMDIST (75, 1500, .05, false)</b>
<b>Pr(X &lt;= 74)</b>	<b>0.483458 = BINOMDIST (74, 1500, .05, true)</b>
<b>Pr(X &gt;= 75)</b>	<b>0.516542 = 1 - BINOMDIST (74, 1500, .05, true)</b>

### Example 4.30

**Infectious Disease** Suppose a group of 100 women ages 60–64 received a new flu vaccine in 2004, and 5 of them died within the next year. Is this event unusual, or can this death rate be expected for people of this age-sex group? Specifically, how likely are at least 5 of 100 60- to 64-year-old women who receive a flu vaccine to die in the next year?

### Solution

We first find the expected annual death rate in 60- to 64-year-old women. From a 2004 U.S. life table, we find that 60- to 64-year-old women have an approximate

probability of death within the next year of .009 [3]. Thus, from the binomial distribution the probability that  $k$  of 100 women will die during the next year is given by  $\binom{100}{k}(.009)^k(.991)^{100-k}$ . We want to know whether 5 deaths in a sample of 100 women

is an “unusual” event. One approach to this problem might be to find the probability of obtaining at least 5 deaths in this group =  $Pr(X \geq 5)$  given that the probability of death for an individual woman is .009. This probability can be expressed as

$$\sum_{k=5}^{100} \binom{100}{k} (.009)^k (.991)^{100-k}$$

Because this sum of 96 probabilities is tedious to compute, we instead compute

$$Pr(X < 5) = \sum_{k=0}^4 \binom{100}{k} (.009)^k (.991)^{100-k}$$

and then evaluate  $Pr(X \geq 5) = 1 - Pr(X < 5)$ . The binomial tables cannot be used because  $n > 20$ . Therefore, the sum of 5 binomial probabilities is evaluated using Excel, as shown in Table 4.7.

We see that

$$Pr(X \leq 4) = .998$$

$$\text{and } Pr(X \geq 5) = 1 - Pr(X \leq 4) = .002$$

**Table 4.7**

**Calculation of the probability of at least 5 deaths among 100 women 60–64 years of age in 2004**

<b>n</b>	<b>100</b>
<b>p</b>	<b>0.009</b>
<b>Pr(X &lt;= 4)</b>	<b>0.99781 = BINOMDIST (4, 100, .009, true)</b>
<b>Pr(X &gt;= 5)</b>	<b>0.00219 = 1 - BINOMDIST (4, 100, .009, true)</b>

Thus at least 5 deaths in 100 is very unusual and would probably be grounds for considering halting use of the vaccine.

## 4.9 Expected Value and Variance of the Binomial Distribution

The expected value and variance of the binomial distribution are important both in terms of our general knowledge about the binomial distribution and for our later work on estimation and hypothesis testing. From Definition 4.5 we know that the general formula for the expected value of a discrete random variable is

$$E(X) = \sum_{i=1}^R x_i Pr(X = x_i)$$

In the special case of a binomial distribution, the only values that take on positive probability are 0, 1, 2, . . . ,  $n$ , and these values occur with probabilities

$$\binom{n}{0} p^0 q^n, \quad \binom{n}{1} p^1 q^{n-1}, \dots$$

$$\text{Thus } E(X) = \sum_{k=0}^n k \binom{n}{k} p^k q^{n-k}$$

This summation reduces to the simple expression  $np$ . Similarly, using Definition 4.6, we can show that

$$\text{Var}(X) = \sum_{k=0}^n (k - np)^2 \binom{n}{k} p^k q^{n-k} = npq$$

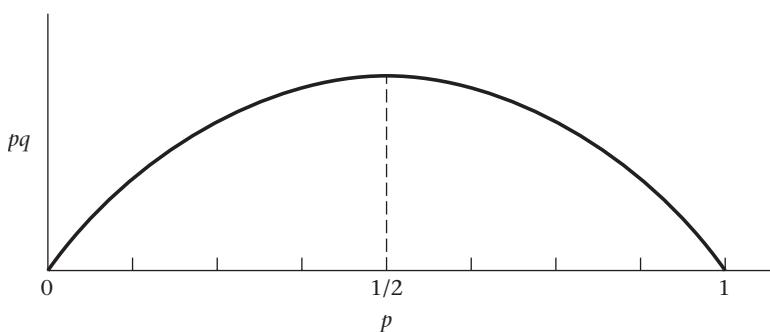
which leads directly to the following result:

**Equation 4.7**

The expected value and the variance of a binomial distribution are  $np$  and  $npq$ , respectively.

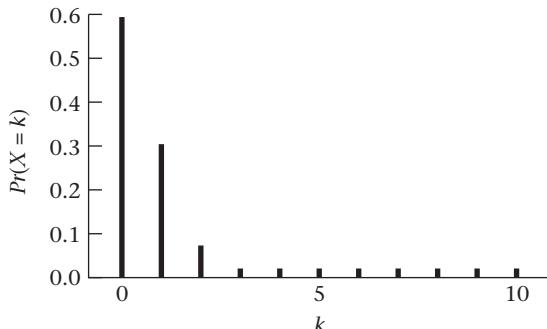
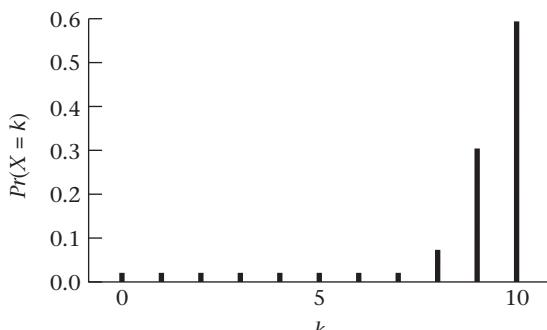
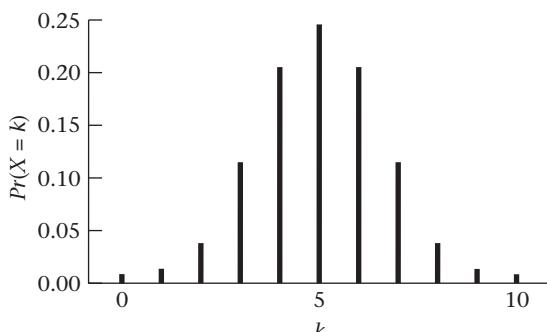
These results make good sense because the expected number of successes in  $n$  trials is simply the probability of success on one trial multiplied by  $n$ , which equals  $np$ . Furthermore, for a given number of trials  $n$ , the binomial distribution has the highest variance when  $p = 1/2$ , as shown in Figure 4.3. The variance of the distribution decreases as  $p$  moves away from  $1/2$  in either direction, becoming 0 when  $p = 0$  or 1. This result makes sense because when  $p = 0$  there must be 0 successes in  $n$  trials and when  $p = 1$  there must be  $n$  successes in  $n$  trials, and there is no variability in either instance. Furthermore, when  $p$  is near 0 or near 1, the distribution of the number of successes is clustered near 0 and  $n$ , respectively, and there is comparatively little variability as compared with the situation when  $p = 1/2$ . This point is illustrated in Figure 4.4.

**Figure 4.3** Plot of  $pq$  versus  $p$



### REVIEW QUESTIONS 4C

- 1 The probability of a woman developing breast cancer over a lifetime is about  $1/9$ .
  - (a) What is the probability that 2 women of 10 will develop breast cancer over a lifetime?
  - (b) What is the probability that at least 2 women of 10 will develop breast cancer over a lifetime?
- 2 Suppose we have 10 subjects and the probability of having a disease at one point in time for 1 subject is .1. What is the probability that exactly 1 of the 10 subjects has the disease? Why is this not the same as .1?

**Figure 4.4** The binomial distribution for various values of  $p$  when  $n = 10$ (a)  $n = 10, p = .05$ (b)  $n = 10, p = .95$ (c)  $n = 10, p = .50$ 

## 4.10 The Poisson Distribution

The Poisson distribution is perhaps the second most frequently used discrete distribution after the binomial distribution. This distribution is usually associated with rare events.

### Example 4.31

**Infectious Disease** Consider the distribution of number of deaths attributed to typhoid fever over a long period of time, for example, 1 year. Assuming the probability of a new death from typhoid fever in any one day is very small and the number of cases reported in any two distinct periods of time are independent random variables, then the number of deaths over a 1-year period will follow a Poisson distribution.

**Example 4.32**

**Bacteriology** The preceding example concerns a rare event occurring over time. Rare events can also be considered not only over time but also on a surface area, such as the distribution of number of bacterial colonies growing on an agar plate. Suppose we have a 100-cm<sup>2</sup> agar plate. The probability of finding any bacterial colonies at any one point  $a$  (or more precisely in a small area around  $a$ ) is very small, and the events of finding bacterial colonies at any two points  $a_1, a_2$  are independent. The number of bacterial colonies over the entire agar plate will follow a Poisson distribution.

Consider Example 4.31. Ask the question: What is the distribution of the number of deaths caused by typhoid fever from time 0 to time  $t$  (where  $t$  is some long period of time, such as 1 year or 20 years)?

Three assumptions must be made about the incidence of the disease. Consider any general *small* subinterval of the time period  $t$ , denoted by  $\Delta t$ .

**Assumption 4.1**

Assume that

- (1) The probability of observing 1 death is directly proportional to the length of the time interval  $\Delta t$ . That is,  $Pr(1 \text{ death}) \approx \lambda \Delta t$  for some constant  $\lambda$ .
- (2) The probability of observing 0 deaths over  $\Delta t$  is approximately  $1 - \lambda \Delta t$ .
- (3) The probability of observing more than 1 death over this time interval is essentially 0.

**Assumption 4.2**

**Stationarity** Assume the number of deaths per unit time is the same throughout the entire time interval  $t$ . Thus an increase in the incidence of the disease as time goes on within the time period  $t$  would violate this assumption. Note that  $t$  should not be overly long because this assumption is less likely to hold as  $t$  increases.

**Assumption 4.3**

**Independence** If a death occurs within one time subinterval, then it has no bearing on the probability of death in the next time subinterval. This assumption would be violated in an epidemic situation because if a new case of disease occurs, then subsequent deaths are likely to build up over a short period of time until after the epidemic subsides.

Given these assumptions, the Poisson probability distribution can be derived:

**Equation 4.8**

The probability of  $k$  events occurring in a time period  $t$  for a Poisson random variable with parameter  $\lambda$  is

$$Pr(X = k) = e^{-\mu} \mu^k / k!, \quad k = 0, 1, 2, \dots$$

where  $\mu = \lambda t$  and  $e$  is approximately 2.71828.

Thus the Poisson distribution depends on a single parameter  $\mu = \lambda t$ . Note that the parameter  $\lambda$  represents the *expected number of events per unit time*, whereas the parameter  $\mu$  represents the *expected number of events over time period t*. One important difference between the Poisson and binomial distributions concerns the numbers of trials and events. For a binomial distribution there are a finite number of trials  $n$ , and the number of events can be no larger than  $n$ . For a Poisson distribution the number of trials is essentially infinite and the number of events (or number of deaths) can be indefinitely large, although the probability of  $k$  events becomes very small as  $k$  increases.

**Example 4.33**

**Infectious Disease** Consider the typhoid-fever example. Suppose the number of deaths from typhoid fever over a 1-year period is Poisson distributed with parameter  $\mu = 4.6$ . What is the probability distribution of the number of deaths over a 6-month period? A 3-month period?

**Solution**

Let  $X$  = the number of deaths in 6 months. Because  $\mu = 4.6$ ,  $t = 1$  year, it follows that  $\lambda = 4.6$  deaths per year. For a 6-month period we have  $\lambda = 4.6$  deaths per year,  $t = .5$  year. Thus  $\mu = \lambda t = 2.3$ . Therefore,

$$Pr(X = 0) = e^{-2.3} = .100$$

$$Pr(X = 1) = \frac{2.3}{1!} e^{-2.3} = .231$$

$$Pr(X = 2) = \frac{2.3^2}{2!} e^{-2.3} = .265$$

$$Pr(X = 3) = \frac{2.3^3}{3!} e^{-2.3} = .203$$

$$Pr(X = 4) = \frac{2.3^4}{4!} e^{-2.3} = .117$$

$$Pr(X = 5) = \frac{2.3^5}{5!} e^{-2.3} = .054$$

$$Pr(X \geq 6) = 1 - (.100 + .231 + .265 + .203 + .117 + .054) = .030$$

Let  $Y$  = the number of deaths in 3 months. For a 3-month period, we have  $\lambda = 4.6$  deaths per year,  $t = .25$  year,  $\mu = \lambda t = 1.15$ . Therefore,

$$Pr(Y = 0) = e^{-1.15} = .317$$

$$Pr(Y = 1) = \frac{1.15}{1!} e^{-1.15} = .364$$

$$Pr(Y = 2) = \frac{1.15^2}{2!} e^{-1.15} = .209$$

$$Pr(Y = 3) = \frac{1.15^3}{3!} e^{-1.15} = .080$$

$$Pr(Y \geq 4) = 1 - (.317 + .364 + .209 + .080) = .030$$

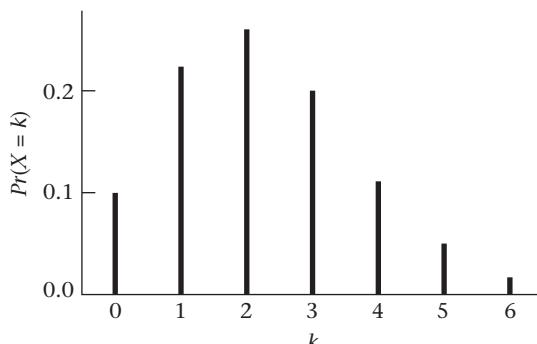
These distributions are plotted in Figure 4.5. Note that the distribution tends to become more symmetric as the time interval increases or, more specifically, as  $\mu$  increases.

The Poisson distribution can also be applied to Example 4.32, in which the distribution of the number of bacterial colonies in an agar plate of area  $A$  is discussed. Assuming that the probability of finding 1 colony in an area the size of  $\Delta A$  at any point on the plate is  $\lambda \Delta A$  for some  $\lambda$  and that the number of bacterial colonies found at 2 different points of the plate are independent random variables, then the probability of finding  $k$  bacterial colonies in an area of size  $A$  is  $e^{-\mu} \mu^k / k!$ , where  $\mu = \lambda A$ .

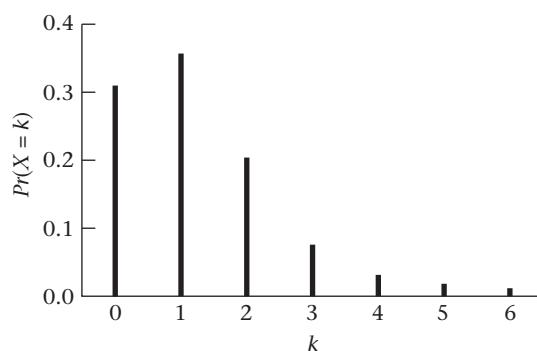
**Example 4.34**

**Bacteriology** If  $A = 100 \text{ cm}^2$  and  $\lambda = .02$  colonies per  $\text{cm}^2$ , calculate the probability distribution of the number of bacterial colonies.

**Figure 4.5** Distribution of the number of deaths attributable to typhoid fever over various time intervals



(a) 6 months



(b) 3 months

**Solution**

We have  $\mu = \lambda A = 100(.02) = 2$ . Let  $X$  = the number of colonies.

$$Pr(X = 0) = e^{-2} = .135$$

$$Pr(X = 1) = e^{-2} 2^1 / 1! = 2e^{-2} = .271$$

$$Pr(X = 2) = e^{-2} 2^2 / 2! = 2e^{-2} = .271$$

$$Pr(X = 3) = e^{-2} 2^3 / 3! = \frac{4}{3} e^{-2} = .180$$

$$Pr(X = 4) = e^{-2} 2^4 / 4! = \frac{2}{3} e^{-2} = .090$$

$$Pr(X \geq 5) = 1 - (.135 + .271 + .271 + .180 + .090) = .053$$

Clearly, the larger  $\lambda$  is, the more bacterial colonies we would expect to find.

## 4.11 Computation of Poisson Probabilities

### Using Poisson Tables

A number of Poisson probabilities for the same parameter  $\mu$  often must be evaluated. This task would be tedious if Equation 4.8 had to be applied repeatedly. Instead, for  $\mu \leq 20$  refer to Table 2 in the Appendix, in which individual Poisson probabilities are specifically calculated. In this table the Poisson parameter  $\mu$  is given in the first row,

the number of events ( $k$ ) is given in the first column, and the corresponding Poisson probability is given in the  $k$  row and  $\mu$  column.

**Example 4.35** Compute the probability of obtaining at least 5 events for a Poisson distribution with parameter  $\mu = 3$ .

**Solution** Refer to Appendix Table 2 under the 3.0 column. Let  $X$  = the number of events.

$$Pr(X = 0) = .0498$$

$$Pr(X = 1) = .1494$$

$$Pr(X = 2) = .2240$$

$$Pr(X = 3) = .2240$$

$$Pr(X = 4) = .1680$$

$$\begin{aligned} \text{Thus } Pr(X \geq 5) &= 1 - Pr(X \leq 4) \\ &= 1 - (.0498 + .1494 + .2240 + .2240 + .1680) \\ &= 1 - .8152 = .1848 \end{aligned}$$

### Electronic Tables for the Poisson Distribution

In many instances we want to evaluate a collection of Poisson probabilities for the same  $\mu$ , but  $\mu$  is not given in Table 2 of the Appendix. For large  $\mu$  ( $\mu \geq 10$ ), a normal approximation, as given in Chapter 5, can be used. Otherwise, an electronic table similar to that presented for the binomial distribution can be used. The POISSON function of Excel 2007 can be used to compute individual and cumulative probabilities for the Poisson distribution (see Companion Website for details).

**Example 4.36** **Infectious Disease** Calculate the probability distribution of deaths caused by typhoid fever over a 1-year period using the information given in Example 4.33.

In this case, we model the number of deaths caused by typhoid fever by a Poisson distribution with  $\mu = 4.6$ . We will use the POISSON function of Excel 2007. The results are given in Table 4.8. We see that 9 or more deaths caused by typhoid fever would be unusual over a 1-year period.

**Table 4.8** Calculation of the probability distribution of the number of deaths caused by typhoid fever over a 1-year period using the POISSON function of Excel 2007

Number of deaths	Probability
0	<code>0.010 = POISSON (0, 4.6, false)</code>
1	<code>0.046 = POISSON (1, 4.6, false)</code>
2	<code>0.106 = POISSON (2, 4.6, false)</code>
3	<code>0.163 = POISSON (3, 4.6, false)</code>
4	<code>0.188 = POISSON (4, 4.6, false)</code>
5	<code>0.173 = POISSON (5, 4.6, false)</code>
6	<code>0.132 = POISSON (6, 4.6, false)</code>
7	<code>0.087 = POISSON (7, 4.6, false)</code>
8	<code>0.050 = POISSON (8, 4.6, false)</code>
<code>&lt;=8</code>	<code>0.955 = POISSON (8, 4.6, true)</code>
<code>&gt;=9</code>	<code>0.045 = 1 - POISSON (8, 4.6, true)</code>

## 4.12 Expected Value and Variance of the Poisson Distribution

In many instances we cannot predict whether the assumptions for the Poisson distribution in Section 4.10 are satisfied. Fortunately, the relationship between the expected value and variance of the Poisson distribution provides an important guideline that helps identify random variables that follow this distribution. This relationship can be stated as follows:

### Equation 4.9

For a Poisson distribution with parameter  $\mu$ , the mean and variance are both equal to  $\mu$ .

This fact is useful, because if we have a data set from a discrete distribution where the *mean and variance are about the same*, then we can preliminarily identify it as a Poisson distribution and use various tests to confirm this hypothesis.

### Example 4.37

**Infectious Disease** The number of deaths attributable to polio during the years 1968–1977 is given in Table 4.9 [4, 5]. Comment on the applicability of the Poisson distribution to this data set.

### Solution

The sample mean and variance of the annual number of deaths caused by polio during the period 1968–1977 are 18.0 and 23.1, respectively. The Poisson distribution will probably fit well here because the variance is approximately the same as the mean.

**Table 4.9**

**Number of deaths attributable to polio during the years 1968–1977**

Year	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
Number of deaths	15	10	19	23	15	17	23	17	26	15

Suppose we are studying a rare phenomenon and want to apply the Poisson distribution. A question that often arises is how to estimate the parameter  $\mu$  of the Poisson distribution in this context. Because the expected value of the Poisson distribution is  $\mu$ ,  $\mu$  can be estimated by the observed mean number of events over time  $t$  (e.g., 1 year), if such data are available. If the data are not available, other data sources can be used to estimate  $\mu$ .

### Example 4.38

**Occupational Health** A public-health issue arose concerning the possible carcinogenic potential of food ingredients containing ethylene dibromide (EDB). In some instances foods were removed from public consumption if they were shown to have excessive quantities of EDB. A previous study had looked at mortality in 161 white male employees of two plants in Texas and Michigan who were exposed to EDB over the time period 1940–1975 [6]. Seven deaths from cancer were observed among these employees. For this time period, 5.8 cancer deaths were expected as calculated from overall mortality rates for U.S. white men. Was the observed number of cancer deaths excessive in this group?

### Solution

Estimate the parameter  $\mu$  from the expected number of cancer deaths from U.S. white male mortality rates; that is,  $\mu = 5.8$ . Then calculate  $Pr(X \geq 7)$ , where  $X$  is a Poisson random variable with parameter  $\mu = 5.8$ . Use the relationship

**Table 4.10** Calculation of the probability distribution of the number of cancer deaths in the EDB example using the POISSON function of Excel 2007

Mean number of deaths	Probability
0	0.003 = POISSON (0, 5.8, false)
1	0.018 = POISSON (1, 5.8, false)
2	0.051 = POISSON (2, 5.8, false)
3	0.098 = POISSON (3, 5.8, false)
4	0.143 = POISSON (4, 5.8, false)
5	0.166 = POISSON (5, 5.8, false)
6	0.160 = POISSON (6, 5.8, false)
<= 6	0.638 = POISSON (6, 5.8, true)
>= 7	0.362 = 1 - POISSON (6, 5.8, true)

$$Pr(X \geq 7) = 1 - Pr(X \leq 6)$$

$$\text{where } Pr(X = k) = e^{-5.8} (5.8)^k / k!$$

Because  $\mu = 5.8$  is not in Table 2 of the Appendix let's use Excel 2007 to perform the calculations. Table 4.10 gives the results.

$$\begin{aligned} \text{Thus } Pr(X \geq 7) &= 1 - Pr(X \leq 6) \\ &= 1 - .638 = .362 \end{aligned}$$

Clearly, the observed number of cancer deaths is not excessive in this group.

### 4.13 Poisson Approximation to the Binomial Distribution

As noted in the preceding section, the Poisson distribution seems to fit well in some applications. Another important use for the Poisson distribution is as an approximation to the binomial distribution. Consider the binomial distribution for large  $n$  and small  $p$ . The mean of this distribution is given by  $np$  and the variance by  $npq$ . Note that  $q \approx 1$  for small  $p$ , and thus  $npq \approx np$ . Therefore the mean and variance of the binomial distribution are almost equal in this case, which suggests the following rule:

#### Equation 4.10

##### Poisson Approximation to the Binomial Distribution

The binomial distribution with large  $n$  and small  $p$  can be accurately approximated by a Poisson distribution with parameter  $\mu = np$ .

The rationale for using this approximation is that the Poisson distribution is easier to work with than the binomial distribution. The binomial distribution involves expressions such as  $\binom{n}{k}$  and  $(1-p)^{n-k}$ , which are cumbersome for large  $n$ .

#### Example 4.39

**Cancer, Genetics** Suppose we are interested in the genetic susceptibility to breast cancer. We find that 4 of 1000 women ages 40–49 whose mothers have had breast

cancer also develop breast cancer over the next year of life. We would expect from large population studies that 1 in 1000 women of this age group will develop a new case of the disease over this period of time. How unusual is this event?

### Solution

The exact binomial probability could be computed by letting  $n = 1000$ ,  $p = 1/1000$ . Hence

$$\begin{aligned} Pr(X \geq 4) &= 1 - Pr(X \leq 3) \\ &= 1 - \left[ \binom{1000}{0} (.001)^0 (.999)^{1000} + \binom{1000}{1} (.001)^1 (.999)^{999} \right. \\ &\quad \left. + \binom{1000}{2} (.001)^2 (.999)^{998} + \binom{1000}{3} (.001)^3 (.999)^{997} \right] \end{aligned}$$

Instead, use the Poisson approximation with  $\mu = 1000(.001) = 1$ , which is obtained as follows:

$$Pr(X \geq 4) = 1 - [Pr(X = 0) + Pr(X = 1) + Pr(X = 2) + Pr(X = 3)]$$

Using Table 2 of the Appendix under the  $\mu = 1.0$  column, we find that

$$Pr(X = 0) = .3679$$

$$Pr(X = 1) = .3679$$

$$Pr(X = 2) = .1839$$

$$Pr(X = 3) = .0613$$

$$\begin{aligned} \text{Thus } Pr(X \geq 4) &= 1 - (.3679 + .3679 + .1839 + .0613) \\ &= 1 - .9810 = .0190 \end{aligned}$$

This event is indeed unusual and suggests a genetic susceptibility to breast cancer among daughters of women who have had breast cancer. For comparative purposes, we have computed the exact binomial probabilities of obtaining 0, 1, 2, and 3 events, which are given by .3677, .3681, .1840, and .0613, respectively. The corresponding exact binomial probability of obtaining 4 or more breast-cancer cases is .0189, which agrees almost exactly with the Poisson approximation of .0190 just given.

How large should  $n$  be or how small should  $p$  be before the approximation is "adequate"? A conservative rule is to use the approximation when  $n \geq 100$  and  $p \leq .01$ . As an example, we give the exact binomial probability and the Poisson approximation for  $n = 100$ ,  $p = .01$ ,  $k = 0, 1, 2, 3, 4, 5$  in Table 4.11. The two probability distributions agree to within .002 in all instances.

**Table 4.11** An example of the Poisson approximation to the binomial distribution for  $n = 100$ ,  $p = .01$ ,  $k = 0, 1, \dots, 5$

$k$	Exact binomial probability	Poisson approximation	$k$	Exact binomial probability	Poisson approximation
0	.366	.368	3	.061	.061
1	.370	.368	4	.015	.015
2	.185	.184	5	.003	.003

**Example 4.40**

**Infectious Disease** An outbreak of poliomyelitis occurred in Finland in 1984 after 20 years without a single case being reported in the country. As a result, an intensive immunization campaign was conducted within 5 weeks between February 9 and March 15, 1985; it covered 94% of the population and was highly successful. During and after the campaign, several patients with Guillain-Barré syndrome (GBS), a rare neurologic disease often resulting in paralysis, were admitted to the neurologic units of hospitals in Finland [7].

The authors provided data on monthly incidence of GBS from April 1984 to October 1985. These data are given in Table 4.12.

**Table 4.12****Monthly incidence of GBS in Finland from April 1984 to October 1985**

Month	Number of GBS cases	Month	Number of GBS cases	Month	Number of GBS cases
April 1984	3	October 1984	2	April 1985	7
May 1984	7	November 1984	2	May 1985	2
June 1984	0	December 1984	3	June 1985	2
July 1984	3	January 1985	3	July 1985	6
August 1984	4	February 1985	8	August 1985	2
September 1984	4	March 1985	14	September 1985	2
				October 1985	6

Determine whether the number of cases in March 1985 is excessive compared with the experience in the other 18 months based on the data in Table 4.12.

**Solution**

If there are  $n$  people in Finland who could get GBS and the monthly incidence of GBS ( $p$ ) is low, then we could model the number of GBS cases in 1 month ( $X$ ) by a binomial distribution with parameters  $n$  and  $p$ . Because  $n$  is large and  $p$  is small, it is reasonable to approximate the distribution of the number of GBS cases in 1 month ( $X$ ) by a Poisson distribution with parameter  $\mu = np$ . To estimate  $\mu$ , we use the average monthly number of GBS cases during the 18-month period from April 1984 to October 1985, excluding the vaccine month of March 1985. The mean number of cases per month =  $(3 + 7 + \dots + 6)/18 = 3.67$ . We now assess whether the number of cases in March 1985 (14) is excessive by computing  $Pr(X \geq 14 | \mu = 3.67)$ . We use Excel 2007 to perform this computation, as shown in Table 4.13.

**Table 4.13****Probability of observing 14 or more cases of GBS in Finland during March 1985**

Mean	3.67
<b>Pr(X &lt;= 13)</b>	<b>0.999969 = POISSON (13, 3.67, true)</b>
<b>Pr(X &gt;= 14)</b>	<b>3.09E-05 = 1 - POISSON (13, 3.67, true)</b>

The results indicate that  $Pr(X \geq 14 | \mu = 3.67) = 3.09 \times 10^{-5}$ . Thus 14 cases in 1 month is very unusual, given the 18-month experience in nonvaccine months, and possibly indicates that the many cases in March 1985 are attributable to the vaccination campaign.

## REVIEW QUESTIONS 4D

- 1 Suppose the number of motor-vehicle fatalities in a city during a week is Poisson-distributed, with an average of 8 fatalities per week.
  - (a) What is the probability that 12 fatalities occur in a specific week?
  - (b) What is the probability that at least 12 fatalities occur during a specific week?
  - (c) How many motor-vehicle fatalities would have to occur during a given week to conclude there are an unusually high number of events in that week?  
*(Hint: Refer to Example 4.36.)*
- 2 Suppose a rare infectious disease occurs at the rate of  $2 \times 10^{-6}$  people per year.
  - (a) What is the probability that in New York City (population about 8 million) exactly 25 cases occur in a given year?
  - (b) What is the probability that at least 25 cases occur in a given year?  
*(Hint: Use the Poisson approximation to the binomial distribution.)*

## 4.14 Summary

In this chapter, random variables were discussed and a distinction between discrete and continuous random variables was made. Specific attributes of random variables, including the notions of probability-mass function (or probability distribution), cdf, expected value, and variance were introduced. These notions were shown to be related to similar concepts for finite samples, as discussed in Chapter 2. In particular, the sample-frequency distribution is a sample realization of a probability distribution, whereas the sample mean ( $\bar{x}$ ) and variance ( $s^2$ ) are sample analogs of the expected value and variance, respectively, of a random variable. The relationship between attributes of probability models and finite samples is explored in more detail in Chapter 6.

Finally, some specific probability models were introduced, focusing on the binomial and Poisson distributions. The binomial distribution was shown to be applicable to binary outcomes, that is, if only two outcomes are possible, where outcomes on different trials are independent. These two outcomes are labeled as “success” and “failure,” where the probability of success is the same for each trial. The Poisson distribution is a classic model used to describe the distribution of rare events.

The study of probability models continues in Chapter 5, where the focus is on continuous random variables.

## PROBLEMS

Let  $X$  be the random variable representing the number of hypertensive adults in Example 3.12.

\*4.1 Derive the probability-mass function for  $X$ .

\*4.2 What is its expected value?

\*4.3 What is its variance?

\*4.4 What is its cumulative-distribution function?

Suppose we want to check the accuracy of self-reported diagnoses of angina by getting further medical records on a subset of the cases.

4.5 If we have 50 reported cases of angina and we want to select 5 for further review, then how many ways can we select these cases if order of selection matters?

4.6 Answer Problem 4.5 if order of selection does not matter.

4.7 Evaluate  $\binom{10}{0}, \binom{10}{1}, \dots, \binom{10}{10}$ .

\*4.8 Evaluate  $9!$ .

4.9 Suppose 6 of 15 students in a grade-school class develop influenza, whereas 20% of grade-school students

nationwide develop influenza. Is there evidence of an excessive number of cases in the class? That is, what is the probability of obtaining at least 6 cases in this class if the nationwide rate holds true?

**4.10** What is the expected number of students in the class who will develop influenza?

\***4.11** What is the probability of obtaining exactly 6 events for a Poisson distribution with parameter  $\mu = 4.0$ ?

**4.12** What is the probability of obtaining at least 6 events for a Poisson distribution with parameter  $\mu = 4.0$ ?

\***4.13** What is the expected value and variance for a Poisson distribution with parameter  $\mu = 4.0$ ?

### Infectious Disease

Newborns were screened for human immunodeficiency virus (HIV) or acquired immunodeficiency syndrome (AIDS) in five Massachusetts hospitals. The data [8] are shown in Table 4.14.

**4.14** If 500 newborns are screened at the inner-city hospital, then what is the exact binomial probability of 5 HIV-positive test results?

**4.15** If 500 newborns are screened at the inner-city hospital, then what is the exact binomial probability of at least 5 HIV-positive test results?

**4.16** Answer Problems 4.14 and 4.15 using an approximation rather than an exact probability.

**4.17** Answer Problem 4.14 for a mixed urban/suburban hospital (hospital C).

**4.18** Answer Problem 4.15 for a mixed urban/suburban hospital (hospital C).

**4.19** Answer Problem 4.16 for a mixed urban/suburban hospital (hospital C).

**4.20** Answer Problem 4.14 for a mixed suburban/rural hospital (hospital E).

**4.21** Answer Problem 4.15 for a mixed suburban/rural hospital (hospital E).

**4.22** Answer Problem 4.16 for a mixed suburban/rural hospital (hospital E).

### Occupational Health

Many investigators have suspected that workers in the tire industry have an unusual incidence of cancer.

\***4.23** Suppose the expected number of deaths from bladder cancer for all workers in a tire plant on January 1, 1964, over the next 20 years (1/1/64–12/31/83) based on U.S. mortality rates is 1.8. If the Poisson distribution is assumed to hold and 6 reported deaths are caused by bladder cancer among the tire workers, how unusual is this event?

\***4.24** Suppose a similar analysis is done for stomach cancer. In this plant, 4 deaths are caused by stomach cancer, whereas 2.5 are expected based on U.S. mortality rates. How unusual is this event?

### Infectious Disease

One hypothesis is that gonorrhea tends to cluster in central cities.

**4.25** Suppose 10 gonorrhea cases are reported over a 3-month period among 10,000 people living in an urban county. The statewide incidence of gonorrhea is 50 per 100,000 over a 3-month period. Is the number of gonorrhea cases in this county unusual for this time period?

### Otolaryngology

Assume the number of episodes per year of otitis media, a common disease of the middle ear in early childhood, follows a Poisson distribution with parameter  $\lambda = 1.6$  episodes per year.

\***4.26** Find the probability of getting 3 or more episodes of otitis media in the first 2 years of life.

\***4.27** Find the probability of not getting any episodes of otitis media in the first year of life.

An interesting question in pediatrics is whether the tendency for children to have many episodes of otitis media is inherited in a family.

\***4.28** What is the probability that 2 siblings will both have 3 or more episodes of otitis media in the first 2 years of life?

\***4.29** What is the probability that exactly 1 sibling will have 3 or more episodes in the first 2 years of life?

**Table 4.14 Seroprevalence of HIV antibody in newborns' blood samples, according to hospital category**

Hospital	Type	Number tested	Number positive	Number positive (per 1000)
A	Inner city	3741	30	8.0
B	Urban/suburban	11,864	31	2.6
C	Urban/suburban	5006	11	2.2
D	Suburban/rural	3596	1	0.3
E	Suburban/rural	6501	8	1.2

\***4.30** What is the probability that neither sibling will have 3 or more episodes in the first 2 years of life?

**4.31** What is the expected number of siblings in a 2-sibling family who will have 3 or more episodes in the first 2 years of life?

### Environmental Health, Obstetrics

Suppose the rate of major congenital malformations in the general population is 2.5 malformations per 100 deliveries. A study is set up to investigate whether offspring of Vietnam-veteran fathers are at special risk for congenital malformations.

\***4.32** If 100 infants are identified in a birth registry as offspring of Vietnam-veteran fathers and 4 have a major congenital malformation, is there an excess risk of malformations in this group?

Using the same birth-registry data, let's look at the effect of maternal use of marijuana on the rate of major congenital malformations.

\***4.33** Of 75 offspring of mothers who used marijuana, 8 have a major congenital malformation. Is there an excess risk of malformations in this group?

### Hypertension

A national study found that treating people appropriately for high blood pressure reduced their overall mortality by 20%. Treating people adequately for hypertension has been difficult because it is estimated that 50% of hypertensives do not know they have high blood pressure, 50% of those who do know are inadequately treated by their physicians, and 50% who are appropriately treated fail to follow this treatment by taking the right number of pills.

**4.34** What is the probability that among 10 true hypertensives at least 50% are being treated appropriately and are complying with this treatment?

**4.35** What is the probability that at least 7 of the 10 hypertensives know they have high blood pressure?

**4.36** If the preceding 50% rates were each reduced to 40% by a massive education program, then what effect would this change have on the overall mortality rate among true hypertensives; that is, would the mortality rate decrease and, if so, what percentage of deaths among hypertensives could be prevented by the education program?

### Renal Disease

The presence of bacteria in a urine sample (bacteriuria) is sometimes associated with symptoms of kidney disease in women. Suppose a determination of bacteriuria has been made over a large population of women at one point in time and 5% of those sampled are positive for bacteriuria.

\***4.37** If a sample size of 5 is selected from this population, what is the probability that 1 or more women are positive for bacteriuria?

**4.38** Suppose 100 women from this population are sampled. What is the probability that 3 or more of them are positive for bacteriuria?

One interesting phenomenon of bacteriuria is that there is a "turnover"; that is, if bacteriuria is measured on the same woman at two different points in time, the results are not necessarily the same. Assume that 20% of all women who are bacteriuric at time 0 are again bacteriuric at time 1 (1 year later), whereas only 4.2% of women who were not bacteriuric at time 0 are bacteriuric at time 1. Let  $X$  be the random variable representing the number of bacteriuric events over the two time periods for 1 woman and still assume that the probability that a woman will be positive for bacteriuria at any one exam is 5%.

\***4.39** What is the probability distribution of  $X$ ?

**4.40** What is the mean of  $X$ ?

**4.41** What is the variance of  $X$ ?

### Pediatrics, Otolaryngology

Otitis media is a disease that occurs frequently in the first few years of life and is one of the most common reasons for physician visits after the routine checkup. A study was conducted to assess the frequency of otitis media in the general population in the first year of life. Table 4.15 gives the number of infants of 2500 infants who were first seen at birth who remained disease-free by the end of the  $i$ th month of life,  $i = 0, 1, \dots, 12$ . (Assume no infants have been lost to follow-up.)

**Table 4.15 Number of infants (of 2500) who remain disease-free at the end of each month during the first year of life**

$i$	Disease-free infants at the end of month $i$
0	2500
1	2425
2	2375
3	2300
4	2180
5	2000
6	1875
7	1700
8	1500
9	1300
10	1250
11	1225
12	1200

\***4.42** What is the probability that an infant will have one or more episodes of otitis media by the end of the sixth month of life? The first year of life?

\***4.43** What is the probability that an infant will have one or more episodes of otitis media by the end of the ninth month of life given that no episodes have been observed by the end of the third month of life?

\***4.44** Suppose an "otitis-prone family" is defined as one in which at least three siblings of five develop otitis media in the first 6 months of life. What proportion of five-sibling families is otitis prone if we assume the disease occurs independently for different siblings in a family?

\***4.45** What is the expected number of otitis-prone families of 100 five-sibling families?

### Cancer, Epidemiology

An experiment is designed to test the potency of a drug on 20 rats. Previous animal studies have shown that a 10-mg dose of the drug is lethal 5% of the time within the first 4 hours; of the animals alive at 4 hours, 10% will die in the next 4 hours.

**4.46** What is the probability that 3 or more rats will die in the first 4 hours?

**4.47** Suppose 2 rats die in the first 4 hours. What is the probability that 2 or fewer rats will die in the next 4 hours?

**4.48** What is the probability that 0 rats will die in the 8-hour period?

**4.49** What is the probability that 1 rat will die in the 8-hour period?

**4.50** What is the probability that 2 rats will die in the 8-hour period?

**4.51** Can you write a general formula for the probability that  $x$  rats will die in the 8-hour period? Evaluate this formula for  $x = 0, 1, \dots, 10$ . (*Hint:* Use the BINOMDIST function of Excel 2007.)

### Environmental Health

An important issue in assessing nuclear energy is whether excess disease risks exist in the communities surrounding nuclear-power plants. A study undertaken in the community surrounding Hanford, Washington, looked at the prevalence of selected congenital malformations in the counties surrounding the nuclear-test facility [9].

\***4.52** Suppose 27 cases of Down's syndrome are found and only 19 are expected based on Birth Defects Monitoring Program prevalence estimates in the states of Washington, Idaho, and Oregon. Are there significant excess cases in the area around the nuclear-power plant?

Suppose 12 cases of cleft palate are observed, whereas only 7 are expected based on Birth Defects Monitoring Program estimates.

\***4.53** What is the probability of observing exactly 12 cases of cleft palate if there is no excess risk of cleft palate in the study area?

\***4.54** Do you feel there is a meaningful excess number of cases of cleft palate in the area surrounding the nuclear-power plant? Explain.

### Health Promotion

A study was conducted among 234 people who had expressed a desire to stop smoking but who had not yet stopped. On the day they quit smoking, their carbon-monoxide level (CO) was measured and the time was noted from the time they smoked their last cigarette to the time of the CO measurement. The CO level provides an "objective" indicator of the number of cigarettes smoked per day during the time immediately before the quit attempt. However, it is known to also be influenced by the time since the last cigarette was smoked. Thus, this time is provided as well as a "corrected CO level," which is adjusted for the time since the last cigarette was smoked. Information is also provided on the age and sex of the participants as well as each participant's self-report of the number of cigarettes smoked per day. The participants were followed for 1 year for the purpose of determining the number of days they remained abstinent. Number of days abstinent ranged from 0 days for those who quit for less than 1 day to 365 days for those who were abstinent for the full year. Assume all people were followed for the entire year.

The data, provided by Dr. Arthur J. Garvey, Boston, Massachusetts, are given in Data Set SMOKE.DAT, on the Companion Website. The format of this file is given in Table 4.16.

**4.55** Develop a life table similar to Table 4.15, giving the number of people who remained abstinent at 1, 2,  $\dots$ , 12 months of life (assume for simplicity that there are 30 days in each of the first 11 months after quitting and 35 days in the 12th month). Plot these data on the computer using the Chart Wizard of Excel 2007 or use some other statistical package. Compute the probability that a person will remain abstinent at 1, 3, 6, and 12 months after quitting.

**4.56** Develop life tables for subsets of the data based on age, sex, number of cigarettes per day, and CO level (one variable at a time). Given these data, do you feel age, sex, number of cigarettes per day, and/or CO level are related to success in quitting? (Methods of analysis for life-table data are discussed in more detail in Chapter 14.)

**Table 4.16 Format of SMOKE.DAT**

Variable	Columns	Code
ID number	1–3	
Age	4–5	
Gender	6	1 = male, 2 = female
Cigarettes/day	7–8	
CO ( $\times 10$ )	9–11	
Minutes elapsed since the last cigarette smoked	12–15	
LogCOAdj <sup>a</sup> ( $\times 1000$ )	16–19	
Days abstinent <sup>b</sup>	20–22	

<sup>a</sup>This variable represents adjusted CO values. CO values were adjusted for minutes elapsed since the last cigarette smoked using the formula,  $\log_{10}\text{CO} (\text{adjusted}) = \log_{10}\text{CO} - (-0.000638) \times (\text{min} - 80)$ , where min is the number of minutes elapsed since the last cigarette smoked.

<sup>b</sup>Those abstinent less than 1 day were given a value of 0.

## Genetics

**4.57** A topic of some interest in the genetic literature over at least the past 30 years has been the study of sex-ratio data. In particular, one suggested hypothesis is that there are enough families with a preponderance of males (females) that the sexes of successive childbirths are not independent random variables but rather are related to each other. This hypothesis has been extended beyond just successive births, so some authors also consider relationships between offspring two birth orders apart (first and third offspring, second and fourth offspring, etc.). Sex-ratio data from the first 5 births in 51,868 families are given in Data Set SEXRAT.DAT (on the Companion Website). The format of this file is given in Table 4.17 [10]. What are your conclusions concerning the preceding hypothesis based on your analysis of these data?

## Infectious Disease

A study considered risk factors for HIV infection among intravenous drug users [11]. It found that 40% of users who had  $\leq 100$  injections per month (light users) and 55% of users who had  $> 100$  injections per month (heavy users) were HIV positive.

**Table 4.17 Format of SEXRAT.DAT**

Variable	Column
Number of children <sup>a</sup>	1
Sex of children <sup>b</sup>	3–7
Number of families	9–12

<sup>a</sup>For families with 5+ children, the sexes of the first 5 children are listed. The number of children is given as 5 for such families.

<sup>b</sup>The sex of successive births is given. Thus, MMMF means the first 3 children were males and the fourth child was a female. There were 484 such families.

**4.58** What is the probability that exactly 3 of 5 light users are HIV positive?

**4.59** What is the probability that at least 3 of 5 light users are HIV positive?

**4.60** Suppose we have a group of 10 light users and 10 heavy users. What is the probability that exactly 3 of the 20 users are HIV positive?

**4.61** What is the probability that at least 4 of the 20 users are HIV positive?

**4.62** Is the distribution of the number of HIV positive among the 20 users binomial? Why or why not?

## Ophthalmology, Diabetes

A recent study [12] of incidence rates of blindness among insulin-dependent diabetics reported that the annual incidence rate of blindness per year was 0.67% among 30- to 39-year-old male insulin-dependent diabetics (IDDM) and 0.74% among 30- to 39-year-old female insulin-dependent diabetics.

**4.63** If a group of 200 IDDM 30- to 39-year-old men is followed, what is the probability that exactly 2 will go blind over a 1-year period?

**4.64** If a group of 200 IDDM 30- to 39-year-old women is followed, what is the probability that at least 2 will go blind over a 1-year period?

**4.65** What is the probability that a 30-year-old IDDM male patient will go blind over the next 10 years?

**4.66** After how many years of follow-up would we expect the cumulative incidence of blindness to be 10% among 30-year-old IDDM females, if the incidence rate remains constant over time?

**4.67** What does cumulative incidence mean, in words, in the context of this problem?

### Cardiovascular Disease

An article was published [13] concerning the incidence of cardiac death attributable to the earthquake in Los Angeles County on January 17, 1994. In the week before the earthquake there were an average of 15.6 cardiac deaths per day in Los Angeles County. On the day of the earthquake, there were 51 cardiac deaths.

**4.68** What is the exact probability of 51 deaths occurring on one day if the cardiac death rate in the previous week continued to hold on the day of the earthquake?

**4.69** Is the occurrence of 51 deaths unusual? (*Hint:* Use the same methodology as in Example 4.30.)

**4.70** What is the maximum number of cardiac deaths that could have occurred on the day of the earthquake to be consistent with the rate of cardiac deaths in the past week? (*Hint:* Use a cutoff probability of .05 to determine the maximum number.)

### Environmental Health

Some previous studies have shown a relationship between emergency-room admissions per day and level of pollution on a given day. A small local hospital finds that the number of admissions to the emergency ward on a single day ordinarily (unless there is unusually high pollution) follows a Poisson distribution with mean = 2.0 admissions per day. Suppose each admitted person to the emergency ward stays there for exactly 1 day and is then discharged.

**4.71** The hospital is planning a new emergency-room facility. It wants enough beds in the emergency ward so that for at least 95% of normal-pollution days it will not need to turn anyone away. What is the smallest number of beds it should have to satisfy this criterion?

**4.72** The hospital also finds that on high-pollution days the number of admissions is Poisson-distributed with mean = 4.0 admissions per day. Answer Problem 4.71 for high-pollution days.

**4.73** On a random day during the year, what is the probability there will be 4 admissions to the emergency ward, assuming there are 345 normal-pollution days and 20 high-pollution days?

**4.74** Answer Problem 4.71 for a random day during the year.

### Women's Health

The number of legal induced abortions per year per 1000 U.S. women ages 15–44 [14] is given in Table 4.18.

For example, of 1000 women ages 15–44 in 1980, 25 had a legal induced abortion during 1980.

**4.75** If we assume (1) no woman has more than 1 abortion and (2) the probability of having an abortion is independent across different years, what is the probability that a 15-year-old woman in 1975 will have an abortion over her 30 years of reproductive life (ages 15–44, or 1975–2004)?

**Table 4.18 Annual incidence of legal induced abortions by time period**

Year	Legal induced abortions per year per 1000 women ages 15–44
1975–1979	21
1980–1984	25
1985–1989	24
1990–1994	24
1995–2004	20

Studies have been undertaken to assess the relationship between abortion and the development of breast cancer. In one study among nurses (the Nurses' Health Study II), there were 16,359 abortions among 2,169,321 person-years of follow-up for women of reproductive age. (*Note:* 1 person-year = 1 woman followed for 1 year.)

**4.76** What is the expected number of abortions among nurses over this time period if the incidence of abortion is 25 per 1000 women per year and no woman has more than 1 abortion?

**4.77** Does the abortion rate among nurses differ significantly from the national experience? Why or why not? (*Hint:* Use the Poisson distribution.) A yes/no answer is not acceptable.

### Endocrinology

**4.78** Consider the Data Set BONEDEN.DAT on the Companion Website. Calculate the difference in bone density of the lumbar spine ( $\text{g}/\text{cm}^2$ ) between the heavier-smoking twin and the lighter-smoking twin (bone density for the heavier-smoking twin minus bone density for the lighter-smoking twin) for each of the 41 twin pairs. Suppose smoking has no relationship to bone density. What would be the expected number of twin pairs with negative difference scores? What is the actual number of twin pairs with negative difference scores? Do you feel smoking is related to bone density of the lumbar spine, given the observed results? Why or why not? A yes/no answer is not acceptable. (*Hint:* Use the binomial distribution.)

**4.79** Sort the differences in smoking between members of a twin pair (expressed in pack-years). Identify the subgroup of 20 twin pairs with the largest differences in smoking. Answer Problem 4.78 based on this subgroup of 20 twin pairs.

**4.80** Answer Problem 4.78 for bone density of the femoral neck.

**4.81** Answer Problem 4.79 for bone density of the femoral neck.

**4.82** Answer Problem 4.78 for bone density of the femoral shaft.

**4.83** Answer Problem 4.79 for bone density of the femoral shaft.

## SIMULATION

An attractive feature of modern statistical packages such as MINITAB or Excel is the ability to use the computer to simulate random variables on the computer and to compare the characteristics of the observed samples with the theoretical properties of the random variables.

**4.84** Draw 100 random samples from a binomial distribution, each based on 10 trials with probability of success = .05 on each trial. Obtain a frequency distribution of the number of successes over the 100 random samples, and plot the distribution. How does it compare with Figure 4.4(a)?

**4.85** Answer Problem 4.84 for a binomial distribution with parameters  $n = 10$  and  $p = .95$ . Compare your results with Figure 4.4(b).

**4.86** Answer Problem 4.84 for a binomial distribution with parameters  $n = 10$  and  $p = .5$ . Compare your results with Figure 4.4(c).

## Cancer

The two-stage model of carcinogenesis is based on the premise that for a cancer to develop, a normal cell must first undergo a "first hit" and mutate to become a susceptible or intermediate cell. An intermediate cell then must undergo a "second hit" and mutate to become a malignant cell. A cancer develops if at least one cell becomes a malignant cell. This model has been applied to the development of breast cancer in females (Moolgavkar et al. [15]).

Suppose there are  $10^8$  normal breast cells and 0 intermediate or malignant breast cells among 20-year-old females. The probability that a normal breast cell will mutate to become an intermediate cell is  $10^{-7}$  per year.

**4.87** What is the probability that there will be at least 5 intermediate cells by age 21? (*Hint:* Use the Poisson distribution.)

**4.88** What is the expected number of intermediate cells by age 45?

The probability that an intermediate cell will mutate to become a malignant cell is  $5 \times 10^{-7}$  per year.

**4.89** Suppose a woman has 300 intermediate cells by age 45. What is the probability that she develops breast cancer by age 46? By age 50? (*Hint:* Use the Poisson approximation to the binomial distribution.)

**4.90** Under the preceding assumptions, what is the probability that a 20-year-old woman will develop breast cancer by age 45? By age 50? (*Hint:* Use a computer to solve this problem. Check your results by simulation.)

## Dentistry

The data in Table 4.19 were reported by men in the Health Professionals Follow-up Study on the number of teeth lost over a 1-year period (January 1, 1987 to December 31, 1987).

**Table 4.19 Distribution of number of teeth lost from January 1, 1987 to December 31, 1987 among 38,905 men in the Health Professionals Follow-up Study**

Number of teeth lost	Frequency
0	35,763
1	1,978
2	591
3	151
4	163
5–9	106
10+	153
Total	38,905

**4.91** If we assume the average number of teeth lost in the 5–9 group is 7 teeth and the average number of teeth lost in the 10+ group is 12 teeth, what is the best estimate of the average number of teeth lost per year?

**4.92** Suppose that on January 1, 1987, a man is 50 years old, that he will live for 30 more years (until 2016), and that the rate of tooth loss over this 30-year period is the same as in 1987. If a man has 13 teeth remaining on January 1, 1987, what is the probability he will need dentures (have 10 or fewer teeth remaining) during his 30-year lifetime? (*Hint:* Use the Poisson distribution.)

**4.93** Suppose dental practice improves over the 30-year period. We assume the rate of tooth loss per year from 1987–2001 (15 years) is the same as in 1987, whereas the rate of tooth loss per year from 2002–2016 (15 years) is half the 1987 rate. What is the probability that the man in Problem 4.92 will require dentures under these altered assumptions? (*Hint:* Consider a mixture of two Poisson distributions.)

## Genetics, Ophthalmology

Mr. G. has retinitis pigmentosa, a genetic ocular disease with several different types of inheritance patterns. In Mr. G.'s family, the mode of inheritance is unknown. However, the most reasonable possibilities are either a recessive or a sex-linked mode of inheritance. Under a recessive mode of inheritance, there are usually two unaffected parents, but there is a 25% probability that each child will have the disease.

Mr. G. has two unaffected brothers and two unaffected sisters.

**4.94** What is the probability of this phenotype among the five children in this family if the mode of inheritance was recessive?

Under a sex-linked mode of inheritance, each male offspring has a 50% chance of inheriting the disease, while the female offspring have no chance of getting the disease.

**Table 4.20 Relationship between incidence of birth defects and census tract**

Census tract	SES	Number of births/yr	Incidence of birth defects	Census tract	SES	Number of births/yr	Incidence of birth defects
A	High	5000	50/10 <sup>5</sup>	E	Low	7000	100/10 <sup>5</sup>
B	Low	12,000	100/10 <sup>5</sup>	F	Low	20,000	100/10 <sup>5</sup>
C	Low	10,000	100/10 <sup>5</sup>	G	High	5000	50/10 <sup>5</sup>
D	Low	8000	100/10 <sup>5</sup>	H	Low	3000	100/10 <sup>5</sup>
					Total	70,000	

**4.95** What is the probability of this phenotype among the five children in this family if the mode of inheritance was sex-linked?

In the retinitis pigmentosa population as a whole, about 40% of cases are recessive, 10% are sex-linked, and 50% are dominant. The mother and father of Mr. G. are normal and hence it is reasonable to assume that the probability of this phenotype under a dominant mode of inheritance is 0.

**4.96** What is the probability that the mode of inheritance for Mr. G.'s family is recessive? Sex-linked?

### Hospital Epidemiology

Suppose the number of admissions to the emergency room at a small hospital follows a Poisson distribution, but the incidence rate changes on different days of the week. On a weekday there are on average two admissions per day, while on a weekend day there is on average one admission per day.

**4.97** What is the probability of at least one admission on a Wednesday?

**4.98** What is the probability of at least one admission on a Saturday?

**4.99** What is the probability of having 0, 1, and 2+ admissions for an entire week, if the results for different days during the week are assumed to be independent?

### Obstetrics

Suppose the incidence of a specific birth defect in a high socioeconomic status (SES) census tract is 50 cases per 100,000 births.

**4.100** If there are 5000 births in the census tract in 1 year, then what is the probability that there will be exactly 5 cases of the birth defect during the year (census tract A in Table 4.20)?

Suppose the incidence of the same birth defect in a low SES census tract is 100 cases per 100,000 births.

**4.101** If there are 12,000 births in the census tract in 1 year, then what is the probability that there will be at least 8 cases of the birth defect during the year (census tract B in Table 4.20)?

Suppose a city is divided into eight census tracts as shown in Table 4.20.

**4.102** Suppose a child is born with the birth defect, but the address of the mother is unknown. What is the probability that the child comes from a low SES census tract?

**4.103** What is the expected number of cases over 1 year in the city?

### REFERENCES

- [1] Boston Globe, October 7, 1980.
- [2] Rinsky, R. A., Zumwalde, R. O., Waxweiler, R. J., Murray, W. E., Bierbaum, P. J., Landrigan, P. J., Terpilak, M., & Cox, C. (1981, January 31). Cancer mortality at a naval nuclear shipyard. *Lancet*, 231–235.
- [3] National Center for Health Statistics. (2007, December 28). *National vital statistics report*, 56(9).
- [4] National Center for Health Statistics. (1974, June 27). *Monthly vital statistics report, annual summary for the United States* (1973), 22(13).
- [5] National Center for Health Statistics. (1978, December 7). *Monthly vital statistics report, annual summary for the United States* (1977), 26(13).
- [6] Ott, M. G., Scharnweber, H. C., & Langner, R. (1980). Mortality experience of 161 employees exposed to ethylene dibromide in two production units. *British Journal of Industrial Medicine*, 37, 163–168.
- [7] Kinnunen, E., Junntila, O., Haukka, J., & Hovi, T. (1998). Nationwide oral poliovirus vaccination campaign and the incidence of Guillain-Barré syndrome. *American Journal of Epidemiology*, 147(1), 69–73.
- [8] Hoff, R., Berardi, V. P., Weiblen, B. J., Mahoney-Trout, L., Mitchell, M. L., & Grady, G. R. (1988). Sero-prevalence of human immunodeficiency virus among childbearing women. *New England Journal of Medicine*, 318(9), 525–530.

- [9] Sever, L. E., Hessol, N. A., Gilbert, E. S., & McIntyre, J. M. (1988). The prevalence at birth of congenital malformations in communities near the Hanford site. *American Journal of Epidemiology*, 127(2), 243–254.
- [10] Renkonen, K. O., Mäkelä, O., & Lehtovaara, R. (1961). Factors affecting the human sex ratio. *Annales Medicinae Experimentalis et Biologiae Fenniae*, 39, 173–184.
- [11] Schoenbaum, E. E., Hartel, D., Selwyn, P. A., Klein, R. S., Davenny, K., Rogers, M., Feiner, C., & Friedland, G. (1989). Risk factors for human immunodeficiency virus infection in intravenous drug users. *New England Journal of Medicine*, 321(13), 874–879.
- [12] Sjolie, A. K., & Green, A. (1987). Blindness in insulin-treated diabetic patients with age at onset less than 30 years. *Journal of Chronic Disease*, 40(3), 215–220.
- [13] Leor, J., Poole, W. K., & Kloner, R. A. (1996). Sudden cardiac death triggered by an earthquake. *New England Journal of Medicine*, 334(7), 413–419.
- [14] National Center for Health Statistics. (1997, December 5). *Morbidity and mortality weekly report* (1980), 46(48).
- [15] Moolgavkar, S. H., Day, N. E., & Stevens, R. G. (1980). Two-stage model for carcinogenesis: Epidemiology of breast cancer in females. *Journal of the National Cancer Institute*, 65, 559–569.

# 5

## Continuous Probability Distributions

### 5.1 Introduction

This chapter discusses continuous probability distributions. Specifically, the normal distribution—the most widely used distribution in statistical work—is explored in depth.

The normal, or Gaussian or “bell-shaped,” distribution is the cornerstone of most methods of estimation and hypothesis testing developed throughout the rest of this text. Many random variables, such as distribution of birthweights or blood pressures in the general population, tend to follow approximately a normal distribution. In addition, many random variables that are not themselves normal are closely approximated by a normal distribution when summed many times. In such cases, using the normal distribution is desirable because it is easy to use and tables for the normal distribution are more widely available than are tables for many other distributions.

**Example 5.1** **Infectious Disease** The number of neutrophils in a sample of 2 white blood cells is not normally distributed, but the number in a sample of 100 white blood cells is very close to being normally distributed.

### 5.2 General Concepts

We want to develop an analog for a continuous random variable to the concept of a probability-mass function, as was developed for a discrete random variable in Section 4.3. Thus we would like to know which values are more probable than others and how probable they are.

**Example 5.2** **Hypertension** Consider the distribution of diastolic blood-pressure (DBP) measurements in 35- to 44-year-old men. In actual practice, this distribution is discrete because only a finite number of blood-pressure values are possible since the measurement is only accurate to within 2 mm Hg. However, assume there is no measurement error and hence the random variable can take on a continuum of possible values. One consequence of this assumption is that the probabilities of specific blood-pressure measurement values such as 117.3 are 0, and thus the concept of a probability-mass function cannot be used. The proof of this statement is beyond the scope of this text. Instead, we speak in terms of the probability that blood pressure falls within a range of values. Thus the probabilities of DBPs (denoted by  $X$ ) falling

in the ranges of  $90 \leq X < 100$ ,  $100 \leq X < 110$ , and  $X \geq 110$  might be 15%, 5%, and 1%, respectively. People whose blood pressures fall in these ranges may be considered mildly hypertensive, moderately hypertensive, and severely hypertensive, respectively.

Although the probability of exactly obtaining any value is 0, people still have the intuitive notion that certain ranges of values occur more frequently than others. This notion can be quantified using the concept of a probability-density function (pdf).

**Definition 5.1**

The probability-density function of the random variable  $X$  is a function such that the area under the density-function curve between any two points  $a$  and  $b$  is equal to the probability that the random variable  $X$  falls between  $a$  and  $b$ . Thus, the total area under the density-function curve over the entire range of possible values for the random variable is 1.

The pdf has large values in regions of high probability and small values in regions of low probability.

**Example 5.3**

**Hypertension** A pdf for DBP in 35- to 44-year-old men is shown in Figure 5.1. Areas  $A$ ,  $B$ , and  $C$  correspond to the probabilities of being mildly hypertensive, moderately hypertensive, and severely hypertensive, respectively. Furthermore, the most likely range of values for DBP occurs around 80 mm Hg, with the values becoming increasingly less likely as we move farther away from 80.

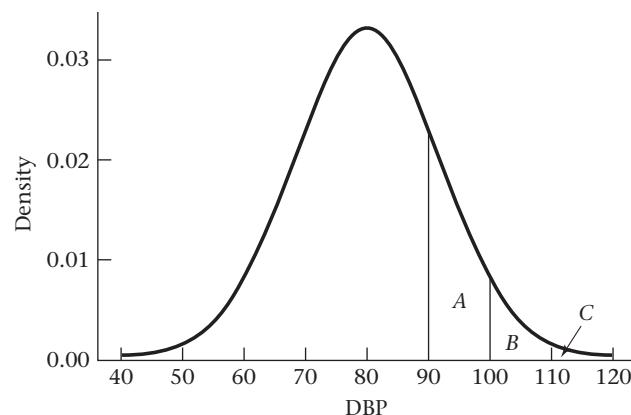
Not all continuous random variables have symmetric bell-shaped distributions as in Figure 5.1.

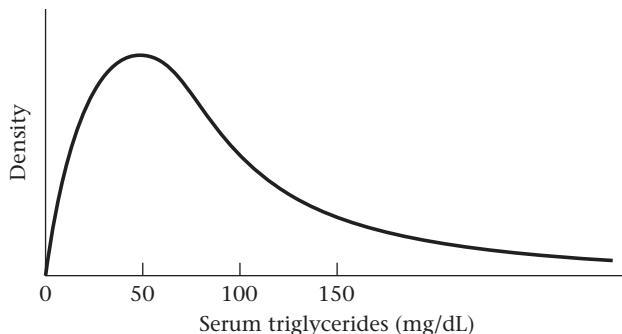
**Example 5.4**

**Cardiovascular Disease** Serum triglyceride level is an asymmetric, positively skewed, continuous random variable whose pdf appears in Figure 5.2.

The cumulative-distribution function (or cdf) is defined similarly to that for a discrete random variable (see Section 4.6).

**Figure 5.1** The pdf of DBP in 35- to 44-year-old men



**Figure 5.2** The pdf for serum triglycerides**Definition 5.2**

The **cumulative-distribution function** for the random variable  $X$  evaluated at the point  $a$  is defined as the probability that  $X$  will take on values  $\leq a$ . It is represented by the area under the pdf to the left of  $a$ .

**Example 5.5**

**Obstetrics** The pdf for the random variable representing the distribution of birthweights in the general population is given in Figure 5.3.

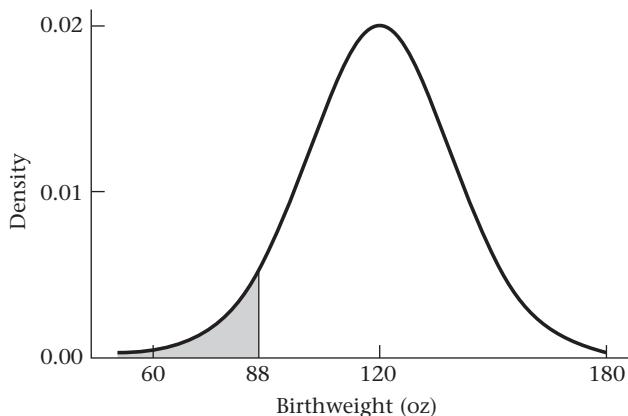
The cdf evaluated at 88 oz =  $Pr(X \leq 88)$  is represented by the area under this curve to the left of 88 oz. The region  $X \leq 88$  oz has a special meaning in obstetrics because 88 oz is the cutoff point obstetricians usually use for identifying low-birthweight infants. Such infants are generally at higher risk for various unfavorable outcomes, such as mortality in the first year of life.

Generally, a distinction is not made between the probabilities  $Pr(X < x)$  and  $Pr(X \leq x)$  when  $X$  is a continuous random variable. This is because they represent the same quantity since the probability of individual values is 0; that is,  $Pr(X = x) = 0$ .

The expected value and variance for continuous random variables have the same meaning as for discrete random variables (see Sections 4.4 and 4.5). However, the mathematical definition of these terms is beyond the scope of this book.

**Definition 5.3**

The **expected value** of a continuous random variable  $X$ , denoted by  $E(X)$ , or  $\mu$ , is the average value taken on by the random variable.

**Figure 5.3** The pdf for birthweight

**Definition 5.4** The **variance** of a continuous random variable  $X$ , denoted by  $\text{Var}(X)$  or  $\sigma^2$ , is the average squared distance of each value of the random variable from its expected value, which is given by  $E(X - \mu)^2$  and can be re-expressed in short form as  $E(X^2) - \mu^2$ . The standard deviation, or  $\sigma$ , is the square root of the variance, that is,  $\sigma = \sqrt{\text{Var}(X)}$ .

**Example 5.6** **Hypertension** The expected value and standard deviation of the distribution of DBP in 35- to 44-year-old men are 80 and 12 mm Hg, respectively.

## 5.3 The Normal Distribution

The normal distribution is the most widely used continuous distribution. It is also frequently called the Gaussian distribution, after the well-known mathematician Karl Friedrich Gauss (Figure 5.4.).

**Example 5.7** **Hypertension** Body weights or DBPs for a group of 35- to 44-year-old men approximately follow a normal distribution.

Many other distributions that are not themselves normal can be made approximately normal by transforming the data onto a different scale.

**Example 5.8** **Cardiovascular Disease** The distribution of serum-triglyceride concentrations from this same group of 35- to 44-year-old men is likely to be positively skewed. However, the log transformation of these measurements usually follows a normal distribution.

Generally speaking, any random variable that can be expressed as a sum of many other random variables can be well approximated by a normal distribution.

For example, many physiologic measures are determined in part by a combination of several genetic and environmental risk factors and can often be well approximated by a normal distribution.

**Figure 5.4** **Karl Friedrich Gauss (1777–1855)**



**Example 5.9**

**Infectious Disease** The number of lymphocytes in a differential of 100 white blood cells (see Example 4.15 for the definition of a differential) tends to be normally distributed because this random variable is a sum of 100 random variables, each representing whether or not an individual cell is a lymphocyte.

Thus, because of its omnipresence the normal distribution is vital to statistical work, and most estimation procedures and hypothesis tests that we will study assume the random variable being considered has an underlying normal distribution.

Another important area of application of the normal distribution is as an approximating distribution to other distributions. The normal distribution is generally more convenient to work with than any other distribution, particularly in hypothesis testing. Thus, if an accurate normal approximation to some other distribution can be found, we often will want to use it.

**Definition 5.5**

The **normal distribution** is defined by its pdf, which is given as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad -\infty < x < \infty$$

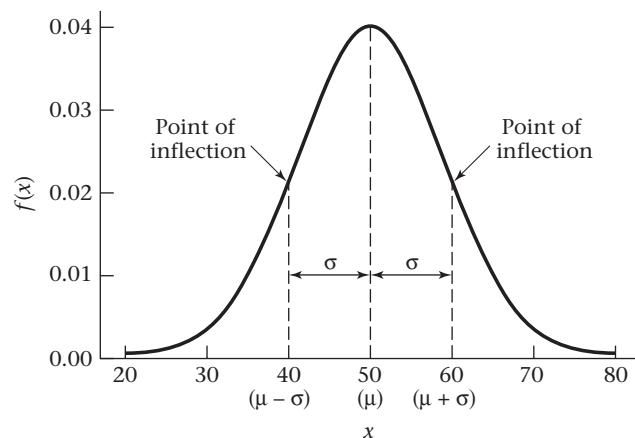
for some parameters  $\mu, \sigma$ , where  $\sigma > 0$ .

The  $\exp$  function merely implies that the quantity to the right in brackets is the power to which “ $e$ ” ( $\approx 2.71828$ ) is raised. This pdf is plotted in Figure 5.5 for a normal distribution with  $\mu = 50$  and  $\sigma^2 = 100$ .

The density function follows a bell-shaped curve, with the mode at  $\mu$  and the most frequently occurring values around  $\mu$ . The curve is symmetric about  $\mu$ , with points of inflection on either side of  $\mu$  at  $\mu - \sigma$  and  $\mu + \sigma$ , respectively. A *point of inflection* is a point at which the slope of the curve changes direction. In Figure 5.5, the slope of the curve increases to the left of  $\mu - \sigma$  and then starts to decrease to the right of  $\mu - \sigma$  and continues to decrease until reaching  $\mu + \sigma$ , after which it starts increasing again. Thus distances from  $\mu$  to points of inflection provide a good visual sense of the magnitude of the parameter  $\sigma$ .

You may wonder why parameters  $\mu$  and  $\sigma^2$  have been used to define the normal distribution when the expected value and variance of an arbitrary distribution were

**Figure 5.5** The pdf for a normal distribution with mean  $\mu$  (50) and variance  $\sigma^2$  (100)



previously defined as  $\mu$  and  $\sigma^2$ . Indeed, from the definition of the normal distribution it can be shown, using calculus methods, that  $\mu$  and  $\sigma^2$  are, respectively, the expected value and variance of this distribution.

**Example 5.10** For DBP the parameters might be  $\mu = 80$  mm Hg,  $\sigma = 12$  mm Hg; for birth weight they might be  $\mu = 120$  oz,  $\sigma = 20$  oz.

Interestingly, the entire shape of the normal distribution is determined by the two parameters  $\mu$  and  $\sigma^2$ . If two normal distributions with the same variance  $\sigma^2$  and different means  $\mu_1, \mu_2$ , where  $\mu_2 > \mu_1$ , are compared, then their density functions will appear as in Figure 5.6, where  $\mu_1 = 50$ ,  $\mu_2 = 62$ , and  $\sigma = 7$ . The heights of the two curves are the same, but one curve is shifted to the right relative to the other curve.

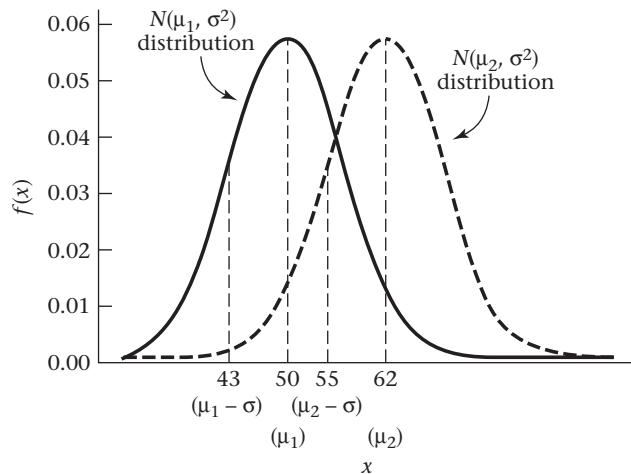
Similarly, two normal distributions with the same mean but different variances ( $\sigma_2^2 > \sigma_1^2$ ) can be compared, as shown in Figure 5.7, with  $\mu = 50$ ,  $\sigma_1 = 5$ , and  $\sigma_2 = 10$ . Thus the  $x$  value corresponding to the highest density ( $x = 50$ ) is the same for each curve, but the curve with the smaller standard deviation ( $\sigma_1 = 5$ ) is higher and has a more concentrated distribution than the curve with the larger standard deviation ( $\sigma_2 = 10$ ). Note that the area under any normal density function must be 1. Thus the two normal distributions shown in Figure 5.7 must cross, because otherwise one curve would remain completely above the other and the areas under both curves could not simultaneously be 1.

**Definition 5.6** A normal distribution with mean  $\mu$  and variance  $\sigma^2$  will generally be referred to as an  $N(\mu, \sigma^2)$  distribution.

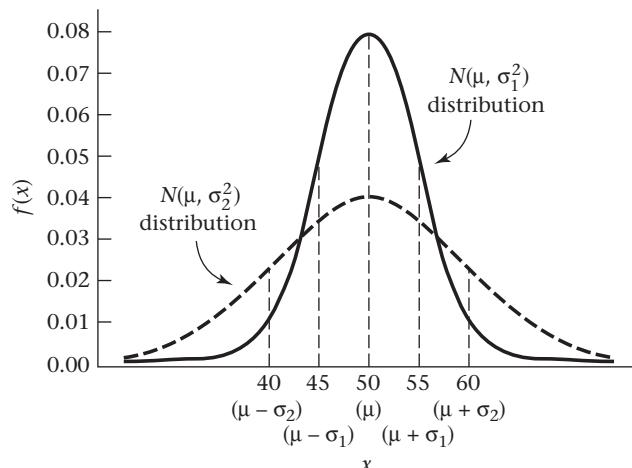
Note that the second parameter is always the variance  $\sigma^2$ , not the standard deviation  $\sigma$ .

Another property of the normal distribution is that the height =  $1/(\sqrt{2\pi}\sigma)$ . Thus the height is inversely proportional to  $\sigma$ . As noted previously, this helps us visualize  $\sigma$ , because the density at the value  $x = \mu$  for an  $N(\mu, \sigma_1^2)$  distribution in Figure 5.7 is larger than for an  $N(\mu, \sigma_2^2)$  distribution.

**Figure 5.6** Comparison of two normal distributions with the same variance and different means



**Figure 5.7** Comparison of two normal distributions with the same means and different variances



**Definition 5.7** A normal distribution with mean 0 and variance 1 is called a **standard**, or **unit**, normal distribution. This distribution is also called an  $N(0,1)$  distribution.

We will see that any information concerning an  $N(\mu, \sigma^2)$  distribution can be obtained from appropriate manipulations of an  $N(0,1)$  distribution.

## 5.4 Properties of the Standard Normal Distribution

To become familiar with the  $N(0,1)$  distribution, let's discuss some of its properties. First, the pdf in this case reduces to

$$\text{Equation 5.1} \quad f(x) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)x^2}, \quad -\infty < x < +\infty$$

This distribution is symmetric about 0, because  $f(x) = f(-x)$ , as shown in Figure 5.8.

### Equation 5.2

It can be shown that about 68% of the area under the standard normal density lies between  $+1$  and  $-1$ , about 95% of the area lies between  $+2$  and  $-2$ , and about 99% lies between  $+2.5$  and  $-2.5$ .

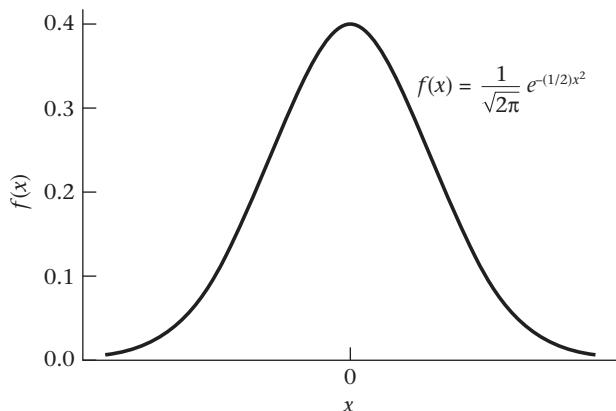
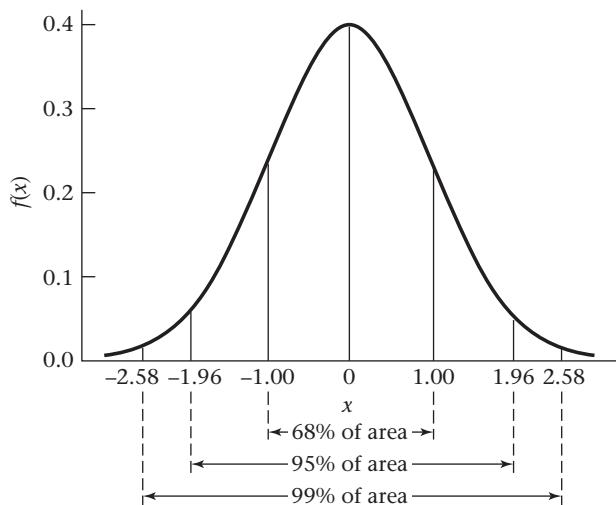
These relationships can be expressed more precisely by saying that

$$Pr(-1 < X < 1) = .6827 \quad Pr(-1.96 < X < 1.96) = .95$$

$$Pr(-2.576 < X < 2.576) = .99$$

Thus the standard normal distribution slopes off very rapidly, and absolute values greater than 3 are unlikely. Figure 5.9 shows these relationships.

Tables of the area under the normal density function, or so-called normal tables, take advantage of the symmetry properties of the normal distribution and generally are concerned with areas for positive values of  $x$ .

**Figure 5.8** The pdf for a standard normal distribution**Figure 5.9** Empirical properties of the standard normal distribution**Definition 5.8**

The cumulative-distribution function (cdf) for a standard normal distribution is denoted by

$$\Phi(x) = \Pr(X \leq x)$$

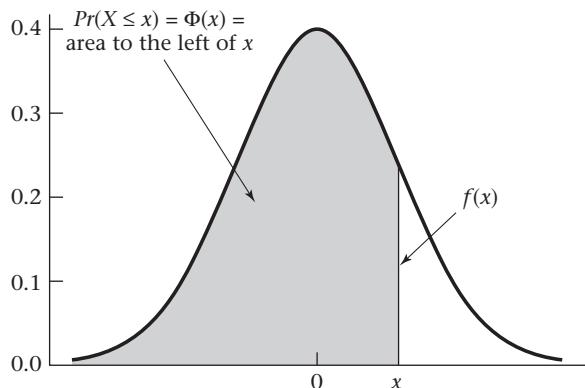
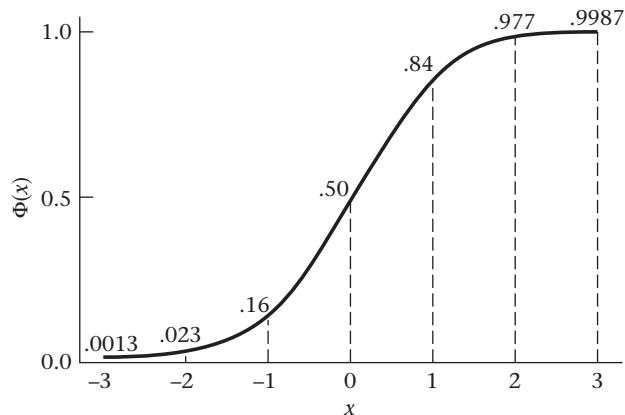
where  $X$  follows an  $N(0,1)$  distribution. This function is shown in Figure 5.10.

**Definition 5.9**

The symbol  $\sim$  is used as shorthand for the phrase “is distributed as.” Thus  $X \sim N(0,1)$  means that the random variable  $X$  is distributed as an  $N(0,1)$  distribution.

## Using Normal Tables

Column A in Table 3 of the Appendix presents  $\Phi(x)$  for various positive values of  $x$  for a standard normal distribution. This cumulative distribution function is illustrated in Figure 5.11. Notice that the area to the left of 0 is .5.

**Figure 5.10** The cdf [ $\Phi(x)$ ] for a standard normal distribution**Figure 5.11** The cdf for a standard normal distribution [ $\Phi(x)$ ]

Furthermore, the area to the left of  $x$  approaches 0 as  $x$  becomes small and approaches 1 as  $x$  becomes large.

The right-hand tail of the standard normal distribution =  $\Pr(X \geq x)$  is given in column B of Appendix Table 3.

**Example 5.11** If  $X \sim N(0,1)$ , then find  $\Pr(X \leq 1.96)$  and  $\Pr(X \leq 1)$ .

**Solution**

From the Appendix, Table 3, column A,

$$\Phi(1.96) = .975 \text{ and } \Phi(1) = .8413$$

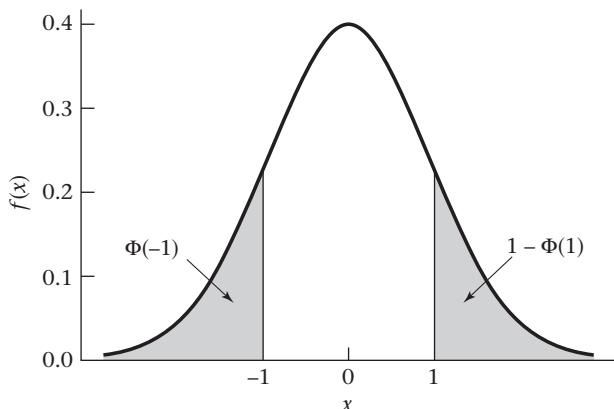
### Equation 5.3

#### Symmetry Properties of the Standard Normal Distribution

From the symmetry properties of the standard normal distribution,

$$\Phi(-x) = \Pr(X \leq -x) = \Pr(X \geq x) = 1 - \Pr(X \leq x) = 1 - \Phi(x)$$

This symmetry property is depicted in Figure 5.12 for  $x = 1$ .

**Figure 5.12** Illustration of the symmetry properties of the normal distribution

**Example 5.12** Calculate  $Pr(X \leq -1.96)$  if  $X \sim N(0,1)$ .

**Solution**  $Pr(X \leq -1.96) = Pr(X \geq 1.96) = .0250$  from column B of Table 3.

Furthermore, for any numbers  $a, b$  we have  $Pr(a \leq X \leq b) = Pr(X \leq b) - Pr(X \leq a)$  and thus we can evaluate  $Pr(a \leq X \leq b)$  for any  $a, b$  from Table 3.

**Example 5.13** Compute  $Pr(-1 \leq X \leq 1.5)$  if  $X \sim N(0,1)$ .

$$\begin{aligned} Pr(-1 \leq X \leq 1.5) &= Pr(X \leq 1.5) - Pr(X \leq -1) \\ &= Pr(X \leq 1.5) - Pr(X \geq 1) = .9332 - .1587 \\ &= .7745 \end{aligned}$$

**Example 5.14**

**Pulmonary Disease** Forced vital capacity (FVC), a standard measure of pulmonary function, is the volume of air a person can expel in 6 seconds. Current research looks at potential risk factors, such as cigarette smoking, air pollution, indoor allergies, or the type of stove used in the home, that may affect FVC in grade-school children. One problem is that age, sex, and height affect pulmonary function, and these variables must be corrected for before considering other risk factors. One way to make these adjustments for a particular child is to find the mean  $\mu$  and standard deviation  $\sigma$  for children of the same age (in 1-year age groups), sex, and height (in 2-in. height groups) from large national surveys and compute a **standardized FVC**, which is defined as  $(X - \mu)/\sigma$ , where  $X$  is the original FVC. The standardized FVC then approximately follows an  $N(0,1)$  distribution, if the distribution of the original FVC values was bell-shaped. Suppose a child is considered in poor pulmonary health if his or her standardized FVC  $< -1.5$ . What percentage of children are in poor pulmonary health?

$$\text{Solution } Pr(X < -1.5) = Pr(X > 1.5) = .0668$$

Thus about 7% of children are in poor pulmonary health.

In many instances we are concerned with tail areas on either side of 0 for a standard normal distribution. For example, the *normal range* for a biological quantity is often defined by a range within  $x$  standard deviations of the mean for some specified value of  $x$ . The probability of a value falling in this range is given by  $Pr(-x \leq X \leq x)$  for a standard normal distribution. This quantity is tabulated in column D of Table 3 in the Appendix for various values of  $x$ .

**Example 5.15** **Pulmonary Disease** Suppose a child is considered to have normal lung growth if his or her standardized FVC is within 1.5 standard deviations of the mean. What proportion of children are within the normal range?

**Solution** Compute  $Pr(-1.5 \leq X \leq 1.5)$ . Under 1.50 in Table 3, column D gives this quantity as .8664. Thus about 87% of children have normal lung growth, according to this definition.

Finally, column C of Table 3 provides the area under the standard normal density from 0 to  $x$  because these areas occasionally prove useful in work on statistical inference.

**Example 5.16** Find the area under the standard normal density from 0 to 1.45.

**Solution** Refer to column C of Table 3 under 1.45. The appropriate area is given by .4265.

Of course, the areas given in columns A, B, C, and D are redundant in that *all* computations concerning the standard normal distribution can be performed using any one of these columns. In particular, we have seen that  $B(x) = 1 - A(x)$ . Also, from the symmetry of the normal distribution we can easily show that  $C(x) = A(x) - .5$ ,  $D(x) = 2 \times C(x) = 2 \times A(x) - 1.0$ . However, this redundancy is deliberate because for some applications one of these columns may be more convenient to use.

### Using Electronic Tables for the Normal Distribution

It is also possible to use “electronic tables” to compute areas under a standard normal distribution. For example, in Excel 2007 the function NORMSDIST( $x$ ) provides the cdf for a standard normal distribution for any value of  $x$ .

**Example 5.17** Using an electronic table, find the area under the standard normal density to the left of 2.824.

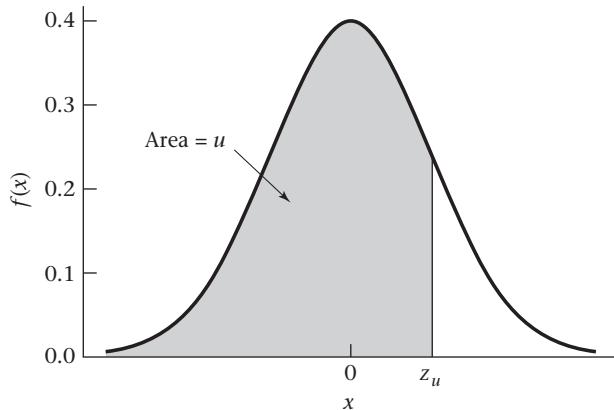
**Solution** We use the Excel 2007 function NORMSDIST evaluated at 2.824 [NORMSDIST(2.824)], with the result as follows:

<b>x</b>	<b>2.824</b>
<b>NORMSDIST(x)</b>	<b>0.997629</b>

The area is .9976.

The percentiles of a normal distribution are often referred to in statistical inference. For example, we might be interested in the upper and lower fifth percentiles of the distribution of FVC in children in order to define a normal range of values.

**Figure 5.13** Graphic display of the  $(100 \times u)$ th percentile of a standard normal distribution ( $z_u$ )



For this purpose, the definition of the percentiles of a standard normal distribution is introduced:

**Definition 5.10**

The  $(100 \times u)$ th percentile of a standard normal distribution is denoted by  $z_u$ . It is defined by the relationship

$$\Pr(X < z_u) = u, \quad \text{where } X \sim N(0,1)$$

Figure 5.13 displays  $z_u$ .

The function  $z_u$  is sometimes referred to as the *inverse normal function*. In previous uses of the normal table, we were given a value  $x$  and have used the normal tables to evaluate the area to the left of  $x$ —that is,  $\Phi(x)$ —for a standard normal distribution.

To obtain  $z_u$ , we perform this operation in reverse. Thus, to evaluate  $z_u$  we must find the area  $u$  in column A of Appendix Table 3 and then find the value  $z_u$  that corresponds to this area. If  $u < .5$ , then we use the symmetry properties of the normal distribution to obtain  $z_u = -z_{1-u}$ , where  $z_{1-u}$  can be obtained from Table 3.

**Example 5.18**

Compute  $z_{.975}$ ,  $z_{.95}$ ,  $z_{.5}$ , and  $z_{.025}$ .

**Solution**

From Table 3 we have

$$\Phi(1.96) = .975$$

$$\Phi(1.645) = .95$$

$$\Phi(0) = .5$$

$$\Phi(-1.96) = 1 - \Phi(1.96) = 1 - .975 = .025$$

$$\text{Thus } z_{.975} = 1.96$$

$$z_{.95} = 1.645$$

$$z_{.5} = 0$$

$$z_{.025} = -1.96$$

where for  $z_{.95}$  we interpolate between 1.64 and 1.65 to obtain 1.645.

**Example 5.19**

Compute the value  $x$  such that the area to the left of  $x$  under a standard normal density = .85.

**Solution**

We use the Excel 2007 function NORMSINV evaluated at .85 [NORMSINV(.85)] with the result given as follows:

<b>x</b>	<b>0.85</b>
<b>NORMSINV(x)</b>	<b>1.036433</b>

Thus the area to the left of 1.036 under a standard normal density is .85.

The percentile  $z_u$  is used frequently in our work on estimation in Chapter 6 and hypothesis testing in Chapters 7–14.

**REVIEW QUESTIONS 5A**

- 1 What is the difference between a probability-density function (pdf) and a probability-mass function?
- 2 Suppose a continuous random variable can only take on values between  $-1$  and  $+1$ . What is the area under the pdf from  $-2$  to  $2$ ?
- 3 What is a standard normal distribution?
- 4 (a) What is the area to the left of  $-0.2$  under a standard normal distribution? What symbol is used to represent this area?  
 (b) What is the area to the right of  $0.3$  under a standard normal distribution? What symbol is used to represent this area?
- 5 (a) What is  $z_{.30}$ ? What does it mean?  
 (b) What is  $z_{.75}$ ? What does it mean?

## 5.5 Conversion from an $N(\mu, \sigma^2)$ Distribution to an $N(0,1)$ Distribution

**Example 5.20**

**Hypertension** Suppose a mild hypertensive is defined as a person whose DBP is between 90 and 100 mm Hg inclusive, and the subjects are 35- to 44-year-old men whose blood pressures are normally distributed with mean 80 and variance 144. What is the probability that a randomly selected person from this population will be a mild hypertensive? This question can be stated more precisely: If  $X \sim N(80, 144)$ , then what is  $Pr(90 < X < 100)$ ?

(The solution is given on page 122.)

More generally, the following question can be asked: If  $X \sim N(\mu, \sigma^2)$ , then what is  $Pr(a < X < b)$  for any  $a, b$ ? To solve this, we convert the probability statement about an  $N(\mu, \sigma^2)$  distribution to an equivalent probability statement about an  $N(0,1)$  distribution. Consider the random variable  $Z = (X - \mu)/\sigma$ . We can show that the following relationship holds:

**Equation 5.4**

If  $X \sim N(\mu, \sigma^2)$  and  $Z = (X - \mu)/\sigma$ , then  $Z \sim N(0,1)$ .

**Equation 5.5****Evaluation of Probabilities for Any Normal Distribution via Standardization**

If  $X \sim N(\mu, \sigma^2)$  and  $Z = (X - \mu)/\sigma$

$$\text{then } Pr(a < X < b) = Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left[\frac{(b-\mu)}{\sigma}\right] - \Phi\left[\frac{(a-\mu)}{\sigma}\right]$$

Because the  $\Phi$  function, which is the cumulative distribution function for a standard normal distribution, is given in column A of Table 3 of the Appendix, probabilities for *any* normal distribution can now be evaluated using the tables in this text. This procedure is shown in Figure 5.14 for  $\mu = 80$ ,  $\sigma = 12$ ,  $a = 90$ ,  $b = 100$ , where the areas in Figure 5.14a and 5.14b are the same.

The procedure in Equation 5.5 is known as **standardization of a normal variable**.

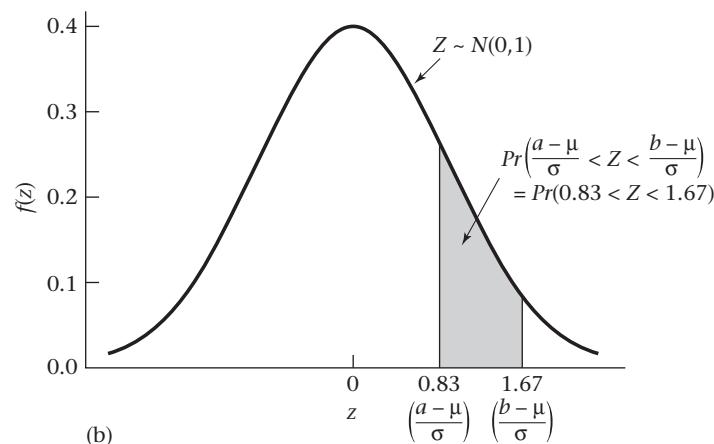
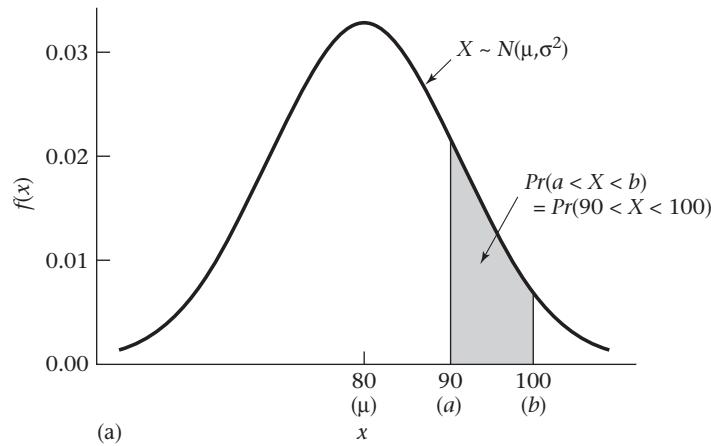
### Equation 5.6

The general principle is that for any probability expression concerning normal random variables of the form  $Pr(a < X < b)$ , the population mean  $\mu$  is subtracted from each boundary point and divided by the standard deviation  $\sigma$  to obtain an equivalent probability expression for the standard normal random variable  $Z$ ,

$$Pr\left[\frac{(a-\mu)}{\sigma} < Z < \frac{(b-\mu)}{\sigma}\right]$$

The standard normal tables are then used to evaluate this latter probability.

**Figure 5.14** Evaluation of probabilities for any normal distribution using standardization



**Solution to Example 5.20**

The probability of being a mild hypertensive among the group of 35- to 44-year-old men can now be calculated.

$$\begin{aligned} Pr(90 < X < 100) &= Pr\left(\frac{90 - 80}{12} < Z < \frac{100 - 80}{12}\right) \\ &= Pr(0.833 < Z < 1.667) = \Phi(1.667) - \Phi(0.833) \\ &= .9522 - .7977 = .155 \end{aligned}$$

Thus about 15.5% of this population will have mild hypertension.

**Example 5.21**

**Botany** Suppose tree diameters of a certain species of tree from some defined forest area are assumed to be normally distributed with mean = 8 in. and standard deviation = 2 in. Find the probability of a tree having an unusually large diameter, which is defined as  $> 12$  in.

**Solution**

We have  $X \sim N(8,4)$  and require

$$\begin{aligned} Pr(X > 12) &= 1 - Pr(X < 12) = 1 - Pr\left(Z < \frac{12 - 8}{2}\right) \\ &= 1 - Pr(Z < 2.0) = 1 - .977 = .023 \end{aligned}$$

Thus 2.3% of trees from this area have an unusually large diameter.

**Example 5.22**

**Cerebrovascular Disease** Diagnosing stroke strictly on the basis of clinical symptoms is difficult. A standard diagnostic test used in clinical medicine to detect stroke in patients is the angiogram. This test has some risks for the patient, and researchers have developed several noninvasive techniques that they hope will be as effective as the angiogram. One such method measures cerebral blood flow (CBF) in the brain because stroke patients tend to have lower CBF levels than normal. Assume that in the general population, CBF is normally distributed with mean = 75 mL/100 g brain tissue and standard deviation = 17 mL/100 g brain tissue. A patient is classified as being at risk for stroke if his or her CBF is lower than 40 mL/100 g brain tissue. What proportion of normal patients will be mistakenly classified as being at risk for stroke?

**Solution**

Let  $X$  be the random variable representing CBF. Then  $X \sim N(75, 17^2) = N(75, 289)$ . We want to find  $Pr(X < 40)$ . We standardize the limit of 40 so as to use the standard normal distribution. The standardized limit is  $(40 - 75)/17 = -2.06$ . Thus, if  $Z$  represents the standardized normal random variable  $= (X - \mu)/\sigma$ , then

$$\begin{aligned} Pr(X < 40) &= Pr(Z < -2.06) \\ &= \Phi(-2.06) = 1 - \Phi(2.06) = 1 - .9803 \approx .020 \end{aligned}$$

Thus about 2.0% of normal patients will be incorrectly classified as being at risk for stroke.

If we use electronic tables, then the pdf, cdf, and inverse normal distribution can be obtained for any normal distribution, and standardization is unnecessary. For example, using Excel 2007, the two functions NORMDIST and NORMINV are available for this purpose. To find the probability  $p$  that an  $N(\mu, \sigma^2)$  distribution is  $\leq x$ , we use the function

$p = \text{NORMDIST}(x, \mu, \sigma, \text{TRUE})$

To find the probability density  $f$  at  $x$ , we use the function

$f = \text{NORMDIST}(x, \mu, \sigma, \text{FALSE})$

To find the value  $x$  such that the cdf for an  $N(\mu, \sigma^2)$  distribution is equal to  $p$ , we use the function

$x = \text{NORMINV}(p, \mu, \sigma)$

More details and examples of using these functions are provided on the Companion Website.

### Equation 5.7

The  $p$ th percentile of a general normal distribution ( $x$ ) can also be written in terms of the percentiles of a standard normal distribution as follows:

$$x = \mu + z_p \sigma$$

### Example 5.23

**Ophthalmology** Glaucoma is an eye disease that is manifested by high intraocular pressure (IOP). The distribution of IOP in the general population is approximately normal with mean = 16 mm Hg and standard deviation = 3 mm Hg. If the normal range for IOP is considered to be between 12 and 20 mm Hg, then what percentage of the general population would fall within this range?

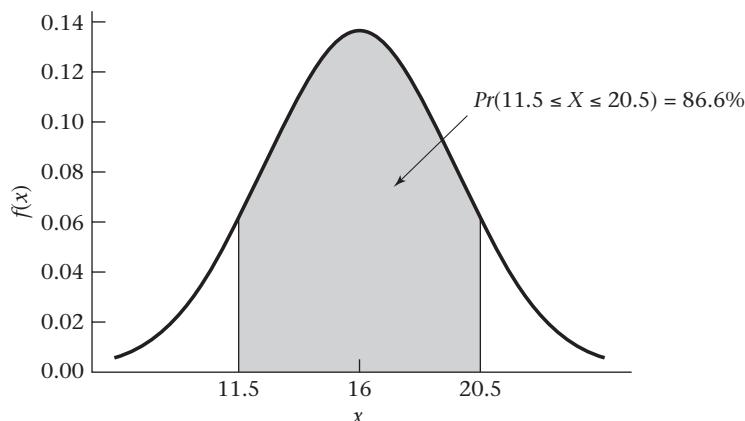
### Solution

Because IOP can only be measured to the nearest integer, we will associate the recorded value of 12 mm Hg with a range of actual IOP values from 11.5 to 12.5 mm Hg. Similarly, we associate a recorded IOP value of 20 mm Hg with a range of actual IOP values from 19.5 to 20.5 mm Hg. Hence, we want to calculate  $Pr(11.5 \leq X \leq 20.5)$ , where  $X \sim N(16, 9)$ , as shown in Figure 5.15. The process of associating a specific observed value (such as 12 mm Hg) with an actual range of value ( $11.5 \leq X \leq 12.5$ ) is called "*incorporating a continuity correction*."

We use the NORMDIST function of Excel 2007 to perform these computations. First, we compute  $p_1 = Pr[X \leq 20.5 | X \sim N(16, 9)]$  given by NORMDIST(20.5, 16, 3, TRUE). Second, we compute  $p_2 = Pr[X \leq 11.5 | X \sim N(16, 9)]$  given by NORMDIST(11.5, 16, 3, TRUE). Thus  $Pr(11.5 \leq X \leq 20.5) = p_1 - p_2 = .866$ . The computations are shown in the following spreadsheet.

Figure 5.15

Calculation of the proportion of people with IOP in the normal range



```

p1=NORMDIST(20.5,16,3,true)      0.933193
p2=NORMDIST(11.5,16,3,true)      0.066807
p=p1-p2                          0.866386

```

Thus, 86.6% of the population has IOP in the normal range.

### Example 5.24

**Hypertension** Suppose the distribution of DBP in 35- to 44-year-old men is normally distributed with mean = 80 mm Hg and variance = 144 mm Hg. Find the upper and lower fifth percentiles of this distribution.

### Solution

We could do this either using Table 3 (Appendix) or using Excel. If we use Table 3 and we denote the upper and lower 5th percentiles by  $x_{.05}$  and  $x_{.95}$ , respectively, then from Equation 5.7 we have

$$\begin{aligned}x_{.05} &= 80 + z_{.05}(12) \\&= 80 - 1.645(12) = 60.3 \text{ mm Hg} \\x_{.95} &= 80 + z_{.95}(12) \\&= 80 + 1.645(12) = 99.7 \text{ mm Hg}\end{aligned}$$

If we use Excel, then we have

$$\begin{aligned}x_{.05} &= \text{NORMINV (.05, 80, 12)} \\x_{.95} &= \text{NORMINV (.95, 80, 12)}\end{aligned}$$

The results denoted by  $x(.05)$  and  $x(.95)$  on the spreadsheet are shown as follows:

$x(.05)$	$\text{NORMINV (.05, 80, 12)}$	60.3
$x(.95)$	$\text{NORMINV (.95, 80, 12)}$	99.7

### REVIEW QUESTIONS 5B

- 1 What is the difference between a standard normal distribution and a general normal distribution?
- 2 What does the principle of standardization mean?
- 3 Suppose the distribution of serum-cholesterol values is normally distributed, with mean = 220 mg/dL and standard deviation = 35 mg/dL.
  - (a) What is the probability that a serum cholesterol level will range from 200 to 250 inclusive (that is, a high normal range)? Assume that cholesterol values can be measured exactly—that is, without the need for incorporating a continuity correction.
  - (b) (1) What is the lowest quintile of serum-cholesterol values (the 20th percentile)?
  - (2) What is the highest quintile of serum-cholesterol values (the 80th percentile)?

## 5.6 Linear Combinations of Random Variables

In work on statistical inference, sums or differences or more complicated linear functions of random variables (either continuous or discrete) are often used. For this reason, the properties of linear combinations of random variables are important to consider.

**Definition 5.11**

A **linear combination**  $L$  of the random variables  $X_1, \dots, X_n$  is defined as any function of the form  $L = c_1X_1 + \dots + c_nX_n$ . A linear combination is sometimes also called a **linear contrast**.

**Example 5.25**

**Renal Disease** Let  $X_1, X_2$  be random variables representing serum-creatinine levels for white and black individuals with end-stage renal disease. Represent the sum, difference, and average of the random variables  $X_1, X_2$  as linear combinations of the random variables  $X_1, X_2$ .

**Solution**

The sum is  $X_1 + X_2$ , where  $c_1 = 1, c_2 = 1$ . The difference is  $X_1 - X_2$ , where  $c_1 = 1, c_2 = -1$ . The average is  $(X_1 + X_2)/2$ , where  $c_1 = 0.5, c_2 = 0.5$ .

It is often necessary to compute the expected value and variance of linear combinations of random variables. To find the expected value of  $L$ , we use the principle that the expected value of the sum of  $n$  random variables is the sum of the  $n$  respective expected values. Applying this principle,

$$\begin{aligned} E(L) &= E(c_1X_1 + \dots + c_nX_n) \\ &= E(c_1X_1) + \dots + E(c_nX_n) = c_1E(X_1) + \dots + c_nE(X_n) \end{aligned}$$

**Equation 5.8****Expected Value of Linear Combinations of Random Variables**

The expected value of the linear combination  $L = \sum_{i=1}^n c_iX_i$  is  $E(L) = \sum_{i=1}^n c_iE(X_i)$ .

**Example 5.26**

**Renal Disease** Suppose the expected values of serum creatinine for the white and the black individuals in Example 5.25 are 1.3 and 1.5, respectively. What is the expected value of the average serum-creatinine level of a single white and a single black individual?

**Solution**

The expected value of the average serum-creatinine level  $= E(0.5X_1 + 0.5X_2) = 0.5E(X_1) + 0.5E(X_2) = 0.65 + 0.75 = 1.4$ .

To compute the variance of linear combinations of random variables, we first assume that the random variables are independent. Under this assumption, it can be shown that the variance of the sum of  $n$  random variables is the sum of the respective variances. Applying this principle,

$$\begin{aligned} Var(L) &= Var(c_1X_1 + \dots + c_nX_n) \\ &= Var(c_1X_1) + \dots + Var(c_nX_n) = c_1^2Var(X_1) + \dots + c_n^2Var(X_n) \end{aligned}$$

because

$$Var(c_iX_i) = c_i^2Var(X_i)$$

**Equation 5.9****Variance of Linear Combinations of Independent Random Variables**

The variance of the linear combination  $L = \sum_{i=1}^n c_iX_i$ , where  $X_1, \dots, X_n$  are independent is  $Var(L) = \sum_{i=1}^n c_i^2Var(X_i)$ .

**Example 5.27**

**Renal Disease** Suppose  $X_1$  and  $X_2$  are defined as in Example 5.26. If we know that  $Var(X_1) = Var(X_2) = 0.25$ , then what is the variance of the average serum-creatinine level over a single white and a single black individual?

**Solution**

We wish to compute  $\text{Var}(0.5X_1 + 0.5X_2)$ . Applying Equation 5.9,

$$\begin{aligned}\text{Var}(0.5X_1 + 0.5X_2) &= (0.5)^2\text{Var}(X_1) + (0.5)^2\text{Var}(X_2) \\ &= 0.25(0.25) + 0.25(0.25) = 0.125\end{aligned}$$

The results for the expected value and variance of linear combinations in Equations 5.8 and 5.9 *do not* depend on assuming normality. However, linear combinations of normal random variables are often of specific concern. It can be shown that any linear combination of independent normal random variables is itself normally distributed. This leads to the following important result:

**Equation 5.10**

If  $X_1, \dots, X_n$  are independent normal random variables with expected values  $\mu_1, \dots, \mu_n$  and variances  $\sigma_1^2, \dots, \sigma_n^2$ , and  $L$  is any linear combination  $= \sum_{i=1}^n c_i X_i$ , then  $L$  is normally distributed with

$$\text{Expected value} = E(L) = \sum_{i=1}^n c_i \mu_i \text{ and variance} = \text{Var}(L) = \sum_{i=1}^n c_i^2 \sigma_i^2$$

**Example 5.28**

**Renal Disease** If  $X_1$  and  $X_2$  are defined as in Examples 5.25–5.27 and are each normally distributed, then what is the distribution of the average  $= 0.5X_1 + 0.5X_2$ ?

**Solution**

Based on the solutions to Examples 5.26 and 5.27, we know that  $E(L) = 1.4$ ,  $\text{Var}(L) = 0.125$ . Therefore,  $(X_1 + X_2)/2 \sim N(1.4, 0.125)$ .

## Dependent Random Variables

In some instances, we are interested in studying linear contrasts of random variables that are not independent.

**Example 5.29**

**Renal Disease** A hypothesis exists that a high-protein diet may aggravate the course of kidney disease among diabetic patients. To test the feasibility of administering a low-protein diet to such patients, a small “pilot” study is set up where 20 diabetic patients are followed for 1 year on the diet. Serum creatinine is a parameter often used to monitor kidney function. Let  $X_1$  be the serum creatinine at baseline and  $X_2$  the serum creatinine after 1 year. We wish to compute the expected value and variance of the change in serum creatinine represented by the random variable  $D = X_1 - X_2$  assuming that  $E(X_1) = E(X_2) = 1.5$  and  $\text{Var}(X_1) = \text{Var}(X_2) = .25$ .

**Solution**

The random variables  $X_1$  and  $X_2$  are not independent because they represent serum-creatinine values on the same subject. However, we can still compute the expected value of  $D$  using Equation 5.8 because the formula for the expected value of a linear contrast is valid whether the random variables involved are independent or dependent. Therefore,  $E(D) = E(X_1) - E(X_2) = 0$ . However, the formula in Equation 5.9 for the variance of a linear contrast is not valid for dependent random variables.

The *covariance* is a measure used to quantify the relationship between two random variables.

**Definition 5.12**

The **covariance** between two random variables  $X$  and  $Y$  is denoted by  $\text{Cov}(X, Y)$  and is defined by

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

which can also be written as  $E(XY) - \mu_x\mu_y$ , where  $\mu_x$  is the average value of  $X$ ,  $\mu_y$  is the average value of  $Y$ , and  $E(XY)$  = average value of the product of  $X$  and  $Y$ .

It can be shown that if the random variables  $X$  and  $Y$  are independent, then the covariance between them is 0. If large values of  $X$  and  $Y$  tend to occur among the same subjects (as well as small values of  $X$  and  $Y$ ), then the covariance is positive. If large values of  $X$  and small values of  $Y$  (or conversely, small values of  $X$  and large values of  $Y$ ) tend to occur among the same subjects, then the covariance is negative.

One issue is that the covariance between two random variables  $X$  and  $Y$  is in the units of  $X$  multiplied by the units of  $Y$ . Thus, it is difficult to interpret the strength of association between two variables from the magnitude of the covariance. To obtain a measure of relatedness independent of the units of  $X$  and  $Y$ , we consider the *correlation coefficient*.

#### Definition 5.13

The **correlation coefficient** between two random variables  $X$  and  $Y$  is denoted by  $\text{Corr}(X, Y)$  or  $\rho$  and is defined by

$$\rho = \text{Corr}(X, Y) = \text{Cov}(X, Y) / (\sigma_x \sigma_y)$$

where  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $X$  and  $Y$ , respectively.

Unlike the covariance, the correlation coefficient is a dimensionless quantity that is independent of the units of  $X$  and  $Y$  and ranges between  $-1$  and  $1$ . For random variables that are approximately linearly related, a correlation coefficient of  $0$  implies independence. A correlation coefficient close to  $1$  implies nearly perfect positive dependence with large values of  $X$  corresponding to large values of  $Y$  and small values of  $X$  corresponding to small values of  $Y$ . An example of a strong positive correlation is between forced expiratory volume (FEV), a measure of pulmonary function, and height (Figure 5.16a). A somewhat weaker positive correlation exists between serum cholesterol and dietary intake of cholesterol (Figure 5.16b). A correlation coefficient close to  $-1$  implies  $\approx$  perfect negative dependence, with large values of  $X$  corresponding to small values of  $Y$  and vice versa, as is evidenced by the relationship between resting pulse rate and age in children under the age of  $10$  (Figure 5.16c). A somewhat weaker negative correlation exists between FEV and number of cigarettes smoked per day in children (Figure 5.16d).

For variables that are not linearly related, it is difficult to infer independence or dependence from a correlation coefficient.

#### Example 5.30

Let  $X$  be the random variable height z-score for 7-year-old children, where height z-score =  $(\text{height} - \mu)/\sigma$ .

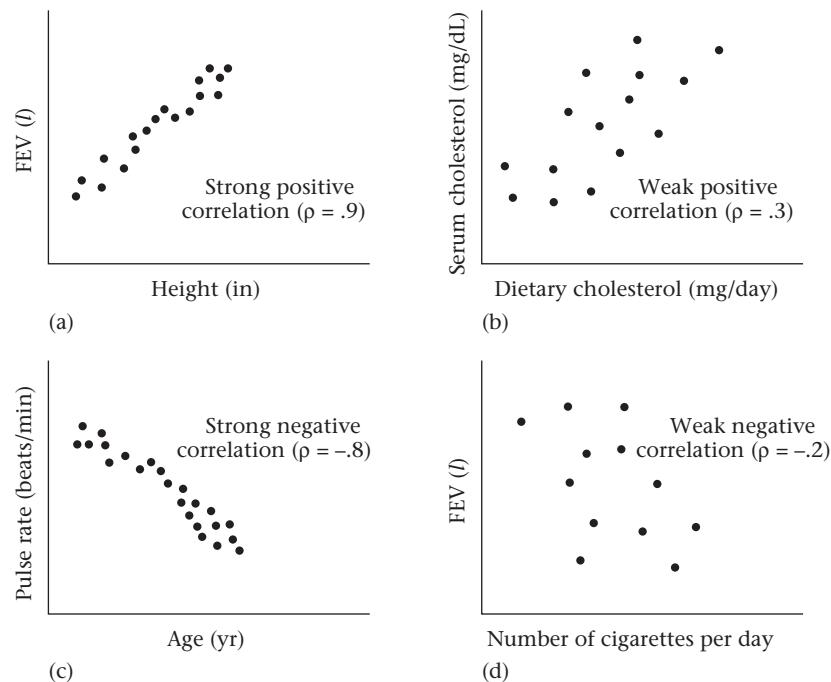
$\mu$  = mean height for 7-year-old children

$\sigma$  = standard deviation of height for 7-year-old children

Let  $Y = \text{height z-score}^2 = X^2$ . Compute the correlation coefficient between  $X$  and  $Y$ , assuming that  $X$  is normally distributed.

#### Solution

Because  $X$  is symmetric about 0 and  $Y$  is the same for positive and negative values of  $X$  with the same absolute value, it is easy to show that  $\text{Corr}(X, Y) = 0$ . However,  $X$  and  $Y$  are totally dependent because if we know  $X$ , then  $Y$  is totally determined. For

**Figure 5.16** Interpretation of various degrees of correlation

example, if  $X = 2$ , then  $Y = 4$ . Thus, it would be a mistake to assume that the random variables  $X$  and  $Y$  are independent if the correlation coefficient between them is 0. This relationship can only be inferred for linearly related variables. We discuss linear and nonlinear relationships between variables in more detail in Chapter 11.

To compute the variance of a linear contrast involving two dependent random variables,  $X_1$  and  $X_2$ , we can generalize Equation 5.9 as follows:

**Equations 5.11**

$$\begin{aligned} \text{Var}(c_1X_1 + c_2X_2) &= c_1^2\text{Var}(X_1) + c_2^2\text{Var}(X_2) + 2c_1c_2\text{Cov}(X_1, X_2) \\ &= c_1^2\text{Var}(X_1) + c_2^2\text{Var}(X_2) + 2c_1c_2\sigma_x\sigma_y\text{Corr}(X_1, X_2) \end{aligned}$$

**Example 5.31**

**Renal Disease** Suppose the correlation coefficient between two determinations of serum creatinine 1 year apart is .5. Compute the variance of the change in serum creatinine over 1 year using the parameters given in Example 5.29.

**Solution**

We have  $\text{Var}(X_1) = \text{Var}(X_2) = .25$  and  $\text{Corr}(X_1, X_2) = .5$ . Also, for the linear contrast  $D = X_1 - X_2$ , we have  $c_1 = 1$  and  $c_2 = -1$ . Therefore, from Equation 5.11 we have

$$\begin{aligned} \text{Var}(D) &= (1)^2(.25) + (-1)^2(.25) + 2(1)(-1)(.5)(.5)(.5) \\ &= .25 + .25 - .25 = .25 \end{aligned}$$

Notice that this variance is much smaller than the variance of the difference in serum creatinine between two different subjects (at the same or different times). In this case, because values for two different subjects are independent,  $\text{Corr}(X_1, X_2) = 0$ , and it follows that

$$\text{Var}(D) = (1)^2(.25) + (-1)^2(.25) + 0 = .50$$

Thus, the rationale for using each person as his or her own control is because it greatly reduces variability. In Chapter 8 we discuss in more detail the paired-sample and independent-sample experimental designs for comparing two groups (such as a treated and a control group).

To compute the variance of a linear contrast involving  $n$  (possibly) dependent random variables, it can be shown that

**Equation 5.12**
**Variance of Linear Combination of Random Variables (General Case)**

The variance of the linear contrast  $L = \sum_{i=1}^n c_i X_i$  is

$$\begin{aligned}\text{Var}(L) &= \sum_{i=1}^n c_i^2 \text{Var}(X_i) + 2 \sum_{\substack{i=1 \\ i < j}}^n \sum_{j=1}^n c_i c_j \text{Cov}(X_i X_j) \\ &= \sum_{i=1}^n c_i^2 \text{Var}(X_i) + 2 \sum_{\substack{i=1 \\ i < j}}^n \sum_{j=1}^n c_i c_j \sigma_i \sigma_j \text{Corr}(X_i, X_j)\end{aligned}$$

We discuss covariance and correlation in more detail in Chapter 11.

Finally, we can generalize Equation 5.10 by stating that

**Equation 5.13**

If  $X_1, \dots, X_n$  are normal random variables with expected values  $\mu_1, \dots, \mu_n$  and variances  $\sigma_1^2, \dots, \sigma_n^2$  and  $L$  is any linear contrast  $= \sum_{i=1}^n c_i X_i$ , then  $L$  is normally distributed with expected value  $= \sum_{i=1}^n c_i \mu_i$  and variance given by

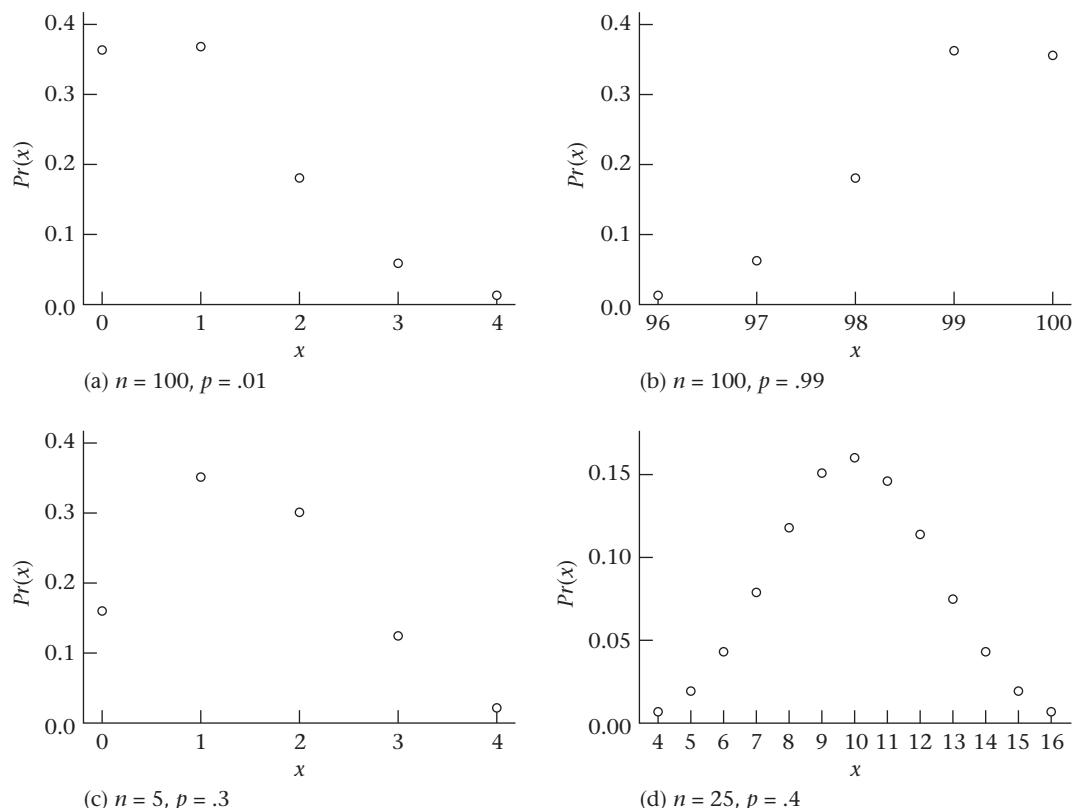
$$\text{Var}(L) = \sum_{i=1}^n c_i^2 \sigma_i^2 + 2 \sum_{\substack{i=1 \\ i < j}}^n \sum_{j=1}^n c_i c_j \sigma_i \sigma_j \rho_{ij}$$

where  $\rho_{ij}$  = correlation between  $X_i$  and  $X_j$ ,  $i \neq j$ , provided that  $L$  is not a constant.

## 5.7 Normal Approximation to the Binomial Distribution

In Chapter 4 we introduced the binomial distribution to assess the probability of  $k$  successes in  $n$  independent trials, where the probability of success ( $p$ ) is the same for each trial. If  $n$  is large, the binomial distribution is very cumbersome to work with and an approximation is easier to use rather than the exact binomial distribution. The normal distribution is often used to approximate the binomial because it is very easy to work with. The key question is: When does the normal distribution provide an accurate approximation to the binomial?

Suppose a binomial distribution has parameters  $n$  and  $p$ . If  $n$  is moderately large and  $p$  is either near 0 or near 1, then the binomial distribution will be very positively or negatively skewed, respectively (Figure 5.17a and 5.17b). Similarly, when  $n$  is small, for any  $p$ , the distribution tends to be skewed (Figure 5.17c). However, if  $n$  is moderately large and  $p$  is not too extreme, then the binomial distribution tends to be symmetric and is well approximated by a normal distribution (Figure 5.17d).

**Figure 5.17 Symmetry properties of the binomial distribution**

We know from Chapter 4 that the mean and variance of a binomial distribution are  $np$  and  $npq$ , respectively. A natural approximation to use is a normal distribution with the *same* mean and variance—that is,  $N(np, npq)$ . Suppose we want to compute  $Pr(a \leq X \leq b)$  for some integers  $a, b$  where  $X$  is binomially distributed with parameters  $n$  and  $p$ . This probability might be approximated by the area under the normal curve from  $a$  to  $b$ . However, we can show empirically that a better approximation to this probability is the area under the normal curve from  $a - \frac{1}{2}$  to  $b + \frac{1}{2}$ . This is generally the case when any discrete distribution is approximated by the normal distribution. Thus the following rule applies:

**Equation 5.14****Normal Approximation to the Binomial Distribution**

If  $X$  is a binomial random variable with parameters  $n$  and  $p$ , then  $Pr(a \leq X \leq b)$  is approximated by the area under an  $N(np, npq)$  curve from  $a - \frac{1}{2}$  to  $b + \frac{1}{2}$ . This rule implies that for the special case  $a = b$ , the binomial probability  $Pr(X = a)$  is approximated by the area under the normal curve from  $a - \frac{1}{2}$  to  $a + \frac{1}{2}$ . The only exception to this rule is that  $Pr(X = 0)$  and  $Pr(X = n)$  are approximated by the area under the normal curve to the left of  $\frac{1}{2}$  and to the right of  $n - \frac{1}{2}$ , respectively.

We saw in Equation 5.10 that if  $X_1, \dots, X_n$  are independent normal random variables, then any linear combination of these variables  $L = \sum_{i=1}^n c_i X_i$  is normally distributed. In particular, if  $c_1 = \dots = c_n = 1$ , then a sum of normal random variables  $L = \sum_{i=1}^n X_i$  is normally distributed.

The normal approximation to the binomial distribution is a special case of a very important statistical principle, the central-limit theorem, which is a generalization of Equation 5.10. Under this principle, for large  $n$ , a sum of  $n$  random variables is approximately normally distributed *even if the individual random variables being summed are not themselves normal*.

### Definition 5.14

Let  $X_i$  be a random variable that takes on the value 1 with probability  $p$  and the value 0 with probability  $q = 1 - p$ . This type of random variable is referred to as a **Bernoulli trial**. This is a special case of a binomial random variable with  $n = 1$ .

We know from the definition of an expected value that  $E(X_i) = 1(p) + 0(q) = p$  and that  $E(X_i^2) = 1^2(p) + 0^2(q) = p$ . Therefore,

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = p - p^2 = p(1 - p) = pq$$

Now consider the random variable

$$X = \sum_{i=1}^n X_i$$

This random variable represents the number of successes among  $n$  trials.

### Example 5.32

Interpret  $X_1, \dots, X_n$  and  $X$  in the case of the number of neutrophils among 100 white blood cells (see Example 4.15).

### Solution

In this case,  $n = 100$  and  $X_i = 1$  if the  $i$ th white blood cell is a neutrophil and  $X_i = 0$  if the  $i$ th white blood cell is not a neutrophil, where  $i = 1, \dots, 100$ .  $X$  represents the number of neutrophils among  $n = 100$  white blood cells.

Given Equations 5.8 and 5.9, we know that

$$E(X) = E\left(\sum_{i=1}^n X_i\right) = p + p + \dots + p = np$$

and

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) = pq + pq + \dots + pq = npq$$

Given the normal approximation to the binomial distribution, we approximate the distribution of  $X$  by a normal distribution with mean =  $np$  and variance =  $npq$ . We discuss the central-limit theorem in more detail in Section 6.5.

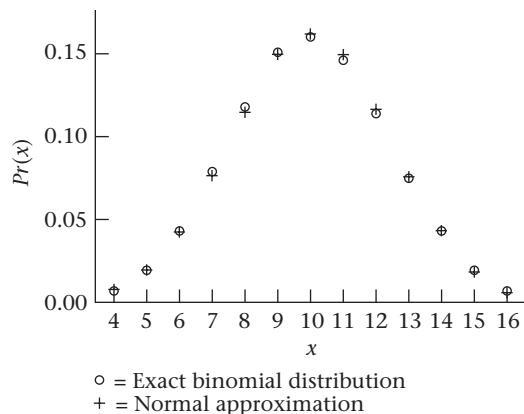
### Example 5.33

Suppose a binomial distribution has parameters  $n = 25$ ,  $p = .4$ . How can  $\Pr(7 \leq X \leq 12)$  be approximated?

### Solution

We have  $np = 25(.4) = 10$ ,  $npq = 25(.4)(.6) = 6.0$ . Thus, this distribution is approximated by a normal random variable  $Y$  with mean 10 and variance 6. We specifically want to compute the area under this normal curve from 6.5 to 12.5. We have

**Figure 5.18** The approximation of the binomial random variable  $X$  with parameters  $n = 25$ ,  $p = .4$  by the normal random variable  $Y$  with mean = 10 and variance = 6



$$\begin{aligned}
 Pr(6.5 \leq Y \leq 12.5) &= \Phi\left(\frac{12.5 - 10}{\sqrt{6}}\right) - \Phi\left(\frac{6.5 - 10}{\sqrt{6}}\right) \\
 &= \Phi(1.021) - \Phi(-1.429) = \Phi(1.021) - [1 - \Phi(1.429)] \\
 &= \Phi(1.021) + \Phi(1.429) - 1 = .8463 + .9235 - 1 = .770
 \end{aligned}$$

This approximation is depicted in Figure 5.18. For comparison, we also computed  $Pr(7 \leq X \leq 12)$  using the BINOMDIST function of Excel and obtained .773, which compares well with the normal approximation of .770.

#### Example 5.34

**Infectious Disease** Suppose we want to compute the probability that between 50 and 75 of 100 white blood cells will be neutrophils, where the probability that any one cell is a neutrophil is .6. These values are chosen as proposed limits to the range of neutrophils in normal people, and we wish to predict what proportion of people will be in the normal range according to this definition.

#### Solution

The exact probability is given by

$$\sum_{k=50}^{75} \binom{100}{k} (.6)^k (.4)^{100-k}$$

The normal approximation is used to approximate the exact probability. The mean of the binomial distribution in this case is  $100(.6) = 60$ , and the variance is  $100(.6)(.4) = 24$ . Thus we find the area between 49.5 and 75.5 for an  $N(60, 24)$  distribution. This area is

$$\begin{aligned}
 \Phi\left(\frac{75.5 - 60}{\sqrt{24}}\right) - \Phi\left(\frac{49.5 - 60}{\sqrt{24}}\right) &= \Phi(3.164) - \Phi(-2.143) \\
 &= \Phi(3.164) + \Phi(2.143) - 1 \\
 &= .9992 + .9840 - 1 = .983
 \end{aligned}$$

Thus 98.3% of the people will be normal.

#### Example 5.35

**Infectious Disease** Suppose a neutrophil count is defined as abnormally high if the number of neutrophils is  $\geq 76$  and abnormally low if the number of neutrophils is  $\leq 49$ . Calculate the proportion of people whose neutrophil counts are abnormally high or low.

**Solution**

The probability of being abnormally high is given by  $\Pr(X \geq 76) \approx \Pr(Y \geq 75.5)$ , where  $X$  is a binomial random variable with parameters  $n = 100$ ,  $p = .6$ , and  $Y \sim N(60, 24)$ . This probability is

$$1 - \Phi\left(\frac{75.5 - 60}{\sqrt{24}}\right) = 1 - \Phi(3.164) = .001$$

Similarly, the probability of being abnormally low is

$$\begin{aligned}\Pr(X \leq 49) &\approx \Pr(Y \leq 49.5) = \Phi\left(\frac{49.5 - 60}{\sqrt{24}}\right) \\ &= \Phi(-2.143) = 1 - \Phi(2.143) \\ &= 1 - .9840 = .016\end{aligned}$$

Thus 0.1% of people will have abnormally high neutrophil counts and 1.6% will have abnormally low neutrophil counts. These probabilities are shown in Figure 5.19.

For comparative purposes, we have also computed (using Excel) the proportion of people who are in the normal range, abnormally high, and abnormally low based on exact binomial probabilities. We obtain  $\Pr(50 \leq X \leq 75) = .983$ ,  $\Pr(X \geq 76) = .0006$ , and  $\Pr(X \leq 49) = .017$ , which corresponds almost exactly to the normal approximations used in Examples 5.34 and 5.35.

Under what conditions should this approximation be used?

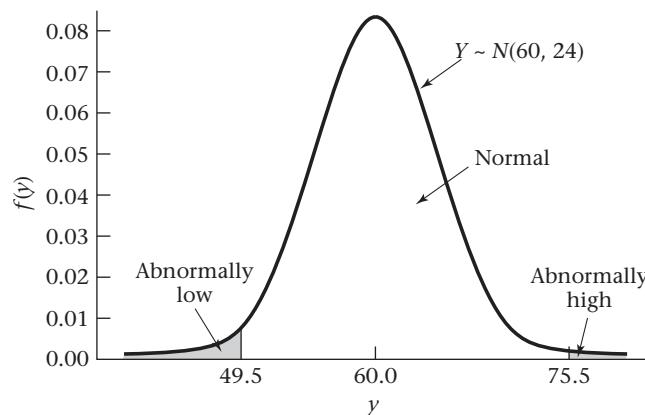
**Equation 5.15**

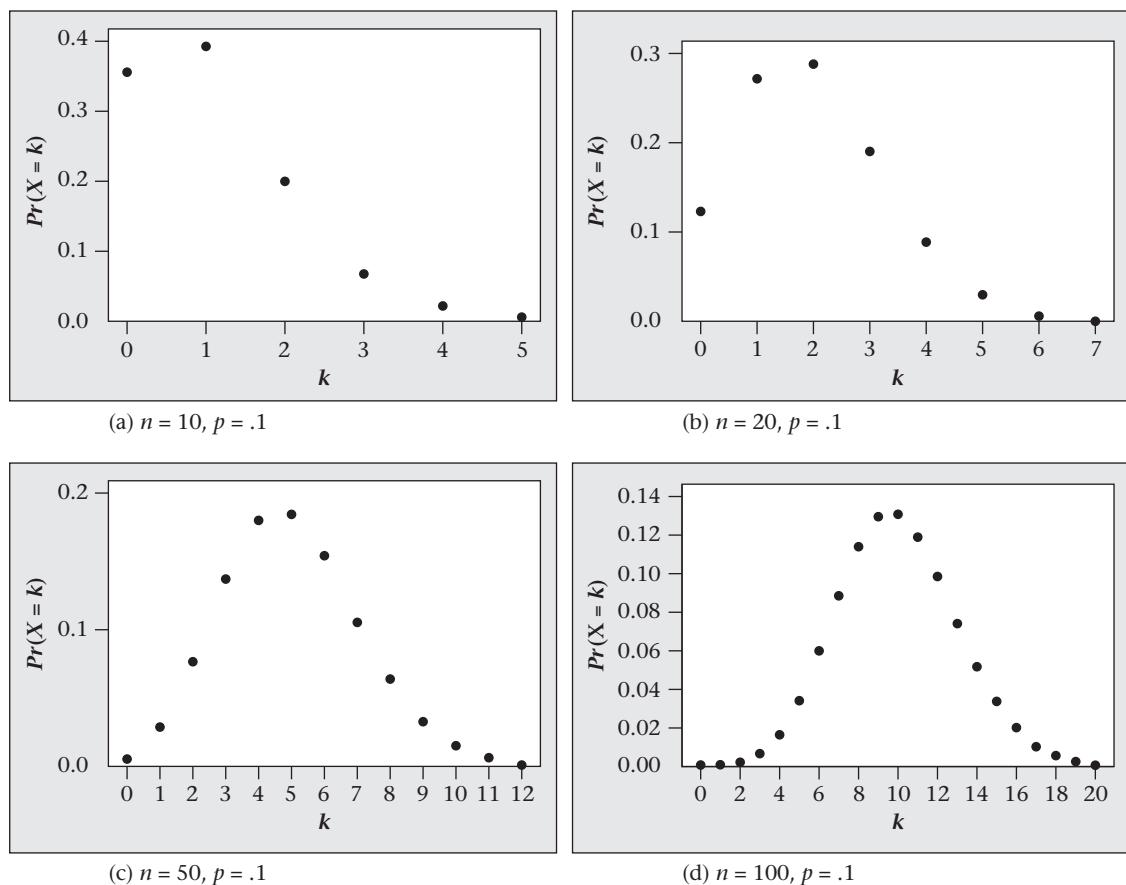
The normal distribution with mean  $np$  and variance  $npq$  can be used to approximate a binomial distribution with parameters  $n$  and  $p$  when  $npq \geq 5$ .

This condition is satisfied if  $n$  is moderately large and  $p$  is not too small or too large. To illustrate this condition, the binomial probability distributions for  $p = .1$ ,  $n = 10, 20, 50$ , and  $100$  are plotted in Figure 5.20 and  $p = .2, n = 10, 20, 50$ , and  $100$  are plotted in Figure 5.21, using MINITAB.

Notice that the normal approximation to the binomial distribution does not fit well in Figure 5.20a,  $n = 10, p = .1$  ( $npq = 0.9$ ) or Figure 5.20b,  $n = 20, p = .1$  ( $npq = 1.8$ ).

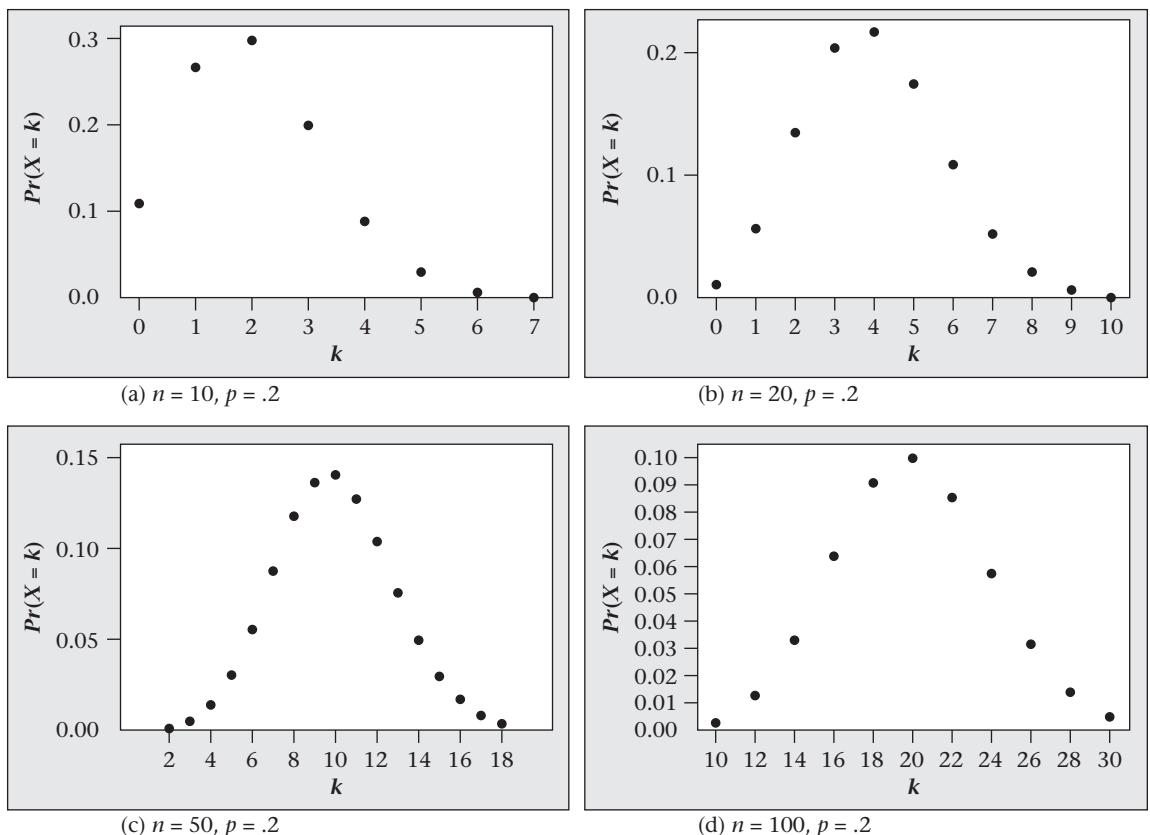
**Figure 5.19** Normal approximation to the distribution of neutrophils



**Figure 5.20** MINITAB plot of binomial distribution,  $n = 10, 20, 50, 100, p = .1$ 

The approximation is marginally adequate in Figure 5.20c,  $n = 50, p = .1$  ( $npq = 4.5$ ), where the right-hand tail is only slightly longer than the left-hand tail. The approximation is quite good in Figure 5.20d,  $n = 100, p = .1$  ( $npq = 9.0$ ), where the distribution appears quite symmetric. Similarly, for  $p = .2$ , although the normal approximation is not good for  $n = 10$  (Figure 5.21a,  $npq = 1.6$ ), it becomes marginally adequate for  $n = 20$  (Figure 5.21b,  $npq = 3.2$ ) and quite good for  $n = 50$  (Figure 5.21c,  $npq = 8.0$ ) and  $n = 100$  (Figure 5.21d,  $npq = 16.0$ ).

Note that the conditions under which the normal approximation to the binomial distribution works well (namely,  $npq \geq 5$ ), which corresponds to  $n$  moderate and  $p$  not too large or too small, are generally *not* the same as the conditions for which the Poisson approximation to the binomial distribution works well [ $n$  large ( $\geq 100$ ) and  $p$  very small ( $p \leq .01$ )]. However, occasionally both these criteria are met. In such cases (for example, when  $n = 1000, p = .01$ ), the two approximations yield about the same results. The normal approximation is preferable because it is easier to apply.

**Figure 5.21** MINITAB plot of binomial distribution,  $n = 10, 20, 50, 100, p = .2$ 

## 5.8 Normal Approximation to the Poisson Distribution

The normal distribution can also be used to approximate discrete distributions other than the binomial distribution, particularly the Poisson distribution. The motivation for this is that the Poisson distribution is cumbersome to use for large values of  $\mu$ .

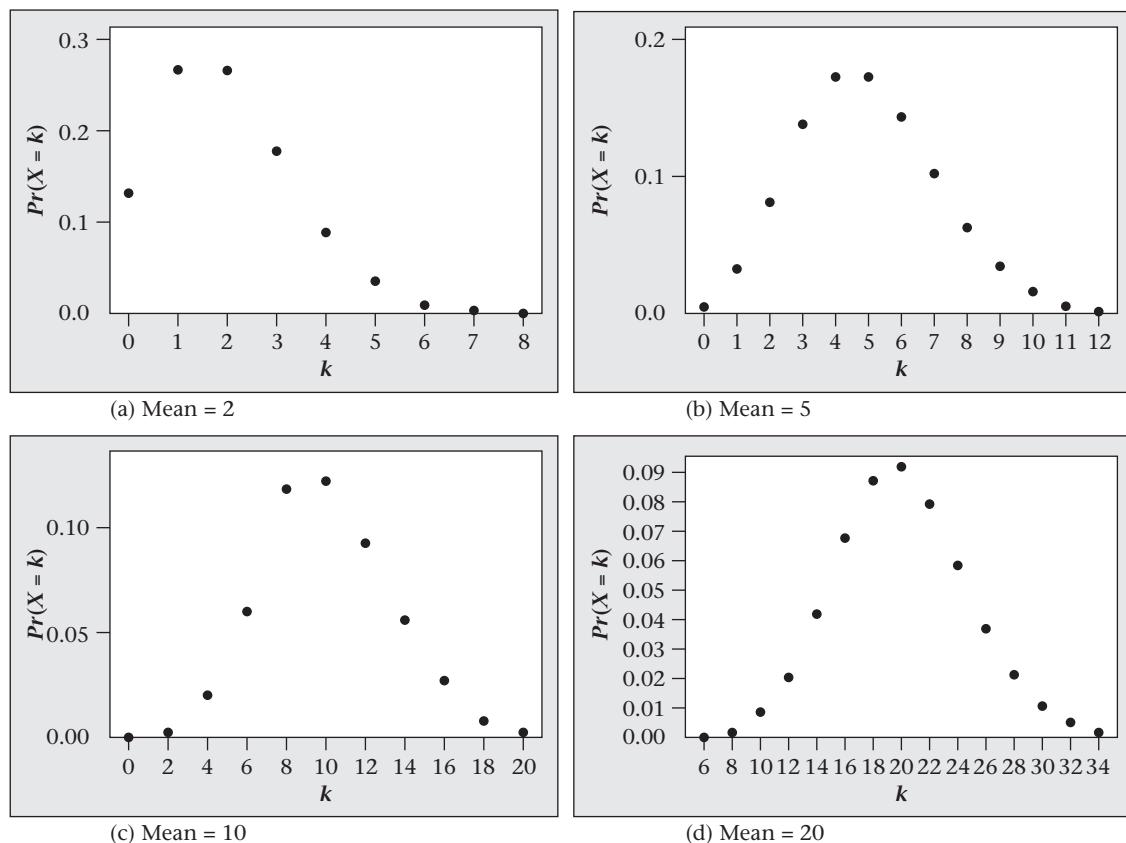
The same technique is used as for the binomial distribution; that is, the mean and variance of the Poisson distribution and the approximating normal distribution are equated.

### Equation 5.16

#### Normal Approximation to the Poisson Distribution

A Poisson distribution with parameter  $\mu$  is approximated by a normal distribution with mean and variance both equal to  $\mu$ .  $Pr(X = x)$  is approximated by the area under an  $N(\mu, \mu)$  density from  $x - \frac{1}{2}$  to  $x + \frac{1}{2}$  for  $x > 0$  or by the area to the left of  $\frac{1}{2}$  for  $x = 0$ . This approximation is used when  $\mu \geq 10$ .

The Poisson distributions for  $\mu = 2, 5, 10$ , and  $20$  are plotted using MINITAB in Figure 5.22. The normal approximation is clearly inadequate for  $\mu = 2$  (Figure 5.22a), marginally adequate for  $\mu = 5$  (Figure 5.22b), and adequate for  $\mu = 10$  (Figure 5.22c) and  $\mu = 20$  (Figure 5.22d).

**Figure 5.22** MINITAB plot of Poisson distribution,  $\mu = 2, 5, 10, 20$ **Example 5.36**

**Bacteriology** Consider again the distribution of the number of bacteria in a Petri plate of area  $A$ . Assume the probability of observing  $x$  bacteria is given exactly by a Poisson distribution with parameter  $\mu = \lambda A$ , where  $\lambda = 0.1$  bacteria/cm<sup>2</sup> and  $A = 100$  cm<sup>2</sup>. Suppose 20 bacteria are observed in this area. How unusual is this event?

**Solution**

The exact distribution of the number of bacteria observed in 100 cm<sup>2</sup> is Poisson with parameter  $\mu = 10$ . We approximate this distribution by a normal distribution with mean = 10 and variance = 10. Therefore, we compute

$$Pr(X \geq 20) \approx Pr(Y \geq 19.5)$$

where  $Y \sim N(\lambda A, \lambda A) = N(10, 10)$

We have

$$\begin{aligned} Pr(Y \geq 19.5) &= 1 - Pr(Y \leq 19.5) = 1 - \Phi\left(\frac{19.5 - 10}{\sqrt{10}}\right) \\ &= 1 - \Phi\left(\frac{9.5}{\sqrt{10}}\right) = 1 - \Phi(3.004) \\ &= 1 - .9987 = .0013 \end{aligned}$$

Thus 20 or more colonies in 100 cm<sup>2</sup> would be expected only 1.3 times in 1000 plates, a rare event indeed. For comparison, we have also computed the exact

Poisson probability of obtaining 20 or more bacteria, using Excel, and obtain  $Pr(X \geq 20 | \mu = 10) = .0035$ . Thus the normal approximation is only fair in this case but does result in the same conclusion that obtaining 20 or more bacteria in  $100 \text{ cm}^2$  is a rare event.

### REVIEW QUESTIONS 5C

- 1** Why do we use the normal approximation to the binomial distribution?
- 2** Which of the following binomial distributions can be well approximated by a normal distribution? A Poisson distribution? Both? Neither?
  - (a)**  $n = 40, p = .05$
  - (b)**  $n = 300, p = .05$
  - (c)**  $n = 500, p = .001$
  - (d)**  $n = 1000, p = .001$
- 3** The prevalence of glaucoma among the elderly in high-risk inner-city populations is about 5%. Suppose an “Eyemobile” is sent to several neighborhoods in the Chicago area to identify subjects for a new glaucoma study. If 500 elderly people (age 65+) are screened by Eyemobile staff, then what is the probability of identifying at least 20 glaucoma cases?
- 4** The number of deaths from heart failure in a hospital is approximately Poisson distributed with mean = 20 cases per year. In 2002, a hospital sees 35 deaths from heart failure. Is this an unusual occurrence? Why or why not?

## 5.9 Summary

In this chapter continuous random variables were discussed. The concept of a probability-density function (pdf), which is the analog to a probability-mass function for discrete random variables, was introduced. In addition, generalizations of the concepts of expected value, variance, and cumulative distribution were presented for continuous random variables.

The normal distribution, the most important continuous distribution, was then studied in detail. The normal distribution is often used in statistical work because many random phenomena follow this probability distribution, particularly those that can be expressed as a sum of many random variables. It was shown that the normal distribution is indexed by two parameters, the mean  $\mu$  and the variance  $\sigma^2$ . Fortunately, all computations concerning any normal random variable can be accomplished using the standard, or unit, normal probability law, which has mean 0 and variance 1. Normal tables were introduced to use when working with the standard normal distribution. Alternatively, electronic tables can be used to evaluate areas and/or percentiles for *any* normal distribution. Also, because the normal distribution is easy to use, it is often employed to approximate other distributions. In particular, we studied the normal approximations to the binomial and Poisson distributions. These are special cases of the central-limit theorem, which is covered in more detail in Chapter 6. Also, to facilitate applications of the central-limit theorem, the properties of linear combinations of random variables were discussed, both for independent and dependent random variables.

In the next three chapters, the normal distribution is used extensively as a foundation for work on statistical inference.

## PROBLEMS

### Cardiovascular Disease

Because serum cholesterol is related to age and sex, some investigators prefer to express it in terms of  $z$ -scores. If  $X$  = raw serum cholesterol, then  $Z = \frac{X - \mu}{\sigma}$ , where  $\mu$  is the mean and  $\sigma$  is the standard deviation of serum cholesterol for a given age–sex group. Suppose  $Z$  is regarded as a standard normal random variable.

\*5.1 What is  $Pr(Z < 0.5)$ ?

\*5.2 What is  $Pr(Z > 0.5)$ ?

\*5.3 What is  $Pr(-1.0 < Z < 1.5)$ ?

Suppose a person is regarded as having high cholesterol if  $Z > 2.0$  and borderline cholesterol if  $1.5 < Z < 2.0$ .

\*5.4 What proportion of people have high cholesterol?

\*5.5 What proportion of people have borderline cholesterol?

### Nutrition

Suppose that total carbohydrate intake in 12- to 14-year-old boys is normally distributed, with mean = 124 g/1000 cal and standard deviation = 20 g/1000 cal.

\*5.6 What percentage of boys in this age range have carbohydrate intake above 140 g/1000 cal?

\*5.7 What percentage of boys in this age range have carbohydrate intake below 90 g/1000 cal?

Suppose boys in this age range who live below the poverty level have a mean carbohydrate intake of 121 g/1000 cal with a standard deviation of 19 g/1000 cal.

5.8 Answer Problem 5.6 for boys in this age range and economic environment.

5.9 Answer Problem 5.7 for boys in this age range and economic environment.

### Pulmonary Disease

Many investigators have studied the relationship between asbestos exposure and death from chronic obstructive pulmonary disease (COPD).

5.10 Suppose that among workers exposed to asbestos in a shipyard in 1980, 33 died over a 10-year period from COPD, whereas only 24 such deaths would be expected based on statewide mortality rates. Is the number of deaths from COPD in this group excessive?

5.11 Twelve cases of leukemia are reported in people living in a particular census tract over a 5-year period. Is this number of cases abnormal if only 6.7 cases would be expected based on national cancer-incidence rates?

### Cardiovascular Disease, Pulmonary Disease

The duration of cigarette smoking has been linked to many diseases, including lung cancer and various forms of heart

disease. Suppose we know that among men ages 30–34 who have ever smoked, the mean number of years they smoked is 12.8 with a standard deviation of 5.1 years. For women in this age group, the mean number of years they smoked is 9.3 with a standard deviation of 3.2.

\*5.12 Assuming that the duration of smoking is normally distributed, what proportion of men in this age group have smoked for more than 20 years?

\*5.13 Answer Problem 5.12 for women.

### Cardiovascular Disease

Serum cholesterol is an important risk factor for coronary disease. We can show that serum cholesterol is approximately normally distributed, with mean = 219 mg/dL and standard deviation = 50 mg/dL.

\*5.14 If the clinically desirable range for cholesterol is  $< 200$  mg/dL, what proportion of people have clinically desirable levels of cholesterol?

\*5.15 Some investigators believe that only cholesterol levels over 250 mg/dL indicate a high-enough risk for heart disease to warrant treatment. What proportion of the population does this group represent?

\*5.16 What proportion of the general population has borderline high-cholesterol levels—that is,  $> 200$  but  $< 250$  mg/dL?

### Hypertension

People are classified as hypertensive if their systolic blood pressure (SBP) is higher than a specified level for their age group, according to the algorithm in Table 5.1.

Assume SBP is normally distributed with mean and standard deviation given in Table 5.1 for age groups 1–14 and 15–44, respectively. Define a *family* as a group of two people in age group 1–14 and two people in age group 15–44. A family is classified as hypertensive if at least one adult and at least one child are hypertensive.

\*5.17 What proportion of 1- to 14-year-olds are hypertensive?

\*5.18 What proportion of 15- to 44-year-olds are hypertensive?

**Table 5.1 Mean and standard deviation of SBP (mm Hg) in specific age groups**

Age group	Mean	Standard deviation	Specified hypertension level
1–14	105.0	5.0	115.0
15–44	125.0	10.0	140.0

**\*5.19** What proportion of families are hypertensive? (Assume that the hypertensive status of different members of a family are independent random variables.)

**\*5.20** Suppose a community has 1000 families living in it. What is the probability that between one and five families are hypertensive?

### Pulmonary Disease

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. FEV is influenced by age, sex, and cigarette smoking. Assume that in 45- to 54-year-old nonsmoking men FEV is normally distributed with mean = 4.0 L and standard deviation = 0.5 L.

In comparably aged currently smoking men FEV is normally distributed, with mean = 3.5 L and standard deviation = 0.6 L.

**5.21** If an FEV of less than 2.5 L is regarded as showing some functional impairment (occasional breathlessness, inability to climb stairs, etc.), then what is the probability that a currently smoking man has functional impairment?

**5.22** Answer Problem 5.21 for a nonsmoking man.

Some people are not functionally impaired now, but their pulmonary function usually declines with age and they eventually will be functionally impaired. Assume that the *decline* in FEV over  $n$  years is normally distributed, with mean =  $0.03n$  L and standard deviation =  $0.02n$  L.

**5.23** What is the probability that a 45-year-old man with an FEV of 4.0 L will be functionally impaired by age 75?

**5.24** Answer Problem 5.23 for a 25-year-old man with an FEV of 4.0 L.

### Infectious Disease

The differential is a standard measurement made during a blood test. It consists of classifying white blood cells into the following five categories: (1) basophils, (2) eosinophils, (3) monocytes, (4) lymphocytes, and (5) neutrophils. The usual practice is to look at 100 randomly selected cells under a microscope and to count the number of cells within each of the five categories. Assume that a normal adult will have the following proportions of cells in each category: basophils, 0.5%; eosinophils, 1.5%; monocytes, 4%; lymphocytes, 34%; and neutrophils, 60%.

**\*5.25** An excess of eosinophils is sometimes consistent with a violent allergic reaction. What is the exact probability that a normal adult will have 5 or more eosinophils?

**\*5.26** An excess of lymphocytes is consistent with various forms of viral infection, such as hepatitis. What is the probability that a normal adult will have 40 or more lymphocytes?

**\*5.27** What is the probability a normal adult will have 50 or more lymphocytes?

**\*5.28** How many lymphocytes would have to appear in the differential before you would feel the "normal" pattern was violated?

**\*5.29** An excess of neutrophils is consistent with several types of bacterial infection. Suppose an adult has  $x$  neutrophils. How large would  $x$  have to be for the probability of a normal adult having  $x$  or more neutrophils to be  $\leq 5\%$ ?

**\*5.30** How large would  $x$  have to be for the probability of a normal adult having  $x$  or more neutrophils to be  $\leq 1\%$ ?

### Blood Chemistry

In pharmacologic research a variety of clinical chemistry measurements are routinely monitored closely for evidence of side effects of the medication under study. Suppose typical blood-glucose levels are normally distributed, with mean = 90 mg/dL and standard deviation = 38 mg/dL.

**5.31** If the normal range is 65–120 mg/dL, then what percentage of values will fall in the normal range?

**5.32** In some studies only values at least 1.5 times as high as the upper limit of normal are identified as abnormal. What percentage of values would fall in this range?

**5.33** Answer Problem 5.32 for values 2.0 times the upper limit of normal.

**5.34** Frequently, tests that yield abnormal results are repeated for confirmation. What is the probability that for a normal person a test will be at least 1.5 times as high as the upper limit of normal on two separate occasions?

**5.35** Suppose that in a pharmacologic study involving 6000 patients, 75 patients have blood-glucose levels at least 1.5 times the upper limit of normal on one occasion. What is the probability that this result could be due to chance?

### Cancer

A treatment trial is proposed to test the efficacy of vitamin E as a preventive agent for cancer. One problem with such a study is how to assess compliance among participants. A small pilot study is undertaken to establish criteria for compliance with the proposed study agents. In this study, 10 patients are given 400 IU/day of vitamin E and 10 patients are given similar-sized tablets of placebo over a 3-month period. Their serum vitamin-E levels are measured before and after the 3-month period, and the change (3-month – baseline) is shown in Table 5.2.

**Table 5.2 Change in serum vitamin E (mg/dL) in pilot study**

Group	Mean	sd	<i>n</i>
Vitamin E	0.80	0.48	10
Placebo	0.05	0.16	10

**\*5.36** Suppose a change of 0.30 mg/dL in serum levels is proposed as a test criterion for compliance; that is, a patient who shows a change of  $\geq 0.30$  mg/dL is considered a compliant vitamin-E taker. If normality is assumed, what percentage of the vitamin-E group would be expected to show a change of at least 0.30 mg/dL?

**\*5.37** Is the measure in Problem 5.36 a measure of sensitivity, specificity, or predictive value?

**\*5.38** What percentage of the placebo group would be expected to show a change of not more than 0.30 mg/dL?

**\*5.39** Is the measure in Problem 5.38 a measure of sensitivity, specificity, or predictive value?

**\*5.40** Suppose a new threshold of change,  $\Delta$  mg/dL, is proposed for establishing compliance. We wish to use a level of  $\Delta$  such that the compliance measures in Problems 5.36 and 5.38 for the patients in the vitamin-E and placebo groups are the same. What should  $\Delta$  be? What would be the compliance in the vitamin-E and placebo groups using this threshold level?

**5.41** Suppose we consider the serum vitamin-E assay as a screening test for compliance with vitamin-E supplementation. Participants whose change in serum vitamin E is  $\geq \Delta$  mg/dL will be considered vitamin-E takers, and participants whose change is  $< \Delta$  mg/dL will be considered placebo takers. Choose several possible values for  $\Delta$ , and construct the receiver operating characteristic (ROC) curve for this test. What is the area under the ROC curve? (*Hint:* The area under the ROC curve can be computed analytically from the properties of linear combinations of normal distributions.)

### Pulmonary Disease

Refer to the pulmonary-function data in the Data Set FEV.DAT on the Companion Website (see Problem 2.23, p. 35). We are interested in whether smoking status is related to level of pulmonary function. However, FEV is affected by age and sex; also, smoking children tend to be older than nonsmoking children. For these reasons, FEV should be standardized for age and sex. To accomplish this, use the z-score approach outlined in Problem 5.1, where the z-scores here are defined by age–sex groups.

**5.42** Plot the distribution of z-scores for smokers and nonsmokers separately. Do these distributions look normal? Do smoking and pulmonary function seem in any way related in these data?

**5.43** Repeat the analyses in Problem 5.42 for the subgroup of children 10+ years of age (because smoking is very rare before this age). Do you reach similar conclusions?

**5.44** Repeat the analyses in Problem 5.43 separately for boys and girls. Are your conclusions the same in the two groups?

(Note: Formal methods for comparing mean FEVs between smokers and nonsmokers are discussed in the material on statistical inference in Chapter 8.)

### Cardiovascular Disease

A clinical trial was conducted to test the efficacy of nifedipine, a new drug for reducing chest pain in patients with angina severe enough to require hospitalization. The duration of the study was 14 days in the hospital unless the patient was withdrawn prematurely from therapy, was discharged from the hospital, or died prior to this time. Patients were randomly assigned to either nifedipine or propranolol and were given the same dosage of each drug in identical capsules at level 1 of therapy. If pain did not cease at this level of therapy or if pain recurred after a period of pain cessation, then the patient progressed to level 2, whereby the dosage of each drug was increased according to a pre-specified schedule. Similarly, if pain continued or recurred at level 2, then the patient progressed to level 3, whereby the dosage of the anginal drug was increased again. Patients randomized to either group received nitrates in any amount deemed clinically appropriate to help control pain.

The main objective of the study was to compare the degree of pain relief with nifedipine vs. propranolol. A secondary objective was to better understand the effects of these agents on other physiologic parameters, including heart rate and blood pressure. Data on these latter parameters are given in Data Set NIFED.DAT (on the Companion Website); the format of this file is shown in Table 5.3.

**5.45** Describe the effect of each treatment regimen on changes in heart rate and blood pressure. Does the distribution of changes in these parameters look normal or not?

**5.46** Compare graphically the effects of the treatment regimens on heart rate and blood pressure. Do you notice any difference between treatments?

**Table 5.3 Format of NIFED.DAT**

Column	Variable	Code
1–2	ID	
4	Treatment group	N = nifedipine/ P = propranolol
6–8	Baseline heart rate <sup>a</sup>	beats/min
10–12	Level 1 heart rate <sup>b</sup>	beats/min
14–16	Level 2 heart rate	beats/min
18–20	Level 3 heart rate	beats/min
22–24	Baseline SBP <sup>a</sup>	mm Hg
26–28	Level 1 SBP <sup>b</sup>	mm Hg
30–32	Level 2 SBP	mm Hg
34–36	Level 3 SBP	mm Hg

<sup>a</sup>Heart rate and SBP immediately before randomization.

<sup>b</sup>Highest heart rate and SBP at each level of therapy.

Note: Missing values indicate one of the following:

- (1) The patient withdrew from the study before entering this level of therapy.
- (2) The patient achieved pain relief before reaching this level of therapy.
- (3) The patient encountered this level of therapy, but this particular piece of data was missing.

(Note: Formal tests for comparing changes in heart rate and blood pressure in the two treatment groups are covered in Chapter 8.)

### Hypertension

Well-known racial differences in blood pressure exist between white and black adults. These differences generally do not exist between white and black children. Because aldosterone levels have been related to blood-pressure levels in adults in previous research, an investigation was performed to look at aldosterone levels among black children and white children [1].

**\*5.47** If the mean plasma-aldosterone level in black children was 230 pmol/L with a standard deviation of 203 pmol/L, then what percentage of black children have levels  $\leq 300$  pmol/L if normality is assumed?

**\*5.48** If the mean plasma-aldosterone level in white children is 400 pmol/L with standard deviation of 218 pmol/L, then what percentage of white children have levels  $\leq 300$  pmol/L if normality is assumed?

**\*5.49** The distribution of plasma-aldosterone concentration in 53 white and 46 black children is shown in Figure 5.23. Does the assumption of normality seem reasonable? Why or why not? (*Hint:* Qualitatively compare the observed number of children who have levels  $\leq 300$  pmol/L with the expected number in each group under the assumption of normality.)

### Hepatic Disease

Suppose we observe 84 alcoholics with cirrhosis of the liver, of whom 29 have hepatomas—that is, liver-cell carcinoma. Suppose we know, based on a large sample, that the risk of hepatoma among alcoholics without cirrhosis of the liver is 24%.

**5.50** What is the probability that we observe exactly 29 alcoholics with cirrhosis of the liver who have hepatomas if the true rate of hepatoma among alcoholics (with or without cirrhosis of the liver) is .24?

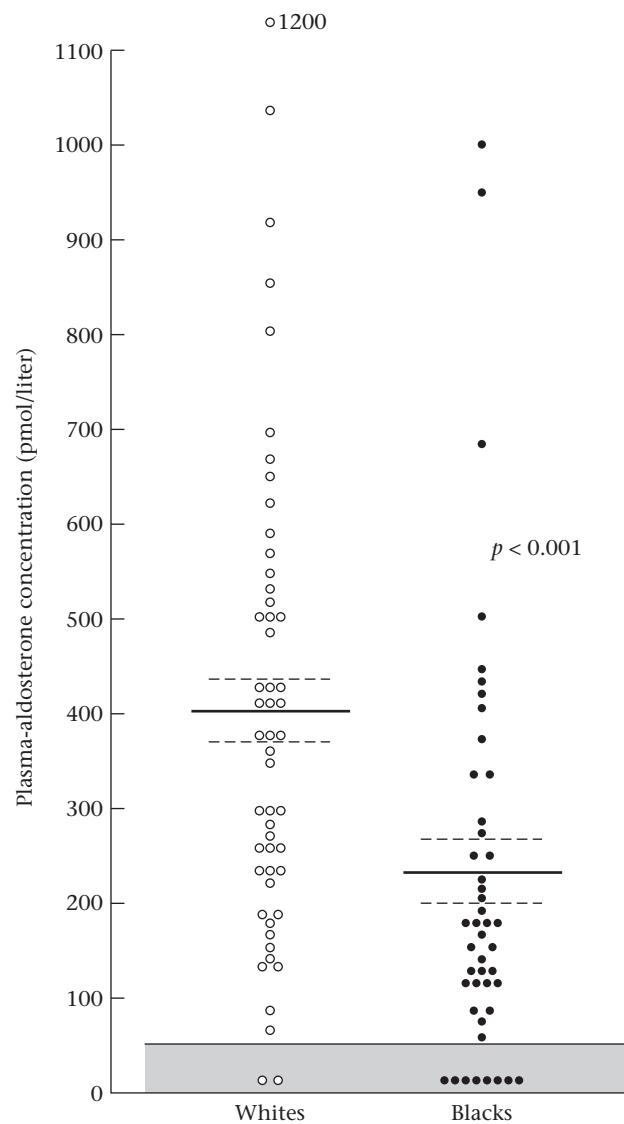
**5.51** What is the probability of observing at least 29 hepatomas among the 84 alcoholics with cirrhosis of the liver under the assumptions in Problem 5.50?

**5.52** What is the smallest number of hepatomas that would have to be observed among the alcoholics with cirrhosis of the liver for the hepatoma experience in this group to differ from the hepatoma experience among alcoholics without cirrhosis of the liver? (*Hint:* Use a 5% probability of getting a result at least as extreme to denote differences between the hepatoma experiences of the two groups.)

### Hypertension

The Fourth Task Force Report on Blood Pressure Control in Children [2] reports blood-pressure norms for children by

**Figure 5.23** Plasma-aldosterone concentrations in 53 white and 46 black children. Values within the shaded area were undetectable ( $< 50$  pmol/L). The solid horizontal lines indicate the mean values, and the broken horizontal lines indicate the mean  $\pm$  se. The concept of standard error (se) is discussed in Chapter 6.



age and sex group. The estimated mean  $\pm$  standard deviation for 17-year-old boys for DBP is  $67.9 \pm 11.6$  mm Hg, based on a large sample.

**5.53** One approach for defining elevated blood pressure is to use 90 mm Hg—the standard for elevated adult DBP—as the cutoff. What percentage of 17-year-old boys would be found to have elevated blood pressure, using this approach?

**5.54** Suppose 2000 17-year-old boys are in the 11th grade, of whom 50 have elevated blood pressure by the criteria in Problem 5.53. Is this an unusually low number of boys with elevated blood pressure? Why or why not?

### Environmental Health

**5.55** A study was conducted relating particulate air pollution and daily mortality in Steubenville, Ohio [3]. On average over the past 10 years there have been 3 deaths per day in Steubenville. Suppose that on 90 high-pollution days—days in which the total suspended particulates are in the highest quartile among all days—the death rate is 3.2 deaths per day, or 288 deaths observed over the 90 high-pollution days. Are there an unusual number of deaths on high-pollution days?

### Nutrition

Refer to Data Set VALID.DAT (on the Companion Website) described in Table 2.16 (p. 36).

**5.56** Consider the nutrients saturated fat, total fat, and total calories. Plot the distribution of each nutrient for both the diet record and the food-frequency questionnaire. Do you think a normal distribution is appropriate for these nutrients?

(Hint: Compute the observed proportion of women who fall within 1.0, 1.5, 2.0, and 2.5 standard deviations of the mean. Compare the observed proportions with the expected proportions based on the assumption of normality.)

**5.57** Answer Problem 5.56 using the  $\ln(\text{nutrient})$  transformation for each nutrient value. Is the normality assumption more appropriate for log-transformed or untransformed values, or neither?

**5.58** A special problem arises for the nutrient alcohol consumption. There are often a large number of nondrinkers (alcohol consumption = 0) and another large group of drinkers (alcohol consumption > 0). The overall distribution of alcohol consumption appears bimodal. Plot the distribution of alcohol consumption for both the diet record and the food frequency questionnaire. Do the distributions appear unimodal or bimodal? Do you think the normality assumption is appropriate for this nutrient?

### Cancer, Neurology

A study concerned the risk of cancer among patients with cystic fibrosis [4]. Given registries of patients with cystic fibrosis in the United States and Canada, cancer incidence among cystic-fibrosis patients between January 1, 1985 and December 31, 1992 was compared with expected cancer-incidence rates based on the Surveillance Epidemiology and End Results program from the National Cancer Institute from 1984 to 1988.

**5.59** Among cystic-fibrosis patients, 37 cancers were observed, whereas 45.6 cancers were expected. What

distribution can be used to model the distribution of the number of cancers among cystic-fibrosis patients?

**5.60** Is there an unusually low number of cancers among cystic-fibrosis patients?

**5.61** In the same study 13 cancers of the digestive tract were observed, whereas only 2 cancers were expected. Is there an unusually high number of digestive cancers among cystic-fibrosis patients?

### Hypertension

A doctor diagnoses a patient as hypertensive and prescribes an antihypertensive medication. To assess the clinical status of the patient, the doctor takes  $n$  replicate blood-pressure measurements before the patient starts the drug (baseline) and  $n$  replicate blood-pressure measurements 4 weeks after starting the drug (follow-up). She uses the average of the  $n$  replicates at baseline minus the average of the  $n$  replicates at follow-up to assess the clinical status of the patient. She knows, from previous clinical experience with the drug, that the mean diastolic blood pressure (DBP) change over a 4-week period over a large number of patients after starting the drug is 5.0 mm Hg with variance  $33/n$ , where  $n$  is the number of replicate measures obtained at both baseline and follow-up.

**5.62** If we assume the change in mean DBP is normally distributed, then what is the probability that a subject will decline by at least 5 mm Hg if 1 replicate measure is obtained at baseline and follow-up?

**5.63** The physician also knows that if a patient is untreated (or does not take the prescribed medication), then the mean DBP over 4 weeks will decline by 2 mm Hg with variance  $33/n$ . What is the probability that an untreated subject will decline by at least 5 mm Hg if 1 replicate measure is obtained at both baseline and follow-up?

**5.64** Suppose the physician is not sure whether the patient is actually taking the prescribed medication. She wants to take enough replicate measurements at baseline and follow-up so that the probability in Problem 5.62 is at least five times the probability in Problem 5.63. How many replicate measurements should she take?

### Endocrinology

A study compared different treatments for preventing bone loss among postmenopausal women younger than 60 years of age [5]. The mean change in bone-mineral density of the lumbar spine over a 2-year period for women in the placebo group was  $-1.8\%$  (a mean decrease), with a standard deviation of  $4.3\%$ . Assume the change in bone-mineral density is normally distributed.

**5.65** If a decline of  $2\%$  in bone-mineral density is considered clinically significant, then what percentage of women in the placebo group can be expected to show a decline of at least this much?

The change in bone-mineral density of the lumbar spine over a 2-year period among women in the alendronate 5-mg group was +3.5% (a mean increase), with a standard deviation of 4.2%.

**5.66** What percentage of women in the alendronate 5-mg group can be expected to have a clinically significant decline in bone-mineral density as defined in Problem 5.65?

**5.67** Suppose 10% of the women assigned to the alendronate 5-mg group are actually not taking their pills (non-compliers). If noncompliers are assumed to have a similar response as women in the placebo group, what percentage of women complying with the alendronate 5-mg treatment would be expected to have a clinically significant decline? (*Hint:* Use the total-probability rule.)

### Cardiovascular Disease

Obesity is an important determinant of cardiovascular disease because it directly affects several established cardiovascular risk factors, including hypertension and diabetes. It is estimated that the average weight for an 18-year-old woman is 123 lb and increases to 142 lb at 50 years of age. Also, let us assume that the average SBP for a 50-year-old woman is 125 mm Hg, with a standard deviation of 15 mm Hg, and that SBP is normally distributed.

**5.68** What proportion of 50-year-old women is hypertensive, if hypertension is defined as  $SBP \geq 140$  mm Hg?

From previous clinical trials, it is estimated that for every 10 lb of weight loss there is, on average, a corresponding reduction in mean SBP of 3 mm Hg.

**5.69** Suppose an average woman did not gain any weight from age 18 to 50. What average SBP for 50-year-old women would be expected under these assumptions?

**5.70** If the standard deviation of SBP under the assumption in Problem 5.69 remained the same (15 mm Hg), and the

distribution of SBP remained normal, then what would be the expected proportion of hypertensive women under the assumption in Problem 5.69?

**5.71** What percentage of hypertension at age 50 is attributable to the weight gain from age 18 to 50?

### SIMULATION

**5.72** Draw 100 random samples from a binomial distribution with parameters  $n = 10$  and  $p = .4$ . Consider an approximation to this distribution by a normal distribution with mean =  $np = 4$  and variance =  $npq = 2.4$ . Draw 100 random samples from the normal approximation. Plot the two frequency distributions on the same graph, and compare the results. Do you think the normal approximation is adequate here?

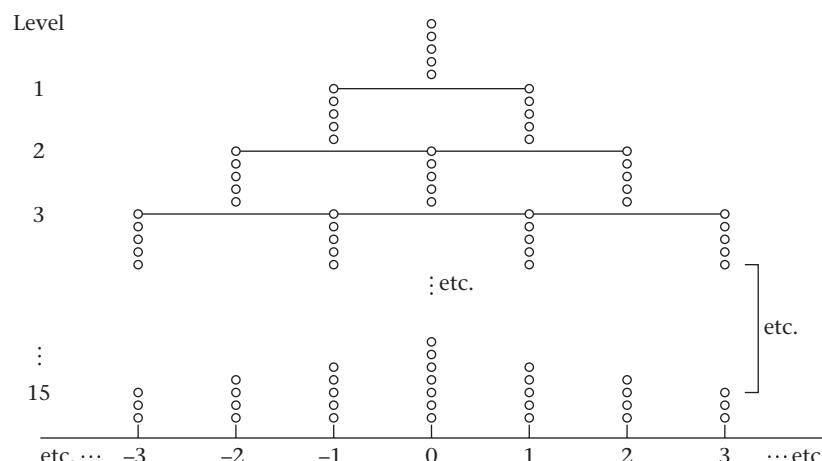
**5.73** Answer the question in Problem 5.72 for a binomial distribution with parameters  $n = 20$  and  $p = .4$  and the corresponding normal approximation.

**5.74** Answer the question in Problem 5.72 for a binomial distribution with parameters  $n = 50$  and  $p = .4$  and the corresponding normal approximation.

### SIMULATION

An apparatus displaces a collection of balls to the top of a stack by suction. At the top level (Level 1) each ball is shifted 1 unit to the left or 1 unit to the right at random with equal probability (see Figure 5.24). The ball then drops down to Level 2. At Level 2, each ball is again shifted 1 unit to the left or 1 unit to the right at random. The process continues for 15 levels; the balls remain at the bottom for a short time and are then forced by suction to the top. (*Note:* A similar apparatus, located in the Museum of Science, Boston, Massachusetts, is displayed in Figure 5.25.)

**Figure 5.24 Apparatus for random displacement of balls**



**Figure 5.25 Probability apparatus at the Museum of Science, Boston**

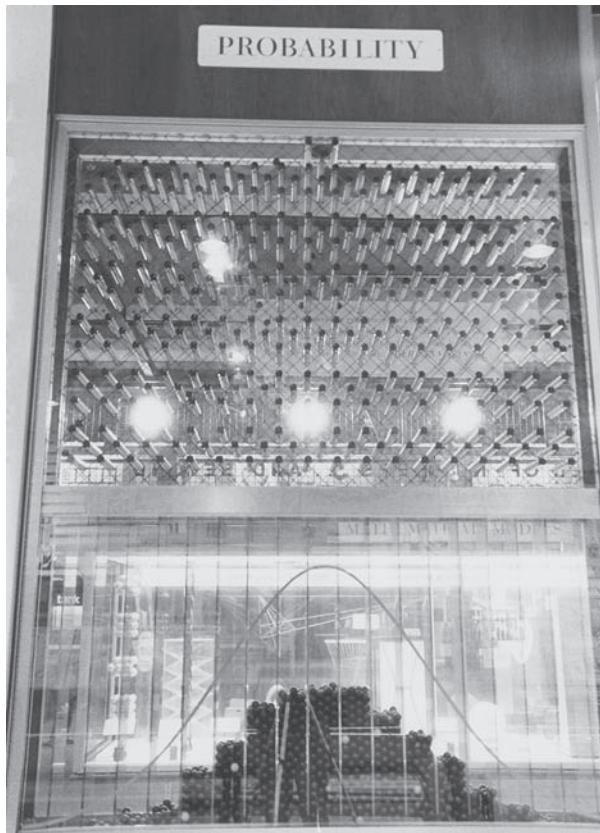


Photo taken by David Rosner, courtesy of the Museum of Science, Boston.

**5.75** What is the exact probability distribution of the position of the balls at the bottom with respect to the entry position (arbitrarily denoted by 0)?

**5.76** Can you think of an approximation to the distribution derived in Problem 5.75?

## SIMULATION

**5.77** Perform a simulation of this process (e.g., using MINITAB or Excel) with 100 balls, and plot the frequency distribution of the position of the balls at the bottom with respect to the entry position. Does the distribution appear to conform to the distributions derived in Problems 5.75 and 5.76?

## Orthopedics

A study was conducted of a diagnostic test (the FAIR test, i.e., hip flexion, adduction, and internal rotation) used to identify people with piriformis syndrome (PS), a pelvic condition that involves malfunction of the piriformis muscle (a deep buttock muscle), which often causes lumbar and buttock pain with sciatica (pain radiating down the leg) [6].

The FAIR test is based on nerve-conduction velocity and is expressed as a difference score (nerve-conduction velocity in an aggravating posture minus nerve-conduction velocity in a neutral posture). It is felt that the larger the FAIR test score, the more likely a participant will be to have PS. Data are given in the Data Set PIRIFORM.DAT for 142 participants without PS (piriform = 1) and 489 participants with PS (piriform = 2) for whom the diagnosis of PS was based on clinical criteria. The FAIR test value is called MAXCHG and is in milliseconds (ms). A cutoff point of  $\geq 1.86$  ms on the FAIR test is proposed to define a positive test.

**5.78** What is the sensitivity of the test for this cutoff point?

**5.79** What is the specificity of the test for this cutoff point?

**5.80** Suppose that 70% of the participants who are referred to an orthopedist who specializes in PS will actually have the condition. If a test score of  $\geq 1.86$  ms is obtained for a participant, then what is the probability that the person has PS?

**5.81** The criterion of  $\geq 1.86$  ms to define a positive test is arbitrary. Using different cutoff points to define positivity, obtain the ROC curve for the FAIR test. What is the area under the ROC curve? What does it mean in this context?

**5.82** Do you think the distribution of FAIR test scores within a group is normally distributed? Why or why not?

## Ophthalmology

Retinitis pigmentosa (RP) is a genetic ocular disease that results in substantial visual loss and in many cases leads to blindness. One measure commonly used to assess the visual function of these patients is the Humphrey 30–2 visual-field total point score. The score is a measure of central vision and is computed as a sum of visual sensitivities over 76 locations, with a higher score indicating better central vision. Normals have an average total point score of 2500 db (decibels), and the average 37-year-old RP patient has a total point score of 900 db. A total point score of  $< 250$  db is often associated with legal blindness. Longitudinal studies have indicated that the change in total point score over  $N$  years of the average RP patient is normally distributed with mean change =  $45N$  and variance of change =  $1225N$ . (Assume the total point score is measured without error; hence, no continuity correction is needed.)

**5.83** What is the probability that a patient will change by  $\geq 200$  db over 5 years?

**5.84** If a 37-year-old RP patient has an initial total point score of 900 db, what is the probability that the patient will become legally blind (that is, have a total point score of  $< 250$  db) by age 50?

Suppose a new treatment is discovered based on ocular implants. The treatment immediately lowers total point score by 50 db. However, the long-term effect is to reduce the mean rate of decline to 25 db per year (from the previous 45 db per year), while maintaining the same variance of

change as previously (that is, variance of change over  $N$  years =  $1225N$ ).

**5.85** If a 37-year-old RP patient has an initial total point score of 900 db and receives the implant treatment, what is the probability that the patient will become legally blind by age 50?

### Diabetes

Physicians recommend that children with type-I (insulin-dependent) diabetes keep up with their insulin shots to minimize the chance of long-term complications. In addition, some diabetes researchers have observed that growth rate of weight during adolescence among diabetic patients is affected by level of compliance with insulin therapy. Suppose 12-year-old type-I diabetic boys who comply with their insulin shots have a weight gain over 1 year that is normally distributed, with mean = 12 lb and variance = 12 lb.

**5.86** What is the probability that compliant type-I diabetic 12-year-old boys will gain at least 15 lb over 1 year?

Conversely, 12-year-old type-I diabetic boys who do not take their insulin shots have a weight gain over 1 year that is normally distributed with mean = 8 lb and variance = 12 lb.

**5.87** Answer the question in Problem 5.86 for noncompliant type-I diabetic 12-year-old boys.

It is generally assumed that 75% of type-I diabetics comply with their insulin regimen. Suppose that a 12-year-old type-I diabetic boy comes to clinic and shows a 5-lb weight gain over 1 year (actually, because of measurement error, assume this is an actual weight gain from 4.5 to 5.5 lb). The boy claims to be taking his insulin medication.

**5.88** What is the probability that he is telling the truth?

### Environmental Health

Some previous studies have shown that mortality rates are higher on days with high pollution levels. In a follow-up on this observation, a group of 50 nonfatal heart attack cases were ascertained over a 1-year period. For each case, the level of pollution (total suspended particulates) was measured on the day of the heart attack (index date) and also 1 month before the heart attack (control date).

The results shown in Table 5.4 were obtained:

**Table 5.4 Comparison of pollution levels on index date vs. control date**

	<i>n</i>
Pollution level on index date > pollution level on control date	30
Pollution level on control date > pollution level on index date	15
Pollution level the same on both days	5
Total	50

**5.89** Suppose the level of pollution has nothing to do with the incidence of heart attack. How many heart attacks would be expected to occur where the pollution level on the index date is higher than the pollution level on the control date? (Ignore cases where the pollution level on the index and control dates are the same.)

**5.90** Given the preceding data, assess whether pollution level acts as a trigger effect in causing heart attack. (*Hint:* Use the normal approximation to the binomial distribution.)

Researchers also analyzed cases occurring in the winter months. They found that on 10 days the pollution level on the index date was higher than on the control date, whereas on 4 days the pollution level on the control date was higher than on the index date. For 2 cases, the pollution level was the same on both days.

**5.91** Answer Problem 5.90 based on cases in winter.

### Ophthalmology

A previous study found that people consuming large quantities of vegetables containing lutein (mainly spinach) were less likely to develop macular degeneration, a common eye disease among older people (age 65+) that causes a substantial loss in visual acuity and in some cases can lead to total blindness. To follow up on this observation, a clinical trial is planned in which participants 65+ years of age without macular degeneration will be assigned to either a high-dose lutein supplement tablet or a placebo tablet taken once per day. To estimate the possible therapeutic effect, a pilot study was conducted in which 9 people 65+ years of age were randomized to placebo and 9 people 65+ years of age were randomized to lutein tablets (active treatment). Their serum lutein level was measured at baseline and again after 4 months of follow-up. From previous studies, people with serum lutein  $\geq 10$  mg/dL are expected to get some protection from macular degeneration. However, the level of serum lutein will vary depending on genetic factors, dietary factors, and study supplements.

**5.92** Suppose that among people randomized to placebo, at a 4-month follow-up mean serum lutein level = 6.4 mg/dL with standard deviation = 3 mg/dL. If we presume a normal distribution for serum lutein, then what percentage of placebo subjects will have serum lutein in the therapeutic range ( $\geq 10$  mg/dL)? (For the following problems, assume that lutein can be measured exactly, so that no continuity correction is necessary.)

**5.93** Suppose that among people randomized to lutein tablets, at a 4-month follow-up the mean serum lutein level = 21 mg/dL with standard deviation = 8 mg/dL. If we presume a normal distribution for serum-lutein values among lutein-treated participants, then what percentage of people randomized to lutein tablets will have serum lutein in the therapeutic range?

Suppose for the sake of simplicity that the incidence of macular degeneration is 1% per year among people

65+ years of age in the therapeutic range ( $\geq 10$  mg/dL) and 2% per year among people 65+ years of age with lower levels of lutein ( $< 10$  mg/dL).

**5.94** What is the expected incidence rate of macular degeneration among lutein-treated participants? (Hint: Use the total-probability rule.)

**5.95** What is the expected relative risk of macular degeneration for lutein-treated participants versus placebo-treated participants in the proposed study?

### Pediatrics

A study was recently published in Western Australia on the relationship between method of conception and prevalence of major birth defects (Hansen et al. [7]).

The prevalence of at least one major birth defect among infants conceived naturally was 4.2%, based on a large sample of infants. Among 837 infants born as a result of in-vitro fertilization (IVF), 75 had at least one major birth defect.

**5.96** How many infants with at least one birth defect would we expect among the 837 IVF infants if the true prevalence of at least one birth defect in the IVF group were the same as for infants conceived naturally?

**5.97** Do an unusual number of infants have at least one birth defect in the IVF group? Why or why not? (Hint: Use an approximation to the binomial distribution.)

In addition, data were also provided regarding specific birth defects. There were 6 chromosomal birth defects among the IVF infants. Also, the prevalence of chromosomal birth defects among infants conceived naturally is 9/4000.

**5.98** Are there an unusual number of chromosomal birth defects in the IVF group? (Hint: Use an approximation to the binomial distribution.)

### Accident Epidemiology

Automobile accidents are a frequent occurrence and one of the leading causes of morbidity and mortality among persons 18–30 years of age. The National Highway & Traffic Safety Administration (NHTSA) has estimated that the average driver in this age group has a 6.5% probability of having at least one police-reported automobile accident over the past year.

Suppose we study a group of medical interns who are on a typical hospital work schedule in which they have to work through the night for at least one of every three nights. Among 20 interns, 5 report having had an automobile accident over the past year while driving to or from work.

Suppose the interns have the same risk of having an automobile accident as a typical person ages 18–30.

**5.99** What is a reasonable probability model for the number of interns with at least one automobile accident over the past year? What are the parameters of this model?

**5.100** Apply the model in Problem 5.99 to assess whether there are an excessive number of automobile accidents

among interns compared with the average 18- to 30-year old. Explain your answer.

The study is expanded to include 50 medical interns, of whom 11 report having had an automobile accident over the past year.

One issue in the above study is that not all people report automobile accidents to the police. The NHTSA estimates that only half of all auto accidents are actually reported. Assume this rate applies to interns.

**5.101** What is an exact probability model for the number of automobile accidents over the past year for the 50 medical interns? (Note: The 11 reported accidents include both police-reported and non-police-reported accidents).

**5.102** Assess whether there are an excessive number of automobile accidents among interns under these altered assumptions. Explain your answer. (Hint: An approximation may be helpful.)

**5.103** What is the 40th percentile of a normal distribution with mean = 5 and variance = 9?

**5.104** What is the sum of the 40th and 60th percentiles of a normal distribution with a mean = 8.2 and variance = 9.5?

**5.105** What is  $z_{.90}$ ?

### Obstetrics

A study was performed of different predictors of low birth-weight deliveries among 32,520 women in the Nurses' Health Study [8].

The data in Table 5.5 were presented concerning the distribution of birthweight in the study:

**Table 5.5 Distribution of birthweight in the Nurses' Health Study**

Category	Birthweight (g)	N	%
A	< 2500	1850	5.7
B	2500–2999	6289	19.3
C	3000–3499	13,537	41.6
D	3500–3999	8572	26.4
E	4000+	2272	7.0
Total		32,520	100.0

**5.106** If 20 women are randomly chosen from the study, what is the probability that exactly 2 will have a low birth-weight delivery (defined as  $< 2500$  g)?

**5.107** What is the probability that at least 2 women will have a low birthweight delivery?

An important risk factor for low birthweight delivery is maternal smoking during pregnancy (MSMOK). The data in Table 5.6 were presented relating MSMOK to birthweight.

**Table 5.6 Association between maternal smoking and birthweight category in the Nurses' Health Study**

Category	Birthweight (g)	% MSMOK = yes
A	< 2500	40
B	2500–2999	34
C	3000–3499	25
D	3500–3999	19
E	4000+	15

**5.108** If 50 women are selected from the < 2500 g group, then what is the probability that at least half of them will have smoked during pregnancy?

**5.109** What is the probability that a woman has a low birth-weight delivery if she smokes during pregnancy? (*Hint:* Use Bayes' rule.)

### Cancer

The Shanghai Women's Health Study (SWHS) was undertaken to determine risk factor for different cancers among Asian women. The women were recruited from urban communities in 1997–2000 and were interviewed every 2 years to obtain health-related information.

One issue is whether risk prediction models derived from American populations are also applicable to Asian women.

**5.110** Suppose the expected number of breast cancer cases among a large number of 45- to 49-year-old women in this study who were followed for 7 years is 149, while the observed number of cases is 107. Is there an unusually small number of cases among Asian women? Why or why not?

Another aspect of the study is to use the SWHS data to predict the long-term incidence of breast cancer in Chinese women. Those incidence data are presented in Table 5.7.

**5.111** What is the predicted cumulative incidence of breast cancer from age 40 to 64 (i.e., over a 25-year period) among Chinese women? (Assume no deaths over this period.)

**5.112** Suppose that in the year 2000 there are 10,000,000 Chinese women age 40 years with no prior breast cancer.

**Table 5.7 Incidence rate of breast cancer by age in the SWHS**

Age	Annual incidence per $10^5$ women
40–44	63.8
45–49	86.6
50–54	92.6
55–59	107.0
60–64	120.9

What is the expected number of breast cancer cases in this group by the year 2025? (Assume no deaths over this period.)

**5.113** What is the difference between a prevalence rate of breast cancer and an incidence rate of breast cancer?

### Diabetes

The Diabetes Prevention Trial (DPT) involved a weight loss trial in which half the subjects received an active intervention and the other half a control intervention. For subjects in the active intervention group, the average reduction in body mass index (BMI, i.e., weight in kg/height<sup>2</sup> in m<sup>2</sup>) over 24 months was 1.9 kg/m<sup>2</sup>. The standard deviation of change in BMI was 6.7 kg/m<sup>2</sup>.

**5.114** If the distribution of BMI change is approximately normal, then what is the probability that a subject in the active group would lose at least 1 BMI unit over 24 months?

In the control group of the Diabetes Prevention Trial, the mean change in BMI was 0 units with a standard deviation of 6 kg/m<sup>2</sup>.

**5.115** What is the probability that a random control group participant would lose at least 1 BMI unit over 24 months?

It was known that only 70% of the subjects in the active group actually complied with the intervention; that is, 30% of subjects either dropped out or did not attend the required group and individual counseling meetings. We will refer to this latter 30% of subjects as dropouts.

**5.116** If we assume that dropouts had the same distribution of change as the subjects in the control group, then what is the probability that an active subject who complied with the intervention lost at least 1 kg/m<sup>2</sup>?

### Ophthalmology, Genetics

Age-related macular degeneration (AMD) is a common eye disease among the elderly that can lead to partial or total loss of vision. It is well known that smoking and excessive weight tend to be associated with higher incidence rates of AMD. More recently, however, several genes have been found to be associated with AMD as well. One gene that has been considered is the Y402H gene. There are three genotypes for the Y402H gene—TT, TC, and CC. The relationship between AMD and the Y402H genotype is as follows:

**Table 5.8 Association between Y402H genotype and prevalence of AMD in a high-risk population**

Y402H	AMD = yes	AMD = no
TT (wild type)	41	380
TC	119	527
CC	121	278
Total	281	1185

**5.117** What is the relative risk for AMD for the CC genotype compared with the TT genotype?

One issue is whether the *Y402H* gene is in Hardy-Weinberg equilibrium (HWE). For a gene to be in HWE, its two alleles must assort independently.

**5.118** Under HWE, what is the expected frequency of the *TC* genotype among the 1185 subjects in the AMD = no group?

**5.119** Are the data consistent with HWE? Specifically, is the number of heterozygotes (*TC*) significantly lower than expected under HWE?

### Hypertension

Blood pressure readings are known to be highly variable. Suppose we have mean SBP for one individual over  $n$  visits with  $k$  readings per visit ( $\bar{X}_{n,k}$ ). The variability of ( $\bar{X}_{n,k}$ ) depends on  $n$  and  $k$  and is given by the formula  $\sigma_w^2 = \sigma_A^2/n + \sigma^2/(nk)$ , where  $\sigma_A^2$  = between visit variability and  $\sigma^2$  = within visit variability. For 30- to 49-year-old white females,  $\sigma_A^2 = 42.9$  and  $\sigma^2 = 12.8$ . For one individual, we also assume that  $\bar{X}_{n,k}$  is normally distributed about their true long-term mean =  $\mu$  with variance =  $\sigma_w^2$ .

**5.120** Suppose a woman is measured at two visits with two readings per visit. If her true long-term SBP = 130 mm Hg, then what is the probability that her observed mean SBP is  $\geq 140$ ? (Ignore any continuity correction.) (Note: By true mean SBP we mean the average SBP over a large number of visits for that subject.)

**5.121** Suppose we want to observe the woman over  $n$  visits, where  $n$  is sufficiently large so that there is less than a 5% chance that her observed mean SBP will not differ from her true mean SBP by more than 5 mm Hg. What is the smallest value of  $n$  to achieve this goal? (Note: Assume two readings per visit.)

It is also known that over a large number of 30- to 49-year-old white women, their true mean SBP is normally distributed with mean = 120 mm Hg and standard deviation = 14 mm Hg. Also, over a large number of black 30- to 49-year-old women, their true mean SBP is normal with mean = 130 mm Hg and standard deviation = 20 mm Hg.

**5.122** Suppose we select a random 30- to 49-year-old white woman and a random 30- to 49-year-old black woman. What is the probability that the black woman has a higher true SBP?

### REFERENCES

- [1] Pratt, J. H., Jones, J. J., Miller, J. Z., Wagner, M. A., & Fineberg, N. S. (1989, October). Racial differences in aldosterone excretion and plasma aldosterone concentrations in children. *New England Journal of Medicine*, 321(17), 1152–1157.
- [2] National High Blood Pressure Group Working on High Blood Pressure in Children and Adolescents. (2004). The fourth report on the diagnosis, evaluations, and treatment of high blood pressure in children and adolescents. *Pediatrics*, 114, 555–576.
- [3] Schwartz, J., & Dockery, D. W. (1992, January). Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology*, 135(1), 12–19.
- [4] Neglia, J. F., Fitzsimmons, S. C., Maisonneuve, P., Schoni, M. H., Schoni-Affolter, F., Corey, M., Lowenfels, A. B., & the Cystic Fibrosis and Cancer Study Group. (1995). The risk of cancer among patients with cystic fibrosis. *New England Journal of Medicine*, 332, 494–499.
- [5] Hosking, D., Chilvers, C. E. D., Christiansen, C., Ravn, P., Wasnick, R., Ross, P., McClung, M., Belske, A., Thompson, D., Daley, M. T., & Yates, A. J. (1998). Prevention of bone loss with alendronate in postmenopausal women under 60 years of age. *New England Journal of Medicine*, 338, 485–492.
- [6] Fishman, L. M., Dombo, G. W., Michaelsen, C., Ringel, S., Rozbruch, J., Rosner, B., & Weber, C. (2002). Piriformis syndrome: Diagnosis, treatment, and outcome—a 10-year study. *Archives of Physical Medicine*, 83, 295–301.
- [7] Hansen, M., Kurinczuk, J. J., Bower, C., & Webb, S. (2002). The risk of major birth defects after intracytoplasmic sperm injection and in vitro fertilization. *New England Journal of Medicine*, 346(10), 725–730.
- [8] Xue, F., Willett, W. C., Rosner, B. A., Forman, M. R., & Michels, K. B. (2008). Parental characteristics as predictors of birthweight. *Human Reproduction*, 23(1), 168–177.

# 6

## Estimation

### 6.1 Introduction

Chapters 3 through 5 explored the properties of different probability models. In doing so, we always assumed the specific probability distributions were known.

**Example 6.1** **Infectious Disease** We assumed the number of neutrophils in a sample of 100 white blood cells was binomially distributed, with parameter  $p = .6$ .

**Example 6.2** **Bacteriology** We assumed the number of bacterial colonies on a 100-cm<sup>2</sup> agar plate was Poisson distributed, with parameter  $\mu = 2$ .

**Example 6.3** **Hypertension** We assumed the distribution of diastolic blood-pressure (DBP) measurements in 35- to 44-year-old men was normal, with mean  $\mu = 80$  mm Hg and standard deviation  $\sigma = 12$  mm Hg.

In general, we have been assuming that the properties of the underlying distributions from which our data are drawn are known and that the only question left is what we can predict about the behavior of the data given an understanding of these properties.

**Example 6.4** **Hypertension** Using the model in Example 6.3, we could predict that about 95% of all DBP measurements from 35- to 44-year-old men should fall between 56 and 104 mm Hg.

The problem addressed in the rest of this text, and the more basic statistical problem, is that we have a data set and we want to **infer** the properties of the underlying distribution from this data set. This inference usually involves **inductive reasoning** rather than **deductive reasoning**; that is, in principle, a variety of different probability models must at least be explored to see which model best “fits” the data.

Statistical inference can be further subdivided into the two main areas of estimation and hypothesis testing. **Estimation** is concerned with estimating the values of specific population parameters; **hypothesis testing** is concerned with testing whether the value of a population parameter is equal to some specific

**value.** Problems of estimation are covered in this chapter, and problems of hypothesis testing are discussed in Chapters 7 through 10.

Some typical problems that involve estimation follow.

**Example 6.5** **Hypertension** Suppose we measure the systolic blood pressure (SBP) of a group of Samoan villagers and we believe the underlying distribution is normal. How can the parameters of this distribution ( $\mu$ ,  $\sigma^2$ ) be estimated? How precise are our estimates?

**Example 6.6** **Infectious Disease** Suppose we look at people living within a low-income census tract in an urban area and we wish to estimate the prevalence of human immunodeficiency virus (HIV) in the community. We assume the number of cases among  $n$  people sampled is binomially distributed, with some parameter  $p$ . How is the parameter  $p$  estimated? How precise is this estimate?

In Examples 6.5 and 6.6, we were interested in obtaining specific values as estimates of our parameters. These values are often referred to as **point estimates**. Sometimes we want to specify a range within which the parameter values are likely to fall. If this range is narrow, then we may feel our point estimate is good. This type of problem involves **interval estimation**.

**Example 6.7** **Ophthalmology** An investigator proposes to screen a group of 1000 people ages 65 or older to identify those with visual impairment—that is, a visual acuity of 20/50 or worse in both eyes, even with the aid of glasses. Suppose we assume the number of people with visual impairment ascertained in this manner is binomially distributed, with parameters  $n = 1000$  and unknown  $p$ . We would like to obtain a point estimate of  $p$  and provide an interval about this point estimate to see how precise our point estimate is. For example, we would feel more confidence in a point estimate of 5% if this interval were .04–.06 than if it were .01–.10.

## 6.2 The Relationship Between Population and Sample

**Example 6.8** **Obstetrics** Suppose we want to characterize the distribution of birthweights of all liveborn infants born in the United States in 2008. Assume the underlying distribution of birthweight has an expected value (or mean)  $\mu$  and variance  $\sigma^2$ . Ideally, we wish to estimate  $\mu$  and  $\sigma^2$  exactly, based on the entire population of U.S. liveborn infants in 2008. But this task is difficult with such a large group. Instead, we decide to select a random sample of  $n$  infants who are *representative* of this large group and use the birthweights  $x_1, \dots, x_n$  from this sample to help us estimate  $\mu$  and  $\sigma^2$ . What is a random sample?

---

**Definition 6.1** A **random sample** is a selection of some members of the population such that each member is independently chosen and has a known nonzero probability of being selected.

---



---

**Definition 6.2** A **simple random sample** is a random sample in which each group member has the same probability of being selected.

---

**Definition 6.3**

The **reference, target, or study population** is the group we want to study. The random sample is selected from the study population.

For ease of discussion, we use the abbreviated term “random sample” to denote a simple random sample.

Although many samples in practice are random samples, this is not the only type of sample used in practice. A popular alternative design is **cluster sampling**.

**Example 6.9**

**Cardiovascular Disease** The Minnesota Heart Study seeks to accurately assess the prevalence and incidence of different types of cardiovascular morbidity (such as heart attack and stroke) in the greater Minneapolis-St. Paul metropolitan area, as well as trends in these rates over time. It is impossible to survey every person in the area. It is also impractical to survey, in person, a random sample of people in the area because that would entail dispersing a large number of interviewers throughout the area. Instead, the metropolitan area is divided into geographically compact regions, or clusters. A random sample of clusters is then chosen for study, and several interviewers go to each cluster selected. The primary goal is to enumerate all households in a cluster and then survey all members of these households, with the secondary goal being to identify all adults age 21 years and older. The interviewers then invite age-eligible individuals to be examined in more detail at a centrally located health site within the cluster. The total sample of all interviewed subjects throughout the metropolitan area is called a *cluster sample*. Similar strategies are also used in many national health surveys. Cluster samples require statistical methods that are beyond the scope of this book. See Cochran [1] for more discussion of cluster sampling.

In this book, we assume that all samples are random samples from a reference population.

**Example 6.10**

**Epidemiology** The Nurses’ Health Study is a large epidemiologic study involving more than 100,000 female nurses residing in 11 large states in the United States. The nurses were first contacted by mail in 1976 and since then have been followed every 2 years by mail. Suppose we want to select a sample of 100 nurses to test a new procedure for obtaining blood samples by mail. One way of selecting the sample is to assign each nurse an ID number and then select the nurses with the lowest 100 ID numbers. This is definitely *not* a random sample because each nurse is not equally likely to be chosen. Indeed, because the first two digits of the ID number are assigned according to state, the 100 nurses with the lowest ID numbers would all come from the same state. An alternative method of selecting the sample is to have a computer generate a set of 100 **random numbers** (from among the numbers 1 to over 100,000), with one number assigned to each nurse in the study. Thus each nurse is equally likely to be included in the sample. This would be a truly random sample. (More details on random numbers are given in Section 6.3.)

In practice, there is rarely an opportunity to enumerate each member of the reference population so as to select a random sample, so the researcher must assume that the sample selected has all the properties of a random sample without formally being a random sample.

In Example 6.8 the reference population is finite and well defined and can be enumerated. In many instances, however, the reference population is effectively infinite and not well defined.

**Example 6.11**

**Cancer** Suppose we want to estimate the 5-year survival rate of women who are initially diagnosed as having breast cancer at the ages of 45–54 and who undergo radical mastectomy at this time. Our reference population is all women who have ever had a first diagnosis of breast cancer when they were 45–54 years old, or whoever will have such a diagnosis in the future when they are 45–54 years old, and who receive radical mastectomies.

This population is effectively infinite. It cannot be formally enumerated, so a truly random sample cannot be selected from it. However, we again assume the sample we have selected behaves as if it were a random sample.

In this text we assume all reference populations discussed are effectively infinite, although, as in Examples 6.8 and 6.10, many are actually very large but finite. Sampling theory is the special branch of statistics that treats statistical inference for finite populations; it is beyond the scope of this text. See Cochran [1] for a good treatment of this subject.

### 6.3 Random-Number Tables

In this section, practical methods for selecting random samples are discussed.

**Example 6.12**

**Hypertension** Suppose we want to study how effective a hypertension treatment program is in controlling the blood pressure of its participants. We have a roster of all 1000 participants in the program but because of limited resources only 20 can be surveyed. We would like the 20 people chosen to be a random sample from the population of all participants in the program. How should we select this random sample?

A computer-generated list of random numbers would probably be used to select this sample.

**Definition 6.4**

A **random number** (or **random digit**) is a random variable  $X$  that takes on the values  $0, 1, 2, \dots, 9$  with equal probability. Thus

$$\Pr(X = 0) = \Pr(X = 1) = \dots = \Pr(X = 9) = \frac{1}{10}$$

**Definition 6.5**

Computer-generated **random numbers** are collections of digits that satisfy the following two properties:

- (1) Each digit  $0, 1, 2, \dots, 9$  is equally likely to occur.
- (2) The value of any particular digit is independent of the value of any other digit selected.

Table 4 in the Appendix lists 1000 random digits generated by a computer algorithm.

**Example 6.13**

Suppose 5 is a particular random digit selected. Does this mean 5's are more likely to occur in the next few digits selected?

**Solution**

No. Each digit either after or before the 5 is still equally likely to be any of the digits 0, 1, 2, . . . , 9 selected.

Computer programs generate large sequences of random digits that approximately satisfy the conditions in Definition 6.5. Thus such numbers are sometimes referred to as **pseudorandom numbers** because they are simulated to approximately satisfy the properties in Definition 6.5.

**Example 6.14**

**Hypertension** How can the random digits in Table 4 be used to select 20 random participants in the hypertension treatment program in Example 6.12?

**Solution**

A roster of the 1000 participants must be compiled, and each participant must then be assigned a number from 000 to 999. Perhaps an alphabetical list of the participants already exists, which would make this task easy. Twenty groups of three digits would then be selected, starting at any position in the random-number table. For example, if we start at the first row of Table 4 we have the numbers listed in Table 6.1.

**Table 6.1**

**Twenty random participants chosen from 1000 participants in the hypertension treatment program**

First 3 rows of random-number table				Actual random numbers chosen				
32924	22324	18125	09077	329	242	232	418	125
54632	90374	94143	49295	090	775	463	290	374
88720	43035	97081	83373	941	434	929	588	720
				430	359	708	183	373

Therefore, our random sample would consist of the people numbered 329, 242, . . . , 373 in the alphabetical list. In this particular case there were no repeats in the 20 three-digit numbers selected. If there had been repeats, then more three-digit numbers would have been selected until 20 unique numbers were selected. This process is called **random selection**.

**Example 6.15**

**Diabetes** Suppose we want to conduct a clinical trial to compare the effectiveness of an oral hypoglycemic agent for diabetes with standard insulin therapy. A small study of this type will be conducted on 10 patients: 5 patients will be randomly assigned to the oral agent and 5 to insulin therapy. How can the table of random numbers be used to make the assignments?

**Solution**

The prospective patients are numbered from 0 to 9, and five unique random digits are selected from some arbitrary position in the random-number table (e.g., from the

28th row). The first five unique digits are 6, 9, 4, 3, 7. Thus the patients numbered 3, 4, 6, 7, 9 are assigned to the oral hypoglycemic agent and the remaining patients (numbered 0, 1, 2, 5, 8) to standard insulin therapy. In some studies the prospective patients are not known in advance and are recruited over time. In this case, if 00 is identified with the 1st patient recruited, 01 with the 2nd patient recruited, . . . , and 09 with the 10th patient recruited, then the oral hypoglycemic agent would be assigned to the 4th ( $3 + 1$ ), 5th ( $4 + 1$ ), 7th ( $6 + 1$ ), 8th ( $7 + 1$ ), and 10th ( $9 + 1$ ) patients recruited and the standard therapy to the 1st ( $0 + 1$ ), 2nd ( $1 + 1$ ), 3rd ( $2 + 1$ ), 6th ( $5 + 1$ ), and 9th ( $8 + 1$ ) patients recruited.

This process is called **random assignment**. It differs from random selection (Example 6.14) in that, typically, the number, in this case of patients, to be assigned to each type of treatment (5) is fixed in advance. The random-number table helps select the 5 patients who are to receive one of the two treatments (oral hypoglycemic agent). By default, the patients not selected for the oral agent are assigned to the alternative treatment (standard insulin therapy). No additional random numbers need be chosen for the second group of 5 patients. If random selection were used instead, then one approach might be to draw a random digit for each patient. If the random digit is from 0 to 4, then the patient is assigned to the oral agent; if the random digit is from 5 to 9, then the patient is assigned to insulin therapy. One problem with this approach is that in a finite sample, equal numbers of patients are not necessarily assigned to each therapy, which is usually the most efficient design. Indeed, referring to the first 10 digits in the 28th row of the random-number table (69644 37198), we see that 4 patients would be assigned to oral therapy (patients 4, 5, 6, and 8) and 6 patients would be assigned to insulin therapy (patients 1, 2, 3, 7, 9, 10) if the method of random selection were used. Random assignment is preferable in this instance because it ensures an equal number of patients assigned to each treatment group.

### Example 6.16

**Obstetrics** The birthweights from 1000 consecutive infants delivered at Boston City Hospital (serving a low-income population) are enumerated in Table 6.2. For this example, consider this population as effectively infinite. Suppose we wish to draw 5 random samples of size 10 from this population using a computer. How can these samples be selected?

### Solution

MINITAB has a function that allows sampling from columns. The user must specify the number of rows to be sampled (the size of the random sample to be selected). Thus, if the 1000 birthweights are stored in a single column (e.g., C1), and we specify 10 rows to be sampled, then we will obtain a random sample of size 10 from this population. This random sample of size 10 can be stored in a different column (e.g., C2). This process can be repeated 5 times and results stored in 5 separate columns. It is also possible to calculate the mean  $\bar{x}$  and standard deviation ( $s$ ) for each random sample. The results are shown in Table 6.3. One issue in obtaining random samples on the computer is whether the samples are obtained with or without replacement. The default option is sampling without replacement, whereby the same data point from the population cannot be selected more than once in a specific sample. In sampling with replacement, repetitions are permissible. Table 6.3 uses sampling without replacement.

**Table 6.2** Sample of birthweights (oz) obtained from 1000 consecutive deliveries at Boston City Hospital

ID Numbers	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
000-019	116	124	119	100	127	103	140	82	107	132	100	92	76	129	138	128	115	133	70	121
020-039	114	114	121	107	120	123	83	96	116	110	71	86	136	118	120	110	107	157	89	71
040-059	98	105	106	52	123	101	111	130	129	94	124	127	128	112	83	95	118	115	86	120
060-079	106	115	100	107	131	114	121	110	115	93	116	76	138	126	143	93	121	135	81	135
080-099	108	152	127	118	110	115	109	133	116	129	118	126	137	110	32	139	132	110	140	119
100-119	109	108	103	88	87	144	105	138	115	104	129	108	92	100	145	93	115	85	124	123
120-139	141	96	146	115	124	113	98	110	153	165	140	132	79	101	127	137	129	144	126	155
140-159	120	128	119	108	113	93	144	124	89	126	87	120	99	60	115	86	143	97	106	148
160-179	113	135	117	129	120	117	92	118	80	132	121	119	57	126	126	77	135	130	102	107
180-199	115	135	112	121	89	135	127	115	133	64	91	126	78	85	106	94	122	111	109	89
200-219	99	118	104	102	94	113	124	118	104	124	133	80	117	112	112	112	102	118	107	104
220-239	90	113	132	122	89	111	118	108	148	103	112	128	86	111	140	126	143	120	124	110
240-259	142	92	132	128	97	132	99	131	120	106	115	101	130	120	130	89	107	152	90	116
260-279	106	111	120	198	123	152	135	83	107	55	131	108	100	104	112	121	102	114	102	101
280-299	118	114	112	133	139	113	77	109	142	144	114	117	97	96	93	120	149	107	107	117
300-319	93	103	121	118	110	89	127	100	156	106	122	105	92	128	124	125	118	113	110	149
320-339	98	98	141	131	92	141	110	134	90	88	111	137	67	95	102	75	108	118	99	79
340-359	110	124	122	104	133	98	108	125	106	128	132	95	114	67	134	136	138	122	103	113
360-379	142	121	125	111	97	127	117	122	120	80	114	126	103	98	108	100	106	98	116	109
380-399	98	97	129	114	102	128	107	119	84	117	119	128	121	113	128	111	112	120	122	91
400-419	117	100	108	101	144	104	110	146	117	107	126	120	104	129	147	111	106	138	97	90
420-439	120	117	94	116	119	108	109	106	134	121	125	105	177	109	109	109	79	118	92	103
440-459	110	95	111	144	130	83	93	81	116	115	131	135	116	97	108	103	134	140	72	112
460-479	101	111	129	128	108	90	113	99	103	41	129	104	144	124	70	106	118	99	85	93
480-499	100	105	104	113	106	88	102	125	132	123	160	100	128	131	49	102	110	106	96	116
500-519	128	102	124	110	129	102	101	119	101	119	141	112	100	105	155	124	67	94	134	123
520-539	92	56	17	135	141	105	133	118	117	112	87	92	104	104	132	121	118	126	114	90
540-559	109	78	117	165	127	122	108	109	119	98	120	101	96	76	143	83	100	128	124	137
560-579	90	129	89	125	131	118	72	121	91	113	91	137	110	137	111	135	105	88	112	104
580-599	102	122	144	114	120	136	144	98	108	130	119	97	142	115	129	125	109	103	114	106
600-619	109	119	89	98	104	115	99	138	122	91	161	96	138	140	32	132	108	92	118	58
620-639	158	127	121	75	112	121	140	80	125	73	115	120	85	104	95	106	100	87	99	113
640-659	95	146	126	58	64	137	69	90	104	124	120	62	83	96	126	155	133	115	97	105
660-679	117	78	105	99	123	86	126	121	109	97	131	133	121	125	120	97	101	92	111	119
680-699	117	80	145	128	140	97	126	109	113	125	157	97	119	103	102	128	116	96	109	112
700-719	67	121	116	126	106	116	77	119	119	122	109	117	127	114	102	75	88	117	99	136
720-739	127	136	103	97	130	129	128	119	22	109	145	129	96	128	122	115	102	127	109	120
740-759	111	114	115	112	146	100	106	137	48	110	97	103	104	107	123	87	140	89	112	123
760-779	130	123	125	124	135	119	78	125	103	55	69	83	106	130	98	81	92	110	112	104
780-799	118	107	117	123	138	130	100	78	146	137	114	61	132	109	133	132	120	116	133	133
800-819	86	116	101	124	126	94	93	132	126	107	98	102	135	59	137	120	119	106	125	122
820-839	101	119	97	86	105	140	89	139	74	131	118	91	98	121	102	115	115	135	100	90
840-859	110	113	136	140	129	117	117	129	143	88	105	110	123	87	97	99	128	128	110	132
860-879	78	128	126	93	148	121	95	121	127	80	109	105	136	141	103	95	140	115	118	117
880-899	114	109	144	119	127	116	103	144	117	131	74	109	117	100	103	123	93	107	113	144
900-919	99	170	97	135	115	89	120	106	141	137	107	132	132	58	113	102	120	98	104	108
920-939	85	115	108	89	88	126	122	107	68	121	113	116	94	85	93	132	146	98	132	104
940-959	102	116	108	107	121	132	105	114	107	121	101	110	137	122	102	125	104	124	121	111
960-979	101	93	93	88	72	142	118	157	121	58	92	114	104	119	91	52	110	116	100	147
980-999	114	99	123	97	79	81	146	92	126	122	72	153	97	89	100	104	124	83	81	129

**Table 6.3** Five random samples of size 10 from the population of infants whose birthweights (oz) appear in Table 6.2

Individual	Sample				
	1	2	3	4	5
1	97	177	97	101	137
2	117	198	125	114	118
3	140	107	62	79	78
4	78	99	120	120	129
5	99	104	132	115	87
6	148	121	135	117	110
7	108	148	118	106	106
8	135	133	137	86	116
9	126	126	126	110	140
10	121	115	118	119	98
$\bar{x}$	116.90	132.80	117.00	106.70	111.90
s	21.70	32.62	22.44	14.13	20.46

## 6.4 Randomized Clinical Trials

An important advance in clinical research is the acceptance of the randomized clinical trial (RCT) as the optimal study design.

### Definition 6.6

A **randomized clinical trial** is a type of research design used for comparing different treatments, in which patients are assigned to a particular treatment by some random mechanism. The process of assigning treatments to patients is called **randomization**. Randomization means the types of patients assigned to different treatment modalities will be similar if the sample sizes are large. However, if the sample sizes are small, then patient characteristics of treatment groups may not be comparable. Thus it is customary to present a table of characteristics of different treatment groups in RCTs to check that the randomization process is working well.

### Example 6.17

**Hypertension** The SHEP (Systolic Hypertension in the Elderly Program) was designed to assess the ability of antihypertensive drug treatment to reduce risk of stroke among people age 60 years or older with isolated systolic hypertension. Isolated systolic hypertension is defined as elevated SBP ( $\geq 160$  mm Hg) but normal DBP ( $< 90$  mm Hg) [2]. Of the 4736 people studied, 2365 were randomly assigned to active drug treatment and 2371 were randomly assigned to placebo. The baseline characteristics of the participants were compared by treatment group to check that the randomization achieved its goal of providing comparable groups of patients in the two treatment groups (see Table 6.4). We see the patient characteristics of the two treatment groups are generally very comparable.

The importance of randomization in modern clinical research cannot be overestimated. Before randomization, comparison of different treatments was often based on selected samples, which are often not comparable.

**Table 6.4 Baseline characteristics of randomized SHEP<sup>a</sup> participants by treatment group**

Characteristic	Active-treatment group	Placebo group	Total
Number randomized	2365	2371	4736
Age, y			
Average <sup>b</sup>	71.6 (6.7)	71.5 (6.7)	71.6 (6.7)
Percentage			
60–69	41.1	41.8	41.5
70–79	44.9	44.7	44.8
≥80	14.0	13.4	13.7
Race-sex, % <sup>c</sup>			
Black men	4.9	4.3	4.6
Black women	8.9	9.7	9.3
White men	38.8	38.4	38.6
White women	47.4	47.7	47.5
Education, y <sup>b</sup>	11.7 (3.5)	11.7 (3.4)	11.7 (3.5)
Blood pressure, mm Hg <sup>b</sup>			
Systolic	170.5 (9.5)	170.1 (9.2)	170.3 (9.4)
Diastolic	76.7 (9.7)	76.4 (9.8)	76.6 (9.7)
Antihypertensive medication at initial contact, %	33.0	33.5	33.3
Smoking, %			
Current smokers	12.6	12.9	12.7
Past smokers	36.6	37.6	37.1
Never smokers	50.8	49.6	50.2
Alcohol use, %			
Never	21.5	21.7	21.6
Formerly	9.6	10.4	10.0
Occasionally	55.2	53.9	54.5
Daily or nearly daily	13.7	14.0	13.8
History of myocardial infarction, %	4.9	4.9	4.9
History of stroke, %	1.5	1.3	1.4
History of diabetes, %	10.0	10.2	10.1
Carotid bruits, %	6.4	7.9	7.1
Pulse rate, beats/min <sup>bd</sup>	70.3 (10.5)	71.3 (10.5)	70.8 (10.5)
Body-mass index, kg/m <sup>2b</sup>	27.5 (4.9)	27.5 (5.1)	27.5 (5.0)
Serum cholesterol, mmol/L <sup>b</sup>			
Total cholesterol	6.1 (1.2)	6.1 (1.1)	6.1 (1.1)
High-density lipoprotein	1.4 (0.4)	1.4 (0.4)	1.4 (0.4)
Depressive symptoms, % <sup>e</sup>	11.1	11.0	11.1
Evidence of cognitive impairment, % <sup>f</sup>	0.3	0.5	0.4
No limitation of activities of daily living, % <sup>d</sup>	95.4	93.8	94.6
Baseline electrocardiographic abnormalities, % <sup>g</sup>	61.3	60.7	61.0

<sup>a</sup>SHEP = Systolic Hypertension in the Elderly Program.<sup>b</sup>Values are mean (sd).<sup>c</sup>Included among the whites were 204 Asians (5% of whites), 84 Hispanics (2% of whites), and 41 classified as "other" (1% of whites).<sup>d</sup>*P* < .05 for the active-treatment group compared with the placebo group.<sup>e</sup>Depressive-symptom-scale score of 7 or greater.<sup>f</sup>Cognitive-impairment-scale score of 4 or greater.<sup>g</sup>One or more of the following Minnesota codes: 1.1 to 1.3 (Q/QS), 3.1 to 3.4 (high R waves), 4.1 to 4.4 (ST depression), 5.1 to 5.4 (T wave changes), 6.1 to 6.8 (AV-conduction defects), 7.1 to 7.8 (ventricular-conduction defects), 8.1 to 8.6 (arrhythmias), and 9.1 to 9.3 and 9.5 (miscellaneous items).

**Example 6.18**

**Infectious Disease** Aminoglycosides are a type of antibiotic that are effective against certain types of gram-negative organisms. They are often given to critically ill patients (such as cancer patients, to prevent secondary infections caused by the treatment received). However, there are also side effects of aminoglycosides, including nephrotoxicity (damage to the kidney) and ototoxicity (temporary hearing loss). For several decades, studies have been performed to compare the efficacy and safety of different aminoglycosides. Many studies have compared the most common aminoglycoside, gentamicin, with other antibiotics in this class (such as tobramycin). The earliest studies were nonrandomized studies. Typically, physicians would compare outcomes for all patients treated with gentamicin in an infectious disease service over a defined period of time with outcomes for all patients treated with another aminoglycoside. No random mechanism was used to assign treatments to patients. The problem is that patients prescribed tobramycin might be sicker than patients prescribed gentamicin, especially if tobramycin is perceived as a more effective antibiotic and is “the drug of choice” for the sickest patient. Ironically, in a nonrandomized study, the more effective antibiotic might actually perform worse because this antibiotic is prescribed more often for the sickest patients. Recent clinical studies are virtually all randomized studies. Patients assigned to different antibiotics tend to be similar in randomized studies, and different types of antibiotics can be compared using comparable patient populations.

### Design Features of Randomized Clinical Trials

The actual method of randomization differs widely in different studies. Random selection, random assignment, or some other random process may be used as the method of randomization. In clinical trials, random assignment is sometimes called **block randomization**.

**Definition 6.7**

**Block randomization** is defined as follows in clinical trials comparing two treatments (treatments A and B). A block size of  $2n$  is determined in advance, where for every  $2n$  patients entering the study,  $n$  patients are randomly assigned to treatment A and the remaining  $n$  patients are assigned to treatment B. A similar approach can be used in clinical trials with more than two treatment groups. For example, if there are  $k$  treatment groups, then the block size might be  $kn$ , where for every  $kn$  patients,  $n$  patients are randomly assigned to the first treatment,  $n$  patients are randomly assigned to the second treatment, . . . ,  $n$  patients are randomly assigned to the  $k$ th treatment.

Thus with two treatment groups under block randomization, for every  $2n$  patients an equal number of patients will be assigned to each treatment. The advantage is that treatment groups will be of equal size in both the short and the long run. Because the eligibility criteria, types of patients entering a trial, or other procedures in a clinical trial sometimes change as a study progresses, this ensures comparability of treatment groups over short periods of time as the study procedures evolve. One disadvantage of blocking is that it may become evident what the randomization scheme is after a while, and physicians may defer entering patients into the study until the treatment they perceive as better is more likely to be selected. To avert this problem, a variable block size is sometimes used. For example, the block size might be 8 for the first block, 6 for the second block, 10 for the third block, and so on.

Another technique that is sometimes used in the randomization process is **stratification**.

**Definition 6.8**

In some clinical studies, patients are subdivided into subgroups, or strata, according to characteristics thought important for patient outcome. Separate randomization lists are maintained for each stratum to ensure comparable patient populations within each stratum. This procedure is called **stratification**. Either random selection (ordinary randomization) or random assignment (block randomization) might be used for each stratum. Typical characteristics used to define strata are age, sex, or overall clinical condition of the patient.

Another important advance in modern clinical research is the use of **blinding**.

**Definition 6.9**

A clinical trial is called **double blind** if neither the physician nor the patient knows what treatment he or she is getting. A clinical trial is called **single blind** if the patient is blinded as to treatment assignment but the physician is not. A clinical trial is **unblinded** if both the physician and patient are aware of the treatment assignment.

Currently, the gold standard of clinical research is the randomized double-blind study, in which patients are assigned to treatments at random and neither the patient nor the physician is aware of the treatment assignment.

**Example 6.19**

**Hypertension** The SHEP study referred to in Example 6.17 was a double-blind study. Neither the patients nor the physicians knew whether the antihypertensive medication was an active drug or a placebo. Blinding is always preferable to prevent biased reporting of outcome by the patient and/or the physician. However, it is not always feasible in all research settings.

**Example 6.20**

**Cerebrovascular Disease** Atrial fibrillation (AF) is a common symptom in the elderly, characterized by a specific type of abnormal heart rhythm. For example, former President George H. W. Bush had this condition while in office. It is well known that the risk of stroke is much higher among people with AF than for other people of comparable age and sex, particularly among the elderly. Warfarin is a drug considered effective in preventing stroke among people with AF. However, warfarin can cause bleeding complications and it is important to determine the optimal dose for each patient in order to maximize the benefit of stroke prevention while minimizing the risk of bleeding. Unfortunately, monitoring the dose requires blood tests every few weeks to assess the prothrombin time (a measure of the clot-forming capacity of blood), after which the dose may be increased, decreased, or kept the same. Because it is usually considered impractical to give control patients regular sham blood tests, the dilemma arises of selecting a good control treatment to compare with warfarin in a clinical-trial setting. In most clinical trials involving warfarin, patients are assigned at random to either warfarin or control treatment, where control is simply nontreatment. However, it is important in this setting that the people making the sometimes subjective determination of whether a stroke has occurred be blind to treatment assignment of individual patients.

Another issue with blinding is that patients may be blind to treatment assignment initially, but the nature of side effects may strongly indicate the actual treatment received.

**Example 6.21**

**Cardiovascular Disease** As part of the Physicians' Health Study, a randomized study was performed comparing aspirin with placebo in preventing cardiovascular disease. One side effect of regular intake of aspirin is gastrointestinal bleeding. The presence of this side effect strongly indicates that the type of treatment received was aspirin.

**REVIEW QUESTIONS 6A**

- 1 What is a random sample?
- 2 What is a randomized clinical trial?
- 3 Why was the use of randomization an important advance in clinical research?

## 6.5 Estimation of the Mean of a Distribution

Now that we have discussed the meaning of a random sample from a population and have explored some practical methods for selecting such samples using computer-generated random numbers, let's move on to estimation. The question remains: How is a specific random sample  $x_1, \dots, x_n$  used to estimate  $\mu$  and  $\sigma^2$ , the mean and variance of the underlying distribution? Estimating the mean is the focus of this section, and estimating the variance is covered in Section 6.7.

### Point Estimation

A natural estimator to use for estimating the population mean  $\mu$  is the sample mean

$$\bar{X} = \sum_{i=1}^n X_i/n$$

What properties of  $\bar{X}$  make it a desirable estimator of  $\mu$ ? We must forget about our particular sample for the moment and consider the set of all possible samples of size  $n$  that could have been selected from the population. The values of  $\bar{X}$  in each of these samples will, in general, be different. These values will be denoted by  $\bar{x}_1, \bar{x}_2$ , and so forth. In other words, we forget about our sample as a unique entity and consider it instead as representative of all possible samples of size  $n$  that could have been drawn from the population. Stated another way,  $\bar{x}$  is a single realization of a random variable  $\bar{X}$  over all possible samples of size  $n$  that could have been selected from the population. In the rest of this text, the symbol  $X$  denotes a random variable, and  $x$  denotes a specific realization of the random variable  $X$  in a sample.

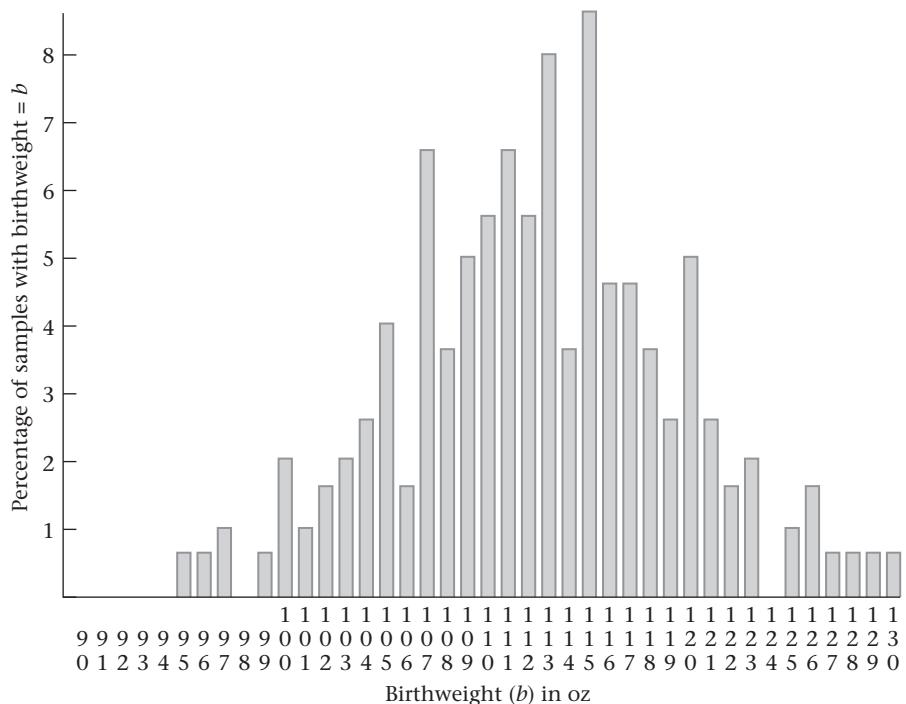
**Definition 6.10**

The **sampling distribution** of  $\bar{X}$  is the distribution of values of  $\bar{x}$  over all possible samples of size  $n$  that could have been selected from the reference population.

Figure 6.1 gives an example of such a sampling distribution. This is a frequency distribution of the sample mean from 200 randomly selected samples of size 10 drawn from the distribution of 1000 birthweights given in Table 6.2, as displayed by the Statistical Analysis System (SAS) procedure PROC CHART.

We can show that the average of these sample means ( $\bar{x}$ 's), when taken over a large number of random samples of size  $n$ , approximates  $\mu$  as the number of samples selected becomes large. In other words, the expected value of  $\bar{X}$  over its sampling distribution is equal to  $\mu$ . This result is summarized as follows:

**Figure 6.1** Sampling distribution of  $\bar{X}$  over 200 samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0–100.9, etc.)



### Equation 6.1

Let  $X_1, \dots, X_n$  be a random sample drawn from some population with mean  $\mu$ .

Then for the sample mean  $\bar{X}$ ,  $E(\bar{X}) = \mu$ .

Note that Equation 6.1 holds for any population regardless of its underlying distribution. In words, we refer to  $\bar{X}$  as an unbiased estimator of  $\mu$ .

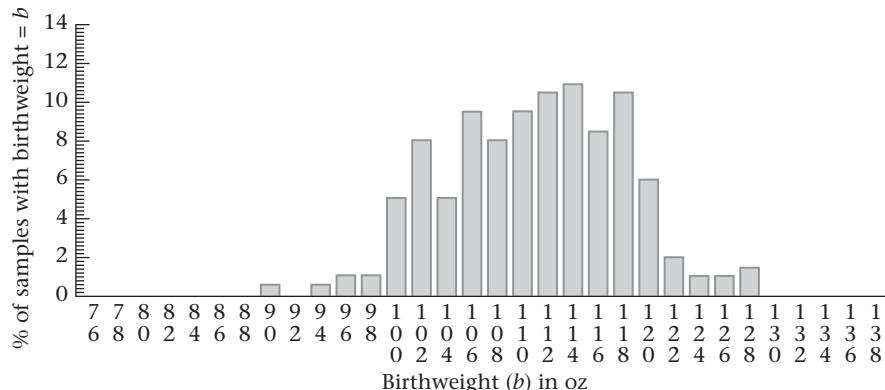
### Definition 6.11

We refer to an estimator of a parameter  $\theta$  as  $\hat{\theta}$ . An **estimator**  $\hat{\theta}$  of a parameter  $\theta$  is **unbiased** if  $E(\hat{\theta}) = \theta$ . This means that the average value of  $\hat{\theta}$  over a large number of repeated samples of size  $n$  is  $\theta$ .

The unbiasedness of  $\bar{X}$  is not sufficient reason to use it as an estimator of  $\mu$ . For symmetric distributions, many unbiased estimators of  $\mu$  exist, including the sample median and the average value of the largest and smallest data points in a sample. Why is  $\bar{X}$  chosen rather than any of the other unbiased estimators? The reason is that if the underlying distribution of the population is normal, then it can be shown that the unbiased estimator with the smallest variance is given by  $\bar{X}$ . Thus  $\bar{X}$  is called the **minimum variance unbiased estimator** of  $\mu$ .

This concept is illustrated in Figure 6.2, where for 200 random samples of size 10 drawn from the population of 1000 birthweights in Table 6.2, the sampling distribution of the sample mean ( $\bar{X}$ ) is plotted in Figure 6.2a, the sample median in Figure 6.2b, and the average of the smallest and largest observations in the sample in Figure 6.2c. Note that the variability of the distribution of sample means is slightly smaller than that of the sample median and considerably smaller than that of the average of the smallest and largest observations.

**Figure 6.2 Sampling distributions of several estimators of  $\mu$  for 200 random samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0–101.9, etc.)**



**Example 6.22**

**Obstetrics** Consider Table 6.3 (p. 156). Notice that the 50 individual birthweights range from 62 to 198 oz and have a sample standard deviation of 23.79 oz. The five sample means range from 106.7 to 132.8 oz and have a sample standard deviation of 9.77 oz. Thus the sample means based on 10 observations are less variable from sample to sample than are the individual observations, which can be considered as sample means from samples of size 1.

Indeed, we would expect the sample means from repeated samples of size 100 to be less variable than those from samples of size 10. We can show this is true. Using the properties of linear combinations of independent random variables given in Equation 5.9,

$$\begin{aligned} \text{Var}(\bar{X}) &= \left(\frac{1}{n^2}\right) \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \left(\frac{1}{n^2}\right) \sum_{i=1}^n \text{Var}(X_i) \end{aligned}$$

However, by definition  $\text{Var}(X_i) = \sigma^2$ . Therefore,

$$\text{Var}(\bar{X}) = \left(1/n^2\right)(\sigma^2 + \sigma^2 + \dots + \sigma^2) = \left(1/n^2\right)(n\sigma^2) = \sigma^2/n$$

The standard deviation ( $sd$ ) =  $\sqrt{\text{variance}}$ ; thus,  $sd(\bar{X}) = \sigma/\sqrt{n}$ . We have the following summary:

**Equation 6.2**

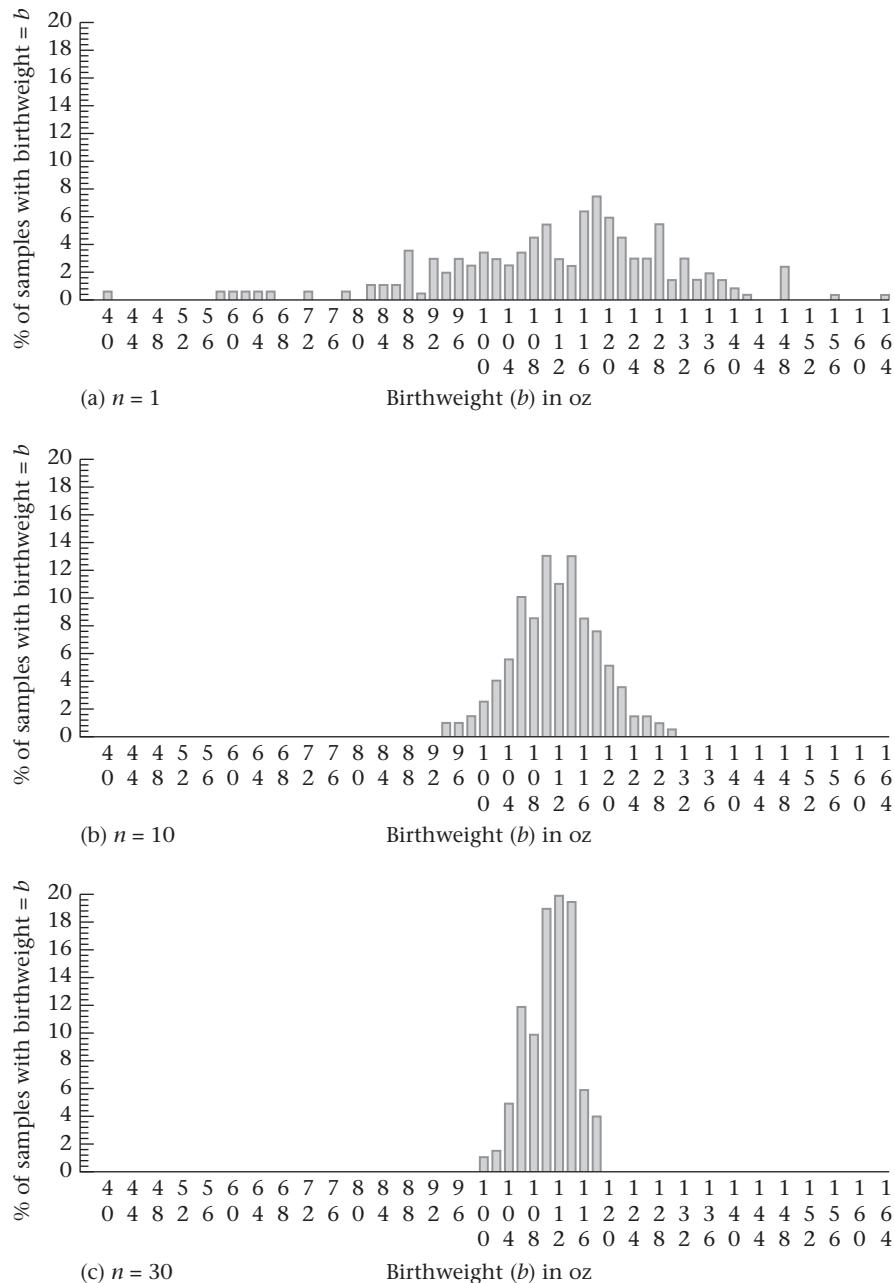
Let  $X_1, \dots, X_n$  be a random sample from a population with underlying mean  $\mu$  and variance  $\sigma^2$ . The set of sample means in repeated random samples of size  $n$  from this population has variance  $\sigma^2/n$ . The standard deviation of this set of sample means is thus  $\sigma/\sqrt{n}$  and is referred to as the *standard error of the mean* or the *standard error*.

In practice, the population variance  $\sigma^2$  is rarely known. We will see in Section 6.7 that a reasonable estimator for the population variance  $\sigma^2$  is the sample variance  $s^2$ , which leads to the following definition:

**Definition 6.12**

The **standard error of the mean (sem)**, or the **standard error (se)**, is given by  $\sigma/\sqrt{n}$  and is estimated by  $s/\sqrt{n}$ . The standard error represents the estimated standard deviation obtained from a set of sample means from repeated samples of size  $n$  from a population with underlying variance  $\sigma^2$ .

Note that the standard error is *not* the standard deviation of an individual observation  $X_i$  but rather of the sample mean  $\bar{X}$ . The standard error of the mean is illustrated in Figure 6.3. In Figure 6.3a, the frequency distribution of the sample mean is plotted for 200 samples of size 1 drawn from the collection of birthweights in Table 6.2. Similar frequency distributions are plotted for 200 sample means from samples of size 10 in Figure 6.3b and from samples of size 30 in Figure 6.3c. Notice that the spread of the frequency distribution in Figure 6.3a, corresponding to  $n = 1$ , is much larger than the spread of the frequency distribution in Figure 6.3b, corresponding to  $n = 10$ . Furthermore, the spread of the frequency distribution in Figure 6.3b, corresponding to  $n = 10$ , is much larger than the spread of the frequency distribution in Figure 6.3c, corresponding to  $n = 30$ .

**Figure 6.3** Illustration of the standard error of the mean ( $100 = 100.0 - 103.9$ , etc.)

**Example 6.23** **Obstetrics** Compute the standard error of the mean for the third sample of birthweights in Table 6.3 (p. 156).

**Solution**

The standard error of the mean is given by

$$s/\sqrt{n} = 22.44/\sqrt{10} = 7.09$$

The standard error is a quantitative measure of the variability of sample means obtained from repeated random samples of size  $n$  drawn from the same population. Notice that the standard error is directly proportional to both  $1/\sqrt{n}$  and to the

population standard deviation  $\sigma$  of individual observations. It justifies the concern with sample size in assessing the precision of our estimate  $\bar{x}$  of the unknown population mean  $\mu$ . The reason it is preferable to estimate  $\mu$  from a sample of size 400 rather than from one of size 100 is that the standard error from the first sample will be half as large as in the second sample. Thus the larger sample should provide a more precise estimate of  $\mu$ . Notice that the precision of our estimate is also affected by the underlying variance  $\sigma^2$  from the population of individual observations, a quantity that is unrelated to the sample size  $n$ . However,  $\sigma^2$  can sometimes be affected by experimental technique. For example, in measuring blood pressure,  $\sigma^2$  can be reduced by better standardization of blood-pressure observers and/or by using additional replicates for individual subjects (for example, using an average of two blood-pressure readings for each subject rather than a single reading).

**Example 6.24**

**Gynecology** Suppose a woman wants to estimate her exact day of ovulation for contraceptive purposes. A theory exists that at the time of ovulation the body temperature rises 0.5 to 1.0°F. Thus changes in body temperature can be used to guess the day of ovulation.

To use this method, we need a good estimate of basal body temperature during a period when ovulation is definitely not occurring. Suppose that for this purpose a woman measures her body temperature on awakening on the first 10 days after menstruation and obtains the following data: 97.2°, 96.8°, 97.4°, 97.4°, 97.3°, 97.0°, 97.1°, 97.3°, 97.2°, 97.3°. What is the best estimate of her underlying basal body temperature ( $\mu$ )? How precise is this estimate?

**Solution**

The best estimate of her underlying body temperature during the nonovulation period ( $\mu$ ) is given by

$$\bar{x} = (97.2 + 96.8 + \dots + 97.3)/10 = 97.20^\circ$$

The standard error of this estimate is given by

$$s/\sqrt{10} = 0.189/\sqrt{10} = 0.06^\circ$$

In our work on confidence intervals (CIs) in this section (p. 168), we show that for many underlying distributions, we can be fairly certain the true mean  $\mu$  is approximately within two standard errors of  $\bar{x}$ . In this case, true mean basal body temperature ( $\mu$ ) is within  $97.20^\circ \pm 2(0.06)^\circ \approx (97.1^\circ - 97.3^\circ)$ . Thus if the temperature is elevated by at least 0.5° above this range on a given day, then it may indicate the woman was ovulating and, for contraceptive purposes, should not have intercourse on that day.

**REVIEW QUESTIONS 6B**

- 1 What is a sampling distribution?
- 2 Why is the sample mean  $\bar{X}$  used to estimate the population mean  $\mu$ ?
- 3 What is the difference between a standard deviation and a standard error?
- 4 Suppose we have a sample of five values of hemoglobin A1c (HgbA1c) obtained from a single diabetic patient. HgbA1c is a serum measure often used to monitor compliance among diabetic patients. The values are 8.5%, 9.3%, 7.9%, 9.2%, and 10.3%.
  - (a) What is the standard deviation for this sample?
  - (b) What is the standard error for this sample?

- 5** Suppose the number of values from the patient in Review Question 6B.4 increases from 5 to 20.
- Would you expect the standard deviation to increase, decrease, or remain the same? Why?
  - Would you expect the standard error to increase, decrease, or remain the same? Why?

## Central-Limit Theorem

If the underlying distribution is normal, then it can be shown that the sample mean is itself normally distributed with mean  $\mu$  and variance  $\sigma^2/n$  (see Section 5.6). In other words,  $\bar{X} \sim N(\mu, \sigma^2/n)$ . If the underlying distribution is *not* normal, we would still like to make some statement about the sampling distribution of the sample mean. This statement is given by the following theorem:

### Equation 6.3

#### Central-Limit Theorem

Let  $X_1, \dots, X_n$  be a random sample from some population with mean  $\mu$  and variance  $\sigma^2$ . Then for large  $n$ ,  $\bar{X} \sim N(\mu, \sigma^2/n)$  even if the underlying distribution of individual observations in the population is not normal. (The symbol  $\sim$  is used to represent “approximately distributed.”)

This theorem is very important because many of the distributions encountered in practice are not normal. In such cases the central-limit theorem can often be applied; this lets us perform statistical inference based on the approximate normality of the sample mean despite the nonnormality of the distribution of individual observations.

### Example 6.25

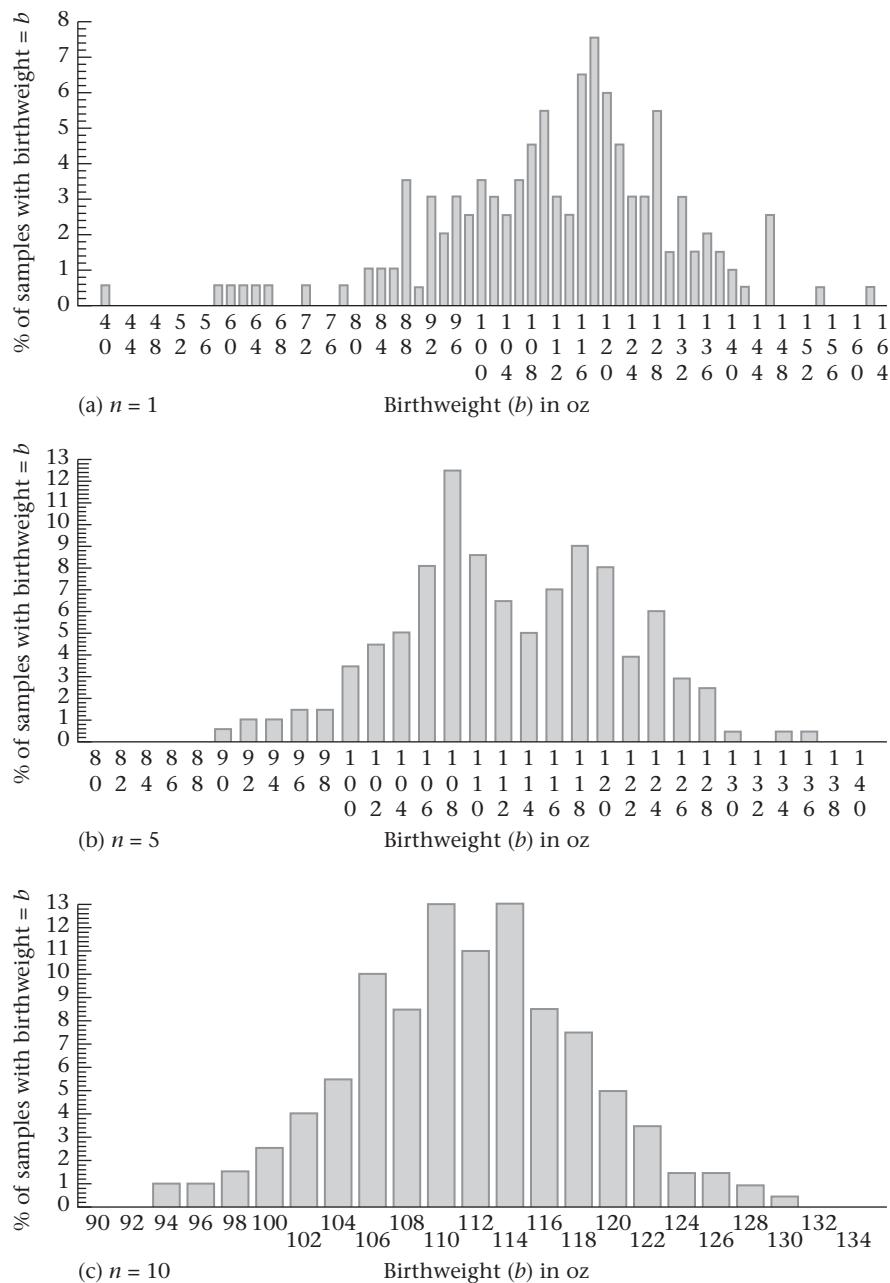
**Obstetrics** The central-limit theorem is illustrated by plotting, in Figure 6.4a, the sampling distribution of mean birthweights obtained by drawing 200 random samples of size 1 from the collection of birthweights in Table 6.2. Similar sampling distributions of sample means are plotted from samples of size 5, in Figure 6.4b, and samples of size 10, in Figure 6.4c. Notice that the distribution of individual birthweights (i.e., sample means from samples of size 1) is slightly skewed to the left. However, the distribution of sample means becomes increasingly closer to bell-shaped as the sample size increases to 5 and 10.

### Example 6.26

**Cardiovascular Disease** Serum triglycerides are an important risk factor for certain types of coronary disease. Their distribution tends to be positively skewed, or skewed to the right, with a few people with very high values, as is shown in Figure 6.5. However, hypothesis tests can be performed based on mean serum triglycerides over moderate samples of people because from the central-limit theorem the distribution of means will be approximately normal, even if the underlying distribution of individual measurements is not. To further ensure normality, the data can also be transformed onto a different scale. For example, if a log transformation is used, then the skewness of the distribution is reduced and the central-limit theorem will be applicable for smaller sample sizes than if the data are kept in the original scale.

### Example 6.27

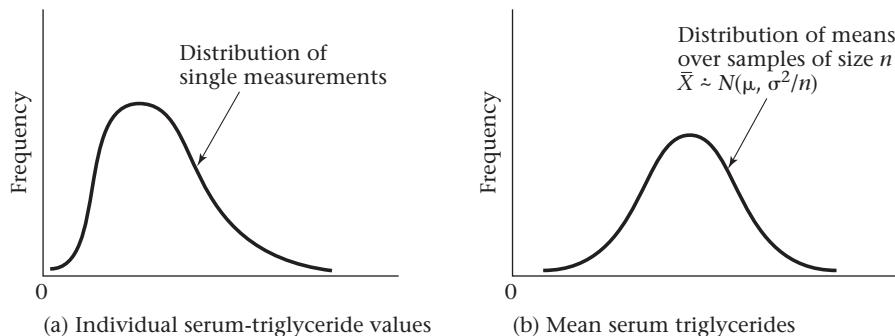
**Obstetrics** Compute the probability that the mean birthweight from a sample of 10 infants from the Boston City Hospital population in Table 6.2 will fall between 98.0 and 126.0 oz (i.e.,  $98 \leq \bar{X} < 126$ ) if the mean birthweight for the 1000 birthweights from the Boston City Hospital population is 112.0 oz with a standard deviation of 20.6 oz.

**Figure 6.4** Illustration of the central-limit theorem:  $100 = 100 - 101.9$ **Solution**

The central-limit theorem is applied, and we assume  $\bar{X}$  follows a normal distribution with mean  $\mu = 112.0$  oz and standard deviation  $\sigma/\sqrt{n} = 20.6/\sqrt{10} = 6.51$  oz. It follows that

$$\begin{aligned}
 Pr(98.0 \leq \bar{X} < 126.0) &= \Phi\left(\frac{126.0 - 112.0}{6.51}\right) - \Phi\left(\frac{98.0 - 112.0}{6.51}\right) \\
 &= \Phi(2.15) - \Phi(-2.15) \\
 &= \Phi(2.15) - [1 - \Phi(2.15)] = 2\Phi(2.15) - 1
 \end{aligned}$$

**Figure 6.5 Distribution of single serum-triglyceride measurements and of means of such measurements over samples of size  $n$**



Refer to Table 3 in the Appendix and obtain

$$Pr(98.0 \leq \bar{X} < 126.0) = 2(.9842) - 1.0 = .968$$

Thus if the central-limit theorem holds, 96.8% of the samples of size 10 would be expected to have mean birthweights between 98 and 126 oz. This value can be checked by referring to Figure 6.2a. Note that the 90 column corresponds to the birthweight interval 90.0–91.9, the 92 column to 92.0–93.9, and so forth. Note that 0.5% of the birthweights are in the 90 column, 0.5% in the 94 column, 1% in the 96 column, 1% in the 126 column, and 1.5% in the 128 column. Thus 2% of the distribution is less than 98.0 oz, and 2.5% of the distribution is 126.0 oz or greater. It follows that  $100\% - 4.5\% = 95.5\%$  of the distribution is actually between 98 and 126 oz. This value corresponds well to the 96.8% predicted by the central-limit theorem, confirming that the central-limit theorem holds approximately for averages from samples of size 10 drawn from this population.

### Interval Estimation

We have been discussing the rationale for using  $\bar{x}$  to estimate the mean of a distribution and have given a measure of variability of this estimate, namely, the standard error. These statements hold for any underlying distribution. However, we frequently wish to obtain an interval of plausible estimates of the mean as well as a best estimate of its precise value. Our interval estimates will hold exactly if the underlying distribution is normal and only approximately if the underlying distribution is not normal, as stated in the central-limit theorem.

#### Example 6.28

**Obstetrics** Suppose the first sample of 10 birthweights given in Table 6.3 has been drawn. Our best estimate of the population mean  $\mu$  would be the sample mean  $\bar{x} = 116.9$  oz. Although 116.9 oz is our best estimate of  $\mu$ , we still are not certain that  $\mu$  is 116.9 oz. Indeed, if the second sample of 10 birthweights had been drawn, a point estimate of 132.8 oz would have been obtained. Our point estimate would certainly have a different meaning if it was highly likely that  $\mu$  was within 1 oz of 116.9 rather than within 1 lb (16 oz).

We have assumed previously that the distribution of birthweights in Table 6.2 was normal with mean  $\mu$  and variance  $\sigma^2$ . It follows from our previous discussion of

the properties of the sample mean that  $\bar{X} \sim N(\mu, \sigma^2/n)$ . Thus, if  $\mu$  and  $\sigma^2$  were known, then the behavior of the set of sample means over a large number of samples of size  $n$  would be precisely known. In particular, 95% of all such sample means will fall within the interval  $(\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n})$ .

**Equation 6.4**

Alternatively, if we re-express  $\bar{X}$  in standardized form by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

then  $Z$  should follow a standard normal distribution. Hence, 95% of the  $Z$  values from repeated samples of size  $n$  will fall between  $-1.96$  and  $+1.96$  because these values correspond to the 2.5th and 97.5th percentiles from a standard normal distribution. However, the assumption that  $\sigma$  is known is somewhat artificial, because  $\sigma$  is rarely known in practice.

## *t* Distribution

Because  $\sigma$  is unknown, it is reasonable to estimate  $\sigma$  by the sample standard deviation  $s$  and to try to construct CIs using the quantity  $(\bar{X} - \mu)/(s/\sqrt{n})$ . The problem is that this quantity is no longer normally distributed.

This problem was first solved in 1908 by a statistician named William Gossett. For his entire professional life, Gossett worked for the Guinness Brewery in Ireland. He chose to identify himself by the pseudonym "Student," and thus the distribution of  $(\bar{X} - \mu)/(s/\sqrt{n})$  is usually referred to as **Student's *t* distribution**. Gossett found that the shape of the distribution depends on the sample size  $n$ . Thus the *t* distribution is not a unique distribution but is instead a family of distributions indexed by a parameter referred to as the **degrees of freedom (*df*)** of the distribution.

**Equation 6.5**

If  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$  and are independent, then  $(\bar{X} - \mu)/(s/\sqrt{n})$  is distributed as a *t* distribution with  $(n - 1)$  *df*.

Once again, Student's *t* distribution is not a unique distribution but is a family of distributions indexed by the degrees of freedom  $d$ . The *t* distribution with  $d$  degrees of freedom is sometimes referred to as the  $t_d$  distribution.

**Definition 6.13**

The  $100 \times u$ th percentile of a *t* distribution with  $d$  degrees of freedom is denoted by  $t_{d,u}$ , that is,

$$Pr(t_d < t_{d,u}) \equiv u$$

**Example 6.29**

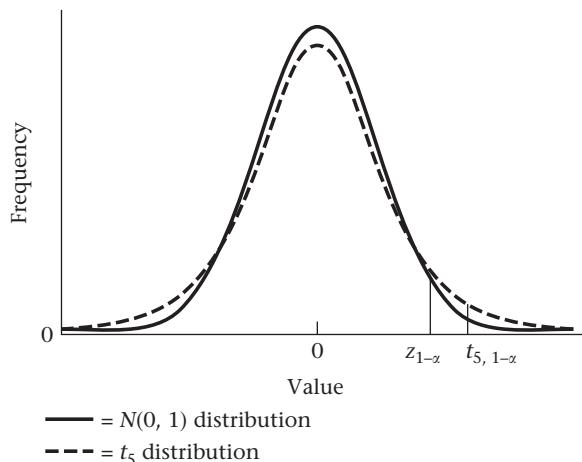
What does  $t_{20,.95}$  mean?

**Solution**

$t_{20,.95}$  is the 95th percentile or the upper 5th percentile of a *t* distribution with 20 degrees of freedom.

It is interesting to compare a *t* distribution with  $d$  degrees of freedom with an  $N(0, 1)$  distribution. The density functions corresponding to these distributions are depicted in Figure 6.6 for the special case where  $d = 5$ .

Notice that the *t* distribution is symmetric about 0 but is more spread out than the  $N(0, 1)$  distribution. It can be shown that for any  $\alpha$ , where  $\alpha > .5$ ,  $t_{d,1-\alpha}$  is always

**Figure 6.6** Comparison of Student's  $t$  distribution with 5 degrees of freedom with an  $N(0, 1)$  distribution

larger than the corresponding percentile for an  $N(0, 1)$  distribution ( $z_{1-\alpha}$ ). This relationship is shown in Figure 6.6. However, as  $d$  becomes large, the  $t$  distribution converges to an  $N(0, 1)$  distribution. An explanation for this principle is that for finite samples the sample variance ( $s^2$ ) is an approximation to the population variance ( $\sigma^2$ ). This approximation gives the statistic  $(\bar{X} - \mu)/(S/\sqrt{n})$  more variability than the corresponding statistic  $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ . As  $n$  becomes large, this approximation gets better and  $S^2$  will converge to  $\sigma^2$ . The two distributions thus get more and more alike as  $n$  increases in size. The upper 2.5th percentile of the  $t$  distribution for various degrees of freedom and the corresponding percentile for the normal distribution are given in Table 6.5.

**Table 6.5** Comparison of the 97.5th percentile of the  $t$  distribution and the normal distribution

$d$	$t_{d, .975}$	$z_{.975}$	$d$	$t_{d, .975}$	$z_{.975}$
4	2.776	1.960	60	2.000	1.960
9	2.262	1.960	$\infty$	1.960	1.960
29	2.045	1.960			

The difference between the  $t$  distribution and the normal distribution is greatest for small values of  $n$  ( $n < 30$ ). Table 5 in the Appendix gives the percentage points of the  $t$  distribution for various degrees of freedom. The degrees of freedom are given in the first column of the table, and the percentiles are given across the first row. The  $u$ th percentile of a  $t$  distribution with  $d$  degrees of freedom is found by reading across the row marked  $d$  and reading down the column marked  $u$ .

**Example 6.30** Find the upper 5th percentile of a  $t$  distribution with 23  $df$ .

**Solution** Find  $t_{23, .95}$ , which is given in row 23 and column .95 of Appendix Table 5 and is 1.714.

Statistical packages such as MINITAB, Excel, SAS, or Stata will also compute exact probabilities associated with the  $t$  distribution. This is particularly useful for values of the degrees of freedom ( $d$ ) that are not given in Table 5.

If  $\sigma$  is unknown, we can replace  $\sigma$  by  $S$  in Equation 6.4 and correspondingly replace the  $z$  statistic by a  $t$  statistic given by

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

The  $t$  statistic should follow a  $t$  distribution with  $n - 1$  df. Hence, 95% of the  $t$  statistics in repeated samples of size  $n$  should fall between the 2.5th and 97.5th percentiles of a  $t_{n-1}$  distribution, or

$$\Pr(t_{n-1,0.025} < t < t_{n-1,0.975}) = 0.95$$

More generally,  $100\% \times (1 - \alpha)$  of the  $t$  statistics should fall between the lower and upper  $\alpha/2$  percentile of a  $t_{n-1}$  distribution in repeated samples of size  $n$  or

$$\Pr(t_{n-1,\alpha/2} < t < t_{n-1,1-\alpha/2}) = 1 - \alpha$$

This inequality can be written in the form of two inequalities:

$$t_{n-1,\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad \text{and} \quad \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1,1-\alpha/2}$$

If we multiply both sides of each inequality by  $(S/\sqrt{n})$  and add  $\mu$  to both sides, we obtain

$$\mu + t_{n-1,\alpha/2} S/\sqrt{n} < \bar{X} \quad \text{and} \quad \bar{X} < t_{n-1,1-\alpha/2} S/\sqrt{n} + \mu$$

Finally, if we subtract  $t_{n-1,\alpha/2} S/\sqrt{n}$  from both sides of the first inequality and  $t_{n-1,1-\alpha/2} S/\sqrt{n}$  from both sides of the second inequality, we get

$$\mu < \bar{X} - t_{n-1,\alpha/2} S/\sqrt{n} \quad \text{and} \quad \bar{X} - t_{n-1,1-\alpha/2} S/\sqrt{n} < \mu$$

Expressed as one inequality, this is

$$\bar{X} - t_{n-1,1-\alpha/2} S/\sqrt{n} < \mu < \bar{X} - t_{n-1,\alpha/2} S/\sqrt{n}$$

From the symmetry of the  $t$  distribution,  $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$ , so this inequality can be rewritten as

$$\bar{X} - t_{n-1,1-\alpha/2} S/\sqrt{n} < \mu < \bar{X} + t_{n-1,1-\alpha/2} S/\sqrt{n}$$

and we can say that

$$\Pr(\bar{X} - t_{n-1,1-\alpha/2} S/\sqrt{n} < \mu < \bar{X} + t_{n-1,1-\alpha/2} S/\sqrt{n}) = 1 - \alpha$$

The interval  $(\bar{x} - t_{n-1,1-\alpha/2} s/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2} s/\sqrt{n})$  is referred to as a  $100\% \times (1 - \alpha)$  CI for  $\mu$ . This can be summarized as follows:

### Equation 6.6

#### Confidence Interval for the Mean of a Normal Distribution

A  $100\% \times (1 - \alpha)$  CI for the mean  $\mu$  of a normal distribution with unknown variance is given by

$$(\bar{x} - t_{n-1,1-\alpha/2} s/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2} s/\sqrt{n})$$

A shorthand notation for the CI is

$$\bar{x} \pm t_{n-1,1-\alpha/2} s/\sqrt{n}$$

**Example 6.31**

Compute a 95% CI for the mean birthweight based on the first sample of size 10 in Table 6.3.

**Solution**

We have  $n = 10$ ,  $\bar{x} = 116.90$ ,  $s = 21.70$ . Because we want a 95% CI,  $\alpha = .05$ . Therefore, from Equation 6.6 the 95% CI is

$$\left[ 116.9 - t_{9,.975}(21.70)/\sqrt{10}, 116.9 + t_{9,.975}(21.70)/\sqrt{10} \right]$$

From Table 5,  $t_{9,.975} = 2.262$ . Therefore, the 95% CI is

$$\begin{aligned} & \left[ 116.9 - 2.262(21.70)/\sqrt{10}, 116.9 + 2.262(21.70)/\sqrt{10} \right] \\ &= (116.9 - 15.5, 116.9 + 15.5) \\ &= (101.4, 132.4) \end{aligned}$$

Note that if the sample size is large (say  $>200$ ), then the percentiles of a  $t$  distribution are virtually the same as for a normal distribution. In this case, a reasonable approximate  $100\% \times (1 - \alpha)$  CI for  $\mu$  is given as follows:

**Equation 6.7****Confidence Interval for the Mean of a Normal Distribution (Large-Sample Case)**

An approximate  $100\% \times (1 - \alpha)$  CI for the mean  $\mu$  of a normal distribution with unknown variance is given by

$$(\bar{x} - z_{1-\alpha/2} s/\sqrt{n}, \bar{x} + z_{1-\alpha/2} s/\sqrt{n})$$

This interval should only be used if  $n > 200$ . In addition, Equation 6.7 can also be used for  $n \leq 200$  if the standard deviation ( $\sigma$ ) is known, by replacing  $s$  with  $\sigma$ .

You may be puzzled at this point as to what a CI is. The parameter  $\mu$  is a fixed unknown constant. How can we state that the probability that it lies within some specific interval is, for example, 95%? The important point to understand is that the boundaries of the interval depend on the sample mean and sample variance and vary from sample to sample. Furthermore, 95% of such intervals that could be constructed from repeated random samples of size  $n$  contain the parameter  $\mu$ .

**Example 6.32**

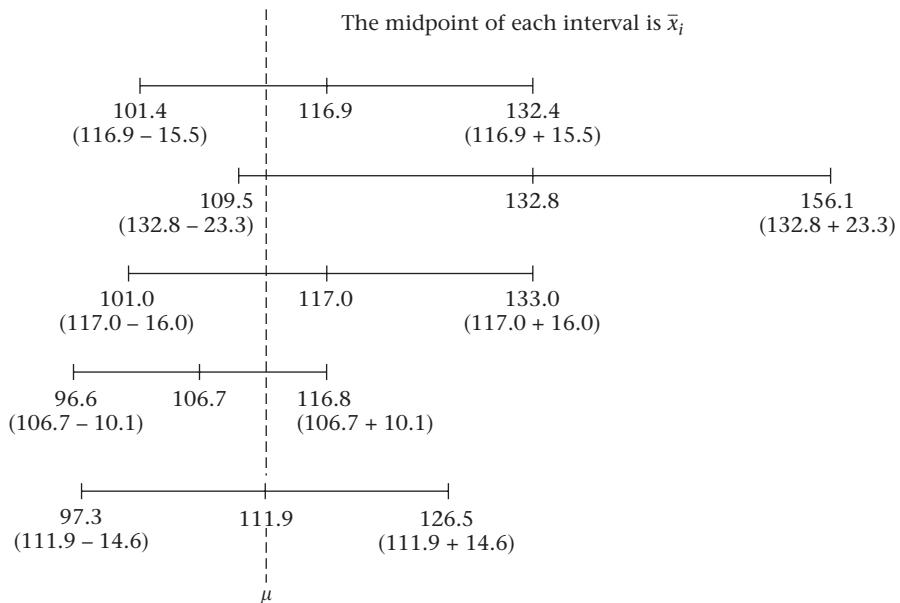
**Obstetrics** Consider the five samples of size 10 from the population of birthweights as shown in Table 6.3 (p. 156). Because  $t_{9,.975} = 2.262$ , the 95% CI is given by

$$\begin{aligned} (\bar{x} - t_{9,.975} s/\sqrt{n}, \bar{x} + t_{9,.975} s/\sqrt{n}) &= \left( \bar{x} - \frac{2.262s}{\sqrt{10}}, \bar{x} + \frac{2.262s}{\sqrt{10}} \right) \\ &= (\bar{x} - 0.715s, \bar{x} + 0.715s) \end{aligned}$$

The interval is different for each sample and is given in Figure 6.7. A dashed line has been added to represent an imaginary value for  $\mu$ . The idea is that over a large number of hypothetical samples of size 10, 95% of such intervals contain the parameter  $\mu$ . Any one interval from a particular sample *may* or *may not* contain the parameter  $\mu$ . In Figure 6.7, by chance all five intervals contain the parameter  $\mu$ . However, with additional random samples this need not be the case.

Therefore, we cannot say there is a 95% chance that the parameter  $\mu$  will fall within a particular 95% CI. However, we can say the following:

**Figure 6.7** A collection of 95% CIs for the mean  $\mu$  as computed from repeated samples of size 10 (see Table 6.3) from the population of birthweights given in Table 6.2



### Equation 6.8

Over the collection of all 95% CIs that could be constructed from repeated random samples of size  $n$ , 95% will contain the parameter  $\mu$ .

The length of the CI gives some idea of the precision of the point estimate  $\bar{x}$ . In this particular case, the length of each CI ranges from 20 to 47 oz, which makes the precision of the point estimate  $\bar{x}$  doubtful and implies that a larger sample size is needed to get a more precise estimate of  $\mu$ .

### Example 6.33

**Gynecology** Compute a 95% CI for the underlying mean basal body temperature using the data in Example 6.24 (p. 165).

#### Solution

The 95% CI is given by

$$\begin{aligned}\bar{x} \pm t_{9,975} s / \sqrt{n} &= 97.2^\circ \pm 2.262(0.189) / \sqrt{10} = 97.2^\circ \pm 0.13^\circ \\ &= (97.07^\circ, 97.33^\circ)\end{aligned}$$

We can also consider CIs with a level of confidence other than 95%.

### Example 6.34

Suppose the first sample in Table 6.3 has been drawn. Compute a 99% CI for the underlying mean birthweight.

#### Solution

The 99% CI is given by

$$(116.9 - t_{9,995}(21.70) / \sqrt{10}, 116.9 + t_{9,995}(21.70) / \sqrt{10})$$

From Table 5 of the Appendix we see that  $t_{9,995} = 3.250$ , and therefore the 99% CI is

$$(116.9 - 3.250(21.70) / \sqrt{10}, 116.9 + 3.250(21.70) / \sqrt{10}) = (94.6, 139.2)$$

Notice that the 99% CI (94.6, 139.2) computed in Example 6.34 is wider than the corresponding 95% CI (101.4, 132.4) computed in Example 6.31. The rationale for this difference is that the higher the level of confidence desired that  $\mu$  lies within an interval, the wider the CI must be. Indeed, for 95% CIs the length was  $2(2.262)s/\sqrt{n}$ ; for 99% CIs, the length was  $2(3.250)s/\sqrt{n}$ . In general, the length of the  $100\% \times (1 - \alpha)$  CI is given by

$$2t_{n-1,1-\alpha/2} s/\sqrt{n}$$

Therefore, we can see the length of a CI is governed by three variables:  $n$ ,  $s$ , and  $\alpha$ .

### Equation 6.9

#### Factors Affecting the Length of a CI

The length of a  $100\% \times (1 - \alpha)$  CI for  $\mu$  equals  $2t_{n-1,1-\alpha/2} s/\sqrt{n}$  and is determined by  $n$ ,  $s$ , and  $\alpha$ .

- $n$  As the sample size ( $n$ ) increases, the length of the CI decreases.
- $s$  As the standard deviation ( $s$ ), which reflects the variability of the distribution of individual observations, increases, the length of the CI increases.
- $\alpha$  As the confidence desired increases ( $\alpha$  decreases), the length of the CI increases.

### Example 6.35

**Gynecology** Compute a 95% CI for the underlying mean basal body temperature using the data in Example 6.24, assuming that the number of days sampled is 100 rather than 10.

#### Solution

The 95% CI is given by

$$\begin{aligned} 97.2^\circ \pm t_{99,975}(0.189)/\sqrt{100} &= 97.2^\circ \pm 1.984(0.189)/10 = 97.2^\circ \pm 0.04^\circ \\ &= (97.16^\circ, 97.24^\circ) \end{aligned}$$

where we use the TINV function of Excel to estimate  $t_{99,975}$  by 1.984. Notice that this interval is much narrower than the corresponding interval  $(97.07^\circ, 97.33^\circ)$  based on a sample of 10 days given in Example 6.33.

### Example 6.36

Compute a 95% CI for the underlying mean basal temperature using the data in Example 6.24, assuming that the standard deviation of basal body temperature is  $0.4^\circ$  rather than  $0.189^\circ$  with a sample size of 10.

#### Solution

The 95% CI is given by

$$97.2^\circ \pm 2.262(0.4)/\sqrt{10} = 97.2^\circ \pm 0.29^\circ = (96.91^\circ, 97.49^\circ)$$

Notice that this interval is much wider than the corresponding interval  $(97.07^\circ, 97.33^\circ)$  based on a standard deviation of  $0.189^\circ$  with a sample size of 10.

Usually only  $n$  and  $\alpha$  can be controlled.  $s$  is a function of the type of variable being studied, although  $s$  itself can sometimes be decreased if changes in technique can reduce the amount of measurement error, day-to-day variability, and so forth. An important way in which  $s$  can be reduced is by obtaining replicate measurements for each individual and using the average of several replicates for an individual, rather than a single measurement.

Up to this point, CIs have been used as descriptive tools for characterizing the precision with which the parameters of a distribution can be estimated. Another use for CIs is in making decisions on the basis of data.

**Example 6.37**

**Cardiovascular Disease, Pediatrics** Suppose we know from large studies that the mean cholesterol level in children ages 2–14 is 175 mg/dL. We wish to see if there is a familial aggregation of cholesterol levels. Specifically, we identify a group of fathers who have had a heart attack and have elevated cholesterol levels ( $\geq 250$  mg/dL) and measure the cholesterol levels of their 2- to 14-year-old offspring.

Suppose we find that the mean cholesterol level in a group of 100 such children is 207.3 mg/dL with standard deviation = 30 mg/dL. Is this value far enough from 175 mg/dL for us to believe that the underlying mean cholesterol level in the population of all children selected in this way is different from 175 mg/dL?

**Solution**

One approach would be to construct a 95% CI for  $\mu$  on the basis of our sample data. We then could use the following decision rule: If the interval contains 175 mg/dL, then we cannot say the underlying mean for this group is any different from the mean for all children (175), because 175 is among the plausible values for  $\mu$  provided by the 95% CI. We would decide there is no demonstrated familial aggregation of cholesterol levels. If the CI does not contain 175, then we would conclude the true underlying mean for this group is different from 175. If the lower bound of the CI is above 175, then there is a demonstrated familial aggregation of cholesterol levels. The basis for this decision rule is discussed in the chapters on hypothesis testing.

The CI in this case is given by

$$207.3 \pm t_{99,975}(30)/\sqrt{100} = 207.3 \pm 6.0 = (201.3, 213.3)$$

Clearly, 175 is far from the lower bound of the interval, and we thus conclude there is familial aggregation of cholesterol.

**REVIEW QUESTIONS 6C**

- 1 What does a 95% CI mean?
- 2 (a) Derive a 95% CI for the underlying mean HgbA1c in Review Question 6B.4.  
(b) Suppose that diabetic patients with an underlying mean HgbA1c <8% are considered in good compliance. How do you evaluate the compliance of the patient in Review Question 6B.4?
- 3 (a) What is the difference between a *t* distribution and a normal distribution?  
(b) What is the 95th percentile of a *t* distribution with 30 *df*? What symbol is used to denote this percentile?
- 4 What is the central-limit theorem? Why is it important in statistics?

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

There were 41 twin pairs in this study. We wish to assess whether there is a relationship between BMD and cigarette smoking. One way to approach this problem is to calculate the difference in BMD between the heavier-smoking twin and the lighter-smoking twin for each twin pair and then calculate the average of these differences over the 41 twin pairs. In this study, there was a mean difference in BMD of  $-0.036 \pm 0.014$  g/cm<sup>2</sup> (mean  $\pm$  *se*) for the 41 twin pairs. We can use CI methodology to address this question. Specifically, the 95% CI for the true mean difference ( $\mu_d$ ) in BMD between the heavier- and lighter-smoking twins is

$$-0.036 \pm t_{40,975} (s/\sqrt{41})$$

However, because  $se = s/\sqrt{41}$ , another way to express this formula is

$$\begin{aligned}-0.036 \pm t_{40,975} (se) &= -0.036 \pm 2.021(0.014) \\ &= -0.036 \pm 0.028 = (-0.064, -0.008)\end{aligned}$$

Because the upper bound of the 95% CI is less than 0, we can be fairly confident that the true mean difference is less than 0. Stated another way, we can be fairly confident the true mean BMD for the heavier-smoking twins is lower than that for the lighter-smoking twins. In statistical terms, we say there is a significant association between BMD and cigarette smoking. We discuss assessment of statistical significance in more detail in Chapter 7.

## 6.7 Estimation of the Variance of a Distribution

### Point Estimation

In Chapter 2, the sample variance was defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

This definition is somewhat counterintuitive because the denominator would be expected to be  $n$  rather than  $n - 1$ . A more formal justification for this definition is now given. If our sample  $x_1, \dots, x_n$  is considered as coming from some population with mean  $\mu$  and variance  $\sigma^2$ , then how can the unknown population variance  $\sigma^2$  be estimated from our sample? The following principle is useful in this regard:

#### Equation 6.10

Let  $X_1, \dots, X_n$  be a random sample from some population with mean  $\mu$  and variance  $\sigma^2$ . The **sample variance  $S^2$**  is an **unbiased estimator** of  $\sigma^2$  over all possible random samples of size  $n$  that could have been drawn from this population; that is,  $E(S^2) = \sigma^2$ .

Therefore, if repeated random samples of size  $n$  are selected from the population, as was done in Table 6.3, and the sample variance  $s^2$  is computed from each sample, then the average of these sample variances over a large number of such samples of size  $n$  is the population variance  $\sigma^2$ . This statement holds for any underlying distribution.

#### Example 6.38

**Gynecology** Estimate the variance of the distribution of basal body temperature using the data in Example 6.24.

#### Solution

We have

$$s^2 = \frac{1}{9} \sum_{i=1}^9 (x_i - \bar{x})^2 = 0.0356$$

which is an unbiased estimate of  $\sigma^2$ .

Note that the intuitive estimator for  $\sigma^2$  with  $n$  in the denominator rather than  $n - 1$ , that is,

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

tends to underestimate the underlying variance  $\sigma^2$  by a factor of  $(n - 1)/n$ . This factor is considerable for small samples but tends to be negligible for large samples. A more complete discussion of the relative merits of different estimators for  $\sigma^2$  is given in [3].

## The Chi-Square Distribution

The problem of interval estimation of the mean of a normal distribution was discussed in Section 6.5. We often want to obtain interval estimates of the variance as well. Once again, as was the case for the mean, the interval estimates will hold exactly only if the underlying distribution is normal. The interval estimates perform much more poorly for the variance than for the mean if the underlying distribution is not normal, and they should be used with caution in this case.

### Example 6.39

**Hypertension** An Arteriosonde machine “prints” blood-pressure readings on a tape so that the measurement can be read rather than heard. A major argument for using such a machine is that the variability of measurements obtained by different observers on the same person will be lower than with a standard blood-pressure cuff.

Suppose we have the data in Table 6.6, consisting of systolic blood pressure (SBP) measurements obtained on 10 people and read by two observers. We use the difference  $d_i$  between the first and second observers to assess interobserver variability. In particular, if we assume the underlying distribution of these differences is normal with mean  $\mu$  and variance  $\sigma^2$ , then it is of primary interest to estimate  $\sigma^2$ . The higher  $\sigma^2$  is, the higher the interobserver variability.

**Table 6.6** SBP measurements (mm Hg) from an Arteriosonde machine obtained from 10 people and read by two observers

Person ( $i$ )	Observer		Difference ( $d_i$ )
	1	2	
1	194	200	-6
2	126	123	+3
3	130	128	+2
4	98	101	-3
5	136	135	+1
6	145	145	0
7	110	111	-1
8	108	107	+1
9	102	99	+3
10	126	128	-2

We have seen previously that an unbiased estimator of the variance  $\sigma^2$  is given by the sample variance  $S^2$ . In this case,

$$\text{Mean difference} = (-6 + 3 + \dots - 2)/10 = -0.2 = \bar{d}$$

$$\begin{aligned} \text{Sample variance} &= S^2 = \sum_{i=1}^n (d_i - \bar{d})^2 / 9 \\ &= [(-6 + 0.2)^2 + \dots + (-2 + 0.2)^2] / 9 = 8.178 \end{aligned}$$

how can an interval estimate for  $\sigma^2$  be obtained?

To obtain an interval estimate for  $\sigma^2$ , a new family of distributions, called chi-square ( $\chi^2$ ) distributions, must be introduced to enable us to find the sampling distribution of  $S^2$  from sample to sample.

---

**Definition 6.14** If  $G = \sum_{i=1}^n X_i^2$

where  $X_1, \dots, X_n \sim N(0,1)$

and the  $X_i$ 's are independent, then  $G$  is said to follow a **chi-square distribution with  $n$  degrees of freedom (df)**. The distribution is often denoted by  $\chi_n^2$ .

---

The chi-square distribution is actually a family of distributions indexed by the parameter  $n$  referred to, again, as the degrees of freedom, as was the case for the  $t$  distribution. Unlike the  $t$  distribution, which is always symmetric about 0 for any degrees of freedom, the chi-square distribution only takes on positive values and is always skewed to the right. The general shape of these distributions is indicated in Figure 6.8.

For  $n = 1, 2$ , the distribution has a mode at 0 [3]. For  $n \geq 3$ , the distribution has a mode greater than 0 and is skewed to the right. The skewness diminishes as  $n$  increases. It can be shown that the expected value of a  $\chi_n^2$  distribution is  $n$  and the variance is  $2n$ .

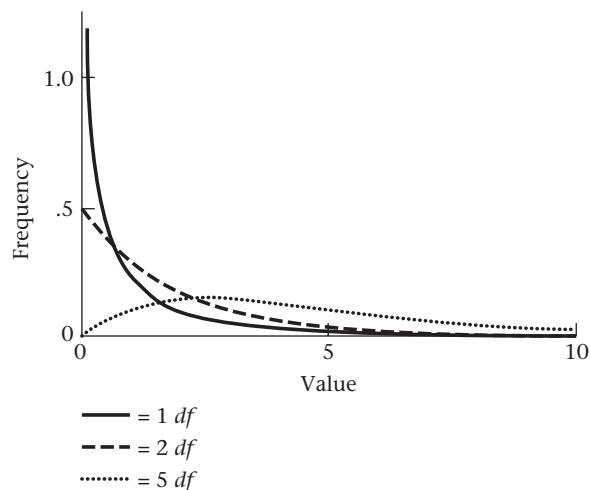
---

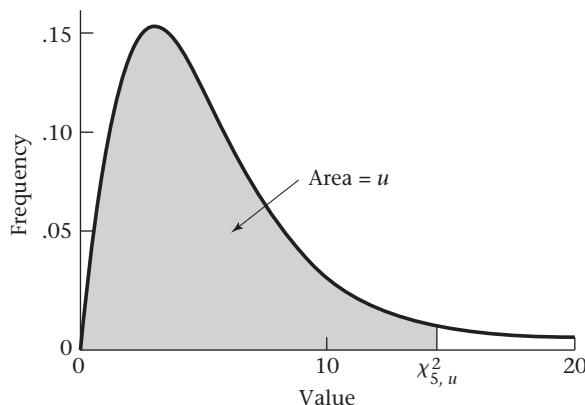
**Definition 6.15** The  **$u$ th percentile of a  $\chi_d^2$  distribution** (i.e., a chi-square distribution with  $d$  df) is denoted by  $\chi_{d,u}^2$  where  $Pr(\chi_d^2 < \chi_{d,u}^2) = u$ . These percentiles are shown in Figure 6.9 for a chi-square distribution with 5 df and appear in Table 6 in the Appendix.

---

Table 6 is constructed like the  $t$  table (Table 5), with the degrees of freedom ( $d$ ) indexed in the first column and the percentile ( $u$ ) indexed in the first row. The main difference between the two tables is that both *lower* ( $u \leq 0.5$ ) and *upper* ( $u > 0.5$ ) percentiles are given for the chi-square distribution, whereas only upper percentiles are

**Figure 6.8 General shape of various  $\chi^2$  distributions with  $d$  df**



**Figure 6.9** Graphic display of the percentiles of a  $\chi^2_5$  distribution

given for the  $t$  distribution. The  $t$  distribution is symmetric about 0, so any lower percentile can be obtained as the negative of the corresponding upper percentile. Because the chi-square distribution is, in general, a skewed distribution, there is no simple relationship between the upper and lower percentiles.

**Example 6.40** Find the upper and lower 2.5th percentiles of a chi-square distribution with 10  $df$ .

**Solution** According to Appendix Table 6, the upper and lower percentiles are given, respectively, by

$$\chi^2_{10,0.975} = 20.48 \quad \text{and} \quad \chi^2_{10,0.025} = 3.25$$

For values of  $d$  not given in Table 6, a computer program, such as MINITAB or Excel or Stata, can be used to obtain percentiles.

### Interval Estimation

To obtain an interval estimate of  $\sigma^2$ , we need to find the sampling distribution of  $S^2$ . Suppose we assume that  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ . Then it can be shown that

$$\text{Equation 6.11} \quad S^2 \sim \frac{\sigma^2 \chi^2_{n-1}}{n-1}$$

To see this, we recall from Section 5.5 that if  $X \sim N(\mu, \sigma^2)$ , then if we standardize  $X$  (that is, we subtract  $\mu$  and divide by  $\sigma$ ), thus creating a new random variable  $Z = (X - \mu)/\sigma$ , then  $Z$  will be normally distributed with mean 0 and variance 1. Thus from Definition 6.14 we see that

$$\text{Equation 6.12} \quad \sum_{i=1}^n Z_i^2 = \sum_{i=1}^n (X_i - \mu)^2 / \sigma^2 \sim \chi^2_n = \text{chi-square distribution with } n \text{ df}$$

Because we usually don't know  $\mu$ , we estimate  $\mu$  by  $\bar{x}$ . However, it can be shown that if we substitute  $\bar{X}$  for  $\mu$  in Equation 6.12, then we lose 1  $df$  [3], resulting in the relationship

**Equation 6.13**

$$\sum_{i=1}^n (X_i - \bar{X})^2 / \sigma^2 \sim \chi_{n-1}^2$$

However, we recall from the definition of a sample variance that  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$ . Thus, multiplying both sides by  $(n-1)$  yields the relationship

$$(n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$$

Substituting into Equation 6.13, we obtain

**Equation 6.14**

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

If we multiply both sides of Equation 6.14 by  $\sigma^2/(n-1)$ , we obtain Equation 6.11,

$$S^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$$

Thus, from Equation 6.11 we see that  $S^2$  follows a chi-square distribution with  $n-1$  df multiplied by the constant  $\sigma^2/(n-1)$ . Manipulations similar to those given in Section 6.5 can now be used to obtain a  $100\% \times (1-\alpha)$  CI for  $\sigma^2$ .

In particular, from Equation 6.11 it follows that

$$Pr\left(\frac{\sigma^2 \chi_{n-1,\alpha/2}^2}{n-1} < S^2 < \frac{\sigma^2 \chi_{n-1,1-\alpha/2}^2}{n-1}\right) = 1 - \alpha$$

This inequality can be represented as two separate inequalities:

$$\frac{\sigma^2 \chi_{n-1,\alpha/2}^2}{n-1} < S^2 \quad \text{and} \quad S^2 < \frac{\sigma^2 \chi_{n-1,1-\alpha/2}^2}{n-1}$$

If both sides of the first inequality are multiplied by  $(n-1)/\chi_{n-1,\alpha/2}^2$  and both sides of the second inequality are multiplied by  $(n-1)/\chi_{n-1,1-\alpha/2}^2$ , then we have

$$\sigma^2 < \frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} \quad \text{and} \quad \frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} < \sigma^2$$

or, on combining these two inequalities,

$$\frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2}$$

It follows that

$$Pr\left[\frac{(n-1)S^2}{\chi_{n-1,1-\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2}\right] = 1 - \alpha$$

Thus the interval  $[(n-1)s^2/\chi_{n-1,1-\alpha/2}^2, (n-1)s^2/\chi_{n-1,\alpha/2}^2]$  is a  $100\% \times (1-\alpha)$  CI for  $\sigma^2$ .

**Equation 6.15**

A  $100\% \times (1-\alpha)$  CI for  $\sigma^2$  is given by

$$[(n-1)s^2/\chi_{n-1,1-\alpha/2}^2, (n-1)s^2/\chi_{n-1,\alpha/2}^2]$$

**Example 6.41**

**Hypertension** We now return to the specific data set in Example 6.39. Suppose we want to construct a 95% CI for the interobserver variability as defined by  $\sigma^2$ .

**Solution**

Because there are 10 people and  $s^2 = 8.178$ , the required interval is given by

$$(9s^2/\chi_{9,975}^2, 9s^2/\chi_{9,025}^2) = [9(8.178)/19.02, 9(8.178)/2.70] = (3.87, 27.26)$$

Similarly, a 95% CI for  $\sigma$  is given by  $(\sqrt{3.87}, \sqrt{27.26}) = (1.97, 5.22)$ . Notice that the CI for  $\sigma^2$  is *not* symmetric about  $s^2 = 8.178$ , in contrast to the CI for  $\mu$ , which *was* symmetric about  $\bar{x}$ . This characteristic is common in CIs for the variance.

We could use the CI for  $\sigma^2$  to make decisions concerning the variability of the Arteriosonde machine if we had a good estimate of the interobserver variability of blood-pressure readings from a standard cuff. For example, suppose we know from previous work that if two people are listening to blood-pressure recordings from a standard cuff, then the interobserver variability as measured by the variance of the set of differences between the readings of two observers is 35. This value is outside the range of the 95% CI for  $\sigma^2(3.87, 27.26)$ , and we thus conclude that the interobserver variability is reduced by using an Arteriosonde machine. Alternatively, if this prior variance were 15, then we could not say that the variances obtained from using the two methods are different.

Note that the CI for  $\sigma^2$  in Equation 6.15 is only valid for normally distributed samples. If the underlying distribution is not normal, then the level of confidence for this interval may not be  $1 - \alpha$  even if the sample size is large. This is different from the CI for  $\mu$  given in Equation 6.6, which will be valid for large  $n$  based on the central-limit theorem, even if the underlying distribution is not normal.

**REVIEW QUESTIONS 6D**

- 1 What is the difference between a  $t$  distribution and a chi-square distribution? When do we use each?
- 2 Suppose we have a normal distribution with mean = 0 and variance = 5. We draw a sample of size 8 from this distribution and compute the sample variance,  $s^2$ . What is the probability that  $s^2 > 10$ ?

## 6.8 Estimation for the Binomial Distribution

### Point Estimation

Point estimation for the parameter  $p$  of a binomial distribution is discussed in this section.

**Example 6.42**

**Cancer** Consider the problem of estimating the prevalence of malignant melanoma in 45- to 54-year-old women in the United States. Suppose a random sample of 5000 women is selected from this age group, of whom 28 are found to have the disease. Let the random variable  $X_i$  represent the disease status for the  $i$ th woman, where  $X_i = 1$  if the  $i$ th woman has the disease and 0 if she does not;  $i = 1, \dots, 5000$ . The random variable  $X_i$  was also defined as a Bernoulli trial in Definition 5.14. Suppose the prevalence of the disease in this age group =  $p$ . How can  $p$  be estimated?

We let  $X = \sum_{i=1}^n X_i$  = the number of women with malignant melanoma among the  $n$  women. From Example 5.32, we have  $E(X) = np$  and  $Var(X) = npq$ . Note that  $X$  can

also be looked at as a binomial random variable with parameters  $n$  and  $p$  because  $X$  represents the number of events in  $n$  independent trials.

Finally, consider the random variable  $\hat{p}$  = sample proportion of events. In our example,  $\hat{p}$  = proportion of women with malignant melanoma. Thus

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = X/n$$

Because  $\hat{p}$  is a sample mean, the results of Equation 6.1 apply and we see that  $E(\hat{p}) = E(X_i) \equiv \mu = p$ . Furthermore, from Equation 6.2 it follows that

$$Var(\hat{p}) = \sigma^2/n = pq/n \quad \text{and} \quad se(\hat{p}) = \sqrt{pq/n}$$

Thus, for any sample of size  $n$  the sample proportion  $\hat{p}$  is an unbiased estimator of the population proportion  $p$ . The standard error of this proportion is given exactly by  $\sqrt{pq/n}$  and is estimated by  $\sqrt{\hat{p}\hat{q}/n}$ . These principles can be summarized as follows:

#### Equation 6.16

##### Point Estimation of the Binomial Parameter $p$

Let  $X$  be a binomial random variable with parameters  $n$  and  $p$ . An unbiased estimator of  $p$  is given by the sample proportion of events  $\hat{p}$ . Its standard error is given exactly by  $\sqrt{pq/n}$  and is estimated by  $\sqrt{\hat{p}\hat{q}/n}$ .

#### Example 6.43

Estimate the prevalence of malignant melanoma in Example 6.42, and provide its standard error.

#### Solution

Our best estimate of the prevalence rate of malignant melanoma among 45- to 54-year-old women is  $28/5000 = .0056$ . Its estimated standard error is

$$\sqrt{.0056(9944)/5000} = .0011$$

### Maximum-Likelihood Estimation

A better justification for the use of  $\hat{p}$  to estimate the binomial parameter  $p$  is that  $\hat{p}$  is a **maximum-likelihood estimator** of  $p$ . To understand this concept, we need to define a **likelihood**.

#### Definition 6.16

Suppose the probability-mass function of a discrete random variable  $X$  is a function of  $k$  parameters  $p_1, \dots, p_k$ , denoted by  $\underline{p}$ . If  $x_1, \dots, x_n$  is a sample of  $n$  independent observations from  $X$ , then the *likelihood* of the sample  $x_1, \dots, x_n$  given  $\underline{p}$  is denoted by  $L(x|\underline{p})$  and represents the probability of obtaining our sample given specified values for the parameters  $p_1, \dots, p_k$ . It is given by

$$L(x|\underline{p}) = Pr(x_1 | \underline{p}) \times \dots \times Pr(x_n | \underline{p}) = \prod_{i=1}^n Pr(x_i | \underline{p})$$

#### Example 6.44

**Diabetes** Suppose we have a sample of 100 men, of whom 30 have diabetes and 70 do not. If the prevalence of diabetes =  $p$ , then what is the likelihood of the sample given  $p$ ?

#### Solution

Each observation is a binary random variable where  $x_i = 1$  if a man has diabetes and = 0 otherwise. In this example  $\underline{p}$  is a single parameter  $p$ . Furthermore,  $Pr(X = 1) = p$  and  $Pr(X = 0) = 1 - p$ . Thus, the likelihood of the sample is  $p^{30}(1 - p)^{70} = L(x|\underline{p})$ . If  $X$  is a continuous random variable, then because  $Pr(X = x) = 0$  for specific values of  $x$ , a slightly different definition of likelihood is used, where the probability of a specific observation given  $\underline{p}$  is replaced by the probability density of that observation given  $\underline{p}$ .

**Definition 6.17**

Suppose the probability-density function ( $f$ ) of a continuous random-variable  $X$  is a function of  $k$  parameters  $p_1, \dots, p_k \equiv \mathbf{p}$ . If  $x_1, \dots, x_n$  is a sample of  $n$  independent observations, and  $\mathbf{x} = (x_1, \dots, x_n)$ , then

$$L(\mathbf{x} | \mathbf{p}) = f(x_1 | \mathbf{p}) \times \dots \times f(x_n | \mathbf{p}) = \prod_{i=1}^n f(x_i | \mathbf{p})$$

**Example 6.45**

Suppose we have  $n$  independent observations  $x_1, \dots, x_n$  from a normal distribution with mean  $= \mu$  and variance  $= \sigma^2$ . What is the likelihood of the sample?

**Solution**

In this case,  $\mathbf{p} = (\mu, \sigma^2)$ . From the definition of a normal density, we have

$$f(x_i) = \left[ 1/(\sqrt{2\pi}\sigma) \right] \exp \left[ -(1/2)(x_i - \mu)^2 \right]$$

Thus the likelihood of the sample is

$$\begin{aligned} L(\mathbf{x} | \mu, \sigma^2) &= \prod_{i=1}^n \left\{ \left[ 1/(\sqrt{2\pi}\sigma) \right] \exp \left[ -(1/2)(x_i - \mu)^2 \right] \right\} \\ &= 1/\left[ (2\pi)^{n/2} \sigma^n \right] \exp \left\{ (-1/2) \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2 \right\} \end{aligned}$$

We have the following definition of a maximum-likelihood estimator.

**Definition 6.18**

Suppose we have a probability-mass function or a probability-density function that is a function of  $k$  parameters  $p_1, \dots, p_k$ . The **maximum-likelihood estimator** (MLE) of the parameters  $p_1, \dots, p_k$  are the values  $p_{1,\text{ML}}, \dots, p_{k,\text{ML}}$  which maximize the likelihood. Heuristically, the MLE can be thought of as the values of the parameters ( $\mathbf{p}$ ) that maximize the probability of the observed data given  $\mathbf{p}$ .

**Example 6.46**

**Diabetes** Find the MLE of the diabetes prevalence  $p$  in Example 6.44.

**Solution**

The likelihood  $L = p^{30} (1-p)^{70}$ . It is usually easier to maximize the log likelihood rather than the likelihood itself. Hence,

$$\log L = 30 \log p + 70 \log(1-p)$$

To maximize  $\log L$ , we take the derivative of  $\log L$  with respect to  $p$  and set the expression to 0. We have

$$d \log L / dp = 30/p - 70/(1-p) = 0$$

or  $30/p = 70/(1-p)$

or  $30(1-p) = 70p$

or  $30 = 100p$ , or  $\hat{p}_{\text{ML}} = 30/100 = .3$

Thus, .3 is the MLE of  $p$ .

In general, if we have a binomial distribution with  $n$  observations of which  $k$  are successes and  $n - k$  are failures, then

$$L(\mathbf{x} | p) = p^k (1-p)^{n-k} \text{ and the MLE of } p = \hat{p}_{\text{ML}} = k/n = \hat{p}$$

The rationale for using the MLE is that in general, for a wide class of distributions, as the sample size gets large the MLE is unbiased and has the smallest variance among all unbiased estimators. Thus it is a useful general method of parameter estimation.

**Example 6.47**

Consider Example 6.45. Find the MLE of  $\mu$ .

**Solution**

From Example 6.45, we have

$$L(\mathbf{x} | \mu, \sigma^2) \equiv L = 1 / [(2\pi)^{n/2} \sigma^n] \exp \left[ (-1/2) \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2 \right]$$

$$\text{or } \log L = (-n/2) \log(2\pi) - n \log(\sigma) - (1/2) \sum_{i=1}^n (x_i - \mu)^2 / \sigma^2$$

We take the derivative  $\log L$  with respect to  $\mu$  and set the expression to 0 as follows:

$$d \log L / d \mu = \sum_{i=1}^n (x_i - \mu) / \sigma^2 = 0$$

or, because  $\sigma^2 > 0$ , we have

$$\sum_{i=1}^n (x_i - \mu) = 0$$

or

$$\sum_{i=1}^n x_i - n \mu = 0$$

or

$$\hat{\mu}_{\text{ML}} = \sum_{i=1}^n x_i / n = \bar{x}$$

Thus the MLE of  $\mu$  is  $\bar{x}$ .

## Interval Estimation—Normal-Theory Methods

Point estimation of the parameter  $p$  of a binomial distribution was covered in the previous two sections. How can an **interval estimate** of the parameter  $p$  be obtained?

**Example 6.48**

**Cancer** Suppose we are interested in estimating the prevalence rate of breast cancer among 50- to 54-year-old women whose mothers have had breast cancer. Suppose that in a random sample of 10,000 such women, 400 are found to have had breast cancer at some point in their lives. We have shown that the best point estimate of the prevalence rate  $p$  is given by the sample proportion  $\hat{p} = 400/10,000 = .040$ . How can an interval estimate of the parameter  $p$  be obtained? (See the solution in Example 6.49.)

Let's assume the normal approximation to the binomial distribution is valid—whereby from Equation 5.14 the number of events  $X$  observed out of  $n$  women will be approximately normally distributed with mean  $np$  and variance  $npq$  or, correspondingly, the proportion of women with events =  $\hat{p} = X/n$  is normally distributed with mean  $p$  and variance  $pq/n$ .

The normal approximation can actually be justified on the basis of the central-limit theorem. Indeed, in the previous section we showed that  $\hat{p}$  could be represented as an average of  $n$  Bernoulli trials, each of which has mean  $p$  and variance  $pq$ . Thus for large  $n$ , from the central-limit theorem, we can see that  $\hat{p} = \bar{X}$  is normally distributed with mean  $\mu = p$  and variance  $\sigma^2/n = pq/n$ , or

**Equation 6.17**

$$\hat{p} \sim N(p, pq/n)$$

Alternatively, because the number of successes in  $n$  Bernoulli trials =  $X = np$  (which is the same as a binomial random variable with parameters  $n$  and  $p$ ), if Equation 6.17 is multiplied by  $n$ ,

$$\text{Equation 6.18} \quad X \sim N(np, npq)$$

This formulation is indeed the same as that for the normal approximation to the binomial distribution, which was given in Equation 5.14. How large should  $n$  be before this approximation can be used? In Chapter 5 we said the normal approximation to the binomial distribution is valid if  $npq \geq 5$ . However, in Chapter 5 we assumed  $p$  was known, whereas here we assume it is unknown. Thus we estimate  $p$  by  $\hat{p}$  and  $q$  by  $\hat{q} = 1 - \hat{p}$  and apply the normal approximation to the binomial if  $n\hat{p}\hat{q} \geq 5$ . Therefore, the results of this section should only be used if  $n\hat{p}\hat{q} \geq 5$ . An approximate 100%  $\times (1 - \alpha)$  CI for  $p$  can now be derived from Equation 6.17 using methods similar to those given in Section 6.5. In particular, from Equation 6.17 we see that

$$Pr(p - z_{1-\alpha/2} \sqrt{pq/n} < \hat{p} < p + z_{1-\alpha/2} \sqrt{pq/n}) = 1 - \alpha$$

This inequality can be written in the form of two inequalities:

$$p - z_{1-\alpha/2} \sqrt{pq/n} < \hat{p} \quad \text{and} \quad \hat{p} < p + z_{1-\alpha/2} \sqrt{pq/n}$$

To explicitly derive a CI based on these inequalities requires solving a quadratic equation for  $p$  in terms of  $\hat{p}$ . To avoid this, it is customary to approximate  $\sqrt{pq/n}$  by  $\sqrt{\hat{p}\hat{q}/n}$  and to rewrite the inequalities in the form

$$p - z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} < \hat{p} \quad \text{and} \quad \hat{p} < p + z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n}$$

We now add  $z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n}$  to both sides of the first inequality and subtract this quantity from both sides of the second inequality, obtaining

$$p < \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} \quad \text{and} \quad \hat{p} - z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} < p$$

Combining these two inequalities, we get

$$\hat{p} - z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} < p < \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n}$$

$$\text{or } Pr(\hat{p} - z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n} < p < \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n}) = 1 - \alpha$$

The approximate 100%  $\times (1 - \alpha)$  CI for  $p$  is given by

$$(\hat{p} - z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n})$$

$$\text{Equation 6.19}$$

#### Normal-Theory Method for Obtaining a CI for the Binomial Parameter $p$ (Wald Method)

An approximate 100%  $\times (1 - \alpha)$  CI for the binomial parameter  $p$  based on the normal approximation to the binomial distribution is given by

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\hat{p}\hat{q}/n}$$

This method of interval estimation should only be used if  $n\hat{p}\hat{q} \geq 5$ .

**Example 6.49**

**Cancer** Using the data in Example 6.48, derive a 95% CI for the prevalence rate of breast cancer among 50- to 54-year-old women whose mothers have had breast cancer.

**Solution**

$$\hat{p} = .040 \quad \alpha = .05 \quad z_{1-\alpha/2} = 1.96 \quad n = 10,000$$

Therefore, an approximate 95% CI is given by

$$\begin{aligned} &[.040 - 1.96\sqrt{.04(.96)/10,000}, .040 + 1.96\sqrt{.04(.96)/10,000}] \\ &= (.040 - .004, .040 + .004) = (.036, .044) \end{aligned}$$

Suppose we know the prevalence rate of breast cancer among all 50- to 54-year-old American women is 2%. Because 2% is less than .036 (the lower confidence limit), we can be quite confident that the underlying rate for the group of women whose mothers have had breast cancer is higher than the rate in the general population.

### Interval Estimation—Exact Methods

The question remains: How is a CI for the binomial parameter  $p$  obtained when either the normal approximation to the binomial distribution is not valid or a more exact CI is desired?

**Example 6.50**

**Cancer, Nutrition** Suppose we want to estimate the rate of bladder cancer in rats that have been fed a diet high in saccharin. We feed this diet to 20 rats and find that 2 develop bladder cancer. In this case, our best point estimate of  $p$  is  $\hat{p} = \frac{2}{20} = .1$ .

However, because

$$n\hat{p}\hat{q} = 20(2/20)(18/20) = 1.8 < 5$$

the normal approximation to the binomial distribution cannot be used and thus normal-theory methods for obtaining CIs are not valid. How can an interval estimate be obtained in this case?

A small-sample method for obtaining confidence limits will be presented.

**Equation 6.20**

#### Exact Method for Obtaining a CI for the Binomial Parameter $p$ (Clopper-Pearson Method)

An exact  $100\% \times (1 - \alpha)$  CI for the binomial parameter  $p$  that is always valid is given by  $(p_1, p_2)$ , where  $p_1, p_2$  satisfy the equations

$$Pr(X \geq x | p = p_1) = \frac{\alpha}{2} = \sum_{k=x}^n \binom{n}{k} p_1^k (1-p_1)^{n-k}$$

$$Pr(X \leq x | p = p_2) = \frac{\alpha}{2} = \sum_{k=0}^x \binom{n}{k} p_2^k (1-p_2)^{n-k}$$

A rationale for this CI is given in our discussion of hypothesis testing for the binomial distribution in Section 7.10 on page 247.

The main problem with using this method is the difficulty in computing expressions such as

$$\sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k}$$

Fortunately, special tables exist for the evaluation of such expressions, one of which is given in Table 7 in the Appendix. This table can be used as follows:

**Equation 6.21****Exact Confidence Limits for Binomial Proportions**

- (1) The sample size ( $n$ ) is given along each curve. There are two curves corresponding to each given sample size. One curve is used to obtain the lower confidence limit and the other to obtain the upper confidence limit.
- (2) If  $0 \leq \hat{p} \leq .5$ , then
  - (a) Refer to the lower horizontal axis, and find the point corresponding to  $\hat{p}$ .
  - (b) Draw a line perpendicular to the horizontal axis, and find the two points where this line intersects the two curves identified in (1).
  - (c) Read across to the left vertical axis; the smaller value corresponds to the lower confidence limit and the larger value to the upper confidence limit.
- (3) If  $.5 < \hat{p} \leq 1.0$ , then
  - (a) Refer to the upper horizontal axis, and find the point corresponding to  $\hat{p}$ .
  - (b) Draw a line perpendicular to the horizontal axis, and find the two points where this line intersects the two curves identified in (1).
  - (c) Read across to the right vertical axis; the smaller value corresponds to the lower confidence limit and the larger value to the upper confidence limit.

**Example 6.51**

**Cancer** Derive an exact 95% CI for the probability of developing bladder cancer, using the data given in Example 6.50.

**Solution**

We refer to Table 7a in the Appendix ( $\alpha = .05$ ) and identify the two curves with  $n = 20$ . Because  $\hat{p} = .1 \leq .5$ , we refer to the lower horizontal axis and draw a vertical line at .10 until it intersects the two curves marked  $n = 20$ . We then read across to the left vertical axis and find the confidence limits of .01 and .32. Thus the exact 95% CI = (.01, .32). Notice that this CI is *not* symmetric about  $\hat{p} = .10$ .

Another approach to solving this problem is to use the BINOMDIST function of Excel. From Equation 6.20, we need to find values of  $p_1$  and  $p_2$  such that

$$Pr(X \geq 2 | p = p_1) = .025 \quad \text{and} \quad Pr(X \leq 2 | p = p_2) = .025$$

However,  $Pr(X \geq 2 | p = p_1) = 1 - Pr(X \leq 1 | p = p_1) = 1 - \text{BINOMDIST}(1, 20, p_1, \text{TRUE})$  and  $Pr(X \leq 2 | p = p_2) = \text{BINOMDIST}(2, 20, p_2, \text{TRUE})$ . Hence we set up a spreadsheet in which the first column has values of  $p_1$  from .01 to 1.0 in increments of .01; the second column has  $1 - \text{BINOMDIST}(1, 20, p_1, \text{TRUE})$ ; the third column has values of  $p_2$  from .01 to 1.0 in increments of .01; and the fourth column has  $\text{BINOMDIST}(2, 20, p_2, \text{TRUE})$ . An excerpt from the spreadsheet is shown in Table 6.7.

Usually with exact confidence limits accurate to a fixed number of decimal places, we cannot exactly satisfy Equation 6.20. Instead, we use a more conservative approach. We find the largest value of  $p_1$  so that  $Pr(X \geq x | p = p_1) \leq \alpha/2$  and the smallest value of  $p_2$  so that  $Pr(X \leq x | p = p_2) \leq \alpha/2$ . Based on Table 6.7 with  $\alpha = .05$ , the values of  $p_1$  and  $p_2$  that satisfy these inequalities are  $p_1 = .01$  and  $p_2 = .32$ . Hence, the 95% CI for  $p$  is (.01, .32).

**Table 6.7** Evaluation of exact binomial confidence limits using Excel, based on the data in Example 6.50

$p_1$	$1 - \text{BINOMDIST}(1, 20, p_1, \text{TRUE})$	$p_2$	$\text{BINOMDIST}(2, 20, p_2, \text{TRUE})$
0.01	0.017	0.25	0.091
0.02	0.060	0.26	0.076
0.03	0.120	0.27	0.064
0.04	0.190	0.28	0.053
0.05	0.264	0.29	0.043
0.06	0.340	0.30	0.035
0.07	0.413	0.31	0.029
0.08	0.483	0.32	0.023
0.09	0.548	0.33	0.019
0.1	0.608	0.34	0.015

**Example 6.52**

**Health Promotion** Suppose that as part of a program for counseling patients with many risk factors for heart disease, 100 smokers are identified. Of this group, 10 give up smoking for at least 1 month. After a 1-year follow-up, 6 of the 10 patients are found to have taken up smoking again. The proportion of ex-smokers who start smoking again is called the *recidivism rate*. Derive a 99% CI for the recidivism rate.

**Solution**

Exact binomial confidence limits must be used, because

$$n\hat{p}\hat{q} = 10(.6)(.4) = 2.4 < 5$$

We refer to the upper horizontal axis of the chart marked  $\alpha = .01$  in Appendix Table 7b and note the point  $\hat{p} = .60$ . We then draw a vertical line at .60 until it intersects the two curves marked  $n = 10$ . We then read across to the right vertical axis and find the confidence limits of .19 and .92. Thus the exact 99% CI = (.19, .92).

More extensive and precise exact binomial confidence limits are available in Geigy Scientific Tables [4]. Also, calculation of exact binomial confidence limits are available directly in some statistical packages, including Stata and indirectly using Excel as previously discussed.

For example, we can also use the Stata command cii to obtain exact 99% confidence limits for  $p$ . The general form of this command is

.cii n x, level(%).

where % is the level of confidence, n is the number of trials and x is the number of successes. The results for the recidivism data are as follows:

.cii 10 6, level(99)

Variable	Obs	Mean	Std. Err.	-- Binomial Exact -- [99% Conf. Interval]
	10	.6	.1549193	.1909163 .9232318

We see that the 99% exact binomial confidence interval is (.19, .92), which is the same as we obtained with Table 7b.

**REVIEW QUESTIONS 6E**

- 1 Suppose an experimental weight-loss program is provided for 40 overweight participants. A participant is considered *partially successful* if he or she has lost 5 lb or more after 6 months. Suppose that 10 of the 40 participants are partially successful in losing weight.
  - (a) What is an estimate of the partial-success rate?
  - (b) Derive a 95% CI for the proportion of partial successes.
- 2 A participant is considered *completely successful* if he or she has lost 20 lb or more after 6 months. Suppose that 4 of the 40 participants are completely successful at losing weight.
  - (a) What is an estimate of the complete success rate?
  - (b) Derive a 95% CI for the proportion of participants who were completely successful at losing weight.

**6.9 Estimation for the Poisson Distribution****Point Estimation**

In this section, we discuss point estimation for the parameter  $\lambda$  of a Poisson distribution.

**Example 6.53**

**Cancer, Environmental Health** A study in Woburn, Massachusetts, in the 1970s looked at possible excess cancer risk in children, with a particular focus on leukemia. This study was later portrayed in the book and movie titled *A Civil Action*. An important environmental issue in the investigation concerned the possible contamination of the town's water supply. Specifically, 12 children (<19 years of age) were diagnosed with leukemia in Woburn during the period January 1, 1970 to December 31, 1979. A key statistical issue is whether this represents an excessive number of leukemia cases, assuming that Woburn has had a constant 12,000 child residents ( $\leq$  19) during this period and that the incidence rate of leukemia in children nationally is 5 cases per 100,000 person-years. Can we estimate the incidence rate of childhood leukemia in Woburn during the 1970s and provide a CI about this estimate?

We let  $X$  = the number of children who developed leukemia in Woburn during the 1970s. Because  $X$  represents a rare event, we assume that  $X$  follows a Poisson distribution with parameter  $\mu = \lambda T$ . We know from Chapter 4 that for a Poisson distribution,  $E(X) = \lambda T$ , where  $T$  = time and  $\lambda$  = number of events per unit time.

**Definition 6.19**


---

A **person-year** is a unit of time defined as 1 person being followed for 1 year.

---

This unit of follow-up time is commonly used in longitudinal studies—that is, studies in which the same individual is followed over time.

**Example 6.54**

**Cancer, Environmental Health** How many person-years accumulated in the Woburn study in Example 6.53?

**Solution**

In the Woburn study, 12,000 children were each followed for 10 years. Thus a total of 120,000 person-years accumulated. This is actually an approximation, because the children who developed leukemia over the 10-year period were only followed up to the time they developed the disease. It is also common to curtail follow-up for

other reasons, such as death or the development of other types of cancer. However, the number of children for whom follow-up is curtailed for these reasons is probably small and the approximation is likely to be accurate.

Finally, although children moved in and out of Woburn over the 10-year period, we assume there was no net migration in or out of the area during the 1970s.

We now wish to assess how to estimate  $\lambda$  based on an observed number of events  $X$  over  $T$  person-years.

### Equation 6.22

#### Point Estimation for the Poisson Distribution

Let's assume the number of events  $X$  over  $T$  person-years is Poisson-distributed with parameter  $\mu = \lambda T$ . An unbiased estimator of  $\lambda$  is given by  $\hat{\lambda} = X / T$ , where  $X$  is the observed number of events over  $T$  person-years.

If  $\lambda$  is the incidence rate per person-year,  $T$  = number of person-years of follow-up, and we assume a Poisson distribution for the number of events  $X$  over  $T$  person-years, then the expected value of  $X$  is given by  $E(X) = \lambda T$ . Therefore,

$$\begin{aligned} E(\hat{\lambda}) &= E(X)/T \\ &= \lambda T/T = \lambda \end{aligned}$$

Thus  $\hat{\lambda}$  is an unbiased estimator of  $\lambda$ .

### Example 6.55

**Cancer, Environmental Health** Estimate the incidence rate of childhood leukemia in Woburn during the 1970s based on the data provided in Example 6.53.

#### Solution

There were 12 events over 120,000 person-years, so the estimated incidence rate =  $12/120,000 = 1/10,000 = 0.0001$  events per person-year. Because cancer incidence rates per person-year are usually very low, it is customary to express such rates per 100,000 (or  $10^5$ ) person-years—that is, to change the unit of time to  $10^5$  person-years. Thus if the unit of time =  $10^5$  person-years, then  $T = 1.2$  and  $\hat{\lambda} = 0.0001 (10^5) = 10$  events per 100,000 person-years.

### Interval Estimation

The question remains as to how to obtain an interval estimate for  $\lambda$ . We use a similar approach as was used to obtain exact confidence limits for the binomial proportion  $p$  in Equation 6.20. For this purpose, it is easier to first obtain a CI for  $\mu$  = expected number of events over time  $T$  of the form  $(\mu_1, \mu_2)$  and then obtain the corresponding CI for  $\lambda$  from  $(\mu_1/T, \mu_2/T)$ . The approach is given as follows:

### Equation 6.23

#### Exact Method for Obtaining a CI for the Poisson Parameter $\lambda$

An exact  $100\% \times (1 - \alpha)$  CI for the Poisson parameter  $\lambda$  is given by  $(\mu_1/T, \mu_2/T)$ , where  $\mu_1, \mu_2$  satisfy the equations

$$\begin{aligned} Pr(X \geq x | \mu = \mu_1) &= \frac{\alpha}{2} = \sum_{k=x}^{\infty} e^{-\mu_1} \mu_1^k / k! \\ &= 1 - \sum_{k=0}^{x-1} e^{-\mu_1} \mu_1^k / k! \end{aligned}$$

$$Pr(X \leq x | \mu = \mu_2) = \frac{\alpha}{2} = \sum_{k=0}^x e^{-\mu_2} \mu_2^k / k!$$

and  $x$  = observed number of events,  $T$  = number of person-years of follow-up.

As in obtaining exact confidence limits for the binomial parameter  $p$ , it is difficult to exactly compute  $\mu_1, \mu_2$  to satisfy Equation 6.23. Table 8 in the Appendix provides the solution to these equations. This table can be used to find 90%, 95%, 98%, 99%, or 99.8% CIs for  $\mu$  if the observed number of events ( $x$ ) is  $\leq 50$ . The observed number of events ( $x$ ) is listed in the first column, and the level of confidence is given in the first row. The CI is obtained by cross-referencing the  $x$  row and the  $1 - \alpha$  column.

**Example 6.56**

Suppose we observe 8 events and assume the number of events is Poisson-distributed with parameter  $\mu$ . Find the 95% CI for  $\mu$ .

**Solution**

We look at Table 8 under the  $x = 8$  row and the 0.95 column to find the 95% CI for  $\mu = (3.45, 15.76)$ .

We see this CI is *not* symmetric about  $x$  (8), because  $15.76 - 8 = 7.76 > 8 - 3.45 = 4.55$ . This is true for all exact CIs based on the Poisson distribution unless  $x$  is very large.

**Example 6.57**

**Cancer, Environmental Health** Compute a 95% CI for both the expected number of childhood leukemias ( $\mu$ ) and the incidence rate of childhood leukemia per  $10^5$  person-years ( $\lambda$ ) in Woburn based on the data provided in Example 6.53.

**Solution**

We observed 12 cases of childhood leukemia over 10 years. Thus, from Table 8, referring to  $x = 12$  and level of confidence 95%, we find that the 95% CI for  $\mu = (6.20, 20.96)$ . Because there were 120,000 person-years =  $T$ , a 95% CI for the incidence rate =  $\left(\frac{6.20}{120,000}, \frac{20.96}{120,000}\right)$  events per person-year or  $\left(\frac{6.20}{120,000} \times 10^5, \frac{20.96}{120,000} \times 10^5\right)$  events per  $10^5$  person-years =  $(5.2, 17.5)$  events per  $10^5$  person-years = 95% CI for  $\lambda$ .

We can also use the Stata cii command to obtain an exact 95% CI for the incidence rate ( $\lambda$ ). The general syntax is

.cii py x, poisson

where py = number of person-years and x = number of events. The results for the leukemia data are as follows:

.cii 120000 12, poisson

Variable	Exposure	Mean	Std. Err.	-- Poisson Exact --	
				[95% Conf. Interval]	
	120000	.0001	.0000289	.0000517	.0001747

We see that the 95% CI for  $\lambda = (5.2/10^5, 17.5/10^5)$ , which agrees with our results from Table 8. Stata cannot be used if we just have available a number of events, without a corresponding number of person-years as in Example 6.56.

**Example 6.58**

**Cancer, Environmental Health** Interpret the results in Example 6.57. Specifically, do you feel there was an excess childhood leukemia risk in Woburn, Massachusetts, relative to expected U.S. incidence rates?

**Solution**

Referring to Example 6.53, we note that the incidence rate of childhood leukemia in the United States during the 1970s was 5 events per  $10^5$  person-years. We denote this rate by  $\lambda_0$ . Referring to Example 6.57, we see that the 95% CI for  $\lambda$  in Woburn =  $(5.2, 17.5)$  events per  $10^5$  person-years. The lower bound of the 95% CI exceeds  $\lambda_0$  ( $= 5$ ),

so we can conclude there was a significant excess of childhood leukemia in Woburn during the 1970s. Another way to express these results is in terms of the standardized morbidity ratio (SMR) defined by

$$\text{SMR} = \frac{\text{incidence rate in Woburn for childhood leukemia}}{\text{U.S. incidence rate for childhood leukemia}} = \frac{10/10^5}{5/10^5} = 2$$

If the U.S. incidence rate is assumed to be known, then a 95% CI for SMR is given by  $\left(\frac{5.2}{5}, \frac{17.5}{5}\right) = (1.04, 3.50)$ . Because the lower bound of the CI for SMR is  $> 1$ , we conclude there is a significant excess risk in Woburn. We pursue a different approach in Chapter 7, addressing this issue in terms of hypothesis testing and  $p$ -values.

Another approach for obtaining exact CIs for the Poisson parameter  $\mu$  is to use the POISSON function of Excel in a similar manner as we used for BINOMDIST in Section 6.8 on page 187. This approach is useful if the observed number of events ( $x$ ) and/or if the desired level of confidence ( $1 - \alpha$ ) does not appear in Table 8. Specifically, Equation 6.23 can be written in the form

$$1 - \text{POISSON}(x - 1, \mu_1, \text{TRUE}) = \alpha/2$$

and

$$\text{POISSON}(x, \mu_2, \text{TRUE}) = \alpha/2$$

To solve these equations, we create columns of possible values for  $\mu_1$  and  $\mu_2$  and find the largest value of  $\mu_1$  and the smallest value of  $\mu_2$  (perhaps to 2 decimal places of accuracy) such that

$$1 - \text{POISSON}(x - 1, \mu_1, \text{TRUE}) \leq \alpha/2$$

and

$$\text{POISSON}(x, \mu_2, \text{TRUE}) \leq \alpha/2$$

The two-sided 100%  $\times (1 - \alpha)$  CI for  $\mu$  is then  $(\mu_1, \mu_2)$ .

In some instances, a random variable representing a rare event over time is assumed to follow a Poisson distribution but the actual amount of person-time is either unknown or is not reported in an article from the literature. In this instance, it is still possible to use Appendix Table 8 to obtain a CI for  $\mu$ , although it is impossible to obtain a CI for  $\lambda$ .

### Example 6.59

**Occupational Health** In Example 4.38, a study was described concerning the possible excess cancer risk among employees with high exposure to ethylene dibromide in two plants in Texas and Michigan. Seven deaths from cancer were reported over the period 1940–1975, whereas only 5.8 cancer deaths were expected based on mortality rates for U.S. white men. Find a 95% CI for the expected number of deaths among the exposed workers, and assess whether their risk differs from that of the general population.

### Solution

In this case, the actual number of person-years used in computing the expected number of deaths was not reported in the original article. Indeed, the computation of the expected number of deaths is complex because

- (1) Each worker is of a different age at the start of follow-up.
- (2) The age of a worker changes over time.
- (3) Mortality rates for men of the same age change over time.

However, we can use Appendix Table 8 to obtain a 95% CI for  $\mu$ . Because  $x = 7$  events, we have a 95% CI for  $\mu = (2.81, 14.42)$ . The expected number of deaths based on U.S.

mortality rates for white males = 5.8, which falls within the preceding interval. Thus we conclude the risk among exposed workers does not differ from the general population.

Table 8 can also be used for applications of the Poisson distribution other than those based specifically on rare events over time.

### Example 6.60

**Bacteriology** Suppose we observe 15 bacteria in a Petri dish and assume the number of bacteria is Poisson-distributed with parameter  $\mu$ . Find a 90% CI for  $\mu$ .

### Solution

We refer to the 15 row and the 0.90 column in Table 8 to obtain the 90% CI (9.25, 23.10).

## 6.10 One-Sided CIs

In the previous discussion of interval estimation, what are known as *two-sided CIs* have been described. Frequently, the following type of problem occurs.

### Example 6.61

**Cancer** A standard treatment exists for a certain type of cancer, and the patients receiving the treatment have a 5-year survival rate of 30%. A new treatment is proposed that has some unknown survival rate  $p$ . We would only be interested in using the new treatment if it were better than the standard treatment. Suppose that 40 out of 100 patients who receive the new treatment survive for 5 years. Can we say the new treatment is better than the standard treatment?

One way to analyze these data is to construct a one-sided CI, where we are interested in only *one* bound of the interval, in this case the lower bound. If 30% is below the lower bound, then it is an unlikely estimate of the 5-year survival rate for patients getting the new treatment. We could reasonably conclude from this that the new treatment is better than the standard treatment in this case.

### Equation 6.24

#### Upper One-Sided CI for the Binomial Parameter $p$ — Normal-Theory Method

An upper one-sided  $100\% \times (1 - \alpha)$  CI is of the form  $p > p_1$  such that

$$Pr(p > p_1) = 1 - \alpha$$

If we assume that the normal approximation to the binomial holds true, then we can show that this CI is given approximately by

$$p > \hat{p} - z_{1-\alpha} \sqrt{\hat{p}\hat{q}/n}$$

This interval estimator should only be used if  $n\hat{p}\hat{q} \geq 5$ .

To see this, note that if the normal approximation to the binomial distribution holds, then  $\hat{p} \sim N(p, pq/n)$ . Therefore, by definition

$$Pr(\hat{p} < p + z_{1-\alpha} \sqrt{pq/n}) = 1 - \alpha$$

We approximate  $\sqrt{pq/n}$  by  $\sqrt{\hat{p}\hat{q}/n}$  and subtract  $z_{1-\alpha} \sqrt{\hat{p}\hat{q}/n}$  from both sides of the equation, yielding

$$\hat{p} - z_{1-\alpha} \sqrt{\hat{p}\hat{q}/n} < p$$

$$\text{or } p > \hat{p} - z_{1-\alpha} \sqrt{\hat{p}\hat{q}/n} \text{ and } Pr(p > \hat{p} - z_{1-\alpha} \sqrt{\hat{p}\hat{q}/n}) = 1 - \alpha$$

Therefore, if the normal approximation to the binomial distribution holds, then  $p > \hat{p} - z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$  is an approximate upper  $100\% \times (1 - \alpha)$  one-sided CI for  $p$ .

Notice that  $z_{1-\alpha}$  is used in constructing one-sided intervals, whereas  $z_{1-\alpha/2}$  was used in constructing two-sided intervals.

**Example 6.62** Suppose a 95% CI for a binomial parameter  $p$  is desired. What percentile of the normal distribution should be used for a one-sided interval? For a two-sided interval?

**Solution** For  $\alpha = .05$ , we use  $z_{1-.05} = z_{.95} = 1.645$  for a one-sided interval and  $z_{1-.05/2} = z_{.975} = 1.96$  for a two-sided interval.

**Example 6.63** **Cancer** Construct an upper one-sided 95% CI for the survival rate based on the cancer-treatment data in Example 6.61.

**Solution** First check that  $n\hat{p}\hat{q} = 100(.4)(.6) = 24 \geq 5$ . The CI is then given by

$$Pr[p > .40 - z_{.95}\sqrt{.4(.6)/100}] = .95$$

$$Pr[p > .40 - 1.645(.049)] = .95$$

$$Pr(p > .319) = .95$$

Because .30 is not within the given interval [that is, (.319, 1.0)], we conclude the new treatment is better than the standard treatment.

If we were interested in 5-year death rates rather than survival rates, then a one-sided interval of the form  $Pr(p < p_2) = 1 - \alpha$  would be appropriate because we would only be interested in the new treatment if its death rate were lower than that of the standard treatment.

### Equation 6.25

#### Lower One-Sided CI for the Binomial Parameter $p$ — Normal-Theory Method

The interval  $p < p_2$  such that

$$Pr(p < p_2) = 1 - \alpha$$

is referred to as a **lower one-sided  $100\% \times (1 - \alpha)$  CI** and is given approximately by

$$p < \hat{p} + z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$$

This expression can be derived in the same manner as in Equation 6.24 by starting with the relationship

$$Pr(\hat{p} > p - z_{1-\alpha}\sqrt{pq/n}) = 1 - \alpha$$

If we approximate  $\sqrt{pq/n}$  by  $\sqrt{\hat{p}\hat{q}/n}$  and add  $z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$  to both sides of the equation, we get

$$Pr(p < \hat{p} + z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}) = 1 - \alpha$$

**Example 6.64** **Cancer** Compute a lower one-sided 95% CI for the death rate using the cancer-treatment data in Example 6.61.

**Solution** We have  $\hat{p} = .6$ . Thus the 95% CI is given by

$$Pr\left[p < .6 + 1.645\sqrt{.6(.4)/100}\right] = .95$$

$$Pr\left[p < .6 + 1.645(.049)\right] = .95$$

$$Pr(p < .681) = .95$$

Because 70% is not within this interval [that is, (0, .681)], we can conclude the new treatment has a lower death rate than the old treatment.

Similar methods can be used to obtain one-sided CIs for the mean and variance of a normal distribution, for the binomial parameter  $p$  using exact methods, and for the Poisson expectation  $\mu$  using exact methods.

## 6.11 Summary

This chapter introduced the concept of a sampling distribution. This concept is crucial to understanding the principles of statistical inference. The fundamental idea is to forget about our sample as a unique entity and instead regard it as a random sample from all possible samples of size  $n$  that could have been drawn from the population under study. Using this concept,  $\bar{X}$  was shown to be an unbiased estimator of the population mean  $\mu$ ; that is, the average of all sample means over all possible random samples of size  $n$  that could have been drawn will equal the population mean. Furthermore, if our population follows a normal distribution, then  $\bar{X}$  has minimum variance among all possible unbiased estimators and is thus called a *minimum-variance unbiased estimator* of  $\mu$ . Another justification for the use of  $\bar{X}$  to estimate  $\mu$  is that  $\bar{X}$  is the *maximum-likelihood estimator* (MLE) of  $\mu$ . An MLE is the value  $\hat{\mu}$ , which maximizes the  $Pr(\text{obtaining our sample} | \hat{\mu})$ . A similar definition holds for obtaining MLEs of parameters in other estimation problems. Finally, if our population follows a normal distribution, then  $\bar{X}$  also follows a normal distribution. However, even if our population is not normal, the sample mean still approximately follows a normal distribution for a sufficiently large sample size. This very important idea, which justifies many of the hypothesis tests we study in the rest of this book, is called the *central-limit theorem*.

The idea of an interval estimate (or CI) was then introduced. Specifically, a 95% CI is defined as an interval that will contain the true parameter for 95% of all random samples that could have been obtained from the reference population. The preceding principles of point and interval estimation were applied to the following:

- (1) Estimating the mean  $\mu$  of a normal distribution
- (2) Estimating the variance  $\sigma^2$  of a normal distribution
- (3) Estimating the parameter  $p$  of a binomial distribution
- (4) Estimating the parameter  $\lambda$  of a Poisson distribution
- (5) Estimating the expected value  $\mu$  of a Poisson distribution

The  $t$  and chi-square distributions were introduced to obtain interval estimates for (1) and (2), respectively.

In Chapters 7 through 14, the discussion of statistical inference continues, focusing primarily on testing hypotheses rather than on parameter estimation. In this regard, some parallels between inference from the points of view of hypothesis testing and CIs are discussed.

## PROBLEMS

### Gastroenterology

Suppose we are asked to construct a list of treatment assignments for patients entering a study comparing different treatments for duodenal ulcer.

**6.1** Anticipating that 20 patients will be entered in the study and two treatments will be used, construct a list of random-treatment assignments starting in the 28th row of the random-number table (Table 4 in the Appendix).

**6.2** Count the number of people assigned to each treatment group. How does this number compare with the expected number in each group?

**6.3** Suppose we change our minds and decide to enroll 40 patients and use four treatment groups. Use a computer program (such as MINITAB or Excel) to construct the list of random-treatment assignments referred to in Problem 6.1.

**6.4** Answer Problem 6.2 for the list of treatment assignments derived in Problem 6.3.

### Pulmonary Disease

The data in Table 6.8 concern the mean triceps skin-fold thickness in a group of normal men and a group of men with chronic airflow limitation [5].

**Table 6.8 Triceps skin-fold thickness in normal men and men with chronic airflow limitation**

Group	Mean	sd	n
Normal	1.35	0.5	40
Chronic airflow limitation	0.92	0.4	32

Source: Reprinted with permission of *Chest*, 85(6), 58S–59S, 1984.

**\*6.5** What is the standard error of the mean for each group?

**6.6** Assume that the central-limit theorem is applicable. What does it mean in this context?

**6.7** Find the upper 1st percentile of a *t* distribution with 16 *df*.

**6.8** Find the lower 10th percentile of a *t* distribution with 28 *df*.

**6.9** Find the upper 2.5th percentile of a *t* distribution with 7 *df*.

**6.10** What are the upper and lower 2.5th percentiles for a chi-square distribution with 2 *df*? What notation is used to denote these percentiles?

Refer to the data in Table 2.11. Regard this hospital as typical of Pennsylvania hospitals.

**6.11** Compute a 95% CI for the mean duration of hospitalization.

**6.12** Compute a 95% CI for the mean white blood count following admission.

**6.13** Answer Problem 6.12 for a 90% CI.

**6.14** What is the relationship between your answers to Problems 6.12 and 6.13?

**\*6.15** What is the best point estimate of the percentage of males among patients discharged from Pennsylvania hospitals?

**\*6.16** What is the standard error of the estimate obtained in Problem 6.15?

**\*6.17** Provide a 95% CI for the percentage of males among patients discharged from Pennsylvania hospitals.

### Microbiology

A nine-laboratory cooperative study was performed to evaluate quality control for susceptibility tests with 30- $\mu\text{g}$  netilmicin disks [6]. Each laboratory tested three standard control strains on a different lot of Mueller-Hinton agar, with 150 tests performed per laboratory. For protocol control, each laboratory also performed 15 additional tests on each of the control strains using the same lot of Mueller-Hinton agar across laboratories. The mean zone diameters for each of the nine laboratories are given in Table 6.9.

**\*6.18** Provide a point and interval estimate (95% CI) for the mean zone diameter across laboratories for each type of control strain, if each laboratory uses different media to perform the susceptibility tests.

**\*6.19** Answer Problem 6.18 if each laboratory uses a common medium to perform the susceptibility tests.

**\*6.20** Provide a point and interval estimate (95% CI) for the interlaboratory standard deviation of mean zone diameters for each type of control strain, if each laboratory uses different media to perform the susceptibility tests.

**\*6.21** Answer Problem 6.20 if each laboratory uses a common medium to perform the susceptibility tests.

**6.22** Are there any advantages to using a common medium versus using different media for performing the susceptibility tests with regard to standardization of results across laboratories?

### Renal Disease

A study of psychological and physiological health in a cohort of dialysis patients with end-stage renal disease was conducted [7]. Psychological and physiological parameters were initially determined at baseline in 102 patients; these parameters were determined again in 69 of the 102 patients at an 18-month follow-up visit. The data in Table 6.10 were reported.

**6.23** Provide a point and interval estimate (95% CI) for the mean of each parameter at baseline and follow-up.

**Table 6.9** Mean zone diameters with 30- $\mu\text{g}$  netilmicin disks tested in nine separate laboratories

Laboratory	Type of control strain					
	<i>E. coli</i>		<i>S. aureus</i>		<i>P. aeruginosa</i>	
	Different media	Common medium	Different media	Common medium	Different media	Common medium
A	27.5	23.8	25.4	23.9	20.1	16.7
B	24.6	21.1	24.8	24.2	18.4	17.0
C	25.3	25.4	24.6	25.0	16.8	17.1
D	28.7	25.4	29.8	26.7	21.7	18.2
E	23.0	24.8	27.5	25.3	20.1	16.7
F	26.8	25.7	28.1	25.2	20.3	19.2
G	24.7	26.8	31.2	27.1	22.8	18.8
H	24.3	26.2	24.3	26.5	19.9	18.1
I	24.9	26.3	25.4	25.1	19.3	19.2

**Table 6.10** Psychological and physiological parameters in patients with end-stage renal disease

Variable	Baseline ( <i>n</i> = 102)		18-month follow-up ( <i>n</i> = 69)	
	Mean	sd	Mean	sd
Serum creatinine (mmol/L)	0.97	0.22	1.00	0.19
Serum potassium (mmol/L)	4.43	0.64	4.49	0.71
Serum phosphate (mmol/L)	1.68	0.47	1.57	0.40
Psychological Adjustment to Illness (PAIS) scale	36.50	16.08	23.27	13.79

**6.24** Do you have any opinion on the physiological and psychological changes in this group of patients? Explain. (Note: A lower score on the PAIS scale indicates worse adjustment to illness.)

### Ophthalmology, Hypertension

A study is conducted to test the hypothesis that people with glaucoma have higher-than-average blood pressure. The study includes 200 people with glaucoma whose mean SBP is 140 mm Hg with a standard deviation of 25 mm Hg.

**6.25** Construct a 95% CI for the true mean SBP among people with glaucoma.

**6.26** If the average SBP for people of comparable age is 130 mm Hg, is there an association between glaucoma and blood pressure?

### Sexually Transmitted Disease

Suppose a clinical trial is conducted to test the efficacy of a new drug, spectinomycin, for treating gonorrhea in females. Forty-six patients are given a 4-g daily dose of the drug and are seen 1 week later, at which time 6 of the patients still have gonorrhea.

\***6.27** What is the best point estimate for  $p$ , the probability of a failure with the drug?

\***6.28** What is a 95% CI for  $p$ ?

\***6.29** Suppose we know penicillin G at a daily dose of 4.8 megaunits has a 10% failure rate. What can be said in comparing the two drugs?

### Pharmacology

Suppose we want to estimate the concentration ( $\mu\text{g}/\text{mL}$ ) of a specific dose of ampicillin in the urine after various periods of time. We recruit 25 volunteers who have received ampicillin and find they have a mean concentration of 7.0  $\mu\text{g}/\text{mL}$  with a standard deviation of 2.0  $\mu\text{g}/\text{mL}$ . Assume the underlying population distribution of concentrations is normally distributed.

\***6.30** Find a 95% CI for the population mean concentration.

\***6.31** Find a 99% CI for the population variance of the concentrations.

\***6.32** How large a sample would be needed to ensure that the length of the CI in Problem 6.30 is 0.5  $\mu\text{g}/\text{mL}$  if we assume the sample standard deviation remains at 2.0  $\mu\text{g}/\text{mL}$ ?

## Environmental Health

Much discussion has taken place concerning possible health hazards from exposure to anesthetic gases. In one study conducted in 1972, 525 Michigan nurse anesthetists were surveyed by mail questionnaires and telephone interviews to determine the incidence rate of cancer [8]. Of this group, 7 women reported having a new malignancy other than skin cancer during 1971.

**6.33** What is the best estimate of the 1971 incidence rate from these data?

**6.34** Provide a 95% CI for the true incidence rate.

A comparison was made between the Michigan report and the 1969 cancer-incidence rates from the Connecticut tumor registry, where the expected incidence rate, based on the age distribution of the Michigan nurses, was determined to be 402.8 per 100,000 person-years.

**6.35** Comment on the comparison between the observed incidence rate and the Connecticut tumor-registry data.

## Obstetrics, Serology

A new assay is developed to obtain the concentration of *M. hominis* mycoplasma in the serum of pregnant women. The developers of this assay want to make a statement on the variability of their laboratory technique. For this purpose, 10 subsamples of 1 mL each are drawn from a large serum sample for one woman, and the assay is performed on each subsample. The concentrations are as follows:  $2^4, 2^3, 2^5, 2^4, 2^5, 2^4, 2^3, 2^4, 2^4, 2^5$ .

**\*6.36** If the distribution of concentrations in the log scale to the base 2 is assumed to be normal, then obtain the best estimate of the variance of the concentrations from these data.

**\*6.37** Compute a 95% CI for the variance of the concentrations.

**\*6.38** Assuming the point estimate in Problem 6.36 is the true population parameter, what is the probability that a particular assay, when expressed in the log scale to the base 2, is no more than 1.5 log units off from its true mean value for a particular woman?

**\*6.39** Answer Problem 6.38 for 2.5 log units.

## Hypertension

Suppose 100 hypertensive people are given an antihypertensive drug and the drug is effective in 20 of them. By effective, we mean their DBP is lowered by at least 10 mm Hg as judged from a repeat blood-pressure measurement 1 month after taking the drug.

**6.40** What is the best point estimate of the probability  $p$  of the drug being effective?

**6.41** Suppose we know that 10% of all hypertensive patients who are given a placebo will have their DBP lowered

by 10 mm Hg after 1 month. Can we carry out some procedure to be sure we are not simply observing the placebo effect?

**6.42** What assumptions have you made to carry out the procedure in Problem 6.41?

Suppose we decide a better measure of the effectiveness of the drug is the mean decrease in blood pressure rather than the measure of effectiveness used previously. Let  $d_i = x_i - y_i, i = 1, \dots, 100$ , where  $x_i$  = DBP for the  $i$ th person before taking the drug and  $y_i$  = DBP for the  $i$ th person 1 month after taking the drug. Suppose the sample mean of the  $d_i$  is +5.3 and the sample variance is 144.0.

**6.43** What is the standard error of  $\bar{d}$ ?

**6.44** What is a 95% CI for the population mean of  $d$ ?

**6.45** Can we make a statement about the effectiveness of the drug?

**6.46** What does a 95% CI mean, in words, in this case?

## SIMULATION

Draw six random samples of size 5 from the data in Table 6.2.

**6.47** Compute the mean birthweight for each of the six samples.

**6.48** Compute the standard deviation based on the sample of six means. What is another name for this quantity?

**6.49** Select the third point from each of the six samples, and compute the sample  $sd$  from the collection of six third points.

**6.50** What theoretical relationship should there be between the standard deviation in Problem 6.48 and the standard deviation in Problem 6.49?

**6.51** How do the actual sample results in Problems 6.48 and 6.49 compare?

## Obstetrics

Figure 6.4b plotted the sampling distribution of the mean from 200 samples of size 5 from the population of 1000 birthweights given in Table 6.2. The mean of the 1000 birthweights in Table 6.2 is 112.0 oz with standard deviation 20.6 oz.

**\*6.52** If the central-limit theorem holds, what proportion of the sample means should fall within 0.5 lb of the population mean (112.0 oz)?

**\*6.53** Answer Problem 6.52 for 1 lb rather than 0.5 lb.

**\*6.54** Compare your results in Problems 6.52 and 6.53 with the actual proportion of sample means that fall in these ranges.

**\*6.55** Do you feel the central-limit theorem is applicable for samples of size 5 from this population? Explain.

## Hypertension, Pediatrics

The etiology of high blood pressure remains a subject of active investigation. One widely accepted hypothesis is that excessive sodium intake adversely affects blood-pressure outcomes. To explore this hypothesis, an experiment was set up to measure responsiveness to the taste of salt and to relate the responsiveness to blood-pressure level. The protocol used involved giving 3-day-old infants in the newborn nursery a drop of various solutions, thus eliciting the sucking response and noting the vigor with which they sucked—denoted by MSB (mean number of sucks per burst of sucking). The content of the solution was changed over 10 consecutive periods: (1) water, (2) water, (3) 0.1 molar salt + water, (4) 0.1 molar salt + water, (5) water, (6) water, (7) 0.3 molar salt + water, (8) 0.3 molar salt + water, (9) water, (10) water. In addition, as a control, the response of the baby to the taste of sugar was also measured after the salt-taste protocol was completed. In this experiment, the sucking response was measured over five different periods with the following stimuli: (1) nonnutritive sucking, that is, a pure sucking response was elicited without using any external substance; (2) water; (3) 5% sucrose + water; (4) 15% sucrose + water; (5) nonnutritive sucking.

The data for the first 100 infants in the study are given in Data Set INFANTBP.DAT, on the Companion Website. The format of the data is given in Data Set INFANTBP.DOC, on the Companion Website.

Construct a variable measuring the response to salt. For example, one possibility is to compute the average MSB for trials 3 and 4 – average MSB for trials 1 and 2 = average MSB when the solution was 0.1 molar salt + water – average MSB when the solution was water. A similar index could be computed comparing trials 7 and 8 with trials 5 and 6.

**6.56** Obtain descriptive statistics and graphic displays for these salt-taste indices. Do the indices appear to be normally distributed? Why or why not? Compute the sample mean for this index, and obtain 95% CIs about the point estimate.

**6.57** Construct indices measuring responsiveness to sugar taste, and provide descriptive statistics and graphical displays for these indices. Do the indices appear normally distributed? Why or why not? Compute the sample mean and associated 95% CIs for these indices.

**6.58** We want to relate the indices to blood-pressure level. Provide a scatter plot relating mean SBP and mean DBP, respectively, to each of the salt-taste and sugar-taste indices. Does there appear to be a relation between the indices and blood-pressure level? We discuss this in more detail in our work on regression analysis in Chapter 11.

## Genetics

Data Set SEXRAT.DAT, on the Companion Website, lists the sexes of children born in over 50,000 families with more than one child.

**6.59** Use interval-estimation methods to determine if the sex of successive births is predictable from the sex of previous births.

## SIMULATION

### Nutrition

Data Set VALID.DAT, on the Companion Website, provides estimated daily consumption of total fat, saturated fat, and alcohol as well as total caloric intake using two different methods of dietary assessment for 173 subjects.

**6.60** Use a computer to draw repeated random samples of size 5 from this population. Does the central-limit theorem seem to hold for these dietary attributes based on samples of size 5?

**6.61** Answer Problem 6.60 for random samples of size 10.

**6.62** Answer Problem 6.60 for random samples of size 20.

**6.63** How do the sampling distributions compare based on samples of size 5, 10, and 20? Use graphic and numeric methods to answer this question.

## Infectious Disease

A cohort of hemophiliacs is followed to elicit information on the distribution of time to onset of AIDS following seroconversion (referred to as *latency time*). All patients who seroconvert become symptomatic within 10 years, according to the distribution in Table 6.11.

**Table 6.11 Latency time to AIDS among hemophiliacs who become HIV positive**

Latency time (years)	Number of patients
0	2
1	6
2	9
3	33
4	49
5	66
6	52
7	37
8	18
9	11
10	4

**6.64** Assuming an underlying normal distribution, compute 95% CIs for the mean and variance of the latency times.

**6.65** Still assuming normality, estimate the probability  $p$  that a patient's latency time will be at least 8 years.

**6.66** Now suppose we are unwilling to assume a normal distribution for latency time. Re-estimate the probability  $p$  that a patient's latency time will be at least 8 years, and provide a 95% CI for  $p$ .

### Environmental Health

We have previously described Data Set LEAD.DAT (on the Companion Website), in which children were classified according to blood-lead level in 1972 and 1973 by the variable lead\_group, where 1 = blood-lead level < 40 µg/100 mL in both 1972 and 1973, 2 = blood-lead level ≥ 40 µg/100 mL in 1973, 3 = blood-lead level > 40 µg/100 mL in 1972 but < 40 µg/100 mL in 1973.

**6.67** Compute the mean, standard deviation, standard error, and 95% CI for the mean verbal IQ for children with specific values of the variable GROUP. Provide a box plot comparing the distribution of verbal IQ for subjects with lead\_group = 1, 2, and 3. Summarize your findings concisely.

**6.68** Answer Problem 6.67 for performance IQ.

**6.69** Answer Problem 6.67 for full-scale IQ.

### Cardiology

Data Set NIFED.DAT (on the Companion Website) was described earlier. We wish to look at the effect of each treatment separately on heart rate and systolic blood pressure (SBP).

**6.70** Provide separate point estimates and 95% CIs for the changes in heart rate and SBP (level 1 to baseline) for the subjects randomized to nifedipine and propranolol, respectively. Also provide box plots of the change scores in the two treatment groups.

**6.71** Answer Problem 6.70 for level 2 to baseline.

**6.72** Answer Problem 6.70 for level 3 to baseline.

**6.73** Answer Problem 6.70 for the last available level to baseline.

**6.74** Answer Problem 6.70 for the average heart rate (or blood pressure) over all available levels to baseline.

### Occupational Health

**\*6.75** Refer to Problem 4.23. Provide a 95% CI for the expected number of deaths from bladder cancer over 20 years among tire workers. Is the number of cases of bladder cancer in this group excessive?

**\*6.76** Refer to Problem 4.24. Provide a 95% CI for the expected number of deaths from stomach cancer over 20 years among tire workers. Is the number of cases of stomach cancer in this group excessive?

### Cancer

The value of mammography as a screening test for breast cancer has been controversial, particularly among young women. A study was recently performed looking at the rate of false positives for repeated screening mammograms among approximately 10,000 women who were members of Harvard Pilgrim Health Care, a large health-maintenance organization in New England [9].

The study reported that of a total of 1996 tests given to 40- to 49-year-old women, 156 yielded false-positive results.

**6.77** What does a false-positive test result mean, in words, in this context?

**6.78** Some physicians feel a mammogram is not cost-effective unless one can be reasonably certain (e.g., 95% certain) that the false-positive rate is less than 10%. Can you address this issue based on the preceding data? (*Hint:* Use a CI approach.)

**6.79** Suppose a woman is given a mammogram every 2 years starting at age 40. What is the probability that she will have at least one false-positive test result among 5 screening tests during her forties? (Assume the repeated screening tests are independent.)

**6.80** Provide a two-sided 95% CI for the probability estimate in Problem 6.79.

### SIMULATION

#### Nutrition

On the computer, we draw 500 random samples of size 5 from the distribution of 173 values of  $\ln(\text{alcohol DR} [\text{diet record}] + 1)$  in the Data Set VALID.DAT, where Alcoh\_dr is the amount of alcohol consumed as reported by diet record by a group of 173 American nurses who recorded each food eaten on a real-time basis, over four 1-week periods spaced approximately 3 months apart over the course of 1 year. For each sample of size 5, we compute the sample mean  $\bar{x}$ , the sample standard deviation  $s$ , and the test statistic  $t$  given by

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where  $n = 5$  and  $\mu_0$  = overall mean of  $\ln(\text{alcohol DR} + 1)$  over the 173 nurses = 1.7973.

**6.81** What distribution should the  $t$ -values follow if the central-limit theorem holds? Assume  $\mu_0$  is the population mean for  $\ln(\text{Alcoh\_dr} + 1)$ .

**6.82** If the central-limit theorem holds, then what percentage of  $t$ -values should exceed 2.776 in absolute value?

**6.83** The actual number of  $t$ -values that exceed 2.776 in absolute value is 38. Do you feel the central-limit theorem is applicable to these data for samples of size 5?

### Cardiovascular Disease

A study was performed to investigate the variability of cholesterol and other lipid measures in children. The reported within-subject standard deviation for cholesterol in children was 7.8 mg/dL [10].

**6.84** Suppose two total cholesterol determinations are obtained from one child, yielding an average value of 200 mg/dL. What is a two-sided 90% CI for the true mean total cholesterol for that child? (*Hint:* Assume the sample

standard deviation of cholesterol for the child is known to be 7.8 mg/dL.)

**6.85** Suppose an average of two total cholesterol determinations is to be used as a screening tool to identify children with high cholesterol. The investigators wish to find a value  $c$ , such that all children whose mean cholesterol values over two determinations are  $\geq c$  will be called back for further screening, whereas children whose mean cholesterol values are  $< c$  will not be followed any further. To determine  $c$ , the investigators want to choose a value  $c$  such that the lower one-sided 90% CI for  $\mu$  if the observed average cholesterol over two determinations =  $c$  would exclude 250 mg/dL. What is the largest value of  $c$  that satisfies this requirement?

### Endocrinology

Refer to Data Set BONEDEN.DAT on the Companion Website.

**6.86** Assess whether there is a relationship between BMD at the femoral neck and cigarette smoking using CI methodology. (*Hint:* Refer to Section 6.6.)

**6.87** Assess whether there is a relationship between BMD at the femoral shaft and cigarette smoking using CI methodology. (*Hint:* Refer to Section 6.6.)

### SIMULATION

**6.88** Using the computer, generate 200 random samples from a binomial distribution with  $n = 10$  and  $p = .6$ . Derive a large sample 90% CI for  $p$  based on each sample.

**6.89** What percentage of the CIs include the parameter  $p$ ?

**6.90** Do you think that the large-sample binomial confidence-limit formula is adequate for this distribution?

**6.91** Answer Problem 6.88 for a binomial distribution with  $n = 20$  and  $p = .6$ .

**6.92** Answer Problem 6.89 for a binomial distribution with  $n = 20$  and  $p = .6$ .

**6.93** Answer Problem 6.90 for a binomial distribution with  $n = 20$  and  $p = .6$ .

**6.94** Answer Problem 6.88 for a binomial distribution with  $n = 50$  and  $p = .6$ .

**6.95** Answer Problem 6.89 for a binomial distribution with  $n = 50$  and  $p = .6$ .

**6.96** Answer Problem 6.90 for a binomial distribution with  $n = 50$  and  $p = .6$ .

### Hypertension

A patient who is taking antihypertensive medication is asked by her doctor to record her blood pressure at home to check that it is in the normotensive range. On each of 10 days, she took an average of two readings, with results as shown in Table 6.12.

**Table 6.12 Home blood pressure recordings for one patient**

Day	SBP (mm Hg)	DBP (mm Hg)
1	121	87.5
2	109	81
3	117.5	91.5
4	125	94
5	125	87.5
6	129	90.5
7	123	90
8	118.5	85.5
9	123.5	87.5
10	127	89
Mean	121.85	88.40
sd	5.75	3.56
n	10	10

The doctor wants to assess whether the underlying mean SBP for this woman is  $< 140$  or  $\geq 140$  mm Hg.

**6.97** Provide a 95% CI for true mean SBP for this patient.

**6.98** Answer the doctor's question given the result in Problem 6.97.

Another issue the doctor wants to study is what the hypertensive status of the patient usually is. A person is classified as hypertensive on any one day if either his or her SBP is  $\geq 140$  mm Hg or his or her DBP is  $\geq 90$  mm Hg.

**6.99** What proportion of days would the woman be classified as hypertensive based on the preceding data?

A person would be classified as hypertensive overall if his or her probability of being hypertensive on an individual day ( $p$ ) is  $\geq 20\%$  based on a large number of days.

**6.100** Develop 95% CI for  $p$  based on your answer to Problem 6.99. (*Hint:* Use Appendix Table 7a.)

**6.101** Would the person be classified as hypertensive overall based on your answer to Problem 6.100? Why or why not? Explain your answer.

### Sports Medicine

Injuries are common in football and may be related to a number of factors, including the type of playing surface, the number of years of playing experience, and whether any previous injury exists. A study of factors affecting injury among Canadian football players was recently reported [11].

The rate of injury to the upper extremity (that is, shoulder to hand) on a dry field consisting of natural grass was 2.73 injuries per 1000 games. Assume this rate is known without error.

**6.102** The study reported 45 injuries to the upper extremity on a dry field consisting of artificial turf over the course of 10,112 games. What procedure can be used to assess

whether the risk of injury is different on artificial turf versus natural grass?

**6.103** Provide a 95% CI for the rate of injury to the upper extremity on artificial turf. (*Hint:* Use the Poisson distribution.) Express each rate as the number of injuries per 1000 games.

### Hypertension

A hypertensive patient has been on antihypertensive medication for several years. Her physician wants to monitor her blood pressure via weekly measurements taken at home. Each week for 6 weeks she takes several blood pressure readings and averages the readings to get a summary blood pressure for the week. The diastolic blood pressure (DBP) results are shown in Table 6.13.

**Table 6.13 Weekly mean DBP readings for an individual patient**

Week	Mean DBP (mm Hg)	Week	Mean DBP (mm Hg)
1	89	4	84
2	88	5	82
3	81	6	89.5
		Mean	85.75
		sd	3.66

**6.104** Her doctor is considering taking her off antihypertensive medications but wants to be fairly certain that her “true” DBP is less than 90 mm Hg. Use a statistical approach to answer this question. (*Hint:* Consider a CI approach.)

The doctor takes the patient off antihypertensive medication and instructs her to measure her blood pressure for 3 consecutive weeks. The doctor will put the patient back on antihypertensive medication if her mean DBP over the 3 weeks is  $\geq 90$  mm Hg.

**6.105** Suppose there is no real change in the patient’s underlying mean blood pressure regardless of whether she is on medication. What is the probability that she will be put back on antihypertensive medication? (*Hint:* Assume that the true mean and standard deviation of DBP for the patient are the same as the measured mean and standard deviation over the 6 weeks while the patient is on antihypertensive medication.)

**6.106** Suppose we observe  $x$  events over  $T$  years for a random variable  $X$  that is Poisson-distributed with expected value =  $\mu$ .

- (a) Show that  $x$  is the MLE of  $\mu$ .
- (b) Suppose  $\lambda$  is the number of events per year. What is the MLE of  $\lambda$ ?

Suppose we have a population with a normal distribution with mean = 50 and standard deviation = 10.

We draw a sample of 13 observations from this distribution.

**6.107** What is the probability that the sample mean will be within 1 unit of the population mean?

**6.108** Suppose we want to choose a large enough sample so that the sample mean is within 1 unit of the population mean 99% of the time. What is the minimum sample size to achieve this goal?

### Radiology, Cancer

A radiologist investigates whether a new (less costly) method for identifying esophageal cancer is as effective as the gold standard.

He obtains the following test results: false positive = 0, true positive = 46, false negative = 1, true negative = 17.

**6.109** What is the sensitivity of the test?

**6.110** Provide a 95% CI for the sensitivity (two decimal place accuracy is sufficient). (*Hint:* The following spreadsheet might be useful.)

p	BINOMDIST(45,47,p,TRUE)	BINOMDIST(46,47,p,TRUE)
0.80	1.000	1.000
0.81	0.999	1.000
0.82	0.999	1.000
0.83	0.998	1.000
0.84	0.997	1.000
0.85	0.996	1.000
0.86	0.993	0.999
0.87	0.988	0.999
0.88	0.982	0.998
0.89	0.972	0.996
0.90	0.956	0.993
0.91	0.933	0.988
0.92	0.899	0.980
0.93	0.850	0.967
0.94	0.782	0.945
0.95	0.688	0.910
0.96	0.566	0.853
0.97	0.414	0.761
0.98	0.242	0.613
0.99	0.080	0.376
0.995	0.023	0.210
0.999	0.001	0.046
0.9995	0.000	0.023

*Note:* These data were provided by Dr. Ori Preis.

### Genetics

The estimation of allele probabilities is essential for the closer quantitative identification of inheritance. It requires the probabilistic formulation of the applied model of inheritance. The hereditary disease phenylketonuria (PKU) is a useful example. PKU follows a recessive form of inheritance.

Suppose there are two alleles at a gene locus denoted by  $a$  and  $A$  where the possible genotypes are  $(aa)$ ,  $(aA)$ , and  $(AA)$ . An individual will only be affected if the genotype  $aa$  appears (i.e., a recessive form of inheritance).

**6.111** Suppose the probability of an  $a$  allele is  $p$ . If people mate randomly, then what is the probability of the  $(aa)$  genotype?

Suppose that on a population level it is impossible to genotype large numbers of individuals. However, it is known that among 10,000 people surveyed in the population, 11 have the PKU clinical phenotype.

**6.112** Provide a point estimate and 95% CI for the probability of having the PKU phenotype.

**6.113** Provide a point estimate and 95% CI for the  $a$  allele frequency  $p$ .

As an experiment, 10,000 people are completely genotyped, of whom 10 have the  $(aa)$  genotype, 630 have the  $(aA)$  genotype [i.e., either  $(aA)$  or  $(Aa)$ ], and 9360 have the  $(AA)$  genotype.

**6.114** If the two alleles of an individual are independent random variables, then provide a point estimate and a 95% CI for the  $a$  allele frequency  $p$ .

**6.115** Does genotyping a population provide more accurate estimates of  $p$  than obtained by only having the clinical phenotype? Why or why not?

## REFERENCES

- [1] Cochran, W. G. (1963). *Sampling techniques* (2nd ed.). New York: Wiley.
- [2] SHEP Cooperative Research Group. (1991). Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension: Final results of the Systolic Hypertension in the Elderly Program (SHEP). *JAMA*, 265(24), 3255–3264.
- [3] Mood, A., & Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.
- [4] *Documenta Geigy scientific tables*, 8th ed., vol. 2. (1982). Basel: Ciba-Geigy.
- [5] Arora, N. S., & Rochester, D. R (1984). Effect of chronic airflow limitation (CAL) on sternocleidomastoid muscle thickness. *Chest*, 85(6), 58S–59S.
- [6] Barry, A. L., Gavan, T. L., & Jones, R. N. (1983). Quality control parameters for susceptibility data with 30 µg netilmicin disks. *Journal of Clinical Microbiology*, 18(5), 1051–1054.
- [7] Oldenburg, B., Macdonald, G. J., & Perkins, R. J. (1988). Prediction of quality of life in a cohort of end-stage renal disease patients. *Journal of Clinical Epidemiology*, 41(6), 555–564.
- [8] Corbett, T. H., Cornell, R. G., Leiding, K., & Endres, J. L. (1973). Incidence of cancer among Michigan nurse-anesthetists. *Anesthesiology*, 38(3), 260–263.
- [9] Elmore, J. G., Barton, M. B., Moceri, V. M., Polk, S., Arena, P. J., & Fletcher, S. W. (1998). Ten year risk of false positive screening mammograms and clinical breast examinations. *New England Journal of Medicine*, 338, 1089–1096.
- [10] Elveback, L. R., Weidman, W. H., & Ellefson, R. D. (1980). Day to day variability and analytic error in determination of lipids in children. *Mayo Clinic Proceedings*, 55, 267–269.
- [11] Hagel, B. E., Fick, G. H., & Meeuwisse, W. H. (2003). Injury risk in men's Canada West University Football. *American Journal of Epidemiology*, 157, 825–833.

# Hypothesis Testing: One-Sample Inference

## 7.1 Introduction

Chapter 6 discussed methods of point and interval estimation for parameters of various distributions. However, researchers often have preconceived ideas about what the values of these parameters might be and wish to test whether the data conform to these ideas.

### Example 7.1

**Cardiovascular Disease, Pediatrics** A current area of research interest is the familial aggregation of cardiovascular risk factors in general and lipid levels in particular. Suppose the “average” cholesterol level in children is 175 mg/dL. A group of men who have died from heart disease within the past year are identified, and the cholesterol levels of their offspring are measured. Two hypotheses are considered:

- (1) The average cholesterol level of these children is 175 mg/dL.
- (2) The average cholesterol level of these children is  $>175$  mg/dL.

This type of question is formulated in a hypothesis-testing framework by specifying two hypotheses—a null and an alternative hypothesis. We wish to compare the relative probabilities of obtaining the sample data under each of these hypotheses. In Example 7.1, the null hypothesis is that the average cholesterol level of the children is 175 mg/dL and the alternative hypothesis is that the average cholesterol level of the children is  $>175$  mg/dL.

Why is hypothesis testing so important? Hypothesis testing provides an objective framework for making decisions using probabilistic methods, rather than relying on subjective impressions. People can form different opinions by looking at data, but a hypothesis test provides a uniform decision-making criterion that is consistent for all people.

In this chapter, some of the basic concepts of hypothesis testing are developed and applied to one-sample problems of statistical inference. In a **one-sample problem**, hypotheses are specified about a single distribution; in a **two-sample problem**, two different distributions are compared.

## 7.2 General Concepts

### Example 7.2

**Obstetrics** Suppose we want to test the hypothesis that mothers with low socio-economic status (SES) deliver babies whose birthweights are lower than “normal.” To test this hypothesis, a list is obtained of birthweights from 100 consecutive, full-term, live-born deliveries from the maternity ward of a hospital in a low-SES area. The mean birthweight ( $\bar{x}$ ) is found to be 115 oz with a sample standard deviation ( $s$ ) of 24 oz. Suppose we know from nationwide surveys based on millions of deliveries

that the mean birthweight in the United States is 120 oz. Can we actually say the underlying mean birthweight from this hospital is lower than the national average?

Assume the 100 birthweights from this hospital come from an underlying normal distribution with unknown mean  $\mu$ . The methods in Section 6.10 could be used to construct a 95% lower one-sided confidence interval (CI) for  $\mu$  based on the sample data—that is, an interval of the form  $\mu < c$ . If this interval contains 120 oz (that is,  $c \geq 120$ ), then the hypothesis that the mean birthweight in this hospital is the same as the national average would be accepted. If the interval does not contain 120 oz ( $c < 120$ ), then the hypothesis that the mean birthweight in this hospital is lower than the national average would be accepted.

Another way of looking at this problem is in terms of hypothesis testing. In particular, the hypotheses being considered can be formulated in terms of null and alternative hypotheses, which can be defined as follows:

### Definition 7.1

The **null hypothesis**, denoted by  $H_0$ , is the hypothesis that is to be tested. The alternative hypothesis, denoted by  $H_1$ , is the hypothesis that in some sense contradicts the null hypothesis.

### Example 7.3

**Obstetrics** In Example 7.2, the null hypothesis ( $H_0$ ) is that the mean birthweight in the low-SES-area hospital ( $\mu$ ) is equal to the mean birthweight in the United States ( $\mu_0$ ). This is the hypothesis we want to test. The alternative hypothesis ( $H_1$ ) is that the mean birthweight in this hospital ( $\mu$ ) is lower than the mean birthweight in the United States ( $\mu_0$ ). We want to compare the relative probabilities of obtaining the sample data under each of these two hypotheses.

We also assume the underlying distribution is normal under either hypothesis. These hypotheses can be written more succinctly in the following form:

### Equation 7.1

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu < \mu_0$$

Suppose the only possible decisions are whether  $H_0$  is true or  $H_1$  is true. Actually, for ease of notation, all outcomes in a hypothesis-testing situation generally refer to the null hypothesis. Hence, if we decide  $H_0$  is true, then we say we accept  $H_0$ . If we decide  $H_1$  is true, then we state that  $H_0$  is not true or, equivalently, that we reject  $H_0$ . Thus four possible outcomes can occur:

- (1) We accept  $H_0$ , and  $H_0$  is in fact true.
- (2) We accept  $H_0$ , and  $H_1$  is in fact true.
- (3) We reject  $H_0$ , and  $H_0$  is in fact true.
- (4) We reject  $H_0$ , and  $H_1$  is in fact true.

These four possibilities are shown in Table 7.1.

**Table 7.1**

**Four possible outcomes in hypothesis testing**

Decision	Truth	
	$H_0$	$H_1$
Accept $H_0$	$H_0$ is true and $H_0$ is accepted	$H_1$ is true and $H_0$ is accepted
Reject $H_0$	$H_0$ is true and $H_0$ is rejected	$H_1$ is true and $H_0$ is rejected

In actual practice, it is impossible, using hypothesis-testing methods, to *prove that the null hypothesis is true*. Thus, in particular, if we *accept  $H_0$* , then we have actually *failed to reject  $H_0$* .

If  $H_0$  is true and  $H_0$  is accepted, or if  $H_1$  is true and  $H_0$  is rejected, then the correct decision has been made. If  $H_0$  is true and  $H_0$  is rejected, or if  $H_1$  is true and  $H_0$  is accepted, then an *error* has been made. The two types of errors are generally treated differently.

#### Definition 7.2

The probability of a **type I error** is the probability of rejecting the null hypothesis when  $H_0$  is true.

#### Definition 7.3

The probability of a **type II error** is the probability of accepting the null hypothesis when  $H_1$  is true. This probability is a function of  $\mu$  as well as other factors.

#### Example 7.4

**Obstetrics** In the context of the birthweight data in Example 7.2, a type I error would be the probability of deciding that the mean birthweight in the hospital was lower than 120 oz when in fact it was 120 oz. A type II error would be the probability of deciding that the mean birthweight was 120 oz when in fact it was lower than 120 oz.

#### Example 7.5

**Cardiovascular Disease, Pediatrics** What are the type I and type II errors for the cholesterol data in Example 7.1?

#### Solution

The type I error is the probability of deciding that offspring of men who died from heart disease have an average cholesterol level higher than 175 mg/dL when in fact their average cholesterol level is 175 mg/dL. The type II error is the probability of deciding that the offspring have normal cholesterol levels when in fact their cholesterol levels are above average.

Type I and type II errors often result in monetary and nonmonetary costs.

#### Example 7.6

**Obstetrics** The birthweight data in Example 7.2 might be used to decide whether a special-care nursery for low-birthweight babies is needed in this hospital. If  $H_1$  were true—that is, if the birthweights in this hospital did tend to be lower than the national average—then the hospital might be justified in having its own special-care nursery. If  $H_0$  were true and the mean birthweight was no different from the U.S. average, then the hospital probably does not need such a nursery. If a type I error is made, then a special-care nursery will be recommended, with all the related extra costs, when in fact it is not needed. If a type II error is made, a special-care nursery will not be funded, when in fact it is needed. The nonmonetary cost of this decision is that some low-birthweight babies may not survive without the unique equipment found in a special-care nursery.

#### Definition 7.4

The probability of a **type I error** is usually denoted by  $\alpha$  and is commonly referred to as the **significance level** of a test.

**Definition 7.5** The probability of a **type II error** is usually denoted by  $\beta$ .

**Definition 7.6** The power of a test is defined as

$$1 - \beta = 1 - \text{probability of a type II error} = Pr(\text{rejecting } H_0 | H_1 \text{ true})$$

**Example 7.7**

**Rheumatology** Suppose a new drug for pain relief is to be tested among patients with osteoarthritis (OA). The measure of pain relief will be the percent change in pain level as reported by the patient after taking the medication for 1 month. Fifty OA patients will participate in the study. What hypotheses are to be tested? What do type I error, type II error, and power mean in this situation?

**Solution**

The hypotheses to be tested are  $H_0: \mu = 0$  vs.  $H_1: \mu > 0$ , where  $\mu$  = mean % change in level of pain over a 1-month period. It is assumed that a positive value for  $\mu$  indicates improvement, whereas a negative value indicates decline.

A type I error is the probability of deciding that the drug is an effective pain reliever based on data from 50 patients, given that the true state of nature is that the drug has no effect on pain relief. The *true state of nature* here means the effect of the drug when tested on a large (infinite) number of patients.

A type II error is the probability of deciding the drug has no effect on pain relief based on data from 50 patients given that the true state of nature is that the drug is an effective pain reliever.

The power of the test is the probability of deciding that the drug is effective as a pain reliever based on data from 50 patients when the true state of nature is that it is effective. It is important to note that the power is not a single number but depends on the true degree of pain relief offered by the drug as measured by the true mean change in pain-relief score ( $\delta$ ). The higher  $\delta$  is, the higher the power will be. In Section 7.5, we present methods for calculating power in more detail.

The general aim in hypothesis testing is to use statistical tests that make  $\alpha$  and  $\beta$  as small as possible. This goal requires compromise because making  $\alpha$  small involves rejecting the null hypothesis less often, whereas making  $\beta$  small involves accepting the null hypothesis less often. These actions are contradictory; that is, as  $\alpha$  decreases,  $\beta$  increases, and as  $\alpha$  increases,  $\beta$  decreases. Our general strategy is to fix  $\alpha$  at some specific level (for example, .10, .05, .01, ...) and to use the test that minimizes  $\beta$  or, equivalently, maximizes the power.

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

Now let's develop the appropriate hypothesis test for the birthweight data in Example 7.2. The statistical model in this case is that the birthweights come from a normal distribution with mean  $\mu$  and unknown variance  $\sigma^2$ . We wish to test the null hypothesis,  $H_0$ , that  $\mu = 120$  oz vs. the alternative hypothesis,  $H_1$ , that  $\mu < 120$  oz. Suppose a more specific alternative, namely  $H_1: \mu = \mu_1 = 110$  oz, is selected.

We will show that the nature of the best test does not depend on the value chosen for  $\mu_1$  provided that  $\mu_1$  is less than 120 oz. We will also fix the  $\alpha$  level at .05 for concreteness.

**Example 7.8**

A very simple test could be used by referring to the table of random digits in Table 4 in the Appendix. Suppose two digits are selected from this table. The null hypothesis is rejected if these two digits are between 00 and 04 inclusive and is accepted if these two digits are between 05 and 99. Clearly, from the properties of the random-number table, the type I error of this test =  $\alpha = Pr(\text{rejecting the null hypothesis } | H_0 \text{ true}) = Pr(\text{drawing two random digits between 00 and 04}) = \frac{5}{100} = .05$ . Thus the proposed test satisfies the  $\alpha$ -level criterion given previously. The problem with this test is that it has very low power. Indeed, the power of the test =  $Pr(\text{rejecting the null hypothesis } | H_1 \text{ true}) = Pr(\text{drawing two random digits between 00 and 04}) = \frac{5}{100} = .05$ .

Note that the outcome of the test has nothing to do with the sample birthweights drawn.  $H_0$  will be rejected just as often when the sample mean birthweight ( $\bar{x}$ ) is 110 oz as when it is 120 oz. Thus this test must be very poor because we would expect to reject  $H_0$  with near certainty if  $\bar{x}$  is small enough and would expect never to reject  $H_0$  if  $\bar{x}$  is large enough.

It can be shown that the best (most powerful) test in this situation is based on the sample mean ( $\bar{x}$ ). If  $\bar{x}$  is sufficiently smaller than  $\mu_0$ , then  $H_0$  is rejected; otherwise,  $H_0$  is accepted. This test is reasonable because if  $H_0$  is true, then the most likely values of  $\bar{x}$  tend to cluster around  $\mu_0$ , whereas if  $H_1$  is true, the most likely values of  $\bar{x}$  tend to cluster around  $\mu_1$ . By "most powerful," we mean that the test based on the sample mean has the highest power among all tests with a given type I error of  $\alpha$ .

**Definition 7.7**

The **acceptance region** is the range of values of  $\bar{x}$  for which  $H_0$  is accepted.

**Definition 7.8**

The **rejection region** is the range of values of  $\bar{x}$  for which  $H_0$  is rejected.

For the birthweight data in Example 7.2, the rejection region consists of small values of  $\bar{x}$  because the underlying mean under the alternative hypothesis ( $\mu_1$ ) is less than the underlying mean under the null hypothesis. This type of test is called a *one-tailed test*.

**Definition 7.9**

A **one-tailed test** is a test in which the values of the parameter being studied (in this case  $\mu$ ) under the alternative hypothesis are allowed to be either greater than or less than the values of the parameter under the null hypothesis ( $\mu_0$ ), *but not both*.

**Example 7.9**

**Cardiovascular Disease, Pediatrics** The hypotheses for the cholesterol data in Example 7.1 are  $H_0: \mu = \mu_0$  vs.  $H_1: \mu > \mu_0$ , where  $\mu$  is the true mean cholesterol level for children of men who have died from heart disease. This test is one-tailed because the alternative mean is only allowed to be greater than the null mean.

In the birthweight example, how small should  $\bar{x}$  be for  $H_0$  to be rejected? This issue can be settled by recalling that the significance level of the test is set at  $\alpha$ .

Suppose  $H_0$  is rejected for all values of  $\bar{x} < c$  and accepted otherwise. The value  $c$  should be selected so that the type I error =  $\alpha$ .

It is more convenient to define test criteria in terms of standardized values rather than in terms of  $\bar{x}$ . Specifically, if we subtract  $\mu_0$  and divide by  $S/\sqrt{n}$ , we obtain the random variable  $t = (\bar{X} - \mu_0)/(S/\sqrt{n})$ , which, based on Equation 6.5, follows a  $t_{n-1}$  distribution under  $H_0$ . We note that under  $H_0$ , based on the definition the percentiles of a  $t$  distribution,  $Pr(t < t_{n-1,\alpha}) = \alpha$ . This leads us to the following test procedure.

### Equation 7.2

#### One-Samplet $t$ Test for the Mean of a Normal Distribution with Unknown Variance (Alternative Mean < Null Mean)

To test the hypothesis  $H_0: \mu = \mu_0, \sigma$  unknown vs.  $H_1: \mu < \mu_0, \sigma$  unknown with a significance level of  $\alpha$ , we compute

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

If  $t < t_{n-1,\alpha}$ , then we reject  $H_0$ .

If  $t \geq t_{n-1,\alpha}$ , then we accept  $H_0$ .

### Definition 7.10

The value  $t$  in Equation 7.2 is called a **test statistic** because the test procedure is based on this statistic.

### Definition 7.11

The value  $t_{n-1,\alpha}$  in Equation 7.2 is called a **critical value** because the outcome of the test depends on whether the test statistic  $t < t_{n-1,\alpha}$  = critical value, whereby we reject  $H_0$  or  $t \geq t_{n-1,\alpha}$ , whereby we accept  $H_0$ .

### Definition 7.12

The general approach in which we compute a test statistic and determine the outcome of a test by comparing the test statistic with a critical value determined by the type I error is called the **critical-value method** of hypothesis testing.

### Example 7.10

**Obstetrics** Use the one-sample  $t$  test to test the hypothesis  $H_0: \mu = 120$  vs.  $H_1: \mu < 120$  based on the birthweight data given in Example 7.2 and using a significance level of .05.

### Solution

We compute the test statistic

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{115 - 120}{24/\sqrt{100}} \\ &= \frac{-5}{2.4} = -2.08 \end{aligned}$$

Using the TINV function of Excel, we see the critical value =  $t_{99, .05} = -TINV(.05, 99) = -1.66$ . Because  $t = -2.08 < -1.66$ , it follows that we can reject  $H_0$  at a significance level of .05.

**Example 7.11**

**Obstetrics** Use the one-sample  $t$  test to test the hypothesis given in Example 7.10 using a significance level of .01.

**Solution**

Using Excel, the critical value is  $t_{99, .01} = -TINV(.01, 99) = -2.36$ . Because  $t = -2.08 > -2.36$ , it follows that we accept  $H_0$  at significance level = .01.

If we use the critical-value method, how do we know what level of  $\alpha$  to use? The actual  $\alpha$  level used should depend on the relative importance of type I and type II errors because the smaller  $\alpha$  is made for a fixed sample size ( $n$ ), the larger  $\beta$  becomes. Most people feel uncomfortable with  $\alpha$  levels much greater than .05. Traditionally, an  $\alpha$  level of exactly .05 is used most frequently.

In general, a number of significance tests could be performed at different  $\alpha$  levels, as was done in Examples 7.10 and 7.11, and whether  $H_0$  would be accepted or rejected in each instance could be noted. This can be somewhat tedious and is unnecessary because, instead, significance tests can be effectively performed *at all  $\alpha$  levels* by obtaining the  $p$ -value for the test.

**Definition 7.13**

The  **$p$ -value** for any hypothesis test is the  $\alpha$  level at which we would be indifferent between accepting or rejecting  $H_0$  given the sample data at hand. That is, the  $p$ -value is the  $\alpha$  level at which the given value of the test statistic (such as  $t$ ) is on the borderline between the acceptance and rejection regions.

According to the test criterion in Equation 7.2, if a significance level of  $p$  is used, then  $H_0$  would be rejected if  $t < t_{n-1, p}$  and accepted if  $t \geq t_{n-1, p}$ . We would be indifferent to the choice between accepting or rejecting  $H_0$  if  $t = t_{n-1, p}$ . We can solve for  $p$  as a function of  $t$  by

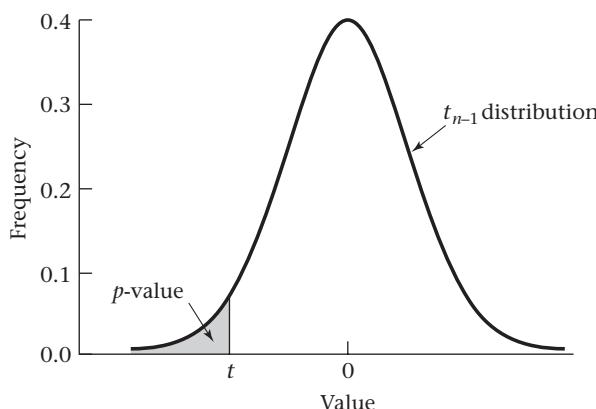
**Equation 7.3**

$$p = Pr(t_{n-1} \leq t)$$

Thus  $p$  is the area to the left of  $t$  under a  $t_{n-1}$  distribution.

The  $p$ -value can be displayed as shown in Figure 7.1.

**Figure 7.1 Graphic display of a  $p$ -value**



**Example 7.12**

**Obstetrics** Compute the *p*-value for the birthweight data in Example 7.2.

**Solution**

From Equation 7.3, the *p*-value is

$$Pr(t_{99} \leq -2.08)$$

Using the TDIST function of Excel, we find this probability is given by TDIST (2.08, 99, 1) = .020, which is the *p*-value. Note that the TDIST function can only be used to obtain right-hand tail areas for positive values of *t*. However, from the symmetry of the *t* distribution,  $Pr(t_{99} \leq -2.08) = Pr(t_{99} \geq 2.08) = TDIST(2.08, 99, 1)$ . The third argument (1) indicates that the *p*-value for a one-tailed test is desired.

An alternative definition of a *p*-value that is useful in other hypothesis-testing problems is as follows:

**Definition 7.14**

The ***p*-value** can also be thought of as the probability of obtaining a test statistic as extreme as or more extreme than the actual test statistic obtained, given that the null hypothesis is true.

We know that under the null hypothesis, the *t* statistic follows a  $t_{n-1}$  distribution. Hence, the probability of obtaining a *t* statistic that is no larger than *t* under the null hypothesis is  $Pr(t_{n-1} \leq t) = p$ -value, as shown in Figure 7.1.

**Example 7.13**

**Cardiology** A topic of recent clinical interest is the possibility of using drugs to reduce infarct size in patients who have had a myocardial infarction within the past 24 hours. Suppose we know that in untreated patients the mean infarct size is 25 ( $ck - g - EQ/m^2$ ). Furthermore, in 8 patients treated with a drug the mean infarct size is 16 with a standard deviation of 10. Is the drug effective in reducing infarct size?

**Solution**

The hypotheses are  $H_0: \mu = 25$  vs.  $H_1: \mu < 25$ . The *p*-value is computed using Equation 7.3. First we compute the *t* statistic given by

$$t = \frac{16 - 25}{10/\sqrt{8}} = -2.55$$

The *p*-value is then given by  $p = Pr(t < -2.55)$ . Referring to Table 5 in the Appendix, we see that  $t_{.975} = 2.365$ , and  $t_{.99} = 2.998$ . Because  $2.365 < 2.55 < 2.998$ , it follows that  $1 - .99 < p < 1 - .975$  or  $.01 < p < .025$ . Using Excel, the exact *p*-value is given by TDIST (2.55, 7, 1) = .019. Thus  $H_0$  is rejected and we conclude that the drug significantly reduces infarct size (all other things being equal).

This can also be interpreted as the probability that mean infarct size among a random sample of 8 patients will be no larger than 16, if the null hypothesis is true. In this example, the null hypothesis is that the drug is ineffective, or in other words, that true mean infarct size for the population of all patients with myocardial infarction who are treated with drug = true mean infarct size for untreated patients = 25.

The *p*-value is important because it tells us *exactly* how significant our results are without performing repeated significance tests at different  $\alpha$  levels. A question typically asked is: How small should the *p*-value be for results to be considered statistically significant? Although this question has no one answer, some commonly used criteria follow.

**Equation 7.4****Guidelines for Judging the Significance of a *p*-Value**

- If  $.01 \leq p < .05$ , then the results are *significant*.  
 If  $.001 \leq p < .01$ , then the results are *highly significant*.  
 If  $p < .001$ , then the results are *very highly significant*.  
 If  $p > .05$ , then the results are considered *not statistically significant* (sometimes denoted by NS).  
 However, if  $.05 \leq p < .10$ , then a trend toward statistical significance is sometimes noted.

Authors frequently do not specify the exact *p*-value beyond giving ranges of the type shown here because whether the *p*-value is .024 or .016 is thought to be unimportant. Other authors give an exact *p*-value even for results that are not statistically significant so that the reader can appreciate how close to statistical significance the results have come. With the advent of statistical packages such as Excel, MINITAB, and Stata, exact *p*-values are easy to obtain. These different approaches lead to the following general principle.

**Equation 7.5****Determination of Statistical Significance for Results from Hypothesis Tests**

Either of the following methods can be used to establish whether results from hypothesis tests are statistically significant:

- (1) The test statistic  $t$  can be computed and compared with the critical value  $t_{n-1, \alpha}$  at an  $\alpha$  level of .05. Specifically, if  $H_0: \mu = \mu_0$  vs.  $H_1: \mu < \mu_0$  is being tested and  $t < t_{n-1, .05}$ , then  $H_0$  is rejected and the results are declared *statistically significant* ( $p < .05$ ). Otherwise,  $H_0$  is accepted and the results are declared *not statistically significant* ( $p \geq .05$ ). We have called this approach the **critical-value method** (see Definition 7.12).
- (2) The exact *p*-value can be computed and, if  $p < .05$ , then  $H_0$  is rejected and the results are declared *statistically significant*. Otherwise, if  $p \geq .05$ , then  $H_0$  is accepted and the results are declared *not statistically significant*. We will refer to this approach as the ***p*-value method**.

These two approaches are equivalent regarding the determination of statistical significance (whether  $p < .05$  or  $p \geq .05$ ). The *p*-value method is somewhat more precise in that it yields an exact *p*-value. The two approaches in Equation 7.5 can also be used to determine statistical significance in other hypothesis-testing problems.

**Example 7.14**

**Obstetrics** Assess the statistical significance of the birthweight data in Example 7.12.

**Solution**

Because the *p*-value is .020, the results would be considered statistically significant and we would conclude that the true mean birthweight is significantly lower in this hospital than in the general population.

**Example 7.15**

**Cardiology** Assess the significance of the infarct-size data in Example 7.13.

**Solution**

The  $p$ -value =  $Pr(t_7 < -2.55)$ . Using the TDIST function of Excel, we found that  $p = .019$ . Thus the results are significant.

In writing up the results of a study, a distinction between scientific and statistical significance should be made because the two terms do not necessarily coincide.

The results of a study can be statistically significant but can still not be scientifically important. This situation would occur if a small difference was found to be statistically significant because of a large sample size. Conversely, some statistically nonsignificant results can be scientifically important, encouraging researchers to perform larger studies to confirm the direction of the findings and possibly reject  $H_0$  with a larger sample size. This statement is true not only for the one-sample  $t$  test but for virtually any hypothesis test.

**Example 7.16**

**Obstetrics** Suppose the mean birthweight in Example 7.2 was 119 oz, based on a sample of size 10,000. Assess the results of the study.

**Solution**

The test statistic would be given by

$$t = \frac{119 - 120}{24/\sqrt{10,000}} = -4.17$$

Thus the  $p$ -value is given by  $Pr(t_{9999} < -4.17)$ . Because a  $t$  distribution with 9999 degrees of freedom ( $df$ ) is virtually the same as an  $N(0,1)$  distribution, we can approximate the  $p$ -value by  $\Phi(-4.17) < .001$ . The results are thus very highly significant but are clearly not very important because of the small difference in mean birthweight (1 oz) between this hospital and the national average.

**Example 7.17**

**Obstetrics** Suppose the mean birthweight in Example 7.2 was 110 oz, based on a sample size of 10. Assess the results of the study.

**Solution**

The test statistic would be given by

$$t = \frac{110 - 120}{24/\sqrt{10}} = -1.32$$

The  $p$ -value is given by  $Pr(t_9 < -1.32)$ . From Appendix Table 5, because  $t_{9,.85} = 1.100$  and  $t_{9,.90} = 1.383$  and  $1.100 < 1.32 < 1.383$ , it follows that  $1 - .90 < p < 1 - .85$  or  $.10 < p < .15$ . Using Excel, the  $p$ -value = TDIST (1.32, 9, 1) = .110. These results are not statistically significant but could be important if the same trends were also apparent in a larger study.

The test criterion in Equation 7.2 was based on an alternative hypothesis that  $\mu < \mu_0$ . In many situations we wish to use an alternative hypothesis that  $\mu > \mu_0$ . In this case  $H_0$  would be rejected if  $\bar{x}$ , or correspondingly our test statistic  $t$ , were large ( $> c$ ) and accepted if  $t$  were small ( $\leq c$ ), where  $c$  is derived next. To ensure a type I error of  $\alpha$ , find  $c$  such that

$$\begin{aligned}\alpha &= Pr(t > c | H_0) = Pr(t > c | \mu = \mu_0) \\ &= 1 - Pr(t \leq c | \mu = \mu_0)\end{aligned}$$

Because  $t$  follows a  $t_{n-1}$  distribution under  $H_0$ , we have

$$\alpha = 1 - Pr(t_{n-1} \leq c) \quad \text{or} \quad 1 - \alpha = Pr(t_{n-1} \leq c)$$

Because  $Pr(t_{n-1} < t_{n-1,1-\alpha}) = 1 - \alpha$ , we have  $c = t_{n-1,1-\alpha}$ . Thus at level  $\alpha$ ,  $H_0$  is rejected if  $t > t_{n-1,1-\alpha}$  and accepted otherwise. The  $p$ -value is the probability of observing a test

statistic at least as large as  $t$  under the null hypothesis. Thus, because  $t$  follows a  $t_{n-1}$  distribution under  $H_0$ , we have

$$p = Pr(t_{n-1} \geq t) = 1 - Pr(t_{n-1} \leq t)$$

The test procedure is summarized as follows.

### Equation 7.6

#### One-Sample $t$ Test for the Mean of a Normal Distribution with Unknown Variance (Alternative Mean > Null Mean)

To test the hypothesis

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu > \mu_0$$

with a significance level of  $\alpha$ , the best test is based on  $t$ , where

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

If  $t > t_{n-1,1-\alpha}$ , then  $H_0$  is rejected

If  $t \leq t_{n-1,1-\alpha}$ , then  $H_0$  is accepted

The  $p$ -value for this test is given by

$$p = Pr(t_{n-1} > t)$$

The  $p$ -value for this test is depicted in Figure 7.2.

### Example 7.18

**Cardiovascular Disease, Pediatrics** Suppose the mean cholesterol level of 10 children whose fathers died from heart disease in Example 7.1 is 200 mg/dL and the sample standard deviation is 50 mg/dL. Test the hypothesis that the mean cholesterol level is higher in this group than in the general population.

#### Solution

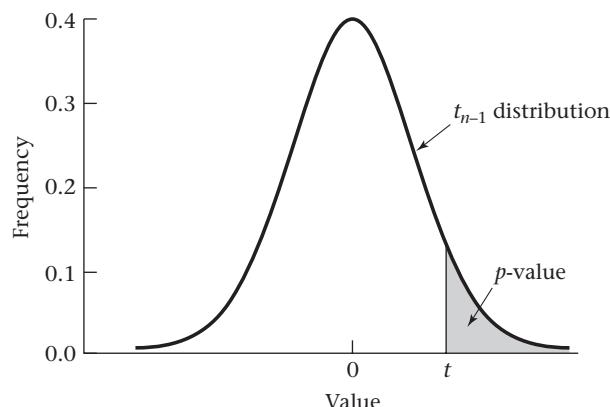
The hypothesis

$$H_0: \mu = 175 \quad \text{vs.} \quad H_1: \mu > 175$$

is tested using an  $\alpha$  level of .05.  $H_0$  is rejected if

$$t > t_{n-1,1-\alpha} = t_{9,.95}$$

**Figure 7.2** *p*-value for the one-sample  $t$  test when the alternative mean ( $\mu_1$ ) > null mean ( $\mu_0$ )



In this case,

$$\begin{aligned} t &= \frac{200 - 175}{50/\sqrt{10}} \\ &= \frac{25}{15.81} = 1.58 \end{aligned}$$

From Table 5, we see that  $t_{9,95} = 1.833$ . Because  $1.833 > 1.58$ , it follows that we accept  $H_0$  at the 5% level of significance.

If we use the  $p$ -value method, the exact  $p$ -value is given by

$$p = Pr(t_9 > 1.58)$$

Using Appendix Table 5, we find  $t_{9,90} = 1.383$  and  $t_{9,95} = 1.833$ . Thus, because  $1.383 < 1.58 < 1.833$ , it follows that  $.05 < p < .10$ . Alternatively, using the TDIST function of Excel, we can get the exact  $p$ -value =  $Pr(t_9 > 1.58) = \text{TDIST}(1.58, 9, 1) = .074$ . Because  $p > .05$ , we conclude that our results are not statistically significant, and the null hypothesis is accepted. Thus the mean cholesterol level of these children does not differ significantly from that of an average child.

### REVIEW QUESTIONS 7A

- 1** What is the difference between a type I and type II error?
- 2** What is the difference between the critical-value method and the  $p$ -value method of hypothesis testing?
- 3** Several studies have shown that women with many children are less likely to get ovarian cancer. In a new study, data are collected from 25 women ages 40–49 with ovarian cancer. The mean parity (number of children) of these women is 1.8 with standard deviation 1.2. Suppose the mean number of children among women in the general population in this age group is 2.5.
  - (a)** What test can be used to test the hypothesis that women with ovarian cancer have fewer children than women in the general population in the same age group?
  - (b)** Perform the test in Review Question 7A.3a using the critical-value method.
  - (c)** What is the  $p$ -value based on the test in Review Question 7A.3a?
  - (d)** What do you conclude from this study?

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

In the previous section the alternative hypothesis was assumed to be in a *specific direction* relative to the null hypothesis.

### Example 7.19

**Obstetrics** Example 7.2 assumed that the mean birthweight of infants from a low-SES-area hospital was either the same as or lower than average. Example 7.1 assumed that the mean cholesterol level of children of men who died from heart disease was either the same as or higher than average.

In most instances this *prior knowledge* is unavailable. If the null hypothesis is not true, then we have no idea in which direction the alternative mean will fall.

**Example 7.20**

**Cardiovascular Disease** Suppose we want to compare fasting serum-cholesterol levels among recent Asian immigrants to the United States with typical levels found in the general U.S. population. Suppose we assume cholesterol levels in women ages 21–40 in the United States are approximately normally distributed with mean 190 mg/dL. It is unknown whether cholesterol levels among recent Asian immigrants are higher or lower than those in the general U.S. population. Let's assume that levels among recent female Asian immigrants are normally distributed with unknown mean  $\mu$ . Hence we wish to test the null hypothesis  $H_0: \mu = \mu_0 = 190$  vs. the alternative hypothesis  $H_1: \mu \neq \mu_0$ . Blood tests are performed on 100 female Asian immigrants ages 21–40, and the mean level ( $\bar{x}$ ) is 181.52 mg/dL with standard deviation = 40 mg/dL. What can we conclude on the basis of this evidence?

The type of alternative given in Example 7.20 is known as a *two-sided* alternative because the alternative mean can be either less than or greater than the null mean.

**Definition 7.15**

A **two-tailed test** is a test in which the values of the parameter being studied (in this case  $\mu$ ) under the alternative hypothesis are allowed to be either *greater than or less than* the values of the parameter under the null hypothesis ( $\mu_0$ ).

The best test here depends on the sample mean  $\bar{x}$  or, equivalently, on the test statistic  $t$ , as it did in the one-sided situation developed in Section 7.3. We showed in Equation 7.2 that to test the hypothesis  $H_0: \mu = \mu_0$  versus  $H_1: \mu < \mu_0$ , the best test was of the form: Reject  $H_0$  if  $t < t_{n-1,\alpha}$  and accept  $H_0$  if  $t \geq t_{n-1,\alpha}$ . This test is clearly only appropriate for alternatives on one side of the null mean, namely  $\mu < \mu_0$ . We also showed in Equation 7.6 that to test the hypothesis

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu > \mu_0$$

the best test was correspondingly of the following form: Reject  $H_0$  if  $t > t_{n-1,1-\alpha}$  and accept  $H_0$  if  $t \leq t_{n-1,1-\alpha}$ .

**Equation 7.7**

A reasonable decision rule to test for alternatives on *either* side of the null mean is to *reject  $H_0$  if  $t$  is either too small or too large*. Another way of stating the rule is that  $H_0$  will be rejected if  $t$  is either  $< c_1$  or  $> c_2$  for some constants  $c_1, c_2$  and  $H_0$  will be accepted if  $c_1 \leq t \leq c_2$ .

The question remains: What are appropriate values for  $c_1$  and  $c_2$ ? These values are again determined by the type I error ( $\alpha$ ). The constants  $c_1, c_2$  should be chosen such that

**Equation 7.8**

$$\begin{aligned} Pr(\text{reject } H_0 | H_0 \text{ true}) &= Pr(t < c_1 \text{ or } t > c_2 | H_0 \text{ true}) \\ &= Pr(t < c_1 | H_0 \text{ true}) + Pr(t > c_2 | H_0 \text{ true}) = \alpha \end{aligned}$$

Half of the type I error is assigned arbitrarily to each of the probabilities on the left side of the second line of Equation 7.8. Thus, we wish to find  $c_1, c_2$  so that

$$\text{Equation 7.9} \quad \Pr(t < c_1 | H_0 \text{ true}) = \Pr(t > c_2 | H_0 \text{ true}) = \alpha/2$$

We know  $t$  follows a  $t_{n-1}$  distribution under  $H_0$ . Because  $t_{n-1,\alpha/2}$  and  $t_{n-1,1-\alpha/2}$  are the lower and upper  $100\% \times \alpha/2$  percentiles of a  $t_{n-1}$  distribution, it follows that

$$\Pr(t < t_{n-1,\alpha/2}) = \Pr(t > t_{n-1,1-\alpha/2}) = \alpha/2$$

Therefore,

$$c_1 = t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2} \quad \text{and} \quad c_2 = t_{n-1,1-\alpha/2}$$

This test procedure can be summarized as follows:

**Equation 7.10**
**One-Sample  $t$  Test for the Mean of a Normal Distribution with Unknown Variance (Two-Sided Alternative)**

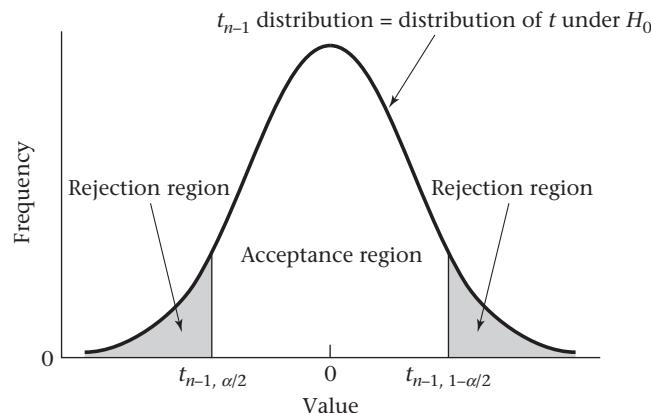
To test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ , with a significance level of  $\alpha$ , the best test is based on  $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ .

If  $|t| > t_{n-1,1-\alpha/2}$   
then  $H_0$  is rejected.

If  $|t| \leq t_{n-1,1-\alpha/2}$   
then  $H_0$  is accepted.

The acceptance and rejection regions for this test are shown in Figure 7.3.

**Figure 7.3** One-sample  $t$  test for the mean of a normal distribution (two-sided alternative)


**Example 7.21**

**Cardiovascular Disease** Test the hypothesis that the mean cholesterol level of recent female Asian immigrants is different from the mean in the general U.S. population, using the data in Example 7.20.

**Solution**

We compute the test statistic

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{181.52 - 190}{40/\sqrt{100}} \\ &= \frac{-8.48}{4} = -2.12 \end{aligned}$$

For a two-sided test with  $\alpha = .05$ , the critical values are  $c_1 = t_{99,.025}$ ,  $c_2 = t_{99,.975}$ .

From Table 5 in the Appendix, because  $t_{99,.975} < t_{60,.975} = 2.000$ , it follows that  $c_2 < 2.000$ . Also, because  $c_1 = -c_2$  it follows that  $c_1 > -2.000$ . Because  $t = -2.12 < -2.000 < c_1$ , it follows that we can reject  $H_0$  at the 5% level of significance. We conclude that the mean cholesterol level of recent Asian immigrants is significantly different from that of the general U.S. population.

Alternatively, we might want to compute a  $p$ -value as we did in the one-sided case. The  $p$ -value is computed in two different ways, depending on whether  $t$  is less than or greater than 0.

#### Equation 7.11

##### **$p$ -Value for the One-Sample $t$ Test for the Mean of a Normal Distribution (Two-Sided Alternative)**

$$\text{Let } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$p = \begin{cases} 2 \times Pr(t_{n-1} \leq t), & \text{if } t \leq 0 \\ 2 \times [1 - Pr(t_{n-1} \leq t)], & \text{if } t > 0 \end{cases}$$

Thus, in words, if  $t \leq 0$ , then  $p = 2$  times the area under a  $t_{n-1}$  distribution to the left of  $t$ ; if  $t > 0$ , then  $p = 2$  times the area under a  $t_{n-1}$  distribution to the right of  $t$ . One way to interpret the  $p$ -value is as follows.

#### Equation 7.12

The  **$p$ -value** is the probability under the null hypothesis of obtaining a test statistic as extreme as or more extreme than the observed test statistic, where, because a two-sided alternative hypothesis is being used, extremeness is measured by the **absolute value** of the test statistic.

Hence if  $t > 0$ , the  $p$ -value is the area to the right of  $t$  plus the area to the left of  $-t$  under a  $t_{n-1}$  distribution.

However, this area simply amounts to twice the right-hand tail area because the  $t$  distribution is symmetric around 0. A similar interpretation holds if  $t < 0$ .

These areas are illustrated in Figure 7.4.

#### Example 7.22

**Cardiovascular Disease** Compute the  $p$ -value for the hypothesis test in Example 7.20.

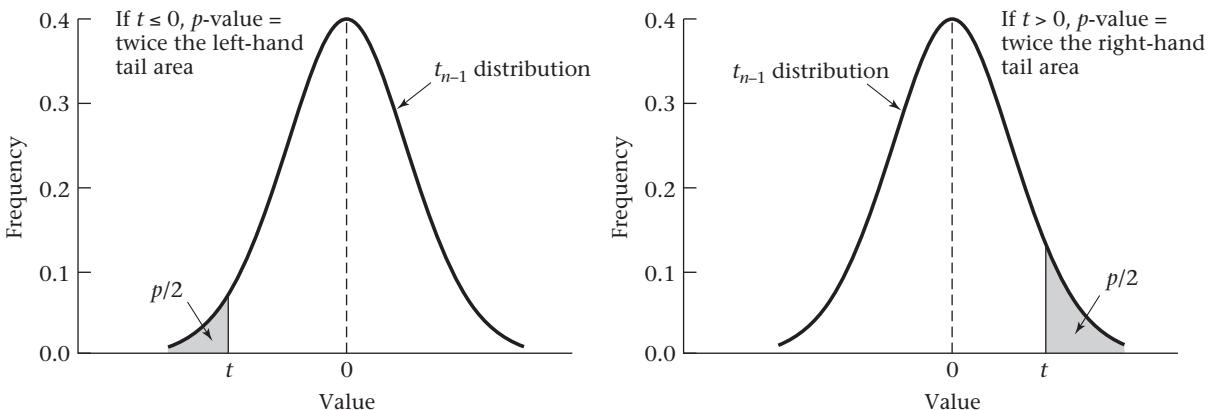
#### Solution

Because  $t = -2.12$ , the  $p$ -value for the test is twice the left-hand tail area, or

$$p = 2 \times Pr(t_{99} < -2.12) = \text{TDIST}(2.12, 99, 2) = .037$$

Hence, the results are statistically significant with a  $p$ -value of .037. Note that if the third argument of TDIST is 2, then a two-sided  $p$ -value is obtained.

Finally, if  $n$  is large (say,  $>200$ ), then the percentiles of the  $t$  distribution ( $t > t_{n-1,1-\alpha/2}$ ) used in determining the critical values in Equation 7.10 can be approximated by the corresponding percentiles of an  $N(0,1)$  distribution ( $z_{1-\alpha/2}$ ). Similarly, in computing  $p$ -values in Equation 7.11, if  $n > 200$ , then  $Pr(t_{n-1} \leq t)$  can be

**Figure 7.4 Illustration of the  $p$ -value for a one-sample  $t$  test for the mean of a normal distribution (two-sided alternative)**

approximated by  $\Pr[N(0,1) < t] = \Phi(t)$ . We used similar approximations in our work on CIs for the mean of a normal distribution with unknown variance in Section 6.5 on page 172 in Chapter 6.

When is a one-sided test more appropriate than a two-sided test? Generally, the sample mean falls in the expected direction from  $\mu_0$  and it is *easier* to reject  $H_0$  using a one-sided test than using a two-sided test. However, this is not necessarily always the case. Suppose we guess from a previous review of the literature that the cholesterol level of Asian immigrants is likely to be lower than that of the general U.S. population because of better dietary habits. In this case, we would use a one-sided test of the form  $H_0: \mu = 190$  vs.  $H_1: \mu < 190$ . From Equation 7.3, the one-sided  $p$ -value =  $\Pr(t_{99} < -2.12) = \text{TDIST}(2.12, 99, 1) = .018 = \frac{1}{2}$  (two-sided  $p$ -value). Alternatively, suppose we guess from a previous literature review that the cholesterol level of Asian immigrants is likely to be higher than that of the general U.S. population because of more stressful living conditions. In this case, we would use a one-sided test of the form  $H_0: \mu = 190$  vs.  $H_1: \mu > 190$ . From Equation 7.6, the  $p$ -value =  $\Pr(t_{99} > -2.12) = .982$ . Thus we would accept  $H_0$  if we use a one-sided test and the sample mean is on the opposite side of the null mean from the alternative hypothesis. Generally, a two-sided test is always appropriate because there can be no question about the conclusions. Also, as just illustrated, a two-sided test can be more conservative because you need not guess the appropriate side of the null hypothesis for the alternative hypothesis. However, in certain situations only alternatives on one side of the null mean are of interest or are possible, and in this case a one-sided test is better because it has more power (that is, it is easier to reject  $H_0$  based on a finite sample if  $H_1$  is actually true) than its two-sided counterpart. In all instances, it is important to decide whether to use a one-sided or a two-sided test *before* data analysis (or preferably before data collection) begins so as not to bias conclusions based on results of hypothesis testing. In particular, do not change from a two-sided to a one-sided test *after* looking at the data.

**Example 7.23 Hypertension** Suppose we are testing the efficacy of a drug to reduce blood pressure. Assume the change in blood pressure (baseline blood pressure minus follow-up blood pressure) is normally distributed with mean  $\mu$  and variance  $\sigma^2$ . An appropriate hypothesis test might be  $H_0: \mu = 0$  vs.  $H_1: \mu > 0$  because the drug is of interest only if it reduces blood pressure, not if it raises blood pressure.

In the rest of this text, we focus primarily on two-sided tests because they are much more widely used in the literature.

In this section and in Section 7.3, we have presented the **one-sample *t* test**, which is used for testing hypotheses concerning the mean of a normal distribution when the variance is unknown. This test is featured in the flowchart in Figure 7.18 (p. 258) where we display techniques for determining appropriate methods of statistical inference. Beginning at the “Start” box, we arrive at the one-sample *t* test box by answering yes to each of the following four questions: (1) one variable of interest? (2) one-sample problem? (3) underlying distribution normal or can central-limit theorem be assumed to hold? and (4) inference concerning  $\mu$ ? and no to question (5)  $\sigma$  known?

## One-Sample *z* Test

In Equations 7.10 and 7.11, the critical values and *p*-values for the one-sample *t* test have been specified in terms of percentiles of the *t* distribution, assuming the underlying variance is unknown. In some applications, the variance may be assumed known from prior studies. In this case, the test statistic *t* can be replaced by the test statistic  $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ . Also, the critical values based on the *t* distribution can be replaced by the corresponding critical values of a standard normal distribution. This leads to the following test procedure:

### Equation 7.13

#### One-Sample *z* Test for the Mean of a Normal Distribution with Known Variance (Two-Sided Alternative)

To test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$  with a significance level of  $\alpha$ , where the underlying standard deviation  $\sigma$  is known, the best test is based on  $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$

If  $z < z_{\alpha/2}$  or  $z > z_{1-\alpha/2}$   
then  $H_0$  is rejected.

If  $z_{\alpha/2} \leq z \leq z_{1-\alpha/2}$   
then  $H_0$  is accepted.

To compute a two-sided *p*-value, we have

$$\begin{aligned} p &= 2\Phi(z) && \text{if } z \leq 0 \\ &= 2[1 - \Phi(z)] && \text{if } z > 0 \end{aligned}$$

### Example 7.24

**Cardiovascular Disease** Consider the cholesterol data in Example 7.21. Assume that the standard deviation is known to be 40 and the sample size is 200 instead of 100. Assess the significance of the results.

#### Solution

The test statistic is

$$\begin{aligned} z &= \frac{181.52 - 190}{40/\sqrt{200}} \\ &= \frac{-8.48}{2.828} = -3.00 \end{aligned}$$

We first use the critical-value method with  $\alpha = 0.05$ . Based on Equation 7.13, the critical values are  $-1.96$  and  $1.96$ . Because  $z = -3.00 < -1.96$ , we can reject  $H_0$  at a 5% level of significance. The two-sided *p*-value is given by  $2 \times \Phi(-3.00) = .003$ .

Similarly, we can consider the one-sample  $z$  test for a one-sided alternative as follows.

**Equation 7.14**
**One-Sample  $z$  Test for the Mean of a Normal Distribution with Known Variance (One-Sided Alternative) ( $\mu_1 < \mu_0$ )**

To test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu < \mu_0$  with a significance level of  $\alpha$ , where the underlying standard deviation  $\sigma$  is known, the best test is based on

$$z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$$

If  $z < z_{\alpha'}$ , then  $H_0$  is rejected; if  $z \geq z_{\alpha'}$ , then  $H_0$  is accepted. The  $p$ -value is given by  $p = \Phi(z)$ .

**Equation 7.15**
**One-Sample  $z$  Test for the Mean of a Normal Distribution with Known Variance (One-Sided Alternative) ( $\mu_1 > \mu_0$ )**

To test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu > \mu_0$  with a significance level of  $\alpha$ , where the underlying standard deviation  $\sigma$  is known, the best test is based on

$$z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$$

If  $z > z_{1-\alpha'}$ , then  $H_0$  is rejected; if  $z \leq z_{1-\alpha'}$ , then  $H_0$  is accepted. The  $p$ -value is given by  $p = 1 - \Phi(z)$ .

In this section we presented the **one-sample  $z$  test**, which is used for testing hypotheses concerning the mean of a normal distribution when the variance is known. Beginning at the “Start” box of the flowchart (Figure 7.18, p. 258), we arrive at the one-sample  $z$  test box by answering yes to each of the following five questions: (1) one variable of interest? (2) one-sample problem? (3) underlying distribution normal or can central-limit theorem be assumed to hold? (4) inference concerning  $\mu$ ? and (5)  $\sigma$  known?

## 7.5 The Power of a Test

The calculation of power is used to plan a study, usually before any data have been obtained, except possibly from a small preliminary study called a pilot study. Also, we usually make a projection concerning the standard deviation without actually having any data to estimate it. Therefore, we assume the standard deviation is known and base power calculations on the one-sample  $z$  test as given in Equations 7.13, 7.14, and 7.15.

### One-Sided Alternatives

**Example 7.25**

**Ophthalmology** A new drug is proposed to prevent the development of glaucoma in people with high intraocular pressure (IOP). A pilot study is conducted with the drug among 10 patients. After 1 month of using the drug, their mean IOP decreases by 5 mm Hg with a standard deviation of 10 mm Hg. The investigators propose to study 100 participants in the main study. Is this a sufficient sample size for the study?

**Solution**

To determine whether 100 participants are enough, we need to do a power calculation. The power of the study is the probability that we will be able to declare a significant difference with a sample of size 100 if the true mean decline in IOP is 5 mm Hg

with a standard deviation of 10 mm Hg. Usually we want a power of at least 80% to perform a study. In this section we examine formulas for computing power and addressing the question just asked.

In Section 7.4 (Equation 7.14) the appropriate hypothesis test was derived to test

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu < \mu_0$$

where the underlying distribution was assumed to be normal and the population variance was assumed to be known. The best test was based on the test statistic  $z$ . In particular, from Equation 7.14 for a type I error of  $\alpha$ ,  $H_0$  is rejected if  $z < z_\alpha$  and  $H_0$  is accepted if  $z \geq z_\alpha$ . The form of the best test *does not depend on the alternative mean chosen ( $\mu_1$ )* as long as the alternative mean is less than the null mean  $\mu_0$ .

Hence, in Example 7.2, where  $\mu_0 = 120$  oz, if we were interested in an alternative mean of  $\mu_1 = 115$  oz rather than  $\mu_1 = 110$  oz, then the same test procedure would still be used. However, what differs for the two alternative means is the power of the test  $= 1 - Pr(\text{type II error})$ . Recall from Definition 7.6 that

$$\begin{aligned}\text{Power} &= Pr(\text{reject } H_0 | H_0 \text{ false}) = Pr(Z < z_\alpha | \mu = \mu_1) \\ &= Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha | \mu = \mu_1\right) \\ &= Pr\left(\bar{X} < \mu_0 + z_\alpha \sigma/\sqrt{n} | \mu = \mu_1\right)\end{aligned}$$

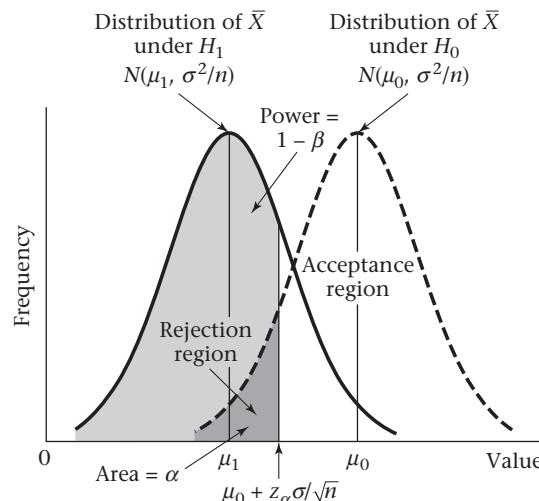
We know that under  $H_1$ ,  $\bar{X} \sim N(\mu_1, \sigma^2/n)$ . Hence, on standardization of limits,

$$\text{Power} = \Phi\left[\left(\mu_0 + z_\alpha \sigma/\sqrt{n} - \mu_1\right)/\left(\sigma/\sqrt{n}\right)\right] = \Phi\left[z_\alpha + \frac{(\mu_0 - \mu_1)}{\sigma}\sqrt{n}\right]$$

This power is depicted graphically in Figure 7.5.

Note that the area to the left of  $\mu_0 + z_\alpha \sigma/\sqrt{n}$  under the  $H_0$  distribution is the significance level  $\alpha$ , whereas the area to the left of  $\mu_0 + z_\alpha \sigma/\sqrt{n}$  under the  $H_1$  distribution is the power  $= 1 - \beta$ .

**Figure 7.5** Illustration of power for the one-sample test for the mean of a normal distribution with known variance ( $\mu_1 < \mu_0$ )



Why should power concern us? The power of a test tells us how likely it is that a statistically significant difference will be detected based on a finite sample size  $n$ , if the alternative hypothesis is true—that is, if the true mean  $\mu$  differs from the mean under the null hypothesis ( $\mu_0$ ). If the power is too low, then there is little chance of finding a significant difference and nonsignificant results are likely even if real differences exist between the true mean  $\mu$  of the group being studied and the null mean  $\mu_0$ . An inadequate sample size is usually the cause of low power to detect a scientifically meaningful difference.

**Example 7.26**

**Obstetrics** Compute the power of the test for the birthweight data in Example 7.2 (p. 204) with an alternative mean of 115 oz and  $\alpha = .05$ , assuming the true standard deviation = 24 oz.

**Solution**

We have  $\mu_0 = 120$  oz,  $\mu_1 = 115$  oz,  $\alpha = .05$ ,  $\sigma = 24$ ,  $n = 100$ . Thus

$$\text{Power} = \Phi\left[z_{.05} + (120 - 115)\sqrt{100}/24\right] = \Phi[-1.645 + 5(10)/24] = \Phi(0.438) = .669$$

Therefore, there is about a 67% chance of detecting a significant difference using a 5% significance level with this sample size.

We have focused on the situation where  $\mu_1 < \mu_0$ . We are also interested in power when testing the hypothesis

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu = \mu_1 > \mu_0$$

as was the case with the cholesterol data in Example 7.1. The best test for this situation was presented in Equation 7.15, where  $H_0$  is rejected if  $z > z_{1-\alpha}$  and accepted if  $z \leq z_{1-\alpha}$ . Notice that if  $z > z_{1-\alpha}$ , then

**Equation 7.16**

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$$

If we multiply both sides of Equation 7.16 by  $\sigma/\sqrt{n}$ , and add  $\mu_0$ , we can re-express the rejection criteria in terms of  $\bar{x}$ , as follows:

**Equation 7.17**

$$\bar{x} > \mu_0 + z_{1-\alpha} \sigma/\sqrt{n}$$

Similarly, the acceptance criteria,  $z \leq z_{1-\alpha}$ , can also be expressed as

**Equation 7.18**

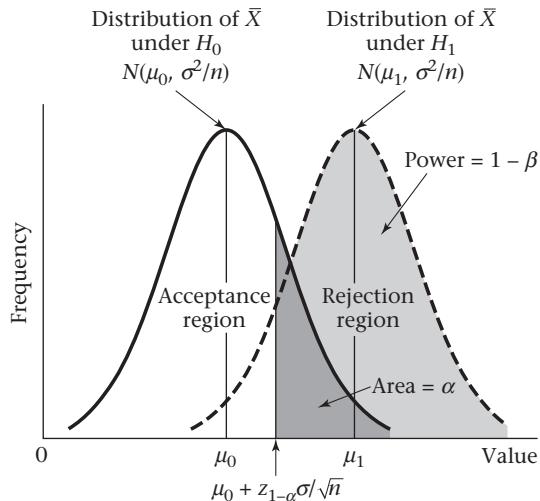
$$\bar{x} \leq \mu_0 + z_{1-\alpha} \sigma/\sqrt{n}$$

Hence the power of the test is given by

$$\begin{aligned} \text{Power} &= Pr\left(\bar{X} > \mu_0 + z_{1-\alpha} \sigma/\sqrt{n} | \mu = \mu_1\right) = 1 - Pr\left(\bar{X} < \mu_0 + z_{1-\alpha} \sigma/\sqrt{n} | \mu = \mu_1\right) \\ &= 1 - \Phi\left(\frac{\mu_0 + z_{1-\alpha} \sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left[z_{1-\alpha} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right] \end{aligned}$$

Using the relationships  $\Phi(-x) = 1 - \Phi(x)$  and  $z_\alpha = -z_{1-\alpha}$ , this expression can be rewritten as

**Figure 7.6** Illustration of power for the one-sample test for the mean of a normal distribution with known variance ( $\mu_1 > \mu_0$ )



$$\Phi\left[-z_{1-\alpha} + \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right] = \Phi\left[z_\alpha + \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right] \text{ if } \mu_1 > \mu_0$$

This power is displayed in Figure 7.6.

### Example 7.27

**Cardiovascular Disease, Pediatrics** Using a 5% level of significance and a sample of size 10, compute the power of the test for the cholesterol data in Example 7.18, with an alternative mean of 190 mg/dL, a null mean of 175 mg/dL, and a standard deviation ( $\sigma$ ) of 50 mg/dL.

#### Solution

We have  $\mu_0 = 175$ ,  $\mu_1 = 190$ ,  $\alpha = .05$ ,  $\sigma = 50$ ,  $n = 10$ . Thus

$$\begin{aligned} \text{Power} &= \Phi\left[-1.645 + (190 - 175)\sqrt{10}/50\right] \\ &= \Phi(-1.645 + 15\sqrt{10}/50) = \Phi(-0.696) \\ &= 1 - \Phi(0.696) = 1 - .757 = .243 \end{aligned}$$

Therefore, the chance of finding a significant difference in this case is only 24%. Thus it is not surprising that a significant difference was not found in Example 7.18 because the sample size was too small.

The power formulas presented in this section can be summarized as follows:

### Equation 7.19

#### Power for the One-Sample z Test for the Mean of a Normal Distribution with Known Variance (One-Sided Alternative)

The power of the test for the hypothesis

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu = \mu_1$$

where the underlying distribution is normal and the population variance ( $\sigma^2$ ) is assumed known is given by

$$\Phi\left(z_\alpha + |\mu_0 - \mu_1|\sqrt{n}/\sigma\right) = \Phi\left(-z_{1-\alpha} + |\mu_0 - \mu_1|\sqrt{n}/\sigma\right)$$

Notice from Equation 7.19 that the power depends on four factors:  $\alpha$ ,  $|\mu_0 - \mu_1|$ ,  $n$ , and  $\sigma$ .

**Equation 7.20****Factors Affecting the Power**

- (1) If the significance level is made smaller ( $\alpha$  decreases),  $z_\alpha$  increases and hence the power decreases.
- (2) If the alternative mean is shifted farther away from the null mean ( $|\mu_0 - \mu_1|$  increases), then the power increases.
- (3) If the standard deviation of the distribution of individual observations increases ( $\sigma$  increases), then the power decreases.
- (4) If the sample size increases ( $n$  increases), then the power increases.

**Example 7.28**

**Cardiovascular Disease, Pediatrics** Compute the power of the test for the cholesterol data in Example 7.27 with a significance level of .01 vs. an alternative mean of 190 mg/dL.

**Solution**

If  $\alpha = .01$ , then the power is given by

$$\begin{aligned}\Phi[z_{.01} + (190 - 175)\sqrt{10}/50] &= \Phi(-2.326 + 15\sqrt{10}/50) \\ &= \Phi(-1.377) = 1 - \Phi(1.377) = 1 - .9158 \approx 8\%\end{aligned}$$

which is lower than the power of 24% for  $\alpha = .05$ , computed in Example 7.27. What does this mean? It means that if the  $\alpha$  level is lowered from .05 to .01, the  $\beta$  error will be higher or, equivalently, the power, which decreases from .24 to .08, will be lower.

**Example 7.29**

**Obstetrics** Compute the power of the test for the birthweight data in Example 7.26 with  $\mu_1 = 110$  oz rather than 115 oz.

**Solution**

If  $\mu_1 = 110$  oz, then the power is given by

$$\Phi[-1.645 + (120 - 110)10/24] = \Phi(2.522) = .994 \approx 99\%$$

which is higher than the power of 67%, as computed in Example 7.26 for  $\mu_1 = 115$  oz. What does this mean? It means that if the alternative mean changes from 115 oz to 110 oz, then the chance of finding a significant difference increases from 67% to 99%.

**Example 7.30**

**Cardiology** Compute the power of the test for the infarct-size data in Example 7.13 with  $\sigma = 10$  and  $\sigma = 15$  using an alternative mean of 20 ( $ck - g - EQ/m^2$ ) and  $\alpha = .05$ .

**Solution**

In Example 7.13,  $\mu_0 = 25$  and  $n = 8$ . Thus, if  $\sigma = 10$ , then

$$\begin{aligned}\text{Power} &= \Phi[-1.645 + (25 - 20)\sqrt{8}/10] = \Phi(-0.23) \\ &= 1 - \Phi(0.23) = 1 - .591 = .409 \approx 41\%\end{aligned}$$

whereas if  $\sigma = 15$ , then

$$\begin{aligned}\text{Power} &= \Phi[-1.645 + (25 - 20)\sqrt{8}/15] = \Phi(-0.702) \\ &= 1 - .759 = .241 \approx 24\%\end{aligned}$$

What does this mean? It means the chance of finding a significant difference declines from 41% to 24% if  $\sigma$  increases from 10 to 15.

**Example 7.31**

**Obstetrics** Assuming a sample size of 10 rather than 100, compute the power for the birthweight data in Example 7.26 with an alternative mean of 115 oz and  $\alpha = .05$ .

**Solution**

We have  $\mu_0 = 120$  oz,  $\mu_1 = 115$  oz,  $\alpha = .05$ ,  $\sigma = 24$ , and  $n = 10$ . Thus

$$\begin{aligned}\text{Power} &= \Phi\left[z_{.05} + (120 - 115)\sqrt{10}/24\right] = \Phi(-1.645 + 5\sqrt{10}/24) \\ &= \Phi(-0.986) = 1 - .838 = .162\end{aligned}$$

What does this mean? It means there is only a 16% chance of finding a significant difference with a sample size of 10, whereas there was a 67% chance with a sample size of 100 (see Example 7.26). These results imply that if 10 infants were sampled, we would have virtually no chance of finding a significant difference and would almost surely report a false-negative result.

For given levels of  $\alpha$  (.05),  $\sigma$  (24 oz),  $n$  (100), and  $\mu_0$  (120 oz), a **power curve** can be drawn for the power of the test for various alternatives  $\mu_1$ . Such a power curve is shown in Figure 7.7 for the birthweight data in Example 7.2. The power ranges from 99% for  $\mu = 110$  oz to about 20% when  $\mu = 118$  oz.

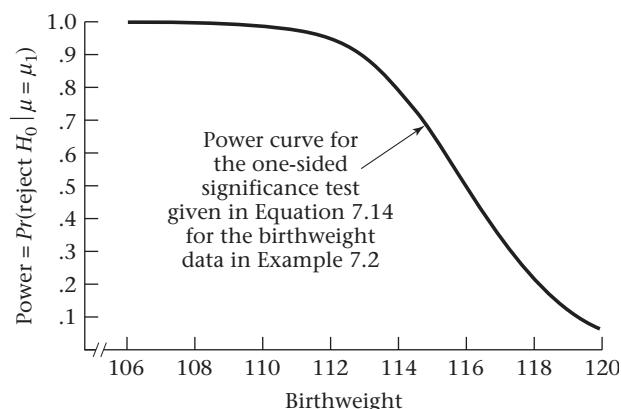
**Two-Sided Alternatives**

The power formula in Equation 7.19 is appropriate for a one-sided significance test at level  $\alpha$  for the mean of a normal distribution with known variance. Using a two-sided test with hypotheses  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ , the following power formula is used.

**Equation 7.21****Power for the One-Sample z Test for the Mean of a Normal Distribution (Two-Sided Alternative)**

The power of the two-sided test  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$  for the specific alternative  $\mu = \mu_1$ , where the underlying distribution is normal and the population variance ( $\sigma^2$ ) is assumed known, is given exactly by

$$\Phi\left[-z_{1-\alpha/2} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right] + \Phi\left[-z_{1-\alpha/2} + \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right]$$

**Figure 7.7** Power curve for the birthweight data in Example 7.2

and approximately by

$$\Phi\left[-z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1|\sqrt{n}}{\sigma}\right]$$

To see this, note that from Equation 7.13 we reject  $H_0$  if

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2} \quad \text{or} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2}$$

If we multiply each inequality by  $\sigma/\sqrt{n}$  and add  $\mu_0$ , we can re-express the rejection criteria in terms of  $\bar{x}$ , as follows:

$$\text{Equation 7.22} \quad \bar{x} < \mu_0 + z_{\alpha/2} \sigma/\sqrt{n} \quad \text{or} \quad \bar{x} > \mu_0 + z_{1-\alpha/2} \sigma/\sqrt{n}$$

The power of the test vs. the specific alternative  $\mu = \mu_1$  is given by

$$\begin{aligned} \text{Equation 7.23} \quad \text{Power} &= Pr(\bar{X} < \mu_0 + z_{\alpha/2} \sigma/\sqrt{n} | \mu = \mu_1) + Pr(\bar{X} > \mu_0 + z_{1-\alpha/2} \sigma/\sqrt{n} | \mu = \mu_1) \\ &= \Phi\left(\frac{\mu_0 + z_{\alpha/2} \sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(\frac{\mu_0 + z_{1-\alpha/2} \sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left[z_{\alpha/2} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right] + 1 - \Phi\left[z_{1-\alpha/2} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right] \end{aligned}$$

Using the relationship  $1 - \Phi(x) = \Phi(-x)$ , the last two terms can be combined as follows:

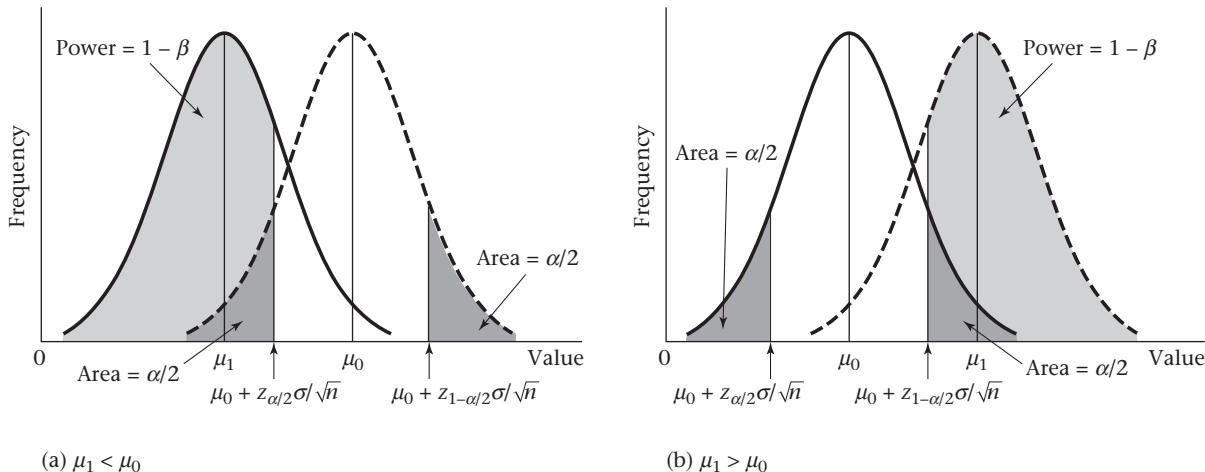
$$\text{Equation 7.24} \quad \text{Power} = \Phi\left[z_{\alpha/2} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right] + \Phi\left[-z_{1-\alpha/2} + \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right]$$

Finally, recalling the relationship  $z_{\alpha/2} = -z_{1-\alpha/2}$ , we have

$$\text{Equation 7.25} \quad \text{Power} = \Phi\left[-z_{1-\alpha/2} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right] + \Phi\left[-z_{1-\alpha/2} + \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right]$$

Equation 7.25 is more tedious to use than is usually necessary. Specifically, if  $\mu_1 < \mu_0$ , then the second term is usually negligible relative to the first term. However, if  $\mu_1 > \mu_0$ , then the first term is usually negligible relative to the second term. Therefore, the approximate power formula in Equation 7.21 is usually used for a two-sided test because it represents the first term in Equation 7.25 if  $\mu_0 > \mu_1$  and the second term in Equation 7.25 if  $\mu_1 > \mu_0$ . The power is displayed in Figure 7.8. Note that the approximate power formula for the two-sided test in Equation 7.21 is the same as the formula for the one-sided test in Equation 7.19, with  $\alpha$  replaced by  $\alpha/2$ .

**Figure 7.8 Illustration of power for a two-sided test for the mean of a normal distribution with known variance**



**Example 7.32**

**Cardiology** A new drug in the class of calcium-channel blockers is to be tested for the treatment of patients with unstable angina, a severe form of angina. The effect this drug will have on heart rate is unknown. Suppose 20 patients are to be studied and the change in heart rate after 48 hours is known to have a standard deviation of 10 beats per minute. What power would such a study have of detecting a significant difference in heart rate over 48 hours if it is hypothesized that the true mean change in heart rate from baseline to 48 hours could be either a mean increase or a decrease of 5 beats per minute?

**Solution**

Use Equation 7.21 with  $\sigma = 10$ ,  $|\mu_0 - \mu_1| = 5$ ,  $\alpha = .05$ ,  $n = 20$ . We have

$$\text{Power} = \Phi(-z_{.05/2} + 5\sqrt{20}/10) = \Phi(-1.96 + 2.236) = \Phi(0.276) = .609 \approx .61$$

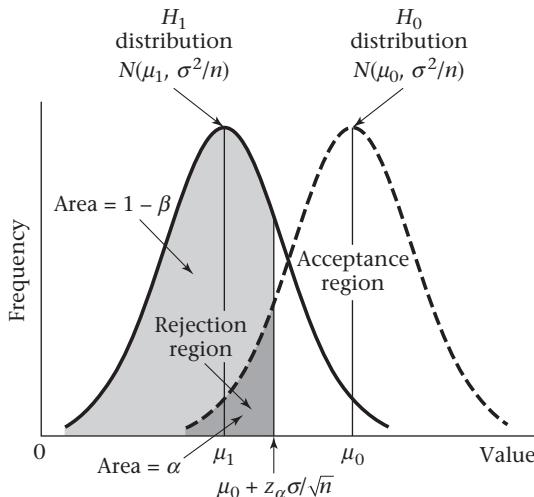
Thus the study would have a 61% chance of detecting a significant difference.

## 7.6 Sample-Size Determination

### One-Sided Alternatives

For planning purposes, we frequently need some idea of an appropriate sample size for investigation before a study actually begins. One possible result of making these calculations is finding out that the appropriate sample size is far beyond the financial means of the investigator(s) and thus abandoning the proposed investigation. Obviously, reaching this conclusion before a study starts is much better than after it is in progress.

What does “an appropriate sample size for investigation” actually mean? Consider the birthweight data in Example 7.2. We are testing the null hypothesis  $H_0: \mu = \mu_0$  vs. the alternative hypothesis  $H_1: \mu = \mu_1$ , assuming that the distribution of birthweights is normal in both cases and that the standard deviation  $\sigma$  is known. We are presumably going to conduct a test with significance level  $\alpha$  and have some idea of what the magnitude of the alternative mean  $\mu_1$  is likely to be. If the test procedure in Equation 7.14 is used, then  $H_0$  would be rejected if  $z < z_\alpha$  or equivalently

**Figure 7.9 Requirements for appropriate sample size**

if  $\bar{x} < \mu_0 + z_\alpha \sigma / \sqrt{n}$  and accepted if  $\bar{x} \geq z_\alpha$  or equivalently if  $\bar{x} \geq \mu_0 + z_\alpha \sigma / \sqrt{n}$ . Suppose the alternative hypothesis is actually true. The investigator should have some idea as to what he or she would like the probability of rejecting  $H_0$  to be in this instance. This probability is, of course, nothing other than the power, or  $1 - \beta$ . Typical values for the desired power are 80%, 90%, ..., and so forth. The problem of determining **sample size** can be summarized as follows: Given that a one-sided significance test will be conducted at level  $\alpha$  and that the true alternative mean is expected to be  $\mu_1$ , what sample size is needed to be able to detect a significant difference with probability  $1 - \beta$ ? The situation is displayed in Figure 7.9.

In Figure 7.9, the underlying sampling distribution of  $\bar{X}$  is shown under the null and alternative hypotheses, respectively, and the critical value  $\mu_0 + z_\alpha \sigma / \sqrt{n}$  has been identified.  $H_0$  will be rejected if  $\bar{x} < \mu_0 + z_\alpha \sigma / \sqrt{n}$ . Hence the area to the left of  $\mu_0 + z_\alpha \sigma / \sqrt{n}$  under the rightmost curve is  $\alpha$ . However, we also want the area to the left of  $\mu_0 + z_\alpha \sigma / \sqrt{n}$  under the leftmost curve, which represents the power, to be  $1 - \beta$ . These requirements will be met if  $n$  is made sufficiently large, because the variance of each curve ( $\sigma^2/n$ ) will decrease as  $n$  increases and thus the curves will separate. From the power formula in Equation 7.19,

$$\text{Power} = \Phi(z_\alpha + |\mu_0 - \mu_1| \sqrt{n} / \sigma) = 1 - \beta$$

We want to solve for  $n$  in terms of  $\alpha$ ,  $\beta$ ,  $|\mu_0 - \mu_1|$ , and  $\sigma$ . To accomplish this, recall that  $\Phi(z_{1-\beta}) = 1 - \beta$  and, therefore,

$$z_\alpha + |\mu_0 - \mu_1| \sqrt{n} / \sigma = z_{1-\beta}$$

Subtract  $z_\alpha$  from both sides of the equation and multiply by  $\sigma / |\mu_0 - \mu_1|$  to obtain

$$\sqrt{n} = \frac{(-z_\alpha + z_{1-\beta})\sigma}{|\mu_0 - \mu_1|}$$

Replace  $-z_\alpha$  by  $z_{1-\alpha}$  and square both sides of the equation to obtain

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

Similarly, if we were to test the hypothesis

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu = \mu_1 > \mu_0$$

as was the case with the cholesterol data in Example 7.1, using a significance level of  $\alpha$  and a power of  $1 - \beta$ , then, from Equation 7.19, the same sample-size formula would hold. This procedure can be summarized as follows.

### Equation 7.26

#### Sample-Size Estimation When Testing for the Mean of a Normal Distribution (One-Sided Alternative)

Suppose we wish to test

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu = \mu_1$$

where the data are normally distributed with mean  $\mu$  and known variance  $\sigma^2$ . The **sample size** needed to conduct a one-sided test with significance level  $\alpha$  and probability of detecting a significant difference  $= 1 - \beta$  is

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{(\mu_0 - \mu_1)^2}$$

### Example 7.33

**Obstetrics** Consider the birthweight data in Example 7.2. Suppose that  $\mu_0 = 120$  oz,  $\mu_1 = 115$  oz,  $\sigma = 24$ ,  $\alpha = .05$ ,  $1 - \beta = .80$ , and we use a one-sided test. Compute the appropriate sample size needed to conduct the test.

#### Solution

$$n = \frac{24^2 (z_{.8} + z_{.95})^2}{25} = 23.04(0.84 + 1.645)^2 = 23.04(6.175) = 142.3$$

The sample size is always rounded up so we can be sure to achieve at least the required level of power (in this case, 80%). Thus a sample size of 143 is needed to have an 80% chance of detecting a significant difference at the 5% level if the alternative mean is 115 oz and a one-sided test is used.

Notice that the sample size is very sensitive to the alternative mean chosen. We see from Equation 7.26 that the sample size is inversely proportional to  $(\mu_0 - \mu_1)^2$ . Thus, if the absolute value of the distance between the null and alternative means is halved, then the sample size needed is four times as large. Similarly, if the distance between the null and alternative means is doubled, then the sample size needed is 1/4 as large.

### Example 7.34

**Obstetrics** Compute the sample size for the birthweight data in Example 7.2 if  $\mu_1 = 110$  oz rather than 115 oz.

#### Solution

The required sample size would be 1/4 as large because  $(\mu_0 - \mu_1)^2 = 100$  rather than 25. Thus  $n = 35.6$ , or 36 people, would be needed.

### Example 7.35

**Cardiovascular Disease, Pediatrics** Consider the cholesterol data in Example 7.1. Suppose the null mean is 175 mg/dL, the alternative mean is 190 mg/dL, the standard deviation is 50, and we wish to conduct a one-sided significance test at the 5% level with a power of 90%. How large should the sample size be?

**Solution**

$$n = \frac{\sigma^2(z_{1-\beta} + z_{1-\alpha})^2}{(\mu_0 - \mu_1)^2} = \frac{50^2(z_{.9} + z_{.95})^2}{(190 - 175)^2}$$

$$= \frac{2500(1.28 + 1.645)^2}{15^2} = \frac{2500(8.556)}{225} = 95.1$$

Thus 96 people are needed to achieve a power of 90% using a 5% significance level. We should not be surprised that we did not find a significant difference with a sample size of 10 in Example 7.18.

Clearly, the required sample size is related to the following four quantities.

**Equation 7.27****Factors Affecting the Sample Size**

- (1) The sample size increases as  $\sigma^2$  increases.
- (2) The sample size increases as the significance level is made smaller ( $\alpha$  decreases).
- (3) The sample size increases as the required power increases ( $1 - \beta$  increases).
- (4) The sample size decreases as the absolute value of the distance between the null and alternative means ( $|\mu_0 - \mu_1|$ ) increases.

**Example 7.36**

**Obstetrics** What would happen to the sample-size estimate in Example 7.33 if  $\sigma$  were increased to 30? If  $\alpha$  were reduced to .01? If the required power were increased to 90%? If the alternative mean were changed to 110 oz (keeping all other parameters the same in each instance)?

**Solution**

From Example 7.33 we see that 143 infants need to be studied to achieve a power of 80% using a 5% significance level with a null mean of 120 oz, an alternative mean of 115 oz, and a standard deviation of 24 oz.

If  $\sigma$  increases to 30, then we need

$$n = 30^2(z_{.8} + z_{.95})^2 / (120 - 115)^2 = 900(0.84 + 1.645)^2 / 25 = 222.3, \text{ or } 223 \text{ infants}$$

If  $\alpha$  were reduced to .01, then we need

$$n = 24^2(z_{.8} + z_{.99})^2 / (120 - 115)^2 = 576(0.84 + 2.326)^2 / 25 = 230.9, \text{ or } 231 \text{ infants}$$

If  $1 - \beta$  were increased to .9, then we need

$$n = 24^2(z_{.9} + z_{.95})^2 / (120 - 115)^2 = 576(1.28 + 1.645)^2 / 25 = 197.1, \text{ or } 198 \text{ infants}$$

If  $\mu_1$  is decreased to 110 or, equivalently, if  $|\mu_0 - \mu_1|$  is increased from 5 to 10, then we need

$$n = 24^2(z_{.8} + z_{.95})^2 / (120 - 110)^2 = 576(0.84 + 1.645)^2 / 100 = 35.6, \text{ or } 36 \text{ infants}$$

Thus the required sample size increases if  $\sigma$  increases,  $\alpha$  decreases, or  $1 - \beta$  increases. The required sample size decreases if the absolute value of the distance between the null and alternative means increases.

One question that arises is how to estimate the parameters necessary to compute sample size. It usually is easy to specify the magnitude of the null mean ( $\mu_0$ ). Similarly, by convention the type I error ( $\alpha$ ) is usually set at .05. What the level of

the power should be is somewhat less clear, although most investigators seem to feel uncomfortable with a power of less than .80. The appropriate values for  $\mu_1$  and  $\sigma^2$  are usually unknown. The parameters  $\mu_1$ ,  $\sigma^2$  may be obtained from previous work, similar experiments, or prior knowledge of the underlying distribution. In the absence of such information, the parameter  $\mu_1$  is sometimes estimated by assessing what a *scientifically important difference*  $|\mu_0 - \mu_1|$  would be in the context of the problem being studied. Conducting a small pilot study is sometimes valuable. Such a study is generally inexpensive, and one of its principal aims is to obtain estimates of  $\mu_1$  and  $\sigma^2$  for the purpose of estimating the sample size needed to conduct the major investigation.

Keep in mind that most sample-size estimates are “ballpark estimates” because of the inaccuracy in estimating  $\mu_1$  and  $\sigma^2$ . These estimates are often used merely to check that the proposed sample size of a study is close to what is actually needed rather than to identify a precise sample size.

### Sample-Size Determination (Two-Sided Alternatives)

The sample-size formula given in Equation 7.26 was appropriate for a one-sided significance test at level  $\alpha$  for the mean of a normal distribution with known variance. If it is not known whether the alternative mean ( $\mu_1$ ) is greater or less than the null mean ( $\mu_0$ ), then a two-sided test is appropriate, and the corresponding sample size needed to conduct a study with power  $1 - \beta$  is given by

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$

To see this, use the approximate power formula in Equation 7.21 and solve for  $n$  in terms of the other parameters, whereby

$$\Phi\left(-z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1| \sqrt{n}}{\sigma}\right) = 1 - \beta$$

or  $-z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1| \sqrt{n}}{\sigma} = z_{1-\beta}$

If  $z_{1-\alpha/2}$  is added to both sides of the equation and the result is multiplied by  $\sigma/|\mu_0 - \mu_1|$ , we get

$$\sqrt{n} = \frac{(z_{1-\beta} + z_{1-\alpha/2})\sigma}{|\mu_0 - \mu_1|}$$

If both sides of the equation are squared, we get

$$n = \frac{(z_{1-\beta} + z_{1-\alpha/2})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

This procedure can be summarized as follows.

#### Equation 7.28

#### Sample-Size Estimation When Testing for the Mean of a Normal Distribution (Two-Sided Alternative)

Suppose we wish to test  $H_0: \mu = \mu_0$  vs.  $H_1: \mu = \mu_1$ , where the data are normally distributed with mean  $\mu$  and known variance  $\sigma^2$ . The sample size needed to conduct a two-sided test with significance level  $\alpha$  and power  $1 - \beta$  is

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$

Note that this sample size is always larger than the corresponding sample size for a one-sided test, given in Equation 7.26, because  $z_{1-\alpha/2}$  is larger than  $z_{1-\alpha}$ .

**Example 7.37**

**Cardiology** Consider a study of the effect of a calcium-channel-blocking agent on heart rate for patients with unstable angina, as described in Example 7.32. Suppose we want at least 80% power for detecting a significant difference if the effect of the drug is to change mean heart rate by 5 beats per minute over 48 hours in either direction and  $\sigma = 10$  beats per minute. How many patients should be enrolled in such a study?

**Solution**

We assume  $\alpha = .05$  and  $\sigma = 10$  beats per minute, as in Example 7.32. We intend to use a two-sided test because we are not sure in what direction the heart rate will change after using the drug. Therefore, the sample size is estimated using the two-sided formulation in Equation 7.28. We have

$$\begin{aligned} n &= \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2} \\ &= \frac{10^2 (z_{.8} + z_{.975})^2}{5^2} = \frac{100(0.84 + 1.96)^2}{25} \\ &= 4(7.84) = 31.36, \text{ or } 32 \text{ patients} \end{aligned}$$

Thus 32 patients must be studied to have at least an 80% chance of finding a significant difference using a two-sided test with  $\alpha = .05$  if the true mean change in heart rate from using the drug is 5 beats per minute. Note that in Example 7.32 the investigators proposed a study with 20 patients, which would provide only 61% power for testing the preceding hypothesis, which would have been inadequate.

If the direction of effect of the drug on heart rate were well known, then a one-sided test might be justified. In this case, the appropriate sample size could be obtained from the one-sided formulation in Equation 7.26, whereby

$$\begin{aligned} n &= \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{(\mu_0 - \mu_1)^2} = \frac{10^2 (z_{.8} + z_{.95})^2}{5^2} \\ &= \frac{100(0.84 + 1.645)^2}{25} = 4(6.175) = 24.7, \text{ or } 25 \text{ patients} \end{aligned}$$

Thus we would need to study only 25 patients for a one-sided test instead of the 32 patients needed for a two-sided test.

### Sample-Size Estimation Based on CI Width

In some instances, it is well known that the treatment has a significant effect on some physiologic parameter. Interest focuses instead on estimating the effect with a given degree of precision.

**Example 7.38**

**Cardiology** Suppose it is well known that propranolol lowers heart rate over 48 hours when given to patients with angina at standard dosage levels. A new study is proposed using a higher dose of propranolol than the standard one. Investigators are interested in estimating the drop in heart rate with high precision. How can this be done?

Suppose we quantify the precision of estimation by the width of the two-sided  $100\% \times (1 - \alpha)$  CI. Based on Equation 6.6, the  $100\% \times (1 - \alpha)$  CI for  $\mu$  = true decline in heart rate is  $\bar{x} \pm t_{n-1,1-\alpha/2} s / \sqrt{n}$ . The width of this CI is  $2t_{n-1,1-\alpha/2} s / \sqrt{n}$ . If we wish this interval to be no wider than  $L$ , then

$$2t_{n-1,1-\alpha/2} s / \sqrt{n} = L$$

We multiply both sides of the equation by  $\sqrt{n}/L$  and obtain

$$2t_{n-1,1-\alpha/2} s / L = \sqrt{n}$$

or, on squaring both sides,

$$n = 4t_{n-1,1-\alpha/2}^2 s^2 / L^2$$

We usually approximate  $t_{n-1,1-\alpha/2}$  by  $z_{1-\alpha/2}$  and obtain the following result:

**Equation 7.29****Sample-Size Estimation Based on CI Width**

Suppose we wish to estimate the mean of a normal distribution with sample variance  $s^2$  and require that the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$  be no wider than  $L$ . The number of subjects needed is approximately

$$n = 4z_{1-\alpha/2}^2 s^2 / L^2$$

**Example 7.39**

**Cardiology** Find the minimum sample size needed to estimate the change in heart rate ( $\mu$ ) in Example 7.38 if we require that the two-sided 95% CI for  $\mu$  be no wider than 5 beats per minute and the sample standard deviation for change in heart rate equals 10 beats per minute.

**Solution**

We have  $\alpha = .05$ ,  $s = 10$ ,  $L = 5$ . Therefore, from Equation 7.29,

$$\begin{aligned} n &= 4(z_{.975})^2 (10)^2 / (5)^2 \\ &= 4(1.96)^2 (100) / 25 = 61.5 \end{aligned}$$

Thus 62 patients need to be studied.

**REVIEW QUESTIONS 7 B**

- 1 In the BMD study referred to in Case Study 2 (p. 30), the mean weight difference between the heavier-smoking twin and the lighter-smoking twin was  $-5.0\% \pm 3.1\%$  (mean  $\pm$  se) based on 41 pairs (expressed as a percentage of the pair mean). Is there a significant difference in weight between the heavier- and the lighter-smoking twin?
- 2 What is the power of a test? What factors affect the power and in what way?
- 3 What factors affect the sample-size estimate for a study? What is the principal difference between a power estimate and a sample-size estimate? When do we use each?

## 7.7 The Relationship Between Hypothesis Testing and Confidence Intervals

A test procedure was presented in Equation 7.10 for testing the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ . Similarly, a method for obtaining a two-sided CI for the parameter  $\mu$  of a normal distribution when the variance is unknown was discussed in Section 6.5. The relationship between these two procedures can be stated as follows.

### Equation 7.30

#### The Relationship Between Hypothesis Testing and Confidence Intervals (Two-Sided Case)

Suppose we are testing  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ .  $H_0$  is rejected with a two-sided level  $\alpha$  test if and only if the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$  does not contain  $\mu_0$ .  $H_0$  is accepted with a two-sided level  $\alpha$  test if and only if the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$  does contain  $\mu_0$ .

Recall that the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu = (c_1, c_2) = \bar{x} \pm t_{n-1,1-\alpha/2} s/\sqrt{n}$ . Suppose we reject  $H_0$  at level  $\alpha$ . Then either  $t < -t_{n-1,1-\alpha/2}$  or  $t > t_{n-1,1-\alpha/2}$ . Suppose that

$$t = (\bar{x} - \mu_0) / (s/\sqrt{n}) < -t_{n-1,1-\alpha/2}$$

We multiply both sides by  $s/\sqrt{n}$  and obtain

$$\bar{x} - \mu_0 < -t_{n-1,1-\alpha/2} s/\sqrt{n}$$

If we add  $\mu_0$  to both sides, then

$$\bar{x} < \mu_0 - t_{n-1,1-\alpha/2} s/\sqrt{n}$$

or

$$\mu_0 > \bar{x} + t_{n-1,1-\alpha/2} s/\sqrt{n} = c_2$$

Similarly, if  $t > t_{n-1,1-\alpha/2}$ , then

$$\bar{x} - \mu_0 > t_{n-1,1-\alpha/2} s/\sqrt{n}$$

or

$$\mu_0 < \bar{x} - t_{n-1,1-\alpha/2} s/\sqrt{n} = c_1$$

Thus, if we reject  $H_0$  at level  $\alpha$  using a two-sided test, then either  $\mu_0 < c_1$  or  $\mu_0 > c_2$ ; that is,  $\mu_0$  must fall outside the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$ . Similarly, it can be shown that if we accept  $H_0$  at level  $\alpha$  using a two-sided test, then  $\mu_0$  must fall within the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$  (or,  $c_1 \leq \mu_0 \leq c_2$ ).

Hence, this relationship is the rationale for using CIs in Chapter 6 to decide on the reasonableness of specific values for the parameter  $\mu$ . If any specific proposed value  $\mu_0$  did not fall in the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$ , then we stated that it was an unlikely value for the parameter  $\mu$ . Equivalently, we could have tested the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$  and rejected  $H_0$  at significance level  $\alpha$ .

Here is another way of expressing this relationship.

**Equation 7.31**

The two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$  contains all values  $\mu_0$  such that we accept  $H_0$  using a two-sided test with significance level  $\alpha$ , where the hypotheses are  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ . Conversely, the  $100\% \times (1 - \alpha)$  CI *does not* contain any value  $\mu_0$  for which we can reject  $H_0$  using a two-sided test with significance level  $\alpha$ , where  $H_0: \mu = \mu_0$  and  $H_1: \mu \neq \mu_0$ .

**Example 7.40**

**Cardiovascular Disease** Consider the cholesterol data in Example 7.20. We have  $\bar{x} = 181.52$  mg/dL,  $s = 40$  mg/dL, and  $n = 100$ . The two-sided 95% CI for  $\mu$  is given by

$$\begin{aligned} & (\bar{x} - t_{99, .975} s / \sqrt{n}, \bar{x} + t_{99, .975} s / \sqrt{n}) \\ &= [\bar{x} - \text{TINV}(.05, 99) s / \sqrt{n}, \bar{x} + \text{TINV}(.05, 99) s / \sqrt{n}] \\ &= \left[ 181.52 - \frac{1.984(40)}{10}, 181.52 + \frac{1.984(40)}{10} \right] \\ &= (181.52 - 7.94, 181.52 + 7.94) = (173.58, 189.46) \end{aligned}$$

This CI contains all values for  $\mu_0$  for which we accept  $H_0: \mu = \mu_0$  and does not contain any value  $\mu_0$  for which we could reject  $H_0$  at the 5% level. Specifically, the 95% CI (173.58, 189.46) does not contain  $\mu_0 = 190$ , which corresponds to the decision in Example 7.21, where we were able to reject  $H_0: \mu = 190$  at the 5% level of significance.

Another way of stating this is that the  $p$ -value computed in Example 7.22 for  $\mu_0 = 190 = .037$ , which is less than .05.

**Example 7.41**

**Cardiovascular Disease** Suppose the sample mean for cholesterol was 185 mg/dL for the cholesterol data in Example 7.20. The 95% CI would be

$$(185 - 7.94, 185 + 7.94) = (177.06, 192.94)$$

which contains the null mean (190). The  $p$ -value for the hypothesis test would be

$$\begin{aligned} p &= 2 \times \Pr[t_{99} < (185 - 190)/4] = 2 \times \Pr(t_{99} < -1.25) \\ &= \text{TDIST}(1.25, 99, 2) \\ &= .214 > .05 \end{aligned}$$

using the TDIST function of Excel. So we can accept  $H_0$  using  $\alpha = .05$ , if  $\mu_0 = 190$ , which is consistent with the statement that 190 falls within the above 95% CI. Thus, the conclusions based on the CI and hypothesis-testing approaches are also the same here.

A similar relationship exists between the one-sided hypothesis test developed in Section 7.3 and the one-sided CI for the parameter  $\mu$  developed in Section 6.10. Equivalent CI statements can also be made about most of the other one-sided or two-sided hypothesis tests covered in this text.

Because the hypothesis-testing and CI approaches yield the same conclusions, is there any advantage to using one method over the other? The  $p$ -value from a hypothesis test tells us precisely how statistically significant the results are. However, often results that are statistically significant are not very important in the context of the subject matter because the actual difference between  $\bar{x}$  and  $\mu_0$

may not be very large, but the results are statistically significant because of a large sample size. A 95% CI for  $\mu$  would give additional information because it would provide a range of values within which  $\mu$  is likely to fall. Conversely, the 95% CI does not contain all the information contained in a  $p$ -value. It does not tell us precisely how significant the results are but merely tells us whether they are significant at the 5% level. Hence it is good practice to compute both a  $p$ -value and a 95% CI for  $\mu$ .

Unfortunately, some researchers have become polarized on this issue, with some statisticians favoring only the hypothesis-testing approach and some epidemiologists favoring only the CI approach. These issues have correspondingly influenced editorial policy, with some journals *requiring* that results be presented in one format or the other. The crux of the issue is that, traditionally, results need to be statistically significant (at the 5% level) to demonstrate the validity of a particular finding. One advantage of this approach is that a uniform statistical standard is provided (the 5% level) for *all* researchers to demonstrate evidence of an association. This protects the research community against scientific claims not based on any statistical or empirical criteria whatsoever (such as solely on the basis of clinical case reports). Advocates of the CI approach contend that the width of the CI provides information on the likely magnitude of the differences between groups, regardless of the level of significance. My opinion is that significance levels and confidence limits provide complementary information and both should be reported, where possible.

### Example 7.42

**Cardiovascular Disease** Consider the cholesterol data in Examples 7.22 and 7.40. The  $p$ -value of .037, computed in Example 7.22, tells us precisely how significant the results are. The 95% CI for  $\mu = (173.58, 189.46)$  computed in Example 7.40 gives a range of likely values that  $\mu$  might assume. The two types of information are complementary.

## 7.8 Bayesian Inference

One limitation of the methods of interval estimation in Section 6.5 or the methods of hypothesis testing in Sections 7.1–7.7 is that it is difficult to make direct statements such as  $Pr(c_1 < \mu < c_2) = 1 - \alpha$ . Instead, we have made statements such as Equation 6.7.

### Equation 7.32

$$Pr\left(\bar{x} - z_{1-\alpha/2} \sigma / \sqrt{n} < \mu < \bar{x} + z_{1-\alpha/2} \sigma / \sqrt{n}\right) = 1 - \alpha$$

Equation 7.32 indicates that  $100\% \times (1 - \alpha)$  of the random intervals depicted in Equation 6.7 will include the unknown parameter  $\mu$ , where we have assumed for simplicity that the standard deviation,  $\sigma$ , is known. Thus we cannot compute the probability that a specific interval of the form  $(\bar{x} - z_{1-\alpha/2} \sigma / \sqrt{n}, \bar{x} + z_{1-\alpha/2} \sigma / \sqrt{n})$  will contain  $\mu$ , which intuitively is what we would desire. However, we can use **Bayesian inference** for this purpose.

To use Bayesian methods, we need to specify a prior distribution for  $\mu$ . As stated in Chapter 3, a *prior distribution* is a generalization of the concept of a prior probability, where a probability distribution of possible values for a parameter (such as  $\mu$ ) is specified before looking at the available sample data. This distribution is then modified after some sample data are acquired to obtain a *posterior distribution* for  $\mu$ .

In the absence of data, we often have no knowledge of  $\mu$ , so all possible values of  $\mu$  are equally likely. This is called a *flat* or *noninformative prior distribution* and is written as

**Equation 7.33**

$$Pr(\mu) \propto c$$

where  $c$  is a constant.

We then need to determine the posterior distribution of  $\mu$  given our sample data  $x_1, \dots, x_n$ , which is written as  $Pr(\mu | x_1, \dots, x_n)$ . It can be shown that if the distribution is normal and  $\sigma$  is known, then all the information in the sample concerning  $\mu$  is contained in  $\bar{x}$ , or in other words that

**Equation 7.34**

$$Pr(\mu | x_1, \dots, x_n) = Pr(\mu | \bar{x})$$

In this case,  $\bar{x}$  is referred to as a *sufficient statistic* for  $\mu$ . Specifying a sufficient statistic(s) in a sample greatly simplifies the task of determining a posterior distribution. To obtain the posterior distribution, we use Bayes' rule.

$$Pr(\mu | \bar{x}) = Pr(\bar{x} | \mu) Pr(\mu) / Pr(\bar{x})$$

Because  $Pr(\bar{x})$  is an expression that does not involve  $\mu$ , we can write the posterior distribution in the form

$$Pr(\mu | \bar{x}) \propto Pr(\bar{x} | \mu) Pr(\mu)$$

$$\text{where } \int_{-\infty}^{\infty} Pr(\mu | \bar{x}) d\mu = 1$$

However, the prior probability of  $\mu$ —that is,  $Pr(\mu)$ —is the same for all values of  $\mu$ . Thus it follows that

**Equation 7.35**

$$Pr(\mu | \bar{x}) \propto Pr(\bar{x} | \mu)$$

In the case of a continuous distribution, we approximate  $Pr(\bar{x} | \mu)$  by  $f(\bar{x} | \mu)$ , where  $f(\bar{x} | \mu)$  is the probability density of  $\bar{x}$  given  $\mu$ . Also, we know from Equation 5.10 that  $\bar{x}$  is normally distributed with mean =  $\mu$  and variance =  $\sigma^2/n$ . Therefore,

**Equation 7.36**

$$\begin{aligned} f(\bar{x} | \mu) &= \left[ 1 / (\sqrt{2\pi} \sigma / \sqrt{n}) \right] \exp \left\{ -(1/2) \left[ (\bar{x} - \mu) / (\sigma / \sqrt{n}) \right]^2 \right\} \\ &= \left[ 1 / (\sqrt{2\pi} \sigma / \sqrt{n}) \right] \exp \left\{ -(1/2) \left[ (\mu - \bar{x}) / (\sigma / \sqrt{n}) \right]^2 \right\} \end{aligned}$$

It follows from Equations 7.35 and 7.36 that

**Equation 7.37**

$$Pr(\mu | \bar{x}) \propto \left[ 1 / (\sqrt{2\pi} \sigma / \sqrt{n}) \right] \exp \left\{ -(1/2) \left[ (\mu - \bar{x}) / (\sigma / \sqrt{n}) \right]^2 \right\}$$

Finally, we can replace the proportionality sign with an equality sign on the right-hand side of Equation 7.37 because the expression on the right-hand side of Equation 7.37 is a probability density whose integral from  $-\infty$  to  $\infty$  must be 1. It follows

from Equation 7.37 that the distribution of  $\mu$  given  $\bar{x}$  is normally distributed with mean =  $\bar{x}$  and variance =  $\sigma^2/n$ , or

$$\text{Equation 7.38} \quad \mu | \bar{x} \sim N(\bar{x}, \sigma^2/n)$$

From Equation 7.38, we can specify a  $100\% \times (1 - \alpha)$  posterior predictive interval for  $\mu$  based on the sample data of the form

$$\text{Equation 7.39} \quad Pr(\mu_1 < \mu < \mu_2) = 1 - \alpha$$

where

$$\mu_1 = \bar{x} - z_{1-\alpha/2} \sigma / \sqrt{n}, \mu_2 = \bar{x} + z_{1-\alpha/2} \sigma / \sqrt{n}$$

### Example 7.43

**Cardiovascular Disease** Consider the cholesterol data in Example 7.21. Assume that the standard deviation is known to be 40 and the sample size is 200 rather than 100. What is the 95% posterior predictive interval for  $\mu$ ?

### Solution

We have  $\bar{x} = 181.52$ ,  $\sigma = 40$ , and  $\alpha = .05$ . Hence, based on Equation 7.38, we have the 95% posterior predictive interval for  $\mu = (\mu_1, \mu_2)$ , where

$$\begin{aligned}\mu_1 &= \bar{x} - z_{1-\alpha/2} \sigma / \sqrt{n} = 181.52 - z_{.975}(40) / \sqrt{200} \\ &= 181.52 - 1.96(40) / \sqrt{200} \\ &= 181.52 - 5.54 = 176.0 \\ \mu_2 &= 181.52 + 5.54 = 187.1\end{aligned}$$

Thus,  $Pr(176.0 < \mu < 187.1) = 95\%$ .

Furthermore, because  $\mu \sim N(\bar{x}, \sigma^2/n) = N(181.52, 40^2/200)$ , it follows that

$$\begin{aligned}Pr(\mu < 190) &= \Phi\left[(190 - 181.52) / (40 / \sqrt{200})\right] \\ &= \Phi(8.48 / 2.83) = \Phi(3.00) = .999\end{aligned}$$

Thus it is highly likely that  $\mu$  is less than 190, where 190 is the mean serum cholesterol for 21- to 40-year-old women in the general population. This type of information cannot be obtained from frequentist inference.

In the case of an uninformative prior, inferences using Bayesian inference are very similar to those using frequentist inference. Specifically, the 95% posterior predictive interval in Example 7.43 is the same as a two-sided 95% CI for  $\mu$  if  $\sigma$  is known. Thus, whether or not  $\mu_0$  falls within a 95% CI or a 95% posterior predictive interval will be the same, and based on Section 7.7, the conclusions we would draw from the data are also the same. However, some cases offer extensive prior information concerning a parameter. In such cases, inferences from frequentist and Bayesian approaches can be different.

### Example 7.44

**Hypertension** Suppose a 40-year-old African-American man has his blood pressure measured at a routine medical visit. Two readings are obtained from the patient, and the mean of the two diastolic blood pressure (DBP) readings ( $\bar{x}$ ) is 95 mm Hg. It usually is recommended that a person be given antihypertensive medication if their “usual” DBP level is  $\geq 90$  mm Hg. We will interpret the usual level of DBP as the underlying average DBP over many visits for that person, where two readings are

obtained at each visit (that is,  $\mu$ ). The question is: What is  $p = Pr(\mu \geq 90 | \bar{x} = 95)$ ? If  $p$  is sufficiently high (for example,  $\geq 80\%$ ), then a patient will be put on treatment; otherwise the patient will come back for additional visits and the mean DBP over more than one visit will be obtained.

One approach to this problem would be to construct a 95% CI for  $\mu$  based on the observed mean DBP. However, this interval will be very wide because a patient usually only comes back for a few (for example,  $\leq 3$ ) visits. In the preceding example, the 95% CI cannot be obtained at all because we only have one visit ( $n = 1$ ) and  $df = n - 1 = 0$ . Thus a Bayesian solution to this problem seems more appropriate. In this case, extensive prior information exists concerning the distribution of DBP in the general population. Specifically, the Hypertension Detection and Follow-Up Program screened 158,955 people in a nationwide study with the purpose of identifying individuals with elevated blood pressure who would be eligible for a subsequent antihypertensive-treatment trial. There were a total of 7811 African-American men ages 30–49 years screened whose mean DBP was 87.8 mm Hg with a standard deviation of 11.9 mm Hg [1]. We will refer to 11.9<sup>2</sup> as the between-person variance ( $\sigma_p^2$ ) and 87.8 as the average underlying mean DBP over many 30- to 49-year-old African-American men (or  $\mu_p$ ). Also, let's assume the distribution of DBP is approximately normal. Thus we have an excellent informative prior for  $\mu$ , namely

$$\mu \sim N(\mu_p, \sigma_p^2) = N(87.8, 11.9^2)$$

Furthermore, in a separate reproducibility study [2], 41 African-American men ages 30–49 years were seen over four visits 1 week apart to determine within-person variability. The within-person variance ( $\sigma_w^2$ ) was 45.9. Therefore, if we assume the DBP distribution over many weeks for the same participant is normal, then  $x_{ij} \sim N(\mu_i, \sigma_w^2) = N(\mu, 45.9)$ , where  $x_{ij}$  = DBP at the  $j$ th visit for the  $i$ th subject and  $\mu_i$  = underlying mean for the  $i$ th subject. Thus, if  $\bar{x}_i$  is the mean DBP over  $n$  visits for the  $i$ th subject, and  $\mu_i$  is the underlying mean DBP for that subject, then

$$f(\bar{x}_i | \mu_i) = \left[ 1 / (\sqrt{2\pi} \sigma_w) \right] \exp \left\{ (-1/2) \left[ (\bar{x}_i - \mu_i) / (\sigma_w / \sqrt{n}) \right]^2 \right\}$$

To obtain the posterior distribution, we note that

$$Pr(\mu_i | \bar{x}_i) \propto f(\bar{x}_i | \mu_i) f(\mu)$$

where

$$f(\bar{x}_i | \mu_i) = \left[ 1 / (\sqrt{2\pi} \sigma_w) \right] \exp \left\{ (-n/2) \left[ (\bar{x}_i - \mu_i) / \sigma_w \right]^2 \right\}$$

and

$$f(\mu_i) = \left[ 1 / (\sqrt{2\pi} \sigma_p) \right] \exp \left\{ (-1/2) \left[ (\mu_i - \mu_p) / \sigma_p \right]^2 \right\}$$

After extensive algebra involving completing the square, the preceding product can be rewritten in the form

$$Pr(\mu_i | \bar{x}_i) = \left[ 1 / (\sqrt{2\pi} \sigma^*) \right] \exp \left\{ (-1/2) \left[ (\mu_i - \mu^*) / \sigma^* \right]^2 \right\}$$

where

$$\mu^* = \left( \bar{x}_i / \sigma_w^2 + \mu_p / \sigma_p^2 \right) / \left( 1 / \sigma_w^2 + 1 / \sigma_p^2 \right)$$

$$\sigma^{2*} = \left( 1 / \sigma_w^2 + 1 / \sigma_p^2 \right)^{-1}$$

Thus the posterior distribution of  $\mu_i$  given  $\bar{x}_i$  is normally distributed with mean given by  $\mu^*$ , which is a weighted average of the participant's observed mean DBP ( $\bar{x}_i$ ) and the overall mean DBP of the population of 30- to 49-year-old African-American men ( $\mu_p$ ), where the weights are inversely proportional to the within-subject and between-subject variance, respectively. The variance  $\sigma^{2*}$  is also a function of both the between-subject ( $\sigma_p^2$ ) and within-subject ( $\sigma_w^2$ ) variance. Therefore, the effect of using the posterior distribution is to "pull back" the observed mean DBP toward the overall population mean. This would explain the well-known phenomenon whereby participants who have a high DBP at a single visit relative to the population mean usually have later measurements closer to the population mean. This phenomenon is called **regression to the mean**.

**Solution to Example 7.44**

In this example,  $\bar{x}_i = 95$ ,  $\mu_p = 87.8$ ,  $\sigma_w^2 = 45.9$  and  $\sigma_p^2 = 11.9^2 = 141.61$ . Therefore,

$$\begin{aligned}\mu^* &= (95/45.9 + 87.8/141.61) / (1/45.9 + 1/141.61) \\ &= 2.6897 / 0.02885 = 93.23 \text{ mm Hg} \\ \sigma^{2*} &= (1/45.9 + 1/141.61)^{-1} = 1/0.0288 = 34.66\end{aligned}$$

Thus the posterior distribution of  $\mu_i$  is normal with mean = 93.23 and variance = 34.66. It follows that

$$\begin{aligned}Pr(\mu_i > 90 | \bar{x}_i = 95) &= 1 - \Phi\left[(90 - 93.23)/\sqrt{34.66}\right] \\ &= 1 - \Phi(-3.23/5.887) \\ &= 1 - \Phi(-0.549) = \Phi(0.549) = .71 = p\end{aligned}$$

Therefore, there is a 71% probability that this participant's true DBP is above 90. In this case, because  $p < .80$ , it probably would be advisable to have the participant come back for another visit to confirm the initial mean value of 95 mm Hg and reassess his level of blood pressure based on the average DBP over two visits.

For all the hypothesis-testing problems we consider in this book, there are both frequentist and Bayesian formulations that are possible. In general, we present only frequentist approaches. Usually the prior distribution is not known in advance, in which case inferences from frequentist and Bayesian approaches are generally similar. An excellent reference providing more coverage of Bayesian methods is [3].

## 7.9 One-Sample $\chi^2$ Test for the Variance of a Normal Distribution

**Example 7.45**

**Hypertension** Consider Example 6.39, concerning the variability of blood-pressure measurements taken on an Arteriosonde machine. We were concerned with the difference between measurements taken by two observers on the same person =  $d_i = x_{1i} - x_{2i}$ , where  $x_{1i}$  = the measurement on the  $i$ th person by the first observer and  $x_{2i}$  = the measurement on the  $i$ th person by the second observer. Let's assume this difference is a good measure of interobserver variability, and we want to compare this variability with the variability using a standard blood-pressure cuff.

We have reason to believe that the variability of the Arteriosonde machine may differ from that of a standard cuff. Intuitively, we think the variability of the new method should be lower. However, because the new method is not as widely used, the observers are probably less experienced in using it; therefore, the variability of the new method could possibly be higher than that of the old method. Thus a two-sided test is used to study this question. Suppose we know from previously published work that  $\sigma^2 = 35$  for  $d_i$  obtained from the standard cuff. We want to test the hypothesis  $H_0: \sigma^2 = \sigma_0^2 = 35$  vs.  $H_1: \sigma^2 \neq \sigma_0^2$ . How should we perform this test?

If  $x_1, \dots, x_n$  are a random sample, then we can reasonably base the test on  $s^2$  because it is an unbiased estimator of  $\sigma^2$ . We know from Equation 6.14 that if  $x_1, \dots, x_n$  are a random sample from an  $N(\mu, \sigma^2)$  distribution, then under  $H_0$ ,

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

Therefore,

$$\Pr(X^2 < \chi_{n-1, \alpha/2}^2) = \alpha/2 = \Pr(X^2 > \chi_{n-1, 1-\alpha/2}^2)$$

Hence, the test procedure is given as follows.

#### Equation 7.40

#### One-Sample $\chi^2$ Test for the Variance of a Normal Distribution (Two-Sided Alternative)

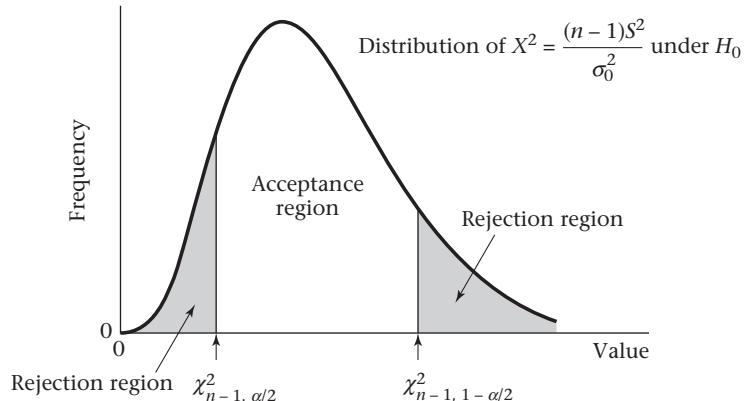
We compute the test statistic  $X^2 = (n-1)s^2/\sigma_0^2$ .

If  $X^2 < \chi_{n-1, \alpha/2}^2$  or  $X^2 > \chi_{n-1, 1-\alpha/2}^2$ , then  $H_0$  is rejected.

If  $\chi_{n-1, \alpha/2}^2 \leq X^2 \leq \chi_{n-1, 1-\alpha/2}^2$ , then  $H_0$  is accepted.

The acceptance and rejection regions for this test are shown in Figure 7.10.

**Figure 7.10** Acceptance and rejection regions for the one-sample  $\chi^2$  test for the variance of a normal distribution (two-sided alternative)



Alternatively, we may want to compute a  $p$ -value for our experiment. The computation of the  $p$ -value will depend on whether  $s^2 \leq \sigma_0^2$  or  $s^2 > \sigma_0^2$ . The rule is given as follows.

**Equation 7.41** **$p$ -Value for a One-Sample  $\chi^2$  Test for the Variance of a Normal Distribution (Two-Sided Alternative)**

Let the test statistic  $X^2 = \frac{(n-1)s^2}{\sigma_0^2}$ .

If  $s^2 \leq \sigma_0^2$ , then  $p$ -value =  $2 \times (\text{area to the left of } X^2 \text{ under a } \chi_{n-1}^2 \text{ distribution})$ .

If  $s^2 > \sigma_0^2$ , then  $p$ -value =  $2 \times (\text{area to the right of } X^2 \text{ under a } \chi_{n-1}^2 \text{ distribution})$ .

The  $p$ -values are illustrated in Figure 7.11.

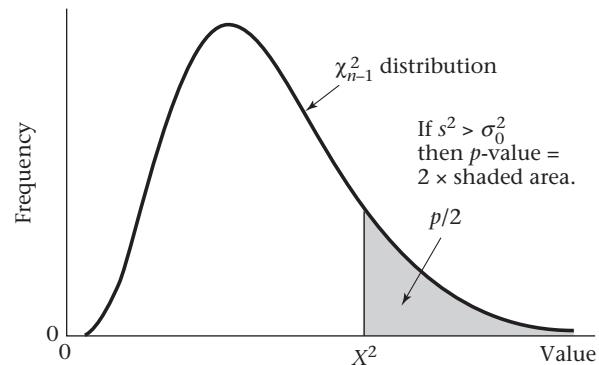
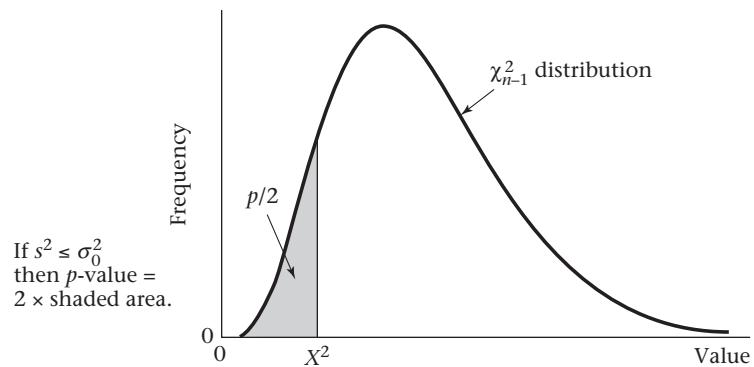
**Example 7.46**

**Hypertension** Assess the statistical significance of the Arteriosonde-machine data in Example 7.45.

**Solution**

We know from Example 6.39 that  $s^2 = 8.178$ ,  $n = 10$ . From Equation 7.40, we compute the test statistic  $X^2$  given by

$$X^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9(8.178)}{35} = 2.103$$

**Figure 7.11****Illustration of the  $p$ -value for a one-sample  $\chi^2$  test for the variance of a normal distribution (two-sided alternative)**

**Table 7.2** Computation of the exact p-value for the Arteriosonde-machine data in Example 7.46 using a one-sample  $\chi^2$  test with Microsoft Excel 2007

Microsoft Excel 2007	
x	2.103
df	9
cdf=CHIDIST(2.103,9)	0.989732
p-value(1-tail) = 1 - CHIDIST(2.103,9)	0.010268
p-value(2-tail) = 2×[1 - CHIDIST(2.103,9)]	0.020536

Under  $H_0$ ,  $X^2$  follows a  $\chi^2$  distribution with 9 degrees of freedom. Thus the critical values are  $\chi^2_{0.025} = 2.70$  and  $\chi^2_{0.975} = 19.02$ . Because  $X^2 = 2.103 < 2.70$ , it follows that  $H_0$  is rejected using a two-sided test with  $\alpha = .05$ . To obtain the p-value, refer to Equation 7.41. Because  $s^2 = 8.178 < 35 = \sigma_0^2$ , the p-value is computed as follows:

$$p = 2 \times Pr(\chi^2_9 < 2.103)$$

From Table 6 of the Appendix we see that

$$\chi^2_{0.025} = 2.70, \chi^2_{0.01} = 2.09$$

Thus, because  $2.09 < 2.103 < 2.70$ , we have  $.01 < p/2 < .025$  or  $.02 < p < .05$ .

To obtain the exact p-value, use Microsoft Excel 2007 to evaluate areas under the  $\chi^2$  distribution. The CHIDIST function computes the area to the right of 2.103 for a  $\chi^2_9$  distribution = .9897. Thus, subtract from 1 and multiply by 2 to obtain the exact two-sided p-value = .021. The details are given in Table 7.2.

Therefore, the results are statistically significant, and we conclude the interobserver variance using the Arteriosonde machine significantly differs from the interobserver variance using the standard cuff. To quantify how different the two variances are, a two-sided 95% CI for  $\sigma^2$  could be obtained, as in Example 6.41. This interval was (3.87, 27.26). Of course, it does not contain 35 because the p-value is less than .05.

In general, the assumption of normality is particularly important for hypothesis testing and CI estimation for variances. If this assumption is not satisfied, then the critical regions and p-values in Equations 7.40 and 7.41 and the confidence limits in Equation 6.15 will not be valid.

In this section, we have presented the **one-sample  $\chi^2$  test for variances**, which is used for testing hypotheses concerning the variance of a normal distribution. Beginning at the “Start” box of the flowchart (Figure 7.18, p. 258), we arrive at the one-sample  $\chi^2$  test for variances by answering yes to each of the following three questions: (1) one variable of interest? (2) one-sample problem? and (3) underlying distribution normal or can central-limit theorem be assumed to hold? and by answering no to (4) inference concerning  $\mu$ ? and yes to (5) inference concerning  $\sigma$ ?

## 7.10 One-Sample Inference for the Binomial Distribution

### Normal-Theory Methods

#### Example 7.47

**Cancer** Consider the breast-cancer data in Example 6.48. In that example we were interested in the effect of having a family history of breast cancer on the incidence of breast cancer. Suppose that 400 of the 10,000 women ages 50–54 sampled whose

mothers had breast cancer had breast cancer themselves at some time in their lives. Given large studies, assume the prevalence rate of breast cancer for U.S. women in this age group is about 2%. The question is: How compatible is the sample rate of 4% with a population rate of 2%?

Another way of asking this question is to restate it in terms of hypothesis testing: If  $p$  = prevalence rate of breast cancer in 50- to 54-year-old women whose mothers have had breast cancer, then we want to test the hypothesis  $H_0: p = .02 = p_0$  vs.  $H_1: p \neq .02$ . How can this be done?

The significance test is based on the sample proportion of cases  $\hat{p}$ . Assume the normal approximation to the binomial distribution is valid. This assumption is reasonable when  $np_0q_0 \geq 5$ . Therefore, from Equation 6.17 we know that under  $H_0$

$$\hat{p} \sim N\left(p_0, \frac{p_0q_0}{n}\right)$$

It is more convenient to standardize  $\hat{p}$ . For this purpose, we subtract the expected value of  $\hat{p}$  under  $H_0 = p_0$  and divide by the standard error of  $\hat{p}$  under  $H_0 = \sqrt{p_0q_0/n}$ , creating the test statistic  $z$  given by

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}}$$

It follows that under  $H_0$ ,  $z \sim N(0, 1)$ . Thus

$$Pr(z < z_{\alpha/2}) = Pr(z > z_{1-\alpha/2}) = \alpha/2$$

Thus the test takes the following form.

#### Equation 7.42

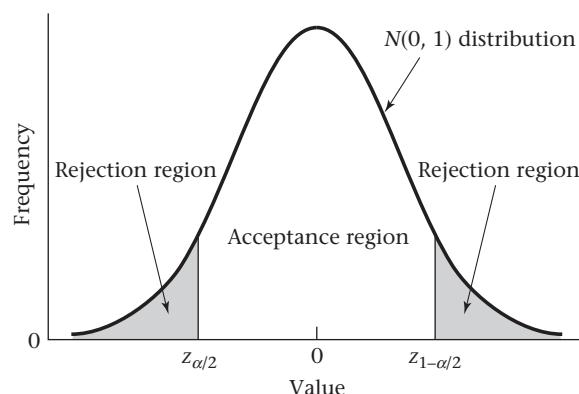
#### One-Sample Test for a Binomial Proportion—Normal-Theory Method (Two-Sided Alternative)

Let the test statistic  $z = (\hat{p} - p_0)/\sqrt{p_0q_0/n}$ .

If  $z < z_{\alpha/2}$  or  $z > z_{1-\alpha/2}$ , then  $H_0$  is rejected. If  $z_{\alpha/2} \leq z \leq z_{1-\alpha/2}$ , then  $H_0$  is accepted. This test should only be used if  $np_0q_0 \geq 5$ .

The acceptance and rejection regions are shown in Figure 7.12.

**Figure 7.12** Acceptance and rejection regions for the one-sample binomial test—normal-theory method (two-sided alternative)



Alternatively, a  $p$ -value could be computed. The computation of the  $p$ -value depends on whether  $\hat{p} \leq p_0$  or  $\hat{p} > p_0$ . If  $\hat{p} \leq p_0$ , then

$$p\text{-value} = 2 \times \text{area to the left of } z \text{ under an } N(0, 1) \text{ curve}$$

If  $\hat{p} > p_0$ , then

$$p\text{-value} = 2 \times \text{area to the right of } z \text{ under an } N(0, 1) \text{ curve}$$

This is summarized as follows.

**Equation 7.43**

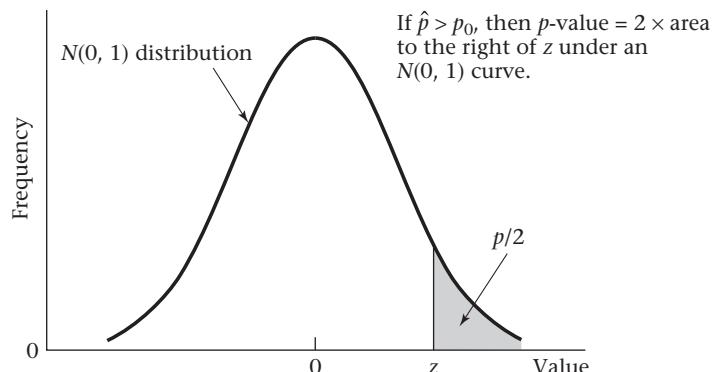
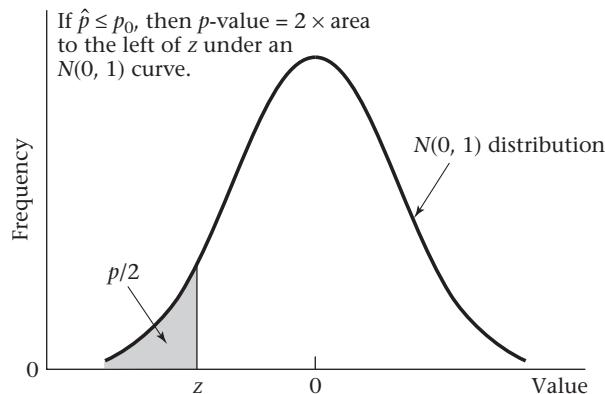
**Computation of the  $p$ -Value for the One-Sample Binomial Test—Normal-Theory Method (Two-Sided Alternative)**

Let the test statistic  $z = (\hat{p} - p_0)/\sqrt{p_0 q_0/n}$ .

If  $\hat{p} \leq p_0$ , then  $p\text{-value} = 2 \times \Phi(z) =$  twice the area to the left of  $z$  under an  $N(0, 1)$  curve. If  $\hat{p} > p_0$ , then  $p\text{-value} = 2 \times [1 - \Phi(z)] =$  twice the area to the right of  $z$  under an  $N(0, 1)$  curve. The calculation of the  $p$ -value is illustrated in Figure 7.13.

These definitions of a  $p$ -value are again compatible with the idea of a  $p$ -value as the probability of obtaining results as extreme as or more extreme than the results in our particular sample.

**Figure 7.13 Illustration of the  $p$ -value for a one-sample binomial test—normal-theory method (two-sided alternative)**



**Example 7.48**

**Cancer** Assess the statistical significance of the data in Example 7.47.

**Solution**

Using the critical-value method, we compute the test statistic

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} \\ &= \frac{.04 - .02}{\sqrt{.02(.98)/10,000}} = \frac{.02}{.0014} = 14.3 \end{aligned}$$

Because  $z_{1-\alpha/2} = z_{.975} = 1.96$ , it follows that  $H_0$  can be rejected using a two-sided test with  $\alpha = .05$ . To compute the  $p$ -value, because  $p = .04 > p_0 = .02$ , it follows that

$$\begin{aligned} p\text{-value} &= 2 \times [1 - \Phi(z)] \\ &= 2 \times [1 - \Phi(14.3)] < .001 \end{aligned}$$

Thus the results are very highly significant.

## Exact Methods

The test procedure presented in Equation 7.42 to test the hypothesis  $H_0: p = p_0$  vs.  $H_1: p \neq p_0$  depends on the assumption that the normal approximation to the binomial distribution is valid. This assumption is only true if  $np_0 q_0 \geq 5$ . How can the preceding hypothesis be tested if this criterion is not satisfied?

We will base our test on *exact* binomial probabilities. In particular, let  $X$  be a binomial random variable with parameters  $n$  and  $p_0$  and let  $\hat{p} = x/n$ , where  $x$  is the observed number of events. The computation of the  $p$ -value depends on whether  $\hat{p} \leq p_0$  or  $\hat{p} > p_0$ . If  $\hat{p} \leq p_0$ , then

$$\begin{aligned} p/2 &= Pr(\leq x \text{ successes in } n \text{ trials} | H_0) \\ &= \sum_{k=0}^x \binom{n}{k} p_0^k (1-p_0)^{n-k} \end{aligned}$$

If  $\hat{p} > p_0$ , then

$$\begin{aligned} p/2 &= Pr(\geq x \text{ successes in } n \text{ trials} | H_0) \\ &= \sum_{k=x}^n \binom{n}{k} p_0^k (1-p_0)^{n-k} \end{aligned}$$

This is summarized as follows.

**Equation 7.44**

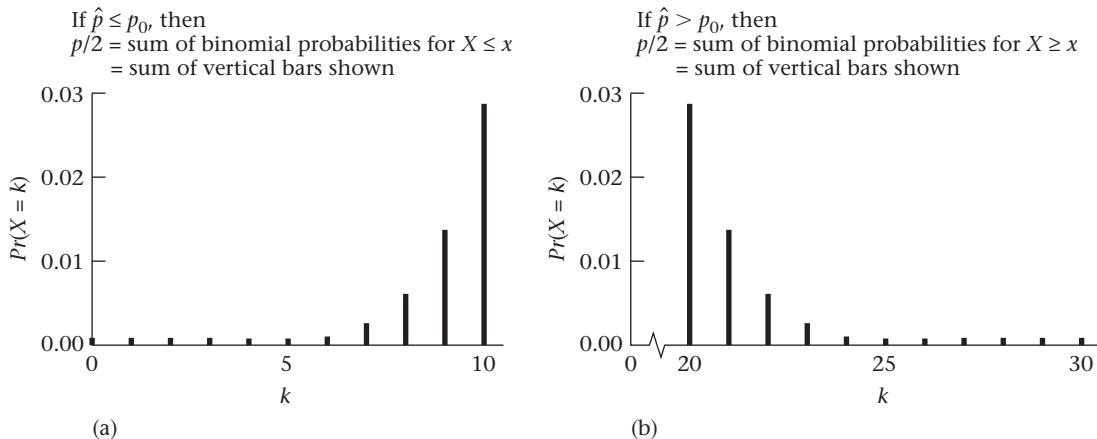
### Computation of the $p$ -Value for the One-Sample Binomial Test—Exact Method (Two-Sided Alternative)

$$\text{If } \hat{p} \leq p_0, p = 2 \times Pr(X \leq x) = \min \left[ 2 \sum_{k=0}^x \binom{n}{k} p_0^k (1-p_0)^{n-k}, 1 \right]$$

$$\text{If } \hat{p} > p_0, p = 2 \times Pr(X \geq x) = \min \left[ 2 \sum_{k=x}^n \binom{n}{k} p_0^k (1-p_0)^{n-k}, 1 \right]$$

The computation of the  $p$ -value is depicted in Figure 7.14, in the case of  $n = 30$ ,  $p_0 = .5$  and  $x = 10$  and  $20$ , respectively.

**Figure 7.14 Illustration of the  $p$ -value for a one-sample binomial test—exact method (two-sided alternative)**



In either case the  $p$ -value corresponds to the sum of the probabilities of all events that are as extreme as or more extreme than the sample result obtained.

**Example 7.49**

**Occupational Health, Cancer** The safety of people who work at or live close to nuclear-power plants has been the subject of widely publicized debate in recent years. One possible health hazard from radiation exposure is an excess of cancer deaths among those exposed. One problem with studying this question is that because the number of deaths attributable to either cancer in general or specific types of cancer is small, reaching statistically significant conclusions is difficult except after long periods of follow-up. An alternative approach is to perform a *proportional-mortality study*, whereby the proportion of deaths attributed to a specific cause in an exposed group is compared with the corresponding proportion in a large population. Suppose, for example, that 13 deaths have occurred among 55- to 64-year-old male workers in a nuclear-power plant and that in 5 of them the cause of death was cancer. Assume, based on vital-statistics reports, that approximately 20% of all deaths can be attributed to some form of cancer. Is this result significant?

**Solution**

We want to test the hypothesis  $H_0: p = .20$  vs.  $H_1: p \neq .20$ , where  $p$  = probability that the cause of death in nuclear-power workers was cancer. The normal approximation to the binomial cannot be used, because

$$np_0 q_0 = 13(.2)(.8) = 2.1 < 5$$

However, the exact procedure in Equation 7.44 can be used:

$$\hat{p} = \frac{5}{13} = .38 > .20$$

$$\text{Therefore, } p = 2 \sum_{k=5}^{13} \binom{13}{k} (.2)^k (.8)^{13-k} = 2 \times \left[ 1 - \sum_{k=0}^4 \binom{13}{k} (.2)^k (.8)^{13-k} \right]$$

From Table 1 in the Appendix, with  $n = 13$ ,  $p = .2$ , we have

$$Pr(0) = .0550$$

$$Pr(1) = .1787$$

$$\begin{aligned}Pr(2) &= .2680 \\Pr(3) &= .2457 \\Pr(4) &= .1535\end{aligned}$$

Therefore,  $p = 2 \times [1 - (.0550 + .1787 + .2680 + .2457 + .1535)]$   
 $= 2 \times (1 - .9009) = .198$

In summary, the proportion of deaths from cancer is not significantly different for nuclear-power-plant workers than for men of comparable age in the general population.

## Power and Sample-Size Estimation

The power of the one-sample binomial test can also be considered using the large-sample test procedure given on page 245. Suppose we are conducting a two-tailed test at level  $\alpha$ , where  $p = p_0$  under the null hypothesis. Under the alternative hypothesis of  $p = p_1$ , the power is given by the following formula.

### Equation 7.45

#### Power for the One-Sample Binomial Test (Two-Sided Alternative)

The power of the one-sample binomial test for the hypothesis

$$H_0: p = p_0 \text{ vs. } H_1: p \neq p_0$$

for the specific alternative  $p = p_1$  is given by

$$\Phi \left[ \sqrt{\frac{p_0 q_0}{p_1 q_1}} \left( Z_{\alpha/2} + \frac{|p_0 - p_1| \sqrt{n}}{\sqrt{p_0 q_0}} \right) \right]$$

To use this formula, we assume that  $np_0 q_0 \geq 5$  so that the normal-theory methods in this section on page 245 are valid.

### Example 7.50

**Cancer** Suppose we wish to test the hypothesis that women with a sister history of breast cancer are at higher risk of developing breast cancer themselves. Suppose we assume, as in Example 7.47, that the prevalence rate of breast cancer is 2% among 50- to 54-year-old U.S. women, whereas it is 5% among women with a sister history. We propose to interview 500 women 50 to 54 years of age with a sister history of the disease. What is the power of such a study assuming that we conduct a two-sided test with  $\alpha = .05$ ?

### Solution

We have  $\alpha = .05$ ,  $p_0 = .02$ ,  $p_1 = .05$ ,  $n = 500$ . The power, as given by Equation 7.45, is

$$\begin{aligned}\text{Power} &= \Phi \left[ \sqrt{\frac{.02(.98)}{.05(.95)}} \left( Z_{.025} + \frac{.03\sqrt{500}}{\sqrt{.02(.98)}} \right) \right] \\&= \Phi[.642(-1.96 + 4.792)] = \Phi(1.819) = .966\end{aligned}$$

Thus there should be a 96.6% chance of finding a significant difference based on a sample size of 500, if the true rate of breast cancer among women with a sister history is 2.5 times as high as that of typical 50- to 54-year-old women.

Similarly, we can consider the issue of appropriate sample size if the one-sample binomial test for a given  $\alpha$ ,  $p_0$ ,  $p_1$ , and power is being used. The sample size is given by the following formula.

**Equation 7.46****Sample-Size Estimation for the One-Sample Binomial Test (Two-Sided Alternative)**

Suppose we wish to test  $H_0: p = p_0$  vs.  $H_1: p \neq p_0$ . The sample size needed to conduct a two-sided test with significance level  $\alpha$  and power  $1 - \beta$  vs. the specific alternative hypothesis  $p = p_1$  is

$$n = \frac{p_0 q_0 \left( z_{1-\alpha/2} + z_{1-\beta} \sqrt{\frac{p_1 q_1}{p_0 q_0}} \right)^2}{(p_1 - p_0)^2}$$

**Example 7.51**

**Cancer** How many women should we interview in the study proposed in Example 7.50 to achieve 90% power if a two-sided significance test with  $\alpha = .05$  is used?

**Solution**

We have  $\alpha = .05$ ,  $1 - \beta = .90$ ,  $p_0 = .02$ ,  $p_1 = .05$ . The sample size is given by Equation 7.46:

$$\begin{aligned} n &= \frac{.02(.98) \left[ z_{.975} + z_{.90} \sqrt{\frac{.05(.95)}{.02(.98)}} \right]^2}{(.03)^2} \\ &= \frac{.0196 [1.96 + 1.28(1.557)]^2}{.0009} = \frac{.0196(15.623)}{.0009} = 340.2, \text{ or } 341 \text{ women} \end{aligned}$$

Thus, 341 women with a sister history of breast cancer must be interviewed in order to have a 90% chance of detecting a significant difference using a two-sided test with  $\alpha = .05$  if the true rate of breast cancer among women with a sister history is 2.5 times as high as that of a typical 50- to 54-year-old woman.

Note that if we wish to perform a one-sided test rather than a two-sided test at level  $\alpha$ , then  $\alpha$  is substituted for  $\alpha/2$  in the power formula in Equation 7.45 and the sample-size formula in Equation 7.46.

In this section, we have presented the **one-sample binomial test**, which is used for testing hypotheses concerning the parameter  $p$  of a binomial distribution. Beginning at the “Start” box of the flowchart (Figure 7.18, p. 258), we arrive at the one-sample binomial test by answering yes to (1) one variable of interest? and (2) one-sample problem? no to (3) underlying distribution normal or can central-limit theorem be assumed to hold? and yes to (4) underlying distribution is binomial?

**REVIEW QUESTIONS 7C**

- 1 A sample of 120 high-school students (60 boys, 60 girls) is weighed in their physical-education class. Of the students, 15% are above the 95th percentile for body-mass index (BMI) [ $\text{wt}(\text{kg})/\text{ht}^2(\text{m}^2)$ ] as determined by national norms. Health educators at the school want to determine whether the obesity profile in the school differs from what is expected.
  - (a) What hypotheses can be used to address this question?
  - (b) What test can be used to test these hypotheses?
  - (c) Write down the test statistic for this test.
  - (d) What is the  $p$ -value of the test?
  - (e) What is your overall conclusion based on your findings?

- 2** The principal at the school also wants to address the questions in Review Question 7C.1 (a)–(e) for specific ethnic groups. Of the 50 Hispanic students at the school, 10 are above the 95th percentile for BMI.
  - (a)** What test can be used to address the question in Review Question 7C.1(a)–(e) among the Hispanic students?
  - (b)** What is the  $p$ -value of the test?
  - (c)** What is your overall conclusion based on your results?
- 3** How much power did the test in Review Question 7C.1 have if the true percentage of students in the school above the 95th percentile for BMI is 15%?

## 7.11 One-Sample Inference for the Poisson Distribution

### Example 7.52

**Occupational Health** Many studies have looked at possible health hazards faced by rubber workers. In one such study, a group of 8418 white male workers ages 40–84 (either active or retired) on January 1, 1964, were followed for 10 years for various mortality outcomes [4]. Their mortality rates were then compared with U.S. white male mortality rates in 1968. In one of the reported findings, 4 deaths due to Hodgkin's disease were observed compared with 3.3 deaths expected from U.S. mortality rates. Is this difference significant?

One problem with this type of study is that workers of different ages in 1964 have very different mortality risks over time. Thus the test procedures in Equations 7.42 and 7.44, which assume a constant  $p$  for all people in the sample, are not applicable. However, these procedures can be generalized to take account of the different mortality risks of different individuals. Let

$$\begin{aligned} X &= \text{total observed number of deaths for members of the study population} \\ p_i &= \text{probability of death for the } i\text{th individual} \end{aligned}$$

Under the null hypothesis that the death rates for the study population are the same as those for the general U.S. population, the expected number of events  $\mu_0$  is given by

$$\mu_0 = \sum_{i=1}^n p_i$$

If the disease under study is rare, then the observed number of events may be considered approximately Poisson-distributed with unknown expected value =  $\mu$ . We wish to test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ .

One approach for significance testing is to use the critical-value method. We know from Section 7.7 that the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$  given by  $(c_1, c_2)$  contains all values  $\mu_0$  for which we would accept  $H_0$  based on the preceding hypothesis test. Thus if  $c_1 \leq \mu_0 \leq c_2$ , then we accept  $H_0$ , whereas if either  $\mu_0 < c_1$  or  $\mu_0 > c_2$ , then we reject  $H_0$ . Table 8 in the Appendix contains exact confidence limits for the Poisson expectation  $\mu$ , and this leads us to the following simple approach for hypothesis testing.

### Equation 7.47

### One-Sample Inference for the Poisson Distribution (Small-Sample Test—Critical-Value Method)

Let  $X$  be a Poisson random variable with expected value =  $\mu$ . To test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$  using a two-sided test with significance level  $\alpha$ ,

- (1) Obtain the two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$  based on the observed value  $x$  of  $X$ . Denote this CI  $(c_1, c_2)$ .
- (2) If  $\mu_0 < c_1$  or  $\mu_0 > c_2$ , then reject  $H_0$ .  
If  $c_1 \leq \mu_0 \leq c_2$ , then accept  $H_0$ .

**Example 7.53****Solution**

**Occupational Health** Test for the significance of the findings in Example 7.52 using the critical-value method with a two-sided significance level of .05.

We wish to test the hypothesis  $H_0: \mu = 3.3$  vs.  $H_1: \mu \neq 3.3$ . We observed 4 events =  $x$ . Hence, referring to Table 8, the two-sided 95% CI for  $\mu$  based on  $x = 4$  is (1.09, 10.24). From Equation 7.47, because  $1.09 \leq 3.3 \leq 10.24$ , we accept  $H_0$  at the 5% significance level.

Another approach to use for significance testing is the  $p$ -value method. We wish to reject  $H_0$  if  $x$  is either much larger or much smaller than  $\mu_0$ . This leads to the following test procedure.

**Equation 7.48****One-Sample Inference for the Poisson Distribution (Small-Sample Test— $p$ -Value Method)**

Let  $\mu$  = expected value of a Poisson distribution. To test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ ,

- (1) Compute

$x$  = observed number of deaths in the study population

- (2) Under  $H_0$ , the random variable  $X$  will follow a Poisson distribution with parameter  $\mu_0$ . Thus, the exact two-sided  $p$ -value is given by

$$\min \left( 2 \times \sum_{k=0}^x \frac{e^{-\mu_0} \mu_0^k}{k!}, 1 \right) \quad \text{if } x < \mu_0$$

$$\min \left[ 2 \times \left( \sum_{k=0}^{x-1} \frac{e^{-\mu_0} \mu_0^k}{k!} \right), 1 \right] \quad \text{if } x \geq \mu_0$$

These computations are shown in Figure 7.15 for the case of  $\mu_0 = 5$ , with  $x = 3$  and 8, respectively.

**Example 7.54****Solution**

**Occupational Health** Test for the significance of the findings in Example 7.52 using the  $p$ -value method.

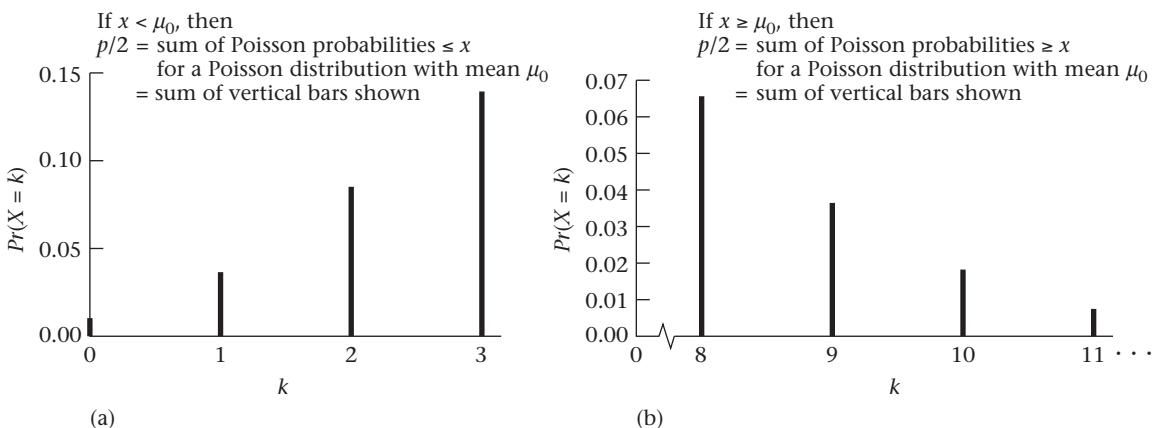
We refer to Equation 7.48. Because  $x = 4 > \mu_0 = 3.3$  the  $p$ -value is given by

$$p = 2 \times \left[ 1 - \sum_{k=0}^3 \frac{e^{-3.3} (3.3)^k}{k!} \right]$$

From the Poisson distribution, we have

$$Pr(0) = e^{-3.3} = .0369$$

$$Pr(1) = \frac{e^{-3.3} (3.3)^1}{1!} = .1217$$

**Figure 7.15 Computation of the exact p-value for the one-sample Poisson test**

$$Pr(2) = \frac{e^{-3.3}(3.3)^2}{2!} = .2008$$

$$Pr(3) = \frac{e^{-3.3}(3.3)^3}{3!} = .2209$$

$$\begin{aligned} \text{Thus } p &= 2 \times [1 - (0.0369 + 0.1217 + 0.2008 + 0.2209)] \\ &= 2 \times (1 - 0.5803) = 0.839 \end{aligned}$$

Thus there is no significant excess or deficit of Hodgkin's disease in this population.

An index frequently used to quantify risk in a study population relative to the general population is the standardized mortality ratio (SMR).

**Definition 7.16**

The **standardized mortality ratio** is defined by  $100\% \times O/E = 100\% \times \text{the observed number of deaths in the study population divided by the expected number of deaths in the study population under the assumption that the mortality rates for the study population are the same as those for the general population.}$  For nonfatal conditions the SMR is sometimes known as the **standardized morbidity ratio**.

Thus,

- If  $SMR > 100\%$ , there is an excess risk in the study population relative to the general population.
- If  $SMR < 100\%$ , there is a reduced risk in the study population relative to the general population.
- If  $SMR = 100\%$ , there is neither an excess nor a deficit of risk in the study population relative to the general population.

**Example 7.55**

**Occupational Health** What is the SMR for Hodgkin's disease using the data in Example 7.52?

**Solution**

$$SMR = 100 \times 4/3.3 = 121\%$$

The test procedures in Equations 7.47 and 7.48 can also be interpreted as tests of whether the SMR is significantly different from 100%.

**Example 7.56**

**Occupational Health** In the rubber-worker data described in Example 7.52, there were 21 bladder cancer deaths and an expected number of events from general-population cancer mortality rates of 18.1. Evaluate the statistical significance of the results.

**Solution**

We refer to the 21 row and the .95 column in Appendix Table 8 and find the 95% CI for  $\mu = (13.00, 32.10)$ . Because  $\mu_0 = \text{expected number of deaths} = 18.1$  is within the 95% CI, we can accept  $H_0$  at the 5% level of significance. To get an exact  $p$ -value, we refer to Equation 7.48 and compute

$$p = 2 \times \left( 1 - \sum_{k=0}^{20} e^{-18.1} (18.1)^k / k! \right)$$

This is a tedious calculation, so we have used the POISSON function of Excel, as shown in Table 7.3. From Excel 2007, we see that  $\Pr(X \leq 20 | \mu = 18.1) = .7227$ . Therefore, the  $p$ -value =  $2 \times (1 - .7227) = .55$ . Thus, there is no significant excess or deficit of bladder cancer deaths in the rubber-worker population. The SMR for bladder cancer =  $100\% \times 21/18.1 = 116\%$ . Another interpretation of the significance tests in Equations 7.47 and 7.48 is that the underlying SMR in the reference population does not significantly differ from 100%.

**Table 7.3**
**Computation of the exact  $p$ -value for the bladder-cancer data in Example 7.56**


---

 Microsoft Excel 2007
 

---

mean	18.1
k	20
<code>Pr(X &lt;= k) = Poisson(20, 18.1, true)</code>	<b>0.722696</b>

---

The test procedures in Equations 7.47 and 7.48 are exact methods. If the expected number of events is large, then the following approximate method can be used.

**Equation 7.49**
**One-Sample Inference for the Poisson Distribution (Large-Sample Test)**

Let  $\mu$  = expected value of a Poisson random variable. To test the hypothesis  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ ,

- (1) Compute  $x$  = observed number of events in the study population.
- (2) Compute the test statistic

$$X^2 = \frac{(x - \mu_0)^2}{\mu_0} = \mu_0 \left( \frac{\text{SMR}}{100} - 1 \right)^2 \sim \chi_1^2 \text{ under } H_0$$

- (3) For a two-sided test at level  $\alpha$ ,  $H_0$  is rejected if

$$X^2 > \chi_{1,1-\alpha}^2$$

and  $H_0$  is accepted if  $X^2 \leq \chi_{1,1-\alpha}^2$

- (4) The exact  $p$ -value is given by  $\Pr(\chi_1^2 > X^2)$ .

(5) This test should only be used if  $\mu_0 \geq 10$ .

The acceptance and rejection regions for this test are depicted in Figure 7.16.  
The computation of the exact  $p$ -value is given in Figure 7.17.

(6) In addition, an approximate  $100\% \times (1 - \alpha)$  CI for  $\mu$  is given by  $x \pm z_{1-\alpha/2} \sqrt{x}$ .

### Example 7.57

**Occupational Health** Assess the statistical significance of the bladder-cancer data in Example 7.56 using the large-sample test.

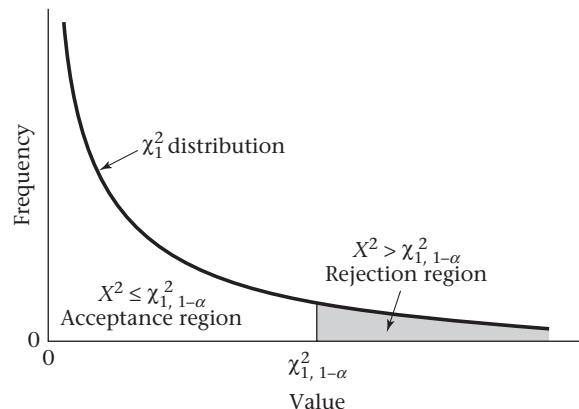
### Solution

We wish to test the hypothesis  $H_0: \mu = 18.1$  vs.  $H_1: \mu \neq 18.1$ . In this case,  $x = 21$  and  $SMR = 100\% \times 21/18.1 = 116\%$ . Hence we have the test statistic

$$\begin{aligned} X^2 &= \frac{(21-18.1)^2}{18.1} \quad \text{or } 18.1 \times (1.16-1)^2 \\ &= \frac{8.41}{18.1} = 0.46 \sim \chi^2_1 \text{ under } H_0 \end{aligned}$$

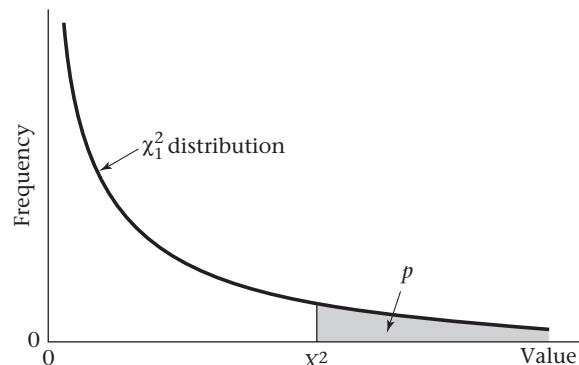
**Figure 7.16**

Acceptance and rejection regions for the one-sample Poisson test (large-sample test)



**Figure 7.17**

Computation of the  $p$ -value for the one-sample Poisson test (large-sample test)



Because  $\chi^2_{1,95} = 3.84 > X^2$ ,  $p > .05$  and  $H_0$  is accepted. Furthermore, from Table 6 in the Appendix we note that  $\chi^2_{1,50} = 0.45$ ,  $\chi^2_{1,75} = 1.32$ , and  $0.45 < X^2 < 1.32$ . Thus  $1 - .75 < p < 1 - .50$ , or  $.25 < p < .50$ . Therefore the rubber workers in this plant do not have a significantly increased risk of bladder cancer mortality relative to the general population. Using MINITAB to compute the  $p$ -value for the large-sample test yields a  $p$ -value =  $Pr(\chi^2_1 > 0.46) = .50$  to two decimal places. In addition, an approximate 95% CI for  $\mu$  is given by  $21 \pm 1.96\sqrt{21} = (12.0, 30.0)$ . From Example 7.56, the exact  $p$ -value based on the Poisson distribution = .55 and the exact 95% CI = (13.0, 32.1). In general, exact methods are strongly preferred for inference concerning the Poisson distribution.

In this section, we have presented the **one-sample Poisson test**, which is used for testing hypotheses concerning the parameter  $\mu$  of a Poisson distribution. Beginning at the “Start” box of the flowchart (Figure 7.18, p. 258), we arrive at the one-sample Poisson test by answering yes to (1) one variable of interest? and (2) one-sample problem? no to (3) underlying distribution normal or can central-limit theorem be assumed to hold? and (4) underlying distribution is binomial? and yes to (5) underlying distribution is Poisson?

## 7.12 Case Study: Effects of Tobacco Use on Bone-Mineral Density in Middle-Aged Women

In Chapter 6, we compared the bone-mineral density (BMD) of the lumbar spine between heavier- and lighter-smoking twins using CI methodology. We now want to consider a similar issue based on hypothesis-testing methods.

### Example 7.58

**Endocrinology** The mean difference in BMD at the lumbar spine between the heavier- and lighter-smoking twins when expressed as a percentage of the twin pair mean was  $-5.0\% \pm 2.0\%$  (mean  $\pm$   $se$ ) based on 41 twin pairs. Assess the statistical significance of the results.

### Solution

We will use the one-sample  $t$  test to test the hypothesis  $H_0: \mu = 0$  vs.  $H_1: \mu \neq 0$ , where  $\mu$  = underlying mean difference in BMD between the heavier- and lighter-smoking twins. Using Equation 7.10, we have the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Because  $\mu_0 = 0$  and  $s/\sqrt{n} = se$ , it follows that

$$t = \frac{\bar{x}}{se} = \frac{-5.0}{2.0} = -2.5 \sim t_{40} \text{ under } H_0$$

Using Table 5 in the Appendix, we see that  $t_{40,.99} = 2.423$ ,  $t_{40,.995} = 2.704$ . Because  $2.423 < 2.5 < 2.704$ , it follows that  $1 - .995 < p/2 < 1 - .99$  or  $.005 < p/2 < .01$  or  $.01 < p < .02$ . The exact  $p$ -value from Excel =  $2 \times Pr(t_{40} < -2.5) = TDIST(2.5, 40, 2) = .017$ . Hence, there is a significant difference in mean BMD between the heavier- and lighter-smoking twins, with the heavier-smoking twins having lower mean BMD.

## 7.13 Summary

In this chapter some of the fundamental ideas of hypothesis testing were introduced: (1) specification of the null ( $H_0$ ) and alternative ( $H_1$ ) hypotheses; (2) type I error ( $\alpha$ ), type II error ( $\beta$ ), and power ( $1 - \beta$ ) of a hypothesis test; (3) the  $p$ -value of a hypothesis test; and (4) the distinction between one-sided and two-sided tests. Methods for estimating the appropriate sample size for a proposed study as determined by the prespecified null and alternative hypotheses and type I and type II errors were also discussed.

These general concepts were applied to several one-sample hypothesis-testing situations:

- (1) The mean of a normal distribution with unknown variance (one-sample  $t$  test)
- (2) The mean of a normal distribution with known variance (one-sample  $z$  test)
- (3) The variance of a normal distribution (one-sample  $\chi^2$  test)
- (4) The parameter  $p$  of a binomial distribution (one-sample binomial test)
- (5) The expected value  $\mu$  of a Poisson distribution (one-sample Poisson test)

Each of the hypothesis tests can be conducted in one of two ways:

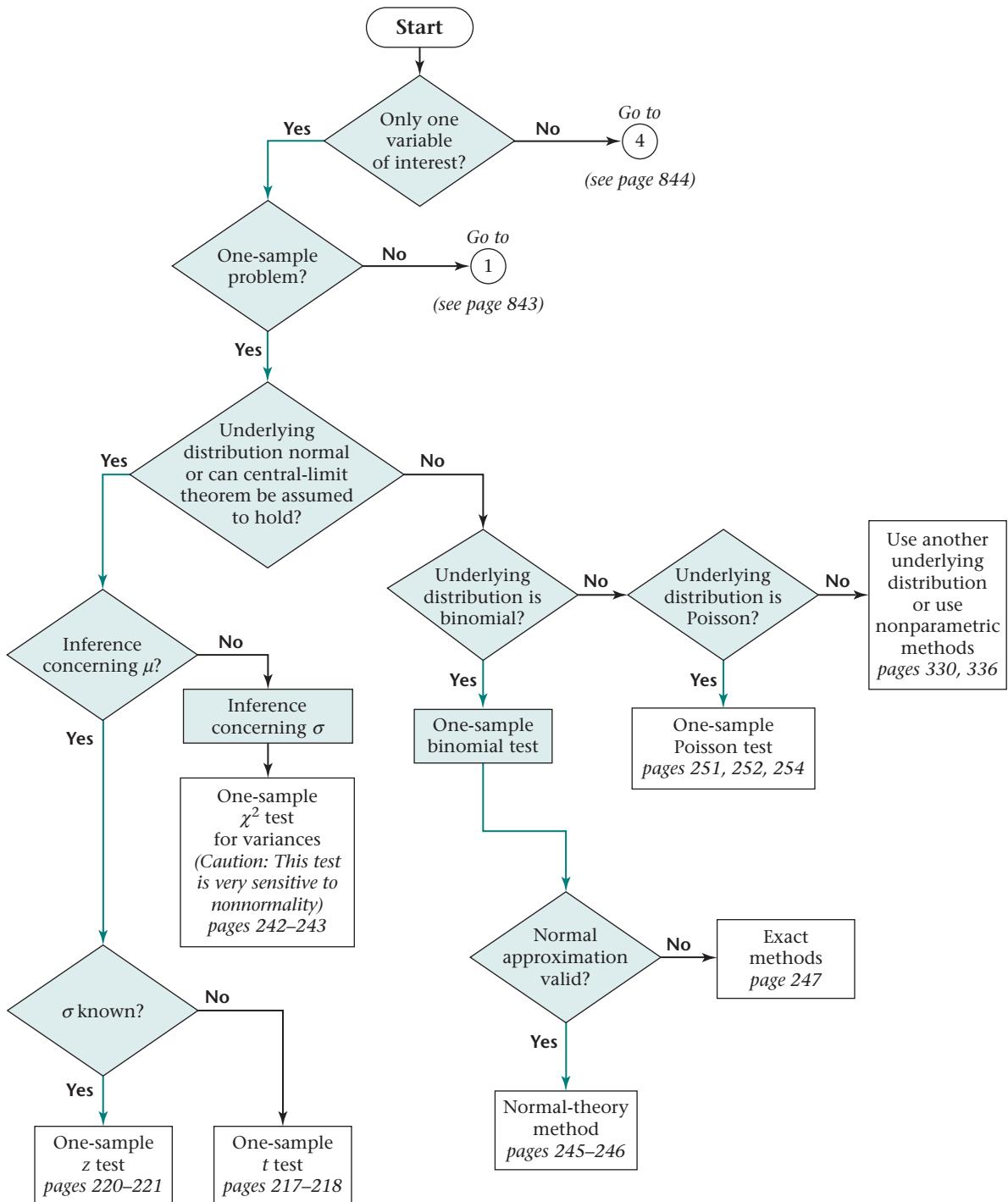
- (1) Specify critical values to determine the acceptance and rejection regions (critical-value method) based on a specified type I error  $\alpha$ .
- (2) Compute  $p$ -values ( $p$ -value method).

These methods were shown to be equivalent in the sense that they yield the same inferences regarding acceptance and rejection of the null hypothesis.

Furthermore, the relationship between the hypothesis-testing methods in this chapter and the CI methods in Chapter 6 was explored. We showed that the inferences that can be drawn from using these methods are usually the same. Finally, methods of Bayesian inference were introduced, both (a) when no prior information exists concerning a parameter and (b) when a substantial amount of prior information is available.

Many hypothesis tests are covered in this book. A master flowchart (pp. 841–846) is provided at the back of the book to help clarify the decision process in selecting the appropriate test. The flowchart can be used to choose the proper test by answering a series of yes/no questions. The specific hypothesis tests covered in this chapter have been presented in an excerpt from the flowchart shown in Figure 7.18 and have been referred to in several places in this chapter. For example, if we are interested in performing hypothesis tests concerning the mean of a normal distribution with known variance, then, beginning at the “Start” box of the flowchart, we would answer *yes* to each of the following questions: (1) only one variable of interest? (2) one-sample problem? (3) underlying distribution normal or can central-limit theorem be assumed to hold? (4) inference concerning  $\mu$ ? (5)  $\sigma$  known? The flowchart leads us to the box on the lower left of the figure, indicating that the one-sample  $z$  test should be used. In addition, the page number(s) where a specific hypothesis test is discussed is also provided in the appropriate box of the flowchart. The boxes marked “Go to 1” and “Go to 4” refer to other parts of the master flowchart in the back of the book.

The study of hypothesis testing is extended in Chapter 8 to situations in which two different samples are compared. This topic corresponds to the answer *yes* to the question (1) only one variable of interest? and no to (2) one-sample problem?

**Figure 7.18** Flowchart for appropriate methods of statistical inference

**PROBLEMS****Renal Disease**

The mean serum-creatinine level measured in 12 patients 24 hours after they received a newly proposed antibiotic was 1.2 mg/dL.

**\*7.1** If the mean and standard deviation of serum creatinine in the general population are 1.0 and 0.4 mg/dL, respectively, then, using a significance level of .05, test whether the mean serum-creatinine level in this group is different from that of the general population.

**\*7.2** What is the *p*-value for the test?

**7.3** Suppose  $\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = -1.52$  and a one-sample *t* test is

performed based on seven subjects. What is the two-tailed *p*-value?

**\*7.4** Suppose the sample standard deviation of serum creatinine in Problem 7.1 is 0.6 mg/dL. Assume that the standard deviation of serum creatinine is not known, and perform the hypothesis test in Problem 7.1. Report a *p*-value.

**\*7.5** Compute a two-sided 95% CI for the true mean serum-creatinine level in Problem 7.4.

**\*7.6** How does your answer to Problem 7.5 relate to your answer to Problem 7.4?

**Diabetes**

Plasma-glucose levels are used to determine the presence of diabetes. Suppose the mean ln (plasma-glucose) concentration (mg/dL) in 35- to 44-year-olds is 4.86 with standard deviation = 0.54. A study of 100 sedentary people in this age group is planned to test whether they have a higher or lower level of plasma glucose than the general population.

**7.7** If the expected difference is 0.10 ln units, then what is the power of such a study if a two-sided test is to be used with  $\alpha = .05$ ?

**7.8** Answer Problem 7.7 if the expected difference is 0.20 ln units.

**7.9** How many people would need to be studied to have 80% power under the assumptions in Problem 7.7?

**Cardiovascular Disease**

Suppose the incidence rate of myocardial infarction (MI) was 5 per 1000 among 45- to 54-year-old men in 1990. To look at changes in incidence over time, 5000 men in this age group were followed for 1 year starting in 2000. Fifteen new cases of MI were found.

**7.10** Using the critical-value method with  $\alpha = .05$ , test the hypothesis that incidence rates of MI changed from 1990 to 2000.

**7.11** Report a *p*-value to correspond to your answer to Problem 7.10.

Suppose that 25% of patients with MI in 1990 died within 24 hours. This proportion is called the 24-hour case-fatality rate.

**7.12** Of the 15 new MI cases in the preceding study, 5 died within 24 hours. Test whether the 24-hour case-fatality rate changed from 1990 to 2000.

**7.13** Suppose we eventually plan to accumulate 50 MI cases during the period 2000–2005. Assume that the 24-hour case-fatality rate is truly 20% during this period. How much power would such a study have in distinguishing between case-fatality rates in 1990 and 2000–2005 if a two-sided test with significance level .05 is planned?

**7.14** How large a sample is needed in Problem 7.13 to achieve 90% power?

**Pulmonary Disease**

Suppose the annual incidence of asthma in the general population among children 0–4 years of age is 1.4% for boys and 1% for girls.

**\*7.15** If 10 cases are observed over 1 year among 500 boys 0–4 years of age with smoking mothers, then test whether there is a significant difference in asthma incidence between this group and the general population using the critical-value method with a two-sided test.

**\*7.16** Report a *p*-value corresponding to your answer to Problem 7.15.

**\*7.17** Suppose that four cases are observed over 1 year among 300 girls 0–4 years of age with smoking mothers. Answer Problem 7.15 based on these data.

**\*7.18** Report a *p*-value corresponding to your answer to Problem 7.17.

**Genetics**

Ribosomal 5S RNA can be represented as a sequence of 120 nucleotides. Each nucleotide can be represented by one of four characters: A (adenine), G (guanine), C (cytosine), or U (uracil). The characters occur with different probabilities for each position. We wish to test whether a new sequence is the same as ribosomal 5S RNA. For this purpose, we replicate the new sequence 100 times and find there are 60 A's in the 20th position.

**7.19** If the probability of an A in the 20th position in ribosomal 5S RNA is .79, then test the hypothesis that the new sequence is the same as ribosomal 5S RNA using the critical-value method.

**7.20** Report a *p*-value corresponding to your results in Problem 7.19.

## Cancer

**7.21** Suppose we identify 50 women 50 to 54 years of age who have both a mother *and* a sister with a history of breast cancer. Five of these women themselves have developed breast cancer at some time in their lives. If we assume that the expected prevalence rate of breast cancer in women whose mothers have had breast cancer is 4%, does having a sister with the disease add to the risk? Explain.

## Obstetrics

The drug erythromycin has been proposed to possibly lower the risk of premature delivery. A related area of interest is its association with the incidence of side effects during pregnancy. Assume 30% of all pregnant women complain of nausea between weeks 24 and 28 of pregnancy. Furthermore, suppose that of 200 women who are taking erythromycin regularly during the period, 110 complain of nausea.

**\*7.22** Test the hypothesis that the incidence rate of nausea for the erythromycin group is the same for a typical pregnant woman.

## Cardiovascular Disease, Nutrition

Much discussion has appeared in the medical literature in recent years on the role of diet in the development of heart disease. The serum-cholesterol levels of a group of people who eat a primarily macrobiotic diet are measured. Among 24 of them, ages 20–39, the mean cholesterol level was found to be 175 mg/dL with a standard deviation of 35 mg/dL.

**7.23** If the mean cholesterol level in the general population in this age group is 230 mg/dL and the distribution is assumed to be normal, then test the hypothesis that the group of people on a macrobiotic diet have cholesterol levels different from those of the general population.

**7.24** Compute a 95% CI for the true mean cholesterol level in this group.

**7.25** What type of complementary information is provided by the hypothesis test and CI in this case?

## Hypertension

A pilot study of a new antihypertensive agent is performed for the purpose of planning a larger study. Five patients who have a mean DBP of at least 95 mm Hg are recruited for the study and are kept on the agent for 1 month. After 1 month the observed mean decline in DBP in these five patients is 4.8 mm Hg with a standard deviation of 9 mm Hg.

**\*7.26** If  $\mu_d$  = true mean difference in DBP between baseline and 1 month, then how many patients would be needed to have a 90% chance of detecting a significant change in DBP over 1 month using a one-tailed test with a significance level of 5%? Assume that the true mean and standard deviation of the DBP difference were the same as observed in the pilot study.

**\*7.27** Suppose we conduct a study of the preceding hypothesis based on 20 participants. What is the probability we will be able to reject  $H_0$  using a one-sided test at the 5% level if the true mean and standard deviation of the DBP difference are the same as in the pilot study?

## Occupational Health

The proportion of deaths due to lung cancer in males ages 15–64 in England and Wales during the period 1970–1972 was 12%. Suppose that of 20 deaths that occur among male workers in this age group who have worked for at least 1 year in a chemical plant, 5 are due to lung cancer. We wish to determine whether there is a difference between the proportion of deaths from lung cancer in this plant and the proportion in the general population.

**7.28** State the hypotheses to use in answering this question.

**7.29** Is a one-sided or two-sided test appropriate here?

**7.30** Perform the hypothesis test, and report a *p*-value.

After reviewing the results from one plant, the company decides to expand its study to include results from three additional plants. It finds that of 90 deaths occurring among 15- to 64-year-old male workers who have worked for at least 1 year in these four plants, 19 are due to lung cancer.

**7.31** Answer Problem 7.30 using the data from four plants, and report a *p*-value.

One criticism of studies of this type is that they are biased by the “healthy worker” effect. That is, workers in general are healthier than the general population, particularly regarding cardiovascular endpoints, which makes the proportion of deaths from noncardiovascular causes seem abnormally high.

**7.32** If the proportion of deaths from ischemic heart disease (IHD) is 40% for all 15- to 64-year-old men in England and Wales, whereas 18 of the preceding 90 deaths are attributed to IHD, then answer Problem 7.31 if deaths caused by IHD are excluded from the total.

## Nutrition

Iron-deficiency anemia is an important nutritional health problem in the United States. A dietary assessment was performed on 51 boys 9 to 11 years of age whose families were below the poverty level. The mean daily iron intake among these boys was found to be 12.50 mg with standard deviation 4.75 mg. Suppose the mean daily iron intake among a large population of 9- to 11-year-old boys from all income strata is 14.44 mg. We want to test whether the mean iron intake among the low-income group is different from that of the general population.

**\*7.33** State the hypotheses that we can use to consider this question.

**\*7.34** Carry out the hypothesis test in Problem 7.33 using the critical-value method with an  $\alpha$  level of .05, and summarize your findings.

**\*7.35** What is the  $p$ -value for the test conducted in Problem 7.34?

The standard deviation of daily iron intake in the larger population of 9- to 11-year-old boys was 5.56 mg. We want to test whether the standard deviation from the low-income group is comparable to that of the general population.

**\*7.36** State the hypotheses that we can use to answer this question.

**\*7.37** Carry out the test in Problem 7.36 using the critical-value method with an  $\alpha$  level of .05, and summarize your findings.

**\*7.38** What is the  $p$ -value for the test conducted in Problem 7.37?

**\*7.39** Compute a 95% CI for the underlying variance of daily iron intake in the low-income group. What can you infer from this CI?

**7.40** Compare the inferences you made from the procedures in Problems 7.37, 7.38, and 7.39.

## Demography

A study is performed using census data to look at various health parameters for a group of 10,000 Americans of Chinese descent living in the Chinatown areas of New York and San Francisco in 1970. The comparison group for this study is the total U.S. population in 1970. Suppose it is found that 100 of these Chinese-Americans have died over a 1-year period and that this represents a 15% decline from the expected mortality rate based on 1970 U.S. age- and sex-specific mortality rates.

**7.41** Test whether the total mortality experience in this group differs significantly from that in the total U.S. population. Report an exact  $p$ -value.

Suppose eight deaths from tuberculosis occur in this group, which is twice the rate expected based on the total U.S. population in 1970.

**7.42** Test whether the mortality experience from tuberculosis in this group differs significantly from that in the total U.S. population. Report an exact  $p$ -value.

## Occupational Health

The mortality experience of 8146 male employees of a research, engineering, and metal-fabrication plant in Tonawanda, New York, was studied from 1946 to 1981 [5]. Potential workplace exposures included welding fumes, cutting oils, asbestos, organic solvents, and environmental ionizing radiation, as a result of waste disposal during the Manhattan Project of World War II. Comparisons were made for specific causes of death between mortality rates in workers and U.S. white-male mortality rates from 1950 to 1978.

Suppose that 17 deaths from cirrhosis of the liver were observed among workers who were hired prior to 1946 and

who had worked in the plant for 10 or more years, whereas 6.3 were expected based on U.S. white-male mortality rates.

**7.43** What is the SMR for this group?

**7.44** Perform a significance test to assess whether there is an association between long duration of employment and mortality from cirrhosis of the liver in the group hired prior to 1946. Report a  $p$ -value.

**7.45** A similar analysis was performed among workers who were hired after 1945 and who were employed for 10 or more years. It found 4 deaths from cirrhosis of the liver, whereas only 3.4 were expected. What is the SMR for this group?

**7.46** Perform a significance test to assess whether there is an association between mortality from cirrhosis of the liver and duration of employment in the group hired after 1945. Report a  $p$ -value.

## Ophthalmology

Researchers have reported that the incidence rate of cataracts may be elevated among people with excessive exposure to sunlight. To confirm this, a pilot study is conducted among 200 people ages 65–69 who report an excessive tendency to burn on exposure to sunlight. Of the 200 people, 4 develop cataracts over a 1-year period. Suppose the expected incidence rate of cataracts among 65- to 69-year-olds is 1% over a 1-year period.

**\*7.47** What test procedure can be used to compare the 1-year rate of cataracts in this population with that in the general population?

**\*7.48** Implement the test procedure in Problem 7.47, and report a  $p$ -value (two-sided).

The researchers decide to extend the study to a 5-year period and find that 20 of the 200 people develop a cataract over a 5-year period. Suppose the expected incidence of cataracts among 65- to 69-year-olds in the general population is 5% over a 5-year period.

**\*7.49** Test the hypothesis that the 5-year incidence rate of cataracts is different in the excessive-sunlight-exposure group compared with the general population, and report a  $p$ -value (two-sided).

**\*7.50** Construct a 95% CI for the 5-year true rate of cataracts among the excessive-sunlight-exposure group.

## Cancer

A study was performed in Basel, Switzerland, on the relationship between the concentration of plasma antioxidant vitamins and cancer risk [6]. Table 7.4 shows data for plasma vitamin-A concentration for stomach-cancer cases and controls.

**7.51** If we assume that the mean plasma vitamin-A concentration among controls is known without error, then

**Table 7.4 Plasma vitamin-A concentration ( $\mu\text{mol/L}$ ) for stomach-cancer cases and controls**

	Mean	se	<i>n</i>
Stomach-cancer cases	2.65	0.11	20
Controls	2.88		2421

what procedure can be used to test whether the mean concentration is the same for stomach-cancer cases and controls?

**7.52** Perform the test in Problem 7.51, and report a *p*-value (two-sided).

**7.53** How many stomach-cancer cases are needed to achieve 80% power if the mean plasma vitamin-A concentration among controls is known without error, the true difference in mean concentration is  $0.20 \mu\text{mol/L}$ , and a two-sided test is used with  $\alpha = .05$ ?

### Nutrition, Cardiovascular Disease

Previous studies have shown that supplementing the diet with oat bran may lower serum-cholesterol levels. However, it is not known whether the cholesterol is reduced by a direct effect of oat bran or by replacing fatty foods in the diet. To address this question, a study was performed to compare the effect of dietary supplementation with high-fiber oat bran (87 g/day) to dietary supplementation with a low-fiber refined wheat product on the serum cholesterol of 20 healthy participants ages 23–49 years [7]. Each subject had a cholesterol level measured at baseline and then was randomly assigned to receive either a high-fiber or a low-fiber diet for 6 weeks. A 2-week period followed during which no supplements were taken. Participants then took the alternate supplement for a 6-week period. The results are shown in Table 7.5.

**7.54** Test the hypothesis that the high-fiber diet has an effect on cholesterol levels as compared with baseline (report your results as  $p < .05$  or  $p > .05$ ).

**7.55** Test the hypothesis that the low-fiber diet has an effect on cholesterol levels as compared with baseline (report your results as  $p < .05$  or  $p > .05$ ).

**7.56** Test the hypothesis that the high-fiber diet has a differential effect on cholesterol levels compared with a low-fiber diet (report your results as  $p < .05$  or  $p > .05$ ).

**7.57** What is the approximate standard error of the mean for the high-fiber compared with the low-fiber diet (that is, the mean difference in cholesterol level between high- and low-fiber diets)?

**7.58** How many participants would be needed to have a 90% chance of finding a significant difference in mean cholesterol lowering between the high- and low-fiber diets if the high-fiber diet lowers mean cholesterol by 5 mg/dL more than the low-fiber diet and a two-sided test is used with significance level = .05?

### Nutrition

Refer to Data Set VALID.DAT, on the Companion Website.

**7.59** Assess whether reported nutrient consumption (saturated fat, total fat, alcohol consumption, total caloric intake) is comparable for the diet record and the food-frequency questionnaire. Use either hypothesis-testing and/or CI methodology.

**7.60** Answer Problem 7.59 for the percentage of calories from fat (separately for total fat and saturated fat) as reported on the diet record and the food-frequency questionnaire. Assume there are 9 calories from fat for every gram of fat consumed.

### Demography

Refer to Data Set SEXRAT.DAT, on the Companion Website.

**7.61** Apply hypothesis-testing methods to answer the questions posed in Problem 4.57.

### Cardiology

Refer to Data Set NIFED.DAT, on the Companion Website.

**7.62** Use hypothesis-testing methods to assess whether either treatment affects blood pressure or heart rate in patients with severe angina.

### Cancer

The combination of photochemotherapy with oral methoxsalen (psoralen) and ultraviolet A radiation (called PUVA treatment) is an effective treatment for psoriasis. However, PUVA is mutagenic, increases the risk of squamous-cell skin cancer, and can cause irregular, pigmented skin lesions.

**Table 7.5 Serum-cholesterol levels before and during high-fiber and low-fiber supplementation**

	<i>n</i>	Baseline	High fiber	Low fiber	Difference (high fiber – low fiber)	Difference (high fiber – baseline)	Difference (low fiber – baseline)
Total cholesterol (mg/dL)	20	$186 \pm 31$	$172 \pm 28$	$172 \pm 25$	-1 (-8, +7)	-14 (-21, -7)	-13 (-20, -6)

Note: Plus-minus ( $\pm$ ) values are mean  $\pm$  *sd*. Values in parentheses are 95% confidence limits.

Stern et al. [8] performed a study to assess the incidence of melanoma among patients treated with PUVA. The study identified 1380 patients with psoriasis who were first treated with PUVA in 1975 or 1976. Patients were subdivided according to the total number of treatments received ( $<250$  or  $\geq 250$  from 1975 to 1996). Within each group, the observed number of melanomas was determined from 1975 to 1996 and compared with the expected number of melanomas as determined by published U.S. age- and sex-specific melanoma incidence rates. The results were as in Table 7.6.

**Table 7.6 Relationship of PUVA treatment to incidence of melanoma**

	Observed	Expected
<250 treatments	5	3.7
$\geq 250$ treatments	6	1.1

**7.63** Suppose we want to compare the observed and expected number of events among the group with  $<250$  treatments. Perform an appropriate significance test, and report a two-tailed  $p$ -value.

**7.64** Provide a 95% CI for the expected number of events in the group with  $\geq 250$  treatments.

**7.65** Interpret the results for Problems 7.63 and 7.64.

## Cancer

Breast cancer is strongly influenced by a woman's reproductive history. In particular, the longer the length of time from the age at menarche (the age when menstruation begins) to the age at first childbirth, the greater the risk is for breast cancer.

A projection was made based on a mathematical model that the 30-year risk of a woman in the general U.S. population developing breast cancer from age 40 to age 70 is 7%. Suppose a special subgroup of five hundred 40-year-old women without breast cancer was studied whose age at menarche was 17 (compared with an average age at menarche of 13 in the general population) and age at first birth was 20 (compared with an average age at first birth of 25 in the general population). These women were followed for development of breast cancer between ages 40 and 70. The study found that 18 of the women develop breast cancer between age 40 and age 70.

**7.66** Test the hypothesis that the underlying rate of breast cancer is the same or different in this group as in the general population.

**7.67** Provide a 95% CI for the true incidence rate of breast cancer over the period from age 40 to 70 in this special subgroup.

**7.68** Suppose 100 million women in the U.S. population have *not* developed breast cancer by the age of 40. What is your best estimate of the number of breast-cancer cases

that would be prevented from age 40 to 70 if all women in the U.S. population reached menarche at age 17 and gave birth to their first child at age 20? Provide a 95% CI for the number of breast-cancer cases prevented.

## Ophthalmology

An investigator wants to test a new eye drop that is supposed to prevent ocular itching during allergy season. To study the drug she uses a *contralateral design* whereby for each participant one eye is randomized (using a random-number table) to get active drug (A) while the other eye gets placebo (P). The participants use the eye drops three times a day for a 1-week period and then report their degree of itching in each eye on a 4-point scale (1 = none, 2 = mild, 3 = moderate, 4 = severe) without knowing which eye drop is used in which eye. Ten participants are randomized into the study.

**7.69** What is the principal advantage of the contralateral design?

Suppose the randomization assignment is as given in Table 7.7.

**Table 7.7 Randomization assignment**

Subject	Eye <sup>a</sup>		Subject	Eye	
	L	R		L	R
1	A	P	6	A	P
2	P	A	7	A	P
3	A	P	8	P	A
4	A	P	9	A	P
5	P	A	10	A	P

<sup>a</sup>A = active drug, P = placebo.

**7.70** More left eyes seem to be assigned to A than to P, and the investigator wonders whether the assignments are really random. Perform a significance test to assess how well the randomization is working. (*Hint:* Use the binomial tables.)

Table 7.8 gives the itching scores reported by the participants.

**7.71** What test can be used to test the hypothesis that the mean degree of itching is the same for active vs. placebo eyes?

**7.72** Implement the test in Problem 7.71 using a two-sided test (report a  $p$ -value).

## Endocrinology

Refer to the Data Set BONEDEN.DAT on the Companion Website.

**7.73** Perform a hypothesis test to assess whether there are significant differences in mean BMD for the femoral neck between the heavier- and lighter-smoking twins.

**Table 7.8** Itching scores reported by participants

Subject	Eye		Difference <sup>a</sup>
	L	R	
1	1	2	-1
2	3	3	0
3	4	3	1
4	2	4	-2
5	4	1	3
6	2	3	-1
7	2	4	-2
8	3	2	1
9	4	4	0
10	1	2	-1
Mean	2.60	2.80	-0.20
sd	1.17	1.03	1.55
N	10	10	10

<sup>a</sup>Itching score left eye – itching score right eye

**7.74** Answer Problem 7.73 for mean BMD at the femoral shaft.

## SIMULATION

Consider the birthweight data in Example 7.2.

**7.75** Suppose that the true mean birthweight for the low-SES babies is 120 oz, the true standard deviation is 24 oz, and the distribution of birthweights is normal. Generate 100 random samples of size 100 each from this distribution. Perform the appropriate *t* test for each sample to test the hypothesis stated in Example 7.2, and compute the proportion of samples for which we declare a significant difference using a 5% level of significance with a one-tailed test.

**7.76** What should this proportion be for a large number of simulated samples? How do the results in Problem 7.75 compare with this?

**7.77** Now assume that the true mean birthweight is 115 oz and repeat the exercise in Problem 7.75, assuming that the other conditions stated in Problem 7.75 are still correct.

**7.78** For what proportion of samples do you declare a significant difference? What should this proportion be for a large number of simulated samples?

## Cancer

A screening program for neuroblastoma (a type of cancer) was undertaken in Germany among children born between November 1, 1993 and June 30, 2000 who were 9 to 18 months of age between May 1995 and April 2000 [9].

A total of 1,475,773 children participated in the screening program, of whom 204 were diagnosed between 12 and 60

months of age. The researchers expected the incidence rate of neuroblastoma to be 7.3 per 100,000 children during this period in the absence of screening.

**7.79** Test whether the number of cases detected by the screening program is significantly greater than expected. Provide a one-tailed *p*-value. (*Hint:* Use the normal approximation to the binomial distribution.)

**7.80** Provide a 95% CI for the incidence rate of neuroblastoma in the screened population. Express the 95% CI as  $(p_1, p_2)$ , where  $p_1$  and  $p_2$  are in the units of number of cases per 100,000 children. Is  $p_0$  (7.3 cases per 100,000 children) in this interval?

Another issue investigated in this study was the case-fatality rate (number of patients who died from neuroblastoma / number of cases identified by the screening program).

**7.81** Suppose the case-fatality rate from neuroblastoma is usually 1/6. Furthermore, 17 fatal cases occurred among the 204 cases identified in the screening program. Test whether the case-fatality rate under the screening program is different from the usual case-fatality rate. Provide a two-tailed *p*-value.

## Environmental Health, Pulmonary Disease

A clinical epidemiologic study was conducted to determine the long-term health effects of workplace exposure to the process of manufacturing the herbicide (2,4,5 trichlorophenoxy) acetic acid (2,4,5-T), which contains the contaminant dioxin [10]. This study was conducted among active and retired workers of a Nitro, West Virginia, plant who were exposed to the 2,4,5-T process between 1948 and 1969. It is well known that workers exposed to 2,4,5-T have high rates of chloracne (a generalized acneiform eruption). Less well known are other potential effects of 2,4,5-T exposure. One of the variables studied was pulmonary function.

Suppose the researchers expect from general population estimates that 5% of workers have an abnormal forced expiratory volume (FEV); defined as less than 80% of predicted, based on their age and height. They found that 32 of 203 men who were exposed to 2,4,5-T while working at the plant had an abnormal FEV.

**7.82** What hypothesis test can be used to test the hypothesis that the percentage of abnormal FEV values among exposed men differs from the general-population estimates?

**7.83** Implement the test in Problem 7.82, and report a *p*-value (two-tailed).

Another type of outcome reported was fetal deaths. Suppose the investigators expect, given general population pregnancy statistics at the time of the survey, that 1.5% of pregnancies will result in a fetal death. They found that among 586 pregnancies where an exposed worker was the father, 11 resulted in a fetal death.

**7.84** Provide a 95% CI for the underlying fetal death rate among offspring of exposed men. Given the CI, how do you interpret the results of the study?

## Hypertension

A medical practice wants to compare the prevalence of side effects in its patients who take a specific antihypertensive drug with published side-effect rates from the literature. The doctors feel that side-effect rates greater than 20% will not be acceptable to patients. As a test of whether a new drug should be adopted by their practice, they conduct a pilot study among 10 patients in their practice who get the drug. If at least 4 have side effects, then the doctors are reluctant to adopt the drug in their practice. Otherwise, they feel the side-effect prevalence is reasonable and they are willing to use the drug in their practice.

**7.85** If the assessment of this pilot-study experience is represented in a one-sided hypothesis-testing framework where  $H_0: p = .2$  vs.  $H_1: p > .2$ , then what is the type I error of the test? (*Hint:* Use exact binomial tables [Appendix Table 1]).

**7.86** Suppose the actual true prevalence of side effects with the new drug is 50%. What is the power of the test procedure described? (*Hint:* Use exact binomial tables [Appendix Table 1]).

**7.87** The use of 10 participants in the pilot study is arbitrary. How many participants should be enrolled to achieve a power of 99% in Problem 7.86? (*Hint:* Use the normal approximation to the binomial distribution.)

## Ophthalmology

An experiment was performed to assess the efficacy of an eye drop in preventing “dry eye.” A principal objective measure used to assess efficacy is the tear breakup time (TBUT), which is reduced in people with dry eye and which the researchers hope will be increased after use of the eye drop.

In the actual study, participants will be randomized to either active drug or placebo, based on a fairly large sample size. However, a pilot study was first performed based on 14 participants. Under protocol A, the participants had their TBUT measured at baseline and were then instructed to not blink their eyes for 3 seconds, after which time the placebo eye drop was instilled. The TBUT was measured again immediately after instillation as well as at 5, 10, and 15 minutes postinstillation, during which time the study participants were in a controlled-chamber environment with low humidity to exacerbate symptoms of dry eye. On 2 other days participants received protocols B and C. Protocol B was identical to protocol A except that participants were told not to blink for 6 seconds prior to drop instillation. Protocol C was identical to protocol A except that participants were told not to blink their eyes for 10 seconds prior to drop instillation. Note that the same participants were used for each protocol and that for each protocol data are available for each of two eyes. Also, for each eye, two replicate measurements were provided.

The data are available in TEAR.DAT with documentation in TEAR.DOC, on the Companion Website. For each protocol,

TBUT (in seconds) was measured (1) at baseline (before drop instillation), (2) immediately after drop instillation, (3) 5 minutes after instillation, (4) 10 minutes after instillation, and (5) 15 minutes after instillation.

The standard protocol used in previous clinical studies of TBUT is a 6-second nonblink interval (protocol B). All the following questions concern protocol B data. For this purpose, average TBUT over two replicates and over both eyes to find a summary value for each time period for each participant.

**7.88** Is there an immediate effect of the eye drop on TBUT? (*Hint:* Compare mean TBUT immediately postinstillation vs. mean TBUT preinstillation.)

**7.89** Does the effect of the placebo eye drop change over time after drop instillation? (*Hint:* Compare mean TBUT at 5, 10, and 15 minutes postinstillation vs. mean TBUT immediately after drop instillation.)

## Hospital Epidemiology

Medical errors are common in hospitals throughout the world. One possible causal factor is the long work hours of hospital personnel. In a pilot study investigating this issue, medical residents were encouraged to sleep 6–8 hours per night for a 3-week period instead of their usual irregular sleep schedule. The researchers expected, given previous data, that there would be one medical error per resident per day on their usual irregular sleep schedule.

Suppose two residents participate in the program (each for 3 weeks), and chart review finds a total of 20 medical errors made by the two residents combined.

**7.90** What test can be used to test the hypothesis that an increase in amount of sleep will change the number of medical errors per day?

**7.91** Implement the test in Problem 7.90, and report a two-tailed *p*-value.

Suppose the true effect of the intervention is to reduce the number of medical errors per day by 20% (to 0.8 medical errors per day). Suppose that in the actual study 10 residents participate in the program, each for a 3-week period.

**7.92** What would be the power of the type of test used in Problem 7.91 under these assumptions? (*Hint:* Use the normal approximation to the Poisson distribution.)

## Ophthalmology

A study was performed among patients with glaucoma, an important eye disease usually manifested by high intraocular pressure (IOP); left untreated, glaucoma can lead to blindness.

The patients were currently on two medications (A and B) to be taken together for this condition. The purpose of this study was to determine whether the patients could drop medications A and B and be switched to a third medication (medication C) without much change in their IOP. Ten patients were

enrolled in the study. They received medications A + B for 60 days and had their IOP measured at the end of the 60-day period (referred to as  $IOP_{A+B}$ ). They were then taken off medications A and B and switched to medication C, which they took for an additional 60 days. IOP was measured a second time at the end of the 60-day period while the patient was on medication C (referred to as  $IOP_C$ ). The results were as shown in Table 7.9.

**Table 7.9 Effect of medication regimen on IOP among glaucoma patients**

Patient number	$IOP_{A+B}$ (mm Hg)	$IOP_C$ (mm Hg)
1	18.0	14.5
2	16.0	18.0
3	17.0	11.5
4	18.0	18.0
5	20.0	21.0
6	19.0	22.0
7	19.0	24.0
8	12.0	14.0
9	17.0	16.0
10	21.5	19.0

**7.93** What procedure can be used to test the hypothesis that there has been no mean difference in IOP after 60 days between the two drug regimens?

**7.94** Perform the procedure mentioned in Problem 7.93, and report a two-tailed  $p$ -value.

A goal of the study is to establish whether switching to medication C is “equivalent” to the original regimen of medication A and B. “Equivalence” here means that the underlying mean IOP after switching to medication C has not changed by more than 2 mm Hg in either direction.

**7.95** What procedure can be used to establish equivalence? Is equivalence the same as accepting the null hypothesis in Problem 7.93? Why or why not?

**7.96** Implement the procedure in Problem 7.95 to address the question of whether the regimens are equivalent.

## Hypertension

Refer to Example 7.44.

**7.97** Suppose a 38-year-old African-American man has two DBP readings at a single visit with mean = 100 mm Hg. What is the probability ( $p$ ) that his true mean DBP level is  $\geq 90$  mm Hg?

**7.98** What is the minimum DBP level, based on an average of two readings at a single visit, so that the patient would

be recommended for treatment based on the criteria in Example 7.44?

## Endocrinology

Osteoporosis is an important cause of morbidity in middle-aged and elderly women. Several drugs are currently used to prevent fractures in postmenopausal women.

Suppose the incidence rate of fractures over a 4-year period is known to be 5% among untreated postmenopausal women with no previous fractures.

A pilot study conducted among 100 women without previous fractures aims to determine whether a new drug can prevent fractures. It is found that two of the women have developed fractures over a 4-year period.

**7.99** Is there a significant difference between the fracture rate in individual treated women and the fracture rate in the untreated women? (Report a two-tailed  $p$ -value.)

Suppose that 8 of the previous 100 women have developed abdominal pain during the trial, while only 1.5% would be expected to develop abdominal pain based on previous natural history studies.

**7.100** Provide a 95% CI for the rate of abdominal pain among the active treated women. Interpret the results compared with previous natural history studies.

**7.101** Suppose the new drug yields a fracture rate of 2.5% over a 4-year period. How many subjects need to be studied to have an 80% chance of detecting a significant difference between the incidence rate of fractures in treated women and the incidence rate of fractures in untreated women (assumed known from Problem 7.99)?

## Cancer

Patients with breast cancer sometimes die for other reasons. In the Nurses’ Health Study, of 3500 women with breast cancer who were followed for 10 years, 80 died of heart disease.

**7.102** Provide a point estimate and 95% CI for the heart-disease death rate (assume that there were no deaths for any other reason).

**7.103** Suppose the baseline 10-year incidence of heart disease in the general population is 2%. How many women with breast cancer need to be studied to have 80% power to detect a difference from the general population if a two-sided test with  $\alpha = 0.05$  is used and the true 10-year incidence rate of heart disease is 2.5% among women with breast cancer? (Assume that there are no deaths for any other reason.)

## Hypertension, Pediatrics

The Task Force on Blood Pressure Control in Children published its findings in 2004 [11], among which were

percentiles of blood pressure for children according to age, sex, and height. The percentiles presented in the report are based on studies from 1970–2000 using mercury blood-pressure measuring devices. To assess how applicable the percentiles are to current blood pressure levels, which are measured mainly with oscillometric devices, data were collected in 200 children. It was found that 30 of the children are above the 90th percentile according to Task Force standards. The 90th percentile is usually used to denote *prehypertension*.

**7.104** Test whether the percentage of children above the 90th percentile differs significantly from the Task Force standard.

**7.105** Provide a 95% CI for the true percentage of children above the 90th percentile using oscillometric devices.

**7.106** Suppose a new study is considered to compare the percentage of children who are *hypertensive* (above the 95th percentile) based on oscillometric devices with that expected based on Task Force standards (i.e., 5%). How many children need to be studied to have a 90% chance of detecting a significant difference if a two-sided test is used with a type I error of 0.05 and the true percentage of hypertension using an oscillometric device is 10%.

**7.107** Suppose the investigators decide they would be satisfied if the power of the study is 80%. Do more or fewer subjects need to be enrolled with this goal compared with the sample size in Problem 7.106? *Do not calculate the sample size.*

## General

**7.108** What is the 25th percentile of a  $\chi^2$  distribution with 20 degrees of freedom? What symbol is used to denote this value?

**7.109** Suppose we wish to test the hypothesis  $H_0: \mu = 2$  vs.  $H_1: \mu \neq 2$ . We find a two-sided  $p$ -value of .03 and a 95% CI for  $\mu$  of (1.5, 4.0). Are these two results possibly compatible? Why or why not?

## Ophthalmology

**7.110** Suppose we draw a sample of size 3 from a distribution that is non-normal with mean = 16 mm Hg and standard deviation = 5 mm Hg. We assume that the sample size is large enough so that the central-limit theorem holds. This is approximately the distribution of intraocular pressure (IOP) in one individual. A criterion for ocular hypertension is that IOP is 20 mm Hg or higher based on an average of several readings. What is the probability that the sample mean exceeds 20 mm Hg based on an average of three readings?

**7.111** How many readings do we need to obtain so that the probability in Problem 7.110 is  $< .01$ ?

## Cancer

A study was conducted in Sweden to relate the age at surgery for undescended testis to the subsequent risk of testicular cancer [12]. Twelve events were reported in 22,884 person-years of follow-up among men who were 13–15 years at age of surgery.

**7.112** What is the estimated incidence rate of testicular cancer among this group of men? Express the rate per 100,000 person-years.

It was reported that the standardized incidence rate in this group compared with men in the general Swedish population was 5.06.

**7.113** What is the expected number of events in the general population over 22,884 person-years of follow-up?

**7.114** Provide a 95% CI for the number of events among men who were 13–15 years at age of surgery.

**7.115** Is there a significant difference ( $p < .05$ ) between the incidence of testicular cancer in men treated surgically for undescended testis at age 13–15 vs. the general population?

**7.116** What is the lifetime risk (from age 15 to age 70) of testicular cancer for men with age at surgery = 15, assuming that the incidence rate remains constant over time? Assume that all men do not die of any other disease up to age 70.

## Cancer

Data from the Surveillance Epidemiology and End Results (SEER) registry provides incidence data for different cancers according to age and sex for various tumor registries throughout the world. The data in Table 7.10 were obtained for colon cancer in women from the Connecticut Tumor Registry from 1988–1992.

**Table 7.10 Connecticut Tumor Registry data: Annual incidence rates for colon cancer from 1988–1992 for females**

Age	Annual incidence rate (per $10^5$ person-years)
40–44	8
45–49	16
50–54	27
55–59	50

**7.117** What is the probability that a 40-year-old woman will develop colon cancer over the next 5 years?

**7.118** What is the probability that a 40-year-old woman will develop colon cancer over the next 20 years (i.e., from age 40.0 to age 59.9)?

The data in Table 7.11 were obtained from the Nurses' Health Study on colon cancer incidence over the time period 1980–2004.

**7.119** Do the SEER rates provide a good fit with the Nurses' Health Study incidence data? Perform a separate test for each age group and simply report  $p > .05$  or  $p < .05$ . (Hint: Use the Poisson distribution.)

**Table 7.11 Nurses' Health Study colon cancer incidence data from 1980–2004**

Age	Cases	Person-years
40–44	10	139,922
45–49	35	215,399
50–54	79	277,027
55–59	104	321,250

## REFERENCES

- [1] Hypertension Detection and Follow-Up Program Cooperative Group. (1977). Blood pressure studies in 14 communities: A two-stage screen for hypertension. *JAMA*, 237, 2385–2391.
- [2] Rosner, B., & Polk, B. F. (1983). Predictive values of routine blood pressure measurements in screening for hypertension. *American Journal of Epidemiology*, 117(4), 429–442.
- [3] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall.
- [4] Andjelkovic, D., Taulbee, J., & Symons, M. (1976). Mortality experience of a cohort of rubber workers, 1964–1973. *Journal of Occupational Medicine*, 18(6), 387–394.
- [5] Teta, M. J., & Ott, M. G. (1988). A mortality study of a research, engineering and metal fabrication facility in western New York State. *American Journal of Epidemiology*, 127(3), 540–551.
- [6] Stähelin, H. B., Gey, K. F., Eichholzer, M., Ludin, E., Bernasconi, F., Thurneysen, J., & Brubacher, G. (1991). Plasma antioxidant vitamins and subsequent cancer mortality in the 12-year follow-up of the prospective Basel Study. *American Journal of Epidemiology*, 133(8), 766–775.
- [7] Swain, J. F., Rouse, I. L., Curley, C. B., & Sacks, F. M. (1990). Comparison of the effects of oat bran and low-fiber wheat on serum lipoprotein levels and blood pressure. *New England Journal of Medicine*, 322(3), 147–152.
- [8] Stern, R. S., Nichols, K. J., & Vakeva, L. H. (1997). Malignant melanoma in patients treated for psoriasis with methoxsalen (Psoralen) and ultraviolet A radiation (PUVA). The PUVA follow-up study. *New England Journal of Medicine*, 336, 1041–1045.
- [9] Woods, W. G., Gao, R.-N., Shuster, J. J., Robison, L. L., Bernstein, M., Weitzman, S., Bunin, G., Levy, I., Brosard, J., Dougherty, G., Tuchman, M., & Lemieux, B. (2002). Screening of infants and mortality due to neuroblastoma. *New England Journal of Medicine*, 346, 1041–1046.
- [10] Suskind, R. R., & Hertzberg, V. S. (1984). Human health effects of 2,4,5-T and its toxic contaminants. *JAMA*, 251, 2372–2380.
- [11] National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents. (2004). The fourth report on the diagnosis, evaluation and treatment of high blood pressure in children and adolescents. *Pediatrics*, 114, 555–576.
- [12] Pettersson, A., Richiardi, L., Nordenskjold, A., Kaijser, M., & Akre, O. (2007). Age at surgery for undescended testis and risk of testicular cancer. *New England Journal of Medicine*, 356(18), 1835–1841.

# Hypothesis Testing: Two-Sample Inference

## 8.1 Introduction

All the tests introduced in Chapter 7 were one-sample tests, in which the underlying parameters of the population from which the sample was drawn were compared with comparable values from other generally large populations *whose parameters were assumed to be known*.

**Example 8.1** **Obstetrics** In the birthweight data in Example 7.2, the underlying mean birthweight in one hospital was compared with the underlying mean birthweight in the United States, *the value of which was assumed known*.

A more frequently encountered situation is the two-sample hypothesis-testing problem.

**Definition 8.1** In a **two-sample hypothesis-testing problem**, the underlying parameters of two different populations, *neither of whose values is assumed known*, are compared.

**Example 8.2** **Hypertension** Let's say we are interested in the relationship between oral contraceptive (OC) use and blood pressure in women.

Two different experimental designs can be used to assess this relationship. One method involves the following design:

**Equation 8.1**

### Longitudinal Study

- (1) Identify a group of nonpregnant, premenopausal women of childbearing age (16–49 years) who are not currently OC users, and measure their blood pressure, which will be called the *baseline blood pressure*.
- (2) Rescreen these women 1 year later to ascertain a subgroup who have remained nonpregnant throughout the year and have become OC users. This subgroup is the study population.
- (3) Measure the blood pressure of the study population at the follow-up visit.
- (4) Compare the baseline and follow-up blood pressure of the women in the study population to determine the difference between blood pressure levels of women when they *were* using the pill at follow-up and when they *were not* using the pill at baseline.

Another method involves the following design:

#### Equation 8.2

##### Cross-Sectional Study

- (1) Identify both a group of OC users and a group of non-OC users among non-pregnant, premenopausal women of childbearing age (16–49 years), and measure their blood pressure.
- (2) Compare the blood pressure level between the OC users and nonusers.

#### Definition 8.2

In a **longitudinal** or **follow-up study** the same group of people is followed *over time*.

#### Definition 8.3

In a **cross-sectional study**, the participants are seen at only one point in time.

There is another important difference between these two designs. The longitudinal study represents a *paired-sample* design because each woman is used as her own control. The cross-sectional study represents an *independent-sample* design because two completely different groups of women are being compared.

#### Definition 8.4

Two samples are said to be **paired** when each data point in the first sample is matched and is related to a unique data point in the second sample.

#### Example 8.3

The paired samples may represent two sets of measurements on the same people. In this case each person is serving as his or her own control, as in Equation 8.1. The paired samples may also represent measurements on different people who are chosen on an individual basis using matching criteria, such as age and sex, to be very similar to each other.

#### Definition 8.5

Two samples are said to be **independent** when the data points in one sample are unrelated to the data points in the second sample.

#### Example 8.4

The samples in Equation 8.2 are completely independent because the data are obtained from unrelated groups of women.

Which type of study is better in this case? The first type of study is probably more definitive because most other factors that influence a woman's blood pressure at the first screening (called confounders) will also be present at the second screening and will not influence the comparison of blood-pressure levels at the first and second screenings. However, the study would benefit from having a control group of women who remained non-OC users throughout the year. The control group would allow us the chance of ruling out other possible causes of blood pressure change besides changes in OC status. The second type of study, by itself, can only be considered suggestive because other confounding factors may influence blood pressure in the two samples and cause an apparent difference to be found where none is actually present.

For example, OC users are known to weigh less than non-OC users. Low weight tends to be associated with low BP, so the blood-pressure levels of OC users as a group would appear lower than the levels of non-OC users.

However, a follow-up study is more expensive than a cross-sectional study. Therefore, a cross-sectional study may be the only financially feasible way of doing the study.

In this chapter, the appropriate methods of hypothesis testing for both the paired-sample and independent-sample situations are studied.

## 8.2 The Paired *t* Test

Suppose the paired-sample study design in Equation 8.1 is adopted and the sample data in Table 8.1 are obtained. The systolic blood-pressure (SBP) level of the  $i$ th woman is denoted at baseline by  $x_{i1}$  and at follow-up by  $x_{i2}$ .

**Table 8.1** SBP levels (mm Hg) in 10 women while not using (baseline) and while using (follow-up) OCs

$i$	SBP level while not using OCs ( $x_{i1}$ )	SBP level while using OCs ( $x_{i2}$ )	$d_i^*$
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2

\* $d_i = x_{i2} - x_{i1}$

### Equation 8.3

Assume that the SBP of the  $i$ th woman is normally distributed at baseline with mean  $\mu_i$  and variance  $\sigma^2$  and at follow-up with mean  $\mu_i + \Delta$  and variance  $\sigma^2$ .

We are thus assuming that the underlying mean difference in SBP between follow-up and baseline is  $\Delta$ . If  $\Delta = 0$ , then there is no difference between mean baseline and follow-up SBP. If  $\Delta > 0$ , then using OC pills is associated with a raised mean SBP. If  $\Delta < 0$ , then using OC pills is associated with a lowered mean SBP.

We want to test the hypothesis  $H_0: \Delta = 0$  vs.  $H_1: \Delta \neq 0$ . How should we do this? The problem is that  $\mu_i$  is unknown, and we are assuming, in general, that it is different for each woman. However, consider the difference  $d_i = x_{i2} - x_{i1}$ . From Equation 8.3 we know that  $d_i$  is normally distributed with mean  $\Delta$  and variance that we denote by  $\sigma_d^2$ . Thus, although blood pressure levels  $\mu_i$  are different for each woman, the differences in blood pressure between baseline and follow-up have the same underlying mean ( $\Delta$ ) and variance ( $\sigma_d^2$ ) over the entire population of women. The hypothesis-testing problem can thus be considered a *one-sample t test based on the differences* ( $d_i$ ). From our work on the one-sample *t* test in Section 7.4, we know that the best test of the hypothesis  $H_0: \Delta = 0$  vs.  $H_1: \Delta \neq 0$ , when the variance is unknown, is based on the mean difference

$$\bar{d} = (d_1 + d_2 + \dots + d_n)/n$$

Specifically, from Equation 7.10 for a two-sided level  $\alpha$  test, we have the following test procedure, called the paired  $t$  test.

#### Equation 8.4

##### Paired $t$ Test

Denote the test statistic  $\bar{d}/(s_d/\sqrt{n})$  by  $t$ , where  $s_d$  is the sample standard deviation of the observed differences:

$$s_d = \sqrt{\left[ \sum_{i=1}^n d_i^2 - \left( \sum_{i=1}^n d_i \right)^2 / n \right] / (n-1)}$$

$n$  = number of matched pairs

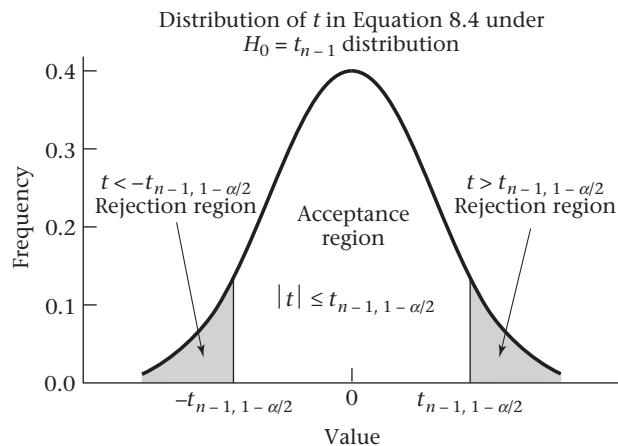
If  $t > t_{n-1, 1-\alpha/2}$  or  $t < -t_{n-1, 1-\alpha/2}$

then  $H_0$  is rejected.

If  $-t_{n-1, 1-\alpha/2} \leq t \leq t_{n-1, 1-\alpha/2}$

then  $H_0$  is accepted. The acceptance and rejection regions for this test are shown in Figure 8.1.

**Figure 8.1** Acceptance and rejection regions for the paired  $t$  test



Similarly, from Equation 7.11, a  $p$ -value for the test can be computed as follows.

#### Equation 8.5

##### Computation of the $p$ -Value for the Paired $t$ Test

If  $t < 0$ ,

$p = 2 \times [\text{the area to the left of } t = \bar{d}/(s_d/\sqrt{n}) \text{ under a } t_{n-1} \text{ distribution}]$

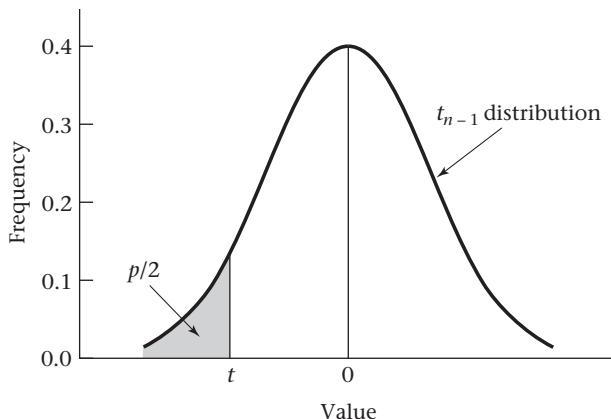
If  $t \geq 0$ ,

$p = 2 \times [\text{the area to the right of } t \text{ under a } t_{n-1} \text{ distribution}]$

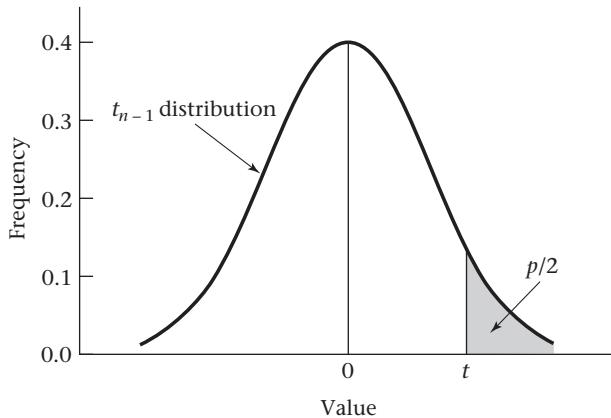
The computation of the  $p$ -value is illustrated in Figure 8.2.

#### Example 8.5

**Hypertension** Assess the statistical significance of the OC–blood pressure data in Table 8.1.

**Figure 8.2** Computation of the *p*-value for the paired *t* test

If  $t = \bar{d}/(s_d/\sqrt{n}) < 0$ , then  $p = 2 \times (\text{area to the left of } t \text{ under a } t_{n-1} \text{ distribution})$ .



If  $t = \bar{d}/(s_d/\sqrt{n}) \geq 0$ , then  $p = 2 \times (\text{area to the right of } t \text{ under a } t_{n-1} \text{ distribution})$ .

**Solution**

$$\bar{d} = (13 + 3 + \dots + 2)/10 = 4.80$$

$$s_d^2 = [(13 - 4.8)^2 + \dots + (2 - 4.8)^2]/9 = 20.844$$

$$s_d = \sqrt{20.844} = 4.566$$

$$t = 4.80/(4.566/\sqrt{10}) = 4.80/1.444 = 3.32$$

The critical-value method is first used to perform the significance test. There are  $10 - 1 = 9$  degrees of freedom (*df*), and from Table 5 in the Appendix we see that  $t_{0.975} = 2.262$ . Because  $t = 3.32 > 2.262$ , it follows from Equation 8.4 that  $H_0$  can be rejected using a two-sided significance test with  $\alpha = .05$ . To compute an approximate *p*-value, refer to Table 5 and note that  $t_{0.9995} = 4.781$ ,  $t_{0.995} = 3.250$ . Thus, because  $3.25 < 3.32 < 4.781$ , it follows that  $.0005 < p/2 < .005$  or  $.001 < p < .01$ . To compute a more exact *p*-value, a computer program must be used. The results in Table 8.2 were obtained using the Microsoft Excel 2007 T-TEST program.

To use the program, the user specifies the arrays being compared on the spreadsheet (B3:B12 and C3:C12), the number of tails for the *p*-value (2), and the type of *t* test (paired *t* test, type 1).

Note from Table 8.2 that the exact two-sided *p*-value = .009. Therefore, we can conclude that starting OC use is associated with a significant change in blood pressure.

**Table 8.2** Use of the Microsoft Excel T-TEST program to analyze the blood-pressure data in Table 8.1

SBP while not using OCs	SBP while using OCs	Paired t test p-value*
115	128	0.008874337
112	115	
107	106	
119	128	
115	122	
138	145	
126	132	
105	109	
104	102	
115	117	

\*TTEST(B3:B12,C3:C12,2,1)

Example 8.5 is a classic example of a paired study because each woman is used as her own control. In many other paired studies, different people are used for the two groups, but they are matched individually on the basis of specific matching characteristics.

### Example 8.6

**Gynecology** A topic of recent clinical interest is the effect of different contraceptive methods on fertility. Suppose we wish to compare how long it takes users of either OCs or diaphragms to become pregnant after stopping contraception. A study group of 20 OC users is formed, and diaphragm users who match each OC user with regard to age (within 5 years), race, parity (number of previous pregnancies), and socioeconomic status (SES) are found. The investigators compute the differences in time to fertility between previous OC and diaphragm users and find that the mean difference  $\bar{d}$  (OC minus diaphragm) in time to fertility is 4 months with a standard deviation ( $s_d$ ) of 8 months. What can we conclude from these data?

### Solution

Perform the paired *t* test. We have

$$t = \bar{d} / (s_d / \sqrt{n}) = 4 / (8 / \sqrt{20}) = 4 / 1.789 = 2.24 \sim t_{19}$$

under  $H_0$ . Referring to Table 5 in the Appendix, we find that

$$t_{19, .975} = 2.093 \quad \text{and} \quad t_{19, .99} = 2.539$$

Thus, because  $2.093 < 2.24 < 2.539$ , it follows that  $.01 < p/2 < .025$  or  $.02 < p < .05$ . Therefore, previous OC users take a significantly longer time to become pregnant than do previous diaphragm users.

In this section, we have introduced the paired *t* test, which is used to compare the mean level of a normally distributed random variable (or a random variable with sample size large enough so that the central-limit theorem can be assumed to hold) between two paired samples. If we refer to the flowchart (Figure 8.13, p. 308), starting from position 1, we answer yes to (1) two-sample problem? (2) underlying distribution normal or can central-limit theorem be assumed to hold? and (3) inferences concerning means? and no to (4) are samples independent? This leads us to the box labeled "Use paired *t* test."

**REVIEW QUESTIONS 8A**

- 1** How do a paired-sample design and an independent-sample design differ?
- 2** A man measures his heart rate before using a treadmill and then after walking on a treadmill for 10 minutes on 7 separate days. His mean heart rate at baseline and 10 minutes after treadmill walking is 85 and 93 beats per minute (bpm), respectively. The mean change from baseline to 10 minutes is 8 bpm with a standard deviation of 6 bpm.
  - (a)** What test can we use to compare pre- and post-treadmill heart rate?
  - (b)** Implement the test in Review Question 8A.2a, and report a two-tailed *p*-value.
  - (c)** Provide a 90% confidence interval (CI) for the mean change in heart rate after using the treadmill for 10 minutes.
  - (d)** What is your overall conclusion concerning the data?

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

In the previous section, methods of hypothesis testing for comparing means from two paired samples were discussed. It is also useful to construct confidence limits for the true mean difference ( $\Delta$ ). The observed difference scores ( $d_i$ ) are normally distributed with mean  $\Delta$  and variance  $\sigma_d^2$ . Thus the sample mean difference ( $\bar{d}$ ) is normally distributed with mean  $\Delta$  and variance  $\sigma_d^2/n$ , where  $\sigma_d^2$  is unknown. The methods of CI estimation in Equation 6.6 can be used to derive a  $100\% \times (1 - \alpha)$  CI for  $\Delta$ , which is given by

$$(\bar{d} - t_{n-1,1-\alpha/2} s_d / \sqrt{n}, \bar{d} + t_{n-1,1-\alpha/2} s_d / \sqrt{n})$$

**Equation 8.6****Confidence Interval for the True Difference ( $\Delta$ ) Between the Underlying Means of Two Paired Samples (Two-Sided)**

A two-sided  $100\% \times (1 - \alpha)$  CI for the true mean difference ( $\Delta$ ) between two paired samples is given by

$$(\bar{d} - t_{n-1,1-\alpha/2} s_d / \sqrt{n}, \bar{d} + t_{n-1,1-\alpha/2} s_d / \sqrt{n})$$

**Example 8.7**

**Hypertension** Using the data in Table 8.1, compute a 95% CI for the true increase in mean SBP after starting OCs.

**Solution**

From Example 8.5 we have  $\bar{d} = 4.80$  mm Hg,  $s_d = 4.566$  mm Hg,  $n = 10$ . Thus, from Equation 8.6, a 95% CI for the true mean SBP change is given by

$$\begin{aligned}\bar{d} \pm t_{n-1,975} s_d / \sqrt{n} &= 4.80 \pm t_{9,975}(1.444) \\ &= 4.80 \pm 2.262(1.444) = 4.80 \pm 3.27 = (1.53, 8.07) \text{ mm Hg}\end{aligned}$$

Thus the true change in mean SBP is most likely between 1.5 and 8.1 mm Hg.

**Example 8.8**

**Gynecology** Using the data in Example 8.6, compute a 95% CI for the true mean difference between OC users and diaphragm users in time to fertility.

**Solution**

From Example 8.6, we have  $\bar{d} = 4$  months,  $s_d = 8$  months,  $n = 20$ . Thus the 95% CI for  $\mu_d$  is given by

$$\begin{aligned}\bar{d} \pm \frac{t_{n-1,975}s_d}{\sqrt{n}} &= 4 \pm \frac{t_{19,975}(8)}{\sqrt{20}} \\ &= 4 \pm \frac{2.093(8)}{\sqrt{20}} = 4 \pm 3.74 = (0.26, 7.74) \text{ months}\end{aligned}$$

Thus the true lag in time to fertility can be anywhere from about 0.25 month to nearly 8 months. A much larger study is needed to narrow the width of this CI.

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

Let's now discuss the question posed in Example 8.2, assuming that the cross-sectional study defined in Equation 8.2 is being used, rather than the longitudinal study defined in Equation 8.1.

**Example 8.9**

**Hypertension** Suppose a sample of eight 35- to 39-year-old nonpregnant, premenopausal OC users who work in a company and have a mean systolic blood pressure (SBP) of 132.86 mm Hg and sample standard deviation of 15.34 mm Hg are identified. A sample of 21 nonpregnant, premenopausal, non-OC users in the same age group are similarly identified who have mean SBP of 127.44 mm Hg and sample standard deviation of 18.23 mm Hg. What can be said about the underlying mean difference in blood pressure between the two groups?

Assume SBP is normally distributed in the first group with mean  $\mu_1$  and variance  $\sigma_1^2$  and in the second group with mean  $\mu_2$  and variance  $\sigma_2^2$ . We want to test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ . Assume in this section that the underlying variances in the two groups are the same (that is,  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ). The means and variances in the two samples are denoted by  $\bar{x}_1, \bar{x}_2, s_1^2, s_2^2$ , respectively.

It seems reasonable to base the significance test on the difference between the two sample means,  $\bar{x}_1 - \bar{x}_2$ . If this difference is far from 0, then  $H_0$  will be rejected; otherwise, it will be accepted. Thus we wish to study the behavior of  $\bar{x}_1 - \bar{x}_2$  under  $H_0$ . We know  $\bar{X}_1$  is normally distributed with mean  $\mu_1$  and variance  $\sigma^2/n_1$  and  $\bar{X}_2$  is normally distributed with mean  $\mu_2$  and variance  $\sigma^2/n_2$ . Hence, from Equation 5.10, because the two samples are independent,  $\bar{X}_1 - \bar{X}_2$  is normally distributed with mean  $\mu_1 - \mu_2$  and variance  $\sigma^2(1/n_1 + 1/n_2)$ . In symbols,

**Equation 8.7**

$$\bar{X}_1 - \bar{X}_2 \sim N\left[\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$$

Under  $H_0$ , we know that  $\mu_1 = \mu_2$ . Thus Equation 8.7 reduces to

**Equation 8.8**

$$\bar{X}_1 - \bar{X}_2 \sim N\left[0, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$$

If  $\sigma^2$  were known, then  $\bar{X}_1 - \bar{X}_2$  could be divided by  $\sigma\sqrt{1/n_1 + 1/n_2}$ . From Equation 8.8,

**Equation 8.9**

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

and the test statistic in Equation 8.9 could be used as a basis for the hypothesis test. Unfortunately,  $\sigma^2$  in general is unknown and must be estimated from the data. How can  $\sigma^2$  be best estimated in this situation?

From the first and second samples, the sample variances are  $s_1^2$ ,  $s_2^2$  respectively, each of which could be used to estimate  $\sigma^2$ . The average of  $s_1^2$  and  $s_2^2$  could simply be used as the estimate of  $\sigma^2$ . However, this average will weight the sample variances equally even if the sample sizes are very different from each other. The sample variances should not be weighted equally because the variance from the larger sample is probably more precise and should be weighted more heavily. The best estimate of the population variance  $\sigma^2$ , which is denoted by  $s^2$ , is given by a weighted average of the two sample variances, where the weights are the number of  $df$  in each sample.

**Equation 8.10**

The pooled estimate of the variance from two independent samples is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

In particular,  $s^2$  will then have  $n_1 - 1$   $df$  from the first sample and  $n_2 - 1$   $df$  from the second sample, or

$$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 \text{ } df$$

overall. Then  $s$  can be substituted for  $\sigma$  in Equation 8.9, and the resulting test statistic can then be shown to follow a  $t$  distribution with  $n_1 + n_2 - 2$   $df$  rather than an  $N(0,1)$  distribution because  $\sigma^2$  is unknown. Thus the following test procedure is used.

**Equation 8.11****Two-Sample  $t$  Test for Independent Samples with Equal Variances**

Suppose we wish to test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$  with a significance level of  $\alpha$  for two normally distributed populations, where  $\sigma^2$  is assumed to be the same for each population.

Compute the test statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where  $s = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)}$

If  $t > t_{n_1 + n_2 - 2, 1-\alpha/2}$  or  $t < -t_{n_1 + n_2 - 2, 1-\alpha/2}$

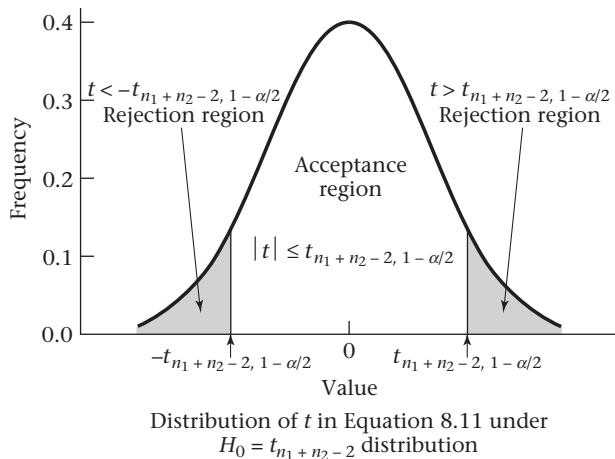
then  $H_0$  is rejected.

If  $-t_{n_1 + n_2 - 2, 1-\alpha/2} \leq t \leq t_{n_1 + n_2 - 2, 1-\alpha/2}$

then  $H_0$  is accepted.

The acceptance and rejection regions for this test are shown in Figure 8.3.

**Figure 8.3** Acceptance and rejection regions for the two-sample  $t$  test for independent samples with equal variances



Similarly, a  $p$ -value can be computed for the test. Computation of the  $p$ -value depends on whether  $\bar{x}_1 \leq \bar{x}_2$  ( $t \leq 0$ ) or  $\bar{x}_1 > \bar{x}_2$  ( $t > 0$ ). In each case, the  $p$ -value corresponds to the probability of obtaining a test statistic at least as extreme as the observed value  $t$ . This is given in Equation 8.12.

### Equation 8.12

#### Computation of the $p$ -Value for the Two-Sample $t$ Test for Independent Samples with Equal Variances

Compute the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where  $s = \sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2} / (n_1 + n_2 - 2)$

If  $t \leq 0$ ,  $p = 2 \times (\text{area to the left of } t \text{ under a } t_{n_1 + n_2 - 2} \text{ distribution})$ .

If  $t > 0$ ,  $p = 2 \times (\text{area to the right of } t \text{ under a } t_{n_1 + n_2 - 2} \text{ distribution})$ .

The computation of the  $p$ -value is illustrated in Figure 8.4.

### Example 8.10

**Hypertension** Assess the statistical significance of the data in Example 8.9.

#### Solution

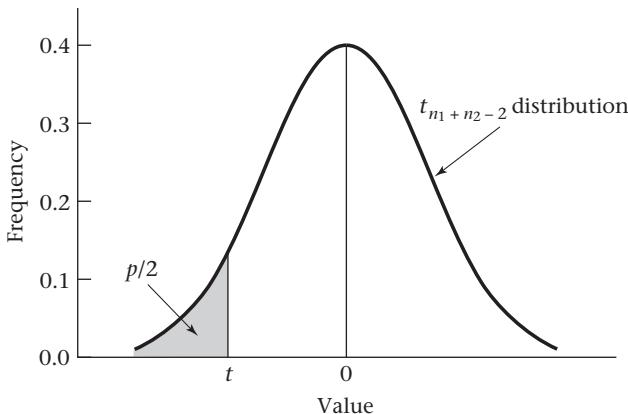
The common variance is first estimated:

$$s^2 = \frac{7(15.34)^2 + 20(18.23)^2}{27} = \frac{8293.9}{27} = 307.18$$

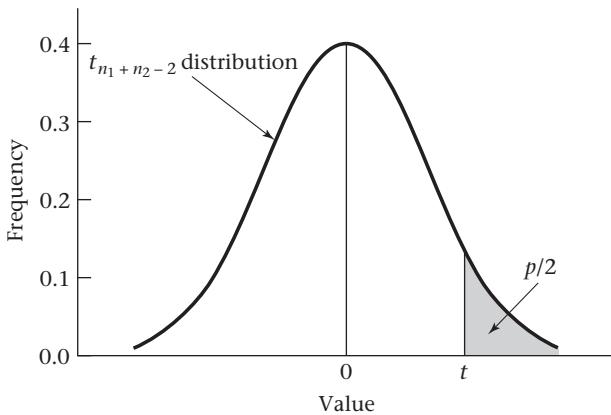
or  $s = 17.527$ . The following test statistic is then computed:

$$t = \frac{132.86 - 127.44}{17.527 \sqrt{1/8 + 1/21}} = \frac{5.42}{17.527 \times 0.415} = \frac{5.42}{7.282} = 0.74$$

**Figure 8.4 Computation of the  $p$ -value for the two-sample  $t$  test for independent samples with equal variances**



If  $t = (\bar{x}_1 - \bar{x}_2) / \left( s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \leq 0$ , then  $p = 2 \times (\text{area to the left of } t \text{ under a } t_{n_1 + n_2 - 2} \text{ distribution})$ .



If  $t = (\bar{x}_1 - \bar{x}_2) / \left( s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) > 0$ , then  $p = 2 \times (\text{area to the right of } t \text{ under a } t_{n_1 + n_2 - 2} \text{ distribution})$ .

If the critical-value method is used, then note that under  $H_0$ ,  $t$  comes from a  $t_{27}$  distribution. Referring to Table 5 in the Appendix, we see that  $t_{27,975} = 2.052$ . Because  $-2.052 \leq 0.74 \leq 2.052$ , it follows that  $H_0$  is accepted using a two-sided test at the 5% level, and we conclude that the mean blood pressures of the OC users and non-OC users do not significantly differ from each other. In a sense, this result shows the superiority of the longitudinal design in Example 8.5. Despite the similarity in the magnitudes of the mean blood-pressure differences between users and nonusers in the two studies, significant differences could be detected in Example 8.5, in contrast to the nonsignificant results that were obtained using the preceding cross-sectional design. The longitudinal design is usually more efficient because it uses people as their own controls.

To compute an approximate  $p$ -value, note from Table 5 that  $t_{27,75} = 0.684$ ,  $t_{27,80} = 0.855$ . Because  $0.684 < 0.74 < 0.855$ , it follows that  $.2 < p/2 < .25$  or  $.4 < p < .5$ . The exact  $p$ -value obtained from MINITAB is  $p = 2 \times P(t_{27} > 0.74) = .46$ .

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

In the previous section, methods of hypothesis testing for the comparison of means from two independent samples were discussed. It is also useful to compute  $100\% \times (1 - \alpha)$  CIs for the true mean difference between the two groups =  $\mu_1 - \mu_2$ . From Equation 8.7, if  $\sigma$  is known, then  $\bar{X}_1 - \bar{X}_2 \sim N[\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)]$  or, equivalently,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

If  $\sigma$  is unknown, then  $\sigma$  is estimated by  $s$  from Equation 8.10 and

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

To construct a two-sided  $100\% \times (1 - \alpha)$  CI, note that

$$Pr \left[ -t_{n_1+n_2-2, 1-\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2, 1-\alpha/2} \right] = 1 - \alpha$$

This can be written in the form of two inequalities:

$$-t_{n_1+n_2-2, 1-\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{and } \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2, 1-\alpha/2}$$

Each inequality is multiplied by  $s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$  and  $\mu_1 - \mu_2$  is added to both sides to obtain

$$\mu_1 - \mu_2 - t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \bar{X}_1 - \bar{X}_2$$

$$\text{and } \bar{X}_1 - \bar{X}_2 \leq \mu_1 - \mu_2 + t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Finally,  $t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$  is added to both sides of the first inequality and subtracted from both sides of the second inequality to obtain

$$\mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\bar{X}_1 - \bar{X}_2 - t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2$$

If these two inequalities are combined, the required CI is obtained.

$$\left( \bar{X}_1 - \bar{X}_2 - t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

If the sample means  $\bar{x}_1, \bar{x}_2$  are substituted for the random variables  $\bar{X}_1, \bar{X}_2$  then this procedure can be summarized as follows.

### Equation 8.13

#### Confidence Interval for the Underlying Mean Difference ( $\mu_1 - \mu_2$ ) Between Two Groups (Two-Sided) ( $\sigma_1^2 = \sigma_2^2$ )

A two-sided  $100\% \times (1 - \alpha)$  CI for the true mean difference  $\mu_1 - \mu_2$  based on two independent samples is given by

$$\left( \bar{x}_1 - \bar{x}_2 - t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{n_1+n_2-2, 1-\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

### Example 8.11

**Hypertension** Using the data in Examples 8.9 and 8.10, compute a 95% CI for the true mean difference in systolic blood pressure (SBP) between 35- to 39-year-old OC users and non-OC users.

### Solution

A 95% CI for the underlying mean difference in SBP between the population of 35- to 39-year-old OC users and non-OC users is given by

$$\begin{aligned} & [5.42 - t_{27, .975}(7.282), 5.42 + t_{27, .975}(7.282)] \\ & = [5.42 - 2.052(7.282), 5.42 + 2.052(7.282)] = (-9.52, 20.36) \end{aligned}$$

This interval is rather wide and indicates that a much larger sample is needed to accurately assess the true mean difference.

In this section, we have introduced the two-sample  $t$  test for independent samples with equal variances. This test is used to compare the mean of a normally distributed random variable (or a random variable with samples large enough so that the central-limit theorem can be assumed to hold) between two independent samples with equal variances. If we refer to the flowchart (Figure 8.13, p. 308), starting from position 1 we answer yes to (1) two-sample problem? (2) underlying distribution normal or can central-limit theorem be assumed to hold? (3) inferences concerning means? (4) are samples independent? and no to (5) are variances of two samples significantly different? (discussed in Section 8.6). This leads us to the box labeled “Use two-sample  $t$  test with equal variances.”

## 8.6 Testing for the Equality of Two Variances

In Section 8.4, when we conducted a two-sample  $t$  test for independent samples, we assumed the underlying variances of the two samples were the same. We then estimated the common variance using a weighted average of the individual sample variances. In this section we develop a significance test to validate this assumption. In particular, we wish to test the hypothesis  $H_0: \sigma_1^2 = \sigma_2^2$  vs.  $H_1: \sigma_1^2 \neq \sigma_2^2$ , where the two samples are assumed to be independent random samples from an  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$  distribution, respectively.

**Example 8.12**

**Cardiovascular Disease, Pediatrics** Consider a problem discussed earlier, namely the familial aggregation of cholesterol levels. In particular, suppose cholesterol levels are assessed in 100 children, 2 to 14 years of age, of men who have died from heart disease and it is found that the mean cholesterol level in the group ( $\bar{x}_1$ ) is 207.3 mg/dL. Suppose the sample standard deviation in this group ( $s_1$ ) is 35.6 mg/dL. Previously, the cholesterol levels in this group of children were compared with 175 mg/dL, which was assumed to be the underlying mean level in children in this age group based on previous large studies.

A better experimental design would be to select a group of control children whose fathers are alive and do not have heart disease and who are from the same census tract as the case children, and then to compare their cholesterol levels with those of the case children. If the case fathers are identified by a search of death records from the census tract, then researchers can select control children who live in the same census tract as the case families but whose fathers have no history of heart disease. The case and control children come from the same census tract but are *not* individually matched. Thus they are considered as two independent samples rather than as two paired samples. The cholesterol levels in these children can then be measured. Suppose the researchers found that among 74 control children, the mean cholesterol level ( $\bar{x}_2$ ) is 193.4 mg/dL with a sample standard deviation ( $s_2$ ) of 17.3 mg/dL. We would like to compare the means of these two groups using the two-sample  $t$  test for independent samples given in Equation 8.11, but we are hesitant to assume equal variances because the sample variance of the case group is about four times as large as that of the control group:

$$35.6^2 / 17.3^2 = 4.23$$

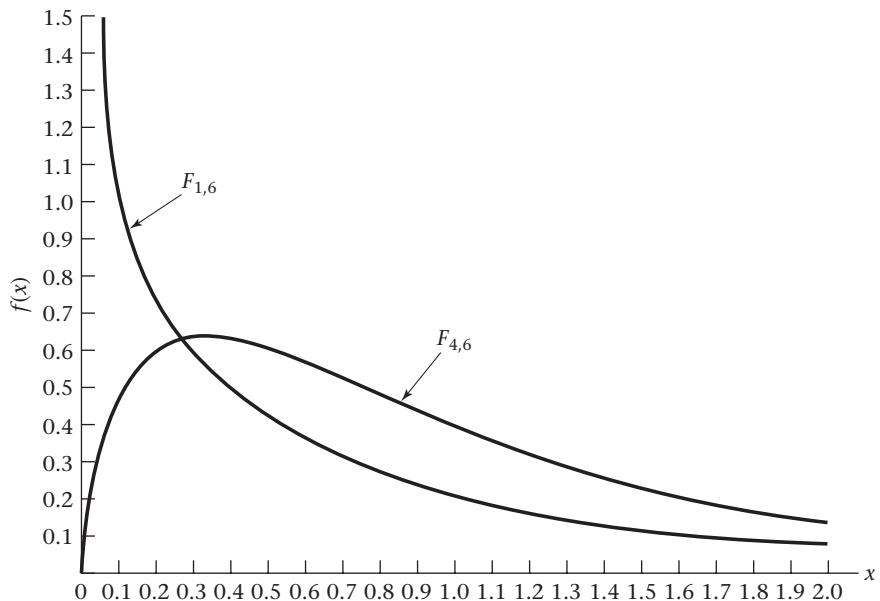
What should we do?

What we need is a significance test to determine if the underlying variances are in fact equal; that is, we want to test the hypothesis  $H_0: \sigma_1^2 = \sigma_2^2$  vs.  $H_1: \sigma_1^2 \neq \sigma_2^2$ . It seems reasonable to base the significance test on the relative magnitudes of the sample variances ( $s_1^2, s_2^2$ ). The best test in this case is based on the ratio of the sample variances ( $s_1^2/s_2^2$ ) rather than on the difference between the sample variances ( $s_1^2 - s_2^2$ ). Thus  $H_0$  would be rejected if the variance ratio is either too large or too small and accepted otherwise. To implement this test, the sampling distribution of  $s_1^2/s_2^2$  under the null hypothesis  $\sigma_1^2 = \sigma_2^2$  must be determined.

## The $F$ Distribution

The distribution of the variance ratio ( $s_1^2/s_2^2$ ) was studied by statisticians R. A. Fisher and G. Snedecor. It can be shown that the variance ratio follows an  **$F$  distribution** under the null hypothesis that  $\sigma_1^2 = \sigma_2^2$ . There is no unique  $F$  distribution but instead a family of  $F$  distributions. This family is indexed by two parameters termed the *numerator* and *denominator degrees of freedom*, respectively. If the sizes of the first and second samples are  $n_1$  and  $n_2$  respectively, then the variance ratio follows an  $F$  distribution with  $n_1 - 1$  (numerator  $df$ ) and  $n_2 - 1$  (denominator  $df$ ), which is called an  $F_{n_1-1, n_2-1}$  distribution.

The  $F$  distribution is generally positively skewed, with the skewness dependent on the relative magnitudes of the two degrees of freedom. If the numerator  $df$  is 1 or 2, then the distribution has a mode at 0; otherwise it has a mode greater than 0. The distribution is illustrated in Figure 8.5. Table 9 in the Appendix gives the percentiles of the  $F$  distribution.

**Figure 8.5** Probability density for the *F* distribution

**Definition 8.6** The  $100 \times p$ th percentile of an *F* distribution with  $d_1$  and  $d_2$  degrees of freedom is denoted by  $F_{d_1, d_2, p}$ . Thus

$$\Pr(F_{d_1, d_2} \leq F_{d_1, d_2, p}) = p$$

The *F* table is organized such that the numerator  $df(d_1)$  is shown in the first row, the denominator  $df(d_2)$  is shown in the first column, and the various percentiles ( $p$ ) are shown in the second column.

**Example 8.13**

Find the upper first percentile of an *F* distribution with 5 and 9 *df*.

**Solution**

$F_{5,9,.99}$  must be found. Look in the 5 column, the 9 row, and the subrow marked .99 to obtain

$$F_{5,9,.99} = 6.06$$

Generally, *F* distribution tables give only upper percentage points because the symmetry properties of the *F* distribution make it possible to derive the lower percentage points of any *F* distribution from the corresponding upper percentage points of an *F* distribution with the degrees of freedom reversed. Specifically, note that under  $H_0$ ,  $S_2^2/S_1^2$  follows an  $F_{d_2, d_1}$  distribution. Therefore,

$$\Pr(S_2^2/S_1^2 \geq F_{d_2, d_1, 1-p}) = p$$

By taking the inverse of each side and reversing the direction of the inequality, we get

$$\Pr\left(\frac{S_1^2}{S_2^2} \leq \frac{1}{F_{d_2, d_1, 1-p}}\right) = p$$

Under  $H_0$ , however,  $S_1^2/S_2^2$  follows an  $F_{d_1, d_2}$  distribution. Therefore

$$\Pr\left(\frac{S_1^2}{S_2^2} \leq F_{d_1, d_2, p}\right) = p$$

It follows from the last two inequalities that

$$F_{d_1, d_2, p} = \frac{1}{F_{d_2, d_1, 1-p}}$$

This principle is summarized as follows.

#### Equation 8.14

##### Computation of the Lower Percentiles of an $F$ Distribution

The lower  $p$ th percentile of an  $F$  distribution with  $d_1$  and  $d_2$  df is the reciprocal of the upper  $p$ th percentile of an  $F$  distribution with  $d_2$  and  $d_1$  df. In symbols,

$$F_{d_1, d_2, p} = 1/F_{d_2, d_1, 1-p}$$

Thus from Equation 8.14 we see that the lower  $p$ th percentile of an  $F$  distribution is the same as the inverse of the upper  $p$ th percentile of an  $F$  distribution with the degrees of freedom reversed.

#### Example 8.14

Estimate  $F_{6,8,05}$ .

#### Solution

From Equation 8.14,  $F_{6,8,05} = 1/F_{8,6,95} = 1/4.15 = 0.241$

## The $F$ Test

We now return to the significance test for the equality of two variances. We want to test the hypothesis  $H_0: \sigma_1^2 = \sigma_2^2$  vs.  $H_1: \sigma_1^2 \neq \sigma_2^2$ . We stated that the test would be based on the variance ratio  $S_1^2/S_2^2$ , which under  $H_0$  follows an  $F$  distribution with  $n_1 - 1$  and  $n_2 - 1$  df. This is a two-sided test, so we want to reject  $H_0$  for both small and large values of  $S_1^2/S_2^2$ . This procedure can be made more specific, as follows.

#### Equation 8.15

##### $F$ Test for the Equality of Two Variances

Suppose we want to conduct a test of the hypothesis  $H_0: \sigma_1^2 = \sigma_2^2$  vs.  $H_1: \sigma_1^2 \neq \sigma_2^2$  with significance level  $\alpha$ .

Compute the test statistic  $F = s_1^2/s_2^2$ .

If  $F > F_{n_1-1, n_2-1, 1-\alpha/2}$  or  $F < F_{n_1-1, n_2-1, \alpha/2}$

then  $H_0$  is rejected.

If  $F_{n_1-1, n_2-1, \alpha/2} \leq F \leq F_{n_1-1, n_2-1, 1-\alpha/2}$

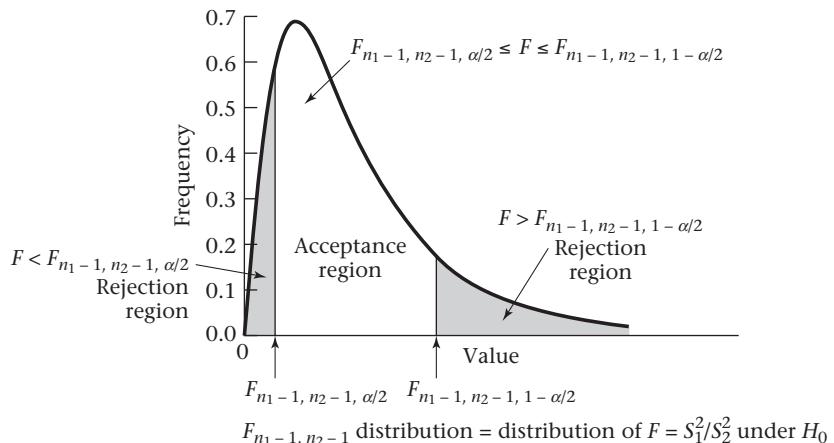
then  $H_0$  is accepted. The acceptance and rejection regions for this test are shown in Figure 8.6.

Alternatively, the exact  $p$ -value is given by Equation 8.16.

#### Equation 8.16

##### Computation of the $p$ -Value for the $F$ Test for the Equality of Two Variances

Compute the test statistic  $F = s_1^2/s_2^2$ .

**Figure 8.6** Acceptance and rejection regions for the  $F$  test for the equality of two variances

$$\begin{aligned} \text{If } F \geq 1, \text{ then } p &= 2 \times Pr(F_{n_1 - 1, n_2 - 1} > F) \\ \text{If } F < 1, \text{ then } p &= 2 \times Pr(F_{n_1 - 1, n_2 - 1} < F) \end{aligned}$$

This computation is illustrated in Figure 8.7.

### Example 8.15

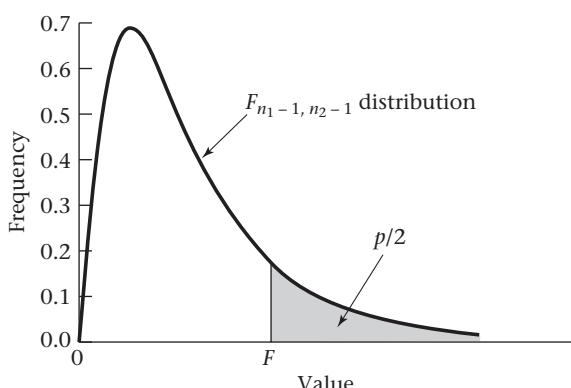
**Cardiovascular Disease, Pediatrics** Test for the equality of the two variances given in Example 8.12.

#### Solution

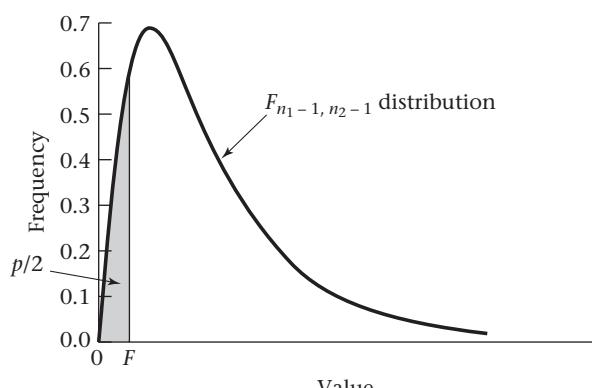
$$F = s_1^2/s_2^2 = 35.6^2/17.3^2 = 4.23$$

Because the two samples have 100 and 74 people, respectively, we know from Equation 8.15 that under  $H_0$ ,  $F \sim F_{99,73}$ . Thus  $H_0$  is rejected if

$$F > F_{99,73,.975} \quad \text{or} \quad F < F_{99,73,.025}$$

**Figure 8.7** Computation of the  $p$ -value for the  $F$  test for the equality of two variances

If  $F = s_1^2/s_2^2 \geq 1$ , then  $p = 2 \times$  (area to the right of  $F$  under an  $F_{n_1 - 1, n_2 - 1}$  distribution)



If  $F = s_1^2/s_2^2 < 1$ , then  $p = 2 \times$  (area to the left of  $F$  under an  $F_{n_1 - 1, n_2 - 1}$  distribution)

Note that neither 99  $df$  nor 73  $df$  appears in Table 9 in the Appendix. One approach is to obtain the percentiles using a computer program. In this example, we want to find the value  $c_1 = F_{99,73,025}$  and  $c_2 = F_{99,73,975}$ , such that

$$\Pr(F_{99,73} \leq c_1) = .025 \quad \text{and} \quad \Pr(F_{99,73} \geq c_2) = .975$$

The result is shown in Table 8.3 using the FINV function of Microsoft Excel 2007, where the first argument of FINV is the desired right-hand tail area and the next two arguments are the numerator and denominator  $df$ , respectively.

**Table 8.3** Computation of critical values for the cholesterol data in Example 8.15 using Excel 2007

Numerator df	99	
Denominator df	73	
<b>Percentile</b>		
0.025	0.6547598	FINV(.975, 99, 73)
0.975	1.54907909	FINV(.025, 99, 73)

Thus  $c_1 = 0.6548$  and  $c_2 = 1.5491$ . Because  $F = 4.23 > c_2$ , it follows that  $p < .05$ . Alternatively, we could compute the exact  $p$ -value. This is given by  $p = 2 \times \Pr(F_{99,73} \geq 4.23)$ .

**Table 8.4** Computation of the exact  $p$ -value in Example 8.15 using Excel 2007

Numerator df	99	
Denominator df	73	
<b>x</b>		
	4.23	
one-tailed $p$ -value	4.41976E-10	FDIST(4.23, 99, 73)
two-tailed $p$ -value	8.83951E-10	2*FDIST(4.23, 99, 73)

Using the Excel 2007 FDIST function, which calculates the right-hand tail area, we see from Table 8.4 that to four decimal places, the  $p$ -value =  $2 \times \Pr(F_{99,73} \geq 4.23) \leq .0001$ . Thus the two sample variances are significantly different. The two-sample  $t$  test with equal variances, as given in Section 8.4, cannot be used, because this test depends on the assumption that the variances are equal.

A question often asked about the  $F$  test is whether it makes a difference which sample is selected as the numerator sample and which is selected as the denominator sample. The answer is that, for a two-sided test, it does *not* make a difference because of the rules for calculating lower percentiles given in Equation 8.14. A variance ratio  $> 1$  is usually more convenient, so there is no need to use Equation 8.14. Thus the larger variance is usually put in the numerator and the smaller variance in the denominator.

### Example 8.16

**Hypertension** Using the data in Example 8.9, test whether the variance of blood pressure is significantly different between OC users and non-OC users.

### Solution

The sample standard deviation of blood pressure for the 8 OC users was 15.34 and for the 21 non-OC users was 18.23. Hence the variance ratio is

$$F = (18.23/15.34)^2 = 1.41$$

Under  $H_0$ ,  $F$  follows an  $F$  distribution with 20 and 7  $df$ , whose percentiles do not appear in Table 9. However, the percentiles of an  $F_{24,7}$  distribution are provided in Table 9. Also, it can be shown that for a specified upper percentile (e.g., the 97.5th

percentile), as either the numerator or denominator *df* increases, the corresponding percentile decreases. Therefore,

$$F_{20,7,.975} \geq F_{24,7,.975} = 4.42 > 1.41$$

It follows that  $p > 2(.025) = .05$ , and the underlying variances of the two samples do not significantly differ from each other. Thus it was correct to use the two-sample *t* test for independent samples with *equal variances* for these data, where the underlying variances were assumed to be the same.

To compute an exact *p*-value, a computer program must be used to evaluate the area under the *F* distribution. The exact *p*-value for Example 8.16 has been evaluated using Excel 2007, with the results given in Table 8.5. The program evaluates the right-hand tail area =  $Pr(F_{20,7} \geq 1.41) = .334$ . The two-tailed *p*-value =  $2 \times Pr(F_{20,7} \geq 1.41) = 2 \times .334 = .669$ .

**Table 8.5** Computation of the exact *p*-value for the blood-pressure data in Example 8.16 using the *F* test for the equality of two variances with the Excel 2007 FDIST program

Numerator df	20
Denominator df	7
x	1.412285883
one-tailed p-value	0.334279505
two-tailed p-value	0.66855901
	FDIST(1.41, 20, 7)
	2*FDIST(1.41, 20, 7)

If the numerator and denominator samples are reversed, then the *F* statistic =  $1/1.41 = 0.71 \sim F_{7,20}$  under  $H_0$ . We use the FDIST program of Excel 2007 to calculate  $Pr(F_{7,20} \geq 0.71)$ . This is given by  $FDIST(0.71, 7, 20) = .666$ . Because  $F < 1$ , we have *p*-value =  $2 \times Pr(F_{7,20} \leq 0.71) = 2 \times (1 - .666) = .669$ , which is the same as the *p*-value in Table 8.5. Thus it was correct to use the two-sample *t* test for independent samples with *equal variances* for these data, where the variances were assumed to be the same.

In this section, we have introduced the *F* test for the equality of two variances. This test is used to compare variance estimates from two normally distributed samples. If we refer to the flowchart (Figure 8.13, p. 308), then starting from position 1 we answer yes to (1) two-sample problem? and (2) underlying distribution normal or can central-limit theorem be assumed to hold? and no to (3) inferences concerning means? and yes to (4) inferences concerning variances? This leads us to the box labeled "Two-sample *F* test to compare variances." Be cautious about using this test with nonnormally distributed samples.

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

The *F* test for the equality of two variances from two independent, normally distributed samples was presented in Equation 8.15. If the two variances are not significantly different, then the two-sample *t* test for independent samples with *equal variances* outlined in Section 8.4 can be used. If the two variances are significantly different, then a two-sample *t* test for independent samples with *unequal variances*, which is presented in this section, should be used.

Specifically, assume there are two normally distributed samples, where the first sample is a random sample of size  $n_1$  from an  $N(\mu_1, \sigma_1^2)$  distribution and the

second sample is a random sample from an  $N(\mu_2, \sigma_2^2)$  distribution, and  $\sigma_1^2 \neq \sigma_2^2$ . We again wish to test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ . Statisticians refer to this problem as the **Behrens-Fisher problem**.

It still makes sense to base the significance test on the difference between the sample means  $\bar{x}_1 - \bar{x}_2$ . Under either hypothesis,  $\bar{X}_1$  is normally distributed with mean  $\mu_1$  and variance  $\sigma_1^2/n_1$  and  $\bar{X}_2$  is normally distributed with mean  $\mu_2$  and variance  $\sigma_2^2/n_2$ . Hence it follows that

**Equation 8.17**

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Under  $H_0$ ,  $\mu_1 - \mu_2 = 0$ . Thus, from Equation 8.17,

**Equation 8.18**

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

If  $\sigma_1^2$  and  $\sigma_2^2$  were known, then the test statistic

**Equation 8.19**

$$z = (\bar{x}_1 - \bar{x}_2) / \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

could be used for the significance test, which under  $H_0$  would be distributed as an  $N(0,1)$  distribution. However,  $\sigma_1^2$  and  $\sigma_2^2$  are usually unknown and are estimated by  $s_1^2$  and  $s_2^2$ , respectively (the sample variances in the two samples). Notice that a pooled estimate of the variance was not computed as in Equation 8.10 because the variances  $(\sigma_1^2, \sigma_2^2)$  are assumed to be different. If  $s_1^2$  is substituted for  $\sigma_1^2$  and  $s_2^2$  for  $\sigma_2^2$  in Equation 8.19, then the following test statistic is obtained:

**Equation 8.20**

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

The exact distribution of  $t$  under  $H_0$  is difficult to derive. However, several approximate solutions have been proposed that have appropriate type I error. The Satterthwaite approximation is presented here. Its advantage is its easy implementation using the ordinary  $t$  tables [1].

**Equation 8.21**

### Two-Sample $t$ Test for Independent Samples with Unequal Variances (Satterthwaite's Method)

- (1) Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- (2) Compute the approximate degrees of freedom  $d'$ , where

$$d' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

(3) Round  $d'$  down to the nearest integer  $d''$ .

$$\text{If } t > t_{d'', 1-\alpha/2} \quad \text{or} \quad t < -t_{d'', 1-\alpha/2}$$

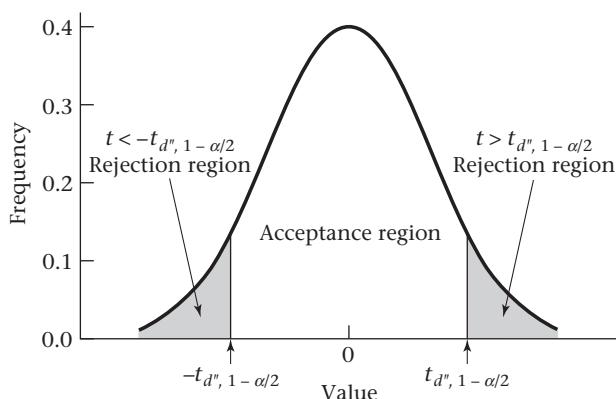
then reject  $H_0$ .

$$\text{If } -t_{d'', 1-\alpha/2} \leq t \leq t_{d'', 1-\alpha/2}$$

then accept  $H_0$ .

The acceptance and rejection regions for this test are illustrated in Figure 8.8.

**Figure 8.8** Acceptance and rejection regions for the two-sample  $t$  test for independent samples with unequal variances



$t_{d''}$  distribution = approximate distribution of  $t$  in Equation 8.21 under  $H_0$

Similarly, the approximate  $p$ -value for the hypothesis test can be computed as follows.

### Equation 8.22

#### Computation of the $p$ -Value for the Two-Sample $t$ Test for Independent Samples with Unequal Variances (Satterthwaite Approximation)

Compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If  $t \leq 0$ , then  $p = 2 \times (\text{area to the left of } t \text{ under a } t_{d''} \text{ distribution})$

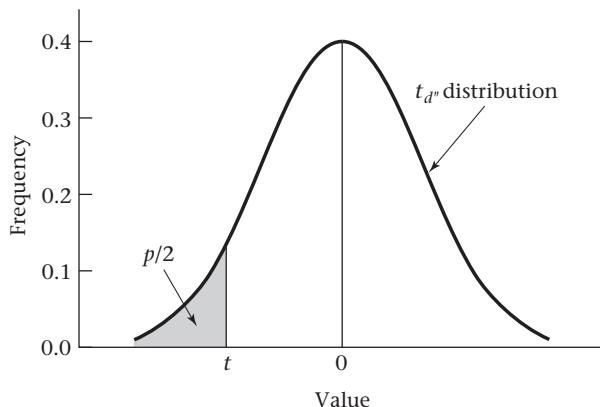
If  $t > 0$ , then  $p = 2 \times (\text{area to the right of } t \text{ under a } t_{d''} \text{ distribution})$   
where  $d''$  is given in Equation 8.21.

Computation of the  $p$ -value is illustrated in Figure 8.9.

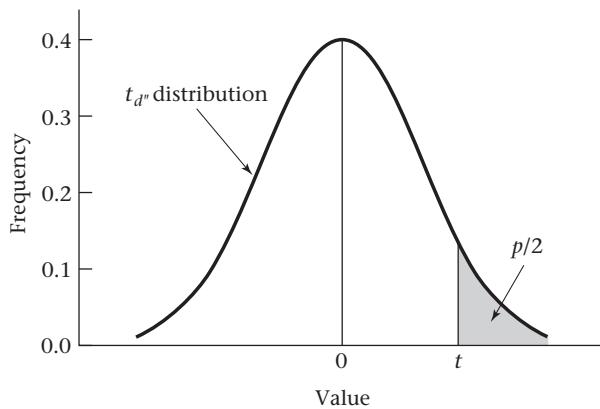
### Example 8.17

**Cardiovascular Disease, Pediatrics** Consider the cholesterol data in Example 8.12. Test for the equality of the mean cholesterol levels of the children whose fathers have died from heart disease vs. the children whose fathers do not have a history of heart disease.

**Figure 8.9 Computation of the  $p$ -value for the two-sample  $t$  test for independent samples with unequal variances**



If  $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2} \leq 0$ , then  $p = 2 \times$   
(area to the left of  $t$  under a  $t_{d''}$  distribution)



If  $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2} > 0$ , then  $p = 2 \times$   
(area to the right of  $t$  under a  $t_{d''}$  distribution)

### Solution

We have already tested for equality of the two variances in Example 8.15 and found them to be significantly different. Thus the two-sample  $t$  test for unequal variances in Equation 8.21 should be used. The test statistic is

$$t = \frac{207.3 - 193.4}{\sqrt{35.6^2/100 + 17.3^2/74}} = \frac{13.9}{4.089} = 3.40$$

The approximate degrees of freedom are now computed:

$$\begin{aligned} d' &= \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1) + \left(s_2^2/n_2\right)^2/(n_2-1)} \\ &= \frac{(35.6^2/100 + 17.3^2/74)^2}{(35.6^2/100)^2/99 + (17.3^2/74)^2/73} = \frac{16.718^2}{1.8465} = 151.4 \end{aligned}$$

Therefore, the approximate degrees of freedom =  $d'' = 151$ . If the critical-value method is used, note that  $t = 3.40 > t_{120, .975} = 1.980 > t_{151, .975}$ . Therefore,  $H_0$  can be

rejected using a two-sided test with  $\alpha = .05$ . Furthermore,  $t = 3.40 > t_{120,9995} = 3.373 > t_{151,9995}$ , which implies that the  $p$ -value  $< 2 \times (1.0 - .9995) = .001$ . To compute the exact  $p$ -value, we use Excel 2007, as shown in Table 8.6.

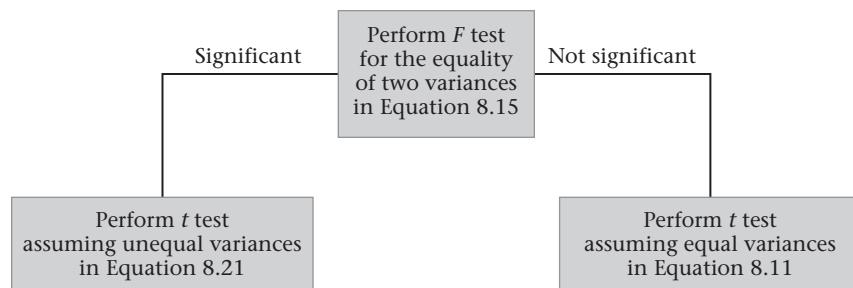
**Table 8.6****Computation of the exact  $p$ -value for Example 8.17 using Excel 2007**

<b>t</b>	<b>3.4</b>
<b>df</b>	<b>151</b>
<b>two-tailed p-value</b>	<b>0.000862 TDIST(3.4,151,2)</b>

We see from Table 8.6 that the  $p$ -value  $= 2 \times [1 - Pr(t_{151} \leq 3.40)] = .0009$ . We conclude that mean cholesterol levels in children whose fathers have died from heart disease are significantly higher than mean cholesterol levels in children of fathers without heart disease. It would be of great interest to identify the cause of this difference; that is, whether it is due to genetic factors, environmental factors such as diet, or both.

In this chapter, two procedures for comparing two means from independent, normally distributed samples have been presented. The first step in this process is to test for the equality of the two variances, using the  $F$  test in Equation 8.15. If this test is not significant, then use the  $t$  test with equal variances; otherwise, use the  $t$  test with unequal variances. This overall strategy is illustrated in Figure 8.10.

**Figure 8.10** **Strategy for testing for the equality of means in two independent, normally distributed samples**



### Example 8.18

**Infectious Disease** Using the data in Table 2.11, compare the mean duration of hospitalization between antibiotic users and nonantibiotic users.

### Solution

Refer to Table 8.7, where the PC-SAS T-TEST program (PROC TTEST) was used to analyze these data. Among the 7 antibiotic users ( $antib = 1$ ), mean duration of hospitalization was 11.57 days with standard deviation 8.81 days; among the 18 nonantibiotic users ( $antib = 2$ ), mean duration of hospitalization was 7.44 days with standard deviation 3.70 days. Both the  $F$  test and the  $t$  test with equal and unequal variances are displayed in this program. Using Figure 8.10, note that the first step in comparing the two means is to perform the  $F$  test for the equality of two variances in order to decide whether to use the  $t$  test with equal or unequal variances. The  $F$  statistic is denoted in Table 8.7 by  $F$  Value = 5.68., with  $p$ -value (labeled  $Pr > F$ ) = .004. Thus the variances differ significantly, and a two-sample  $t$  test with unequal variances should be used. Therefore, refer to the Unequal Variance row, where

**Table 8.7 Use of the PROC TTEST program to analyze the association between antibiotic use and duration of hospitalization (raw data presented in Table 2.11)**

The TTEST Procedure																				
Statistics																				
Variable	antib	Lower CL			Upper CL			Lower CL			Upper CL									
		N	Mean	Mean	Mean	Std Dev	Std Dev	Std Dev	Std Dev	Std Err	Minimum	Maximum								
dur	1	7	3.4234	11.571	19.719	5.6772	8.8102	19.401	3.3299	3	30									
dur	2	18	5.6056	7.4444	9.2833	2.7747	3.6977	5.5434	0.8716	3	17									
dur	Diff		-0.95	4.127	9.2037	4.2821	5.5095	7.7285	2.4541											
	(1-2)																			
T-Tests																				
Variable	Method	Variances			DF		t Value		Pr >  t											
dur	Pooled	Equal			23		1.68		0.1062											
dur	Satterthwaite	Unequal			6.84		1.20		0.2704											
Equality of Variances																				
Variable	Method	Num DF	Den DF	F Value			Pr > F													
dur	Folded F	6	17	5.68			0.0043													

the  $t$  statistic (as given in Equation 8.21) is 1.20 with degrees of freedom  $d''(df) = 6.8$ . The corresponding two-tailed  $p$ -value (labeled  $\text{Pr} > |t|$ ) = .270. Thus no significant difference exists between the mean duration of hospitalization in these two groups.

If the results of the  $F$  test had revealed a nonsignificant difference between the variances of the two samples, then the  $t$  test with equal variances would have been used, which is provided in the Equal Variance row of the SAS output. In this example, considerable differences are present in both the test statistics (1.68 vs. 1.20) and the two-tailed  $p$ -values (.106 vs. .270) resulting from using these two procedures.

Using similar methods to those developed in Section 8.5, we can show that a two-sided  $100\% \times (1 - \alpha)$  CI for the underlying mean difference  $\mu_1 - \mu_2$  in the case of unequal variances is given as follows:

### Equation 8.23

Two-Sided  $100\% \times (1 - \alpha)$  CI for

$$\mu_1 - \mu_2 \left( \sigma_1^2 \neq \sigma_2^2 \right)$$

$$\left( \bar{x}_1 - \bar{x}_2 - t_{d'',1-\alpha/2} \sqrt{s_1^2/n_1 + s_2^2/n_2}, \bar{x}_1 - \bar{x}_2 + t_{d'',1-\alpha/2} \sqrt{s_1^2/n_1 + s_2^2/n_2} \right)$$

where  $d''$  is given in Equation 8.21.

### Example 8.19

**Infectious Disease** Using the data in Table 8.7, compute a 95% CI for the mean difference in duration of hospital stay between patients who do and do not receive antibiotics.

### Solution

Using Table 8.7, the 95% CI is given by

$$\begin{aligned} & \left[ (11.571 - 7.444) - t_{6,.975} \sqrt{8.810^2/7 + 3.698^2/18}, \right. \\ & \quad \left. (11.571 - 7.444) + t_{6,.975} \sqrt{8.810^2/7 + 3.698^2/18} \right] \\ & = [4.127 - 2.447(3.442), 4.127 + 2.447(3.442)] \\ & = (4.127 - 8.423, 4.127 + 8.423) = (-4.30, 12.55) \end{aligned}$$

In this section, we have introduced the two-sample  $t$  test for independent samples with unequal variances. This test is used to compare the mean level of a normally distributed random variable (or a random variable with samples large enough so the central-limit theorem can be assumed to hold) between two independent samples with unequal variances. If we refer to the flowchart (Figure 8.13, p. 308), then starting from position 1 we answer yes to the following five questions: (1) two-sample problem? (2) underlying distribution normal or can central-limit theorem be assumed to hold? (3) inference concerning means? (4) are samples independent? and (5) are variances of two samples significantly different? This leads us to the box labeled “Use two-sample  $t$  test with unequal variances.”

### REVIEW QUESTIONS 8B

- 1** What is an  $F$  distribution used for? How does it differ from a  $t$  distribution?
- 2** Suppose we wish to compare the mean level of systolic blood pressure (SBP) between Caucasian and African-American children. The following data were obtained from the Bogalusa Heart Study for 10- to 14-year-old girls:

**Table 8.8 Comparison of mean SBP of Caucasian and African-American 10- to 14-year-old girls**

	Mean	<i>sd</i>	<i>N</i>
Caucasian	104.4	9.0	1554
African-American	104.7	9.3	927

- (a) What test can be used to compare the means of the two groups?
- (b) Perform the test in Review Question 8B.2a, and report a *p*-value (two-tailed). [Hint:  $F_{926,1553,975} = 1.121$ . Also, for  $d \geq 200$ , assume that a  $t_d$  distribution is the same as an  $N(0,1)$  distribution.]
- (c) What is a 95% CI for the mean difference in SBP between the two ethnic groups?
- 3 The following data comparing SBP between Caucasian and African-American young adult women were obtained from the same study:

**Table 8.9 Comparison of mean SBP of Caucasian and African-American 30- to 34-year-old women**

	Mean	<i>sd</i>	<i>N</i>
Caucasian	107.7	9.5	195
African-American	115.3	14.9	96

Answer the same questions as in Review Question 8B.2a, b, and c. (Note:  $F_{95,194,975} = 1.402$ .)

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

### Example 8.20

**Environmental Health, Pediatrics** In Section 2.9, we described a study performed in El Paso, Texas, that examined the association between lead exposure and developmental features in children [2]. There are different ways to quantify lead exposure. One

method used in the study consisted of defining a control group of children whose blood-lead levels were  $< 40 \mu\text{g}/100 \text{ mL}$  in both 1972 and 1973 ( $n = 78$ ) and an exposed group of children who had blood-lead levels  $\geq 40 \mu\text{g}/100 \text{ mL}$  in either 1972 or 1973 ( $n = 46$ ). Two important outcome variables in the study were the number of finger-wrist taps per 10 seconds in the dominant hand (a measure of neurologic function) as well as the Wechsler full-scale IQ score (a measure of intellectual development). Because only children  $\geq 5$  years of age were given the neurologic tests, we actually have 35 exposed and 64 control children with finger-wrist tapping scores. The distributions of these variables by group were displayed in a box plot in Figures 2.9 and 2.10, respectively. The distributions appeared to be reasonably symmetric, particularly in the exposed group, although there is a hint that a few outliers may be present. (We discuss detection of outliers more formally in Section 8.9.) We also note from these figures that the exposed group seems to have lower levels than the control group for both these variables. How can we confirm whether this impression is correct?

One approach is to use a two-sample  $t$  test to compare the mean level of the exposed group with the mean level of the control group on these variables. We used the PC-SAS TTEST procedure for this purpose, as shown in Tables 8.10 and 8.11. The program actually performs three different significance tests each time the  $t$  test procedure is specified. In Table 8.10, we analyze the mean finger-wrist tapping scores. Following the flowchart in Figure 8.10, we first perform the  $F$  test for equality of two variances. In Table 8.10, the  $F$  statistic (labeled as  $F$  Value) = 1.19 with 34 and 63  $df$ . The  $p$ -value (labeled  $\text{Pr} > F$ ) equals 0.5408, which implies we can accept  $H_0$  that the variances are *not* significantly different. Therefore, following Figure 8.10, we should perform the two-sample  $t$  test with equal variances (Equation 8.11). The  $t$  statistic is in the  $t$  Value column and the Equal row is 2.68 with 97  $df$ . The two-tailed  $p$ -value, found in the column headed  $\text{Pr} > |t|$  and the Equal row is 0.0087, which implies there is a significant difference in mean finger-wrist tapping scores between the exposed and the control groups, with the exposed group having lower mean scores. If there

**Table 8.10 Comparison of mean finger–wrist tapping scores for the exposed vs. control group, using the SAS  $t$  test procedure**

The TTEST Procedure																
Statistics																
Variable	group	Lower CL			Upper CL			Lower CL			Upper CL					
		N	Mean	Mean	Mean	Std Dev	Std Dev	Std Dev	Std Dev	Std Err	Minimum	Maximum				
maxfwt	1	64	51.426	54.438	57.449	10.27	12.057	14.602	1.5071	13	84					
maxfwt	2	35	42.909	47.429	51.948	10.641	13.156	17.237	2.2237	13	83					
maxfwt	Diff(1-2)	1.813	7.0089	12.205	10.92	12.453	14.49	2.618								
T-Tests																
Variable	Method	Variances			DF	$t$ Value			$\text{Pr} >  t $							
maxfwt	Pooled	Equal			97	2.68			0.0087							
maxfwt	Satterthwaite	Unequal			65	2.61			0.0113							
Equality of Variances																
Variable	Method	Num DF	Den DF	$F$ Value			$\text{Pr} > F$									
maxfwt	Folded F	34	63	1.19			0.5408									

**Table 8.11 Comparison of mean full-scale IQ scores for the exposed vs. control group, using the SAS t test procedure**

The TTEST Procedure														
Statistics														
Variable	group	N	Lower CL	Upper CL	Mean	Mean	Lower CL	Upper CL	Std Dev	Std Dev	Std Dev	Std Err	Minimum	Maximum
iqf	1	78	89.425	92.885	96.344	13.257	15.345	18.218	1.7374	50	141			
iqf	2	46	84.397	88.022	91.647	10.125	12.207	15.374	1.7998	46	114			
iqf	Diff(1-2)		-0.388	4.8629	10.114	12.68	14.268	16.313	2.6524					
T-Tests														
Variable	Method		Variances			DF	t Value		Pr >  t					
iqf	Pooled		Equal			122	1.83		0.0692					
iqf	Satterthwaite		Unequal			111	1.94		0.0544					
Equality of Variances														
Variable	Method		Num DF		Den DF		F Value		Pr > F					
iqf	Folded F		77		45		1.58		0.0982					

had been a significant difference between the variances from the  $F$  test—that is, if  $(\text{Pr} > F) < 0.05$ —then we would use the two-sample  $t$  test with unequal variances. The program automatically performs both  $t$  tests and lets the user decide which to use. If a two-sample  $t$  test with unequal variances were used, then referring to the Unequal row, the  $t$  statistic equals 2.61 (as given in Equation 8.21) with 65  $df$  ( $d'$  in Equation 8.21) with a two-sided  $p$ -value equal to 0.0113. The program also provides the mean, standard deviation (Std Dev), and standard error (Std Err) for each group. Referring to Table 8.11, for the analysis of the full-scale IQ scores, we see that the  $p$ -value for the  $F$  test is 0.0982, which is not statistically significant. Therefore, we again use the equal variance  $t$  test. The  $t$  statistic is 1.83 with 122  $df$ , with two-tailed  $p$ -value equal to 0.0692. Thus the mean full-scale IQ scores for the two groups do *not* differ significantly.

## 8.9 The Treatment of Outliers

We saw that the case study in Section 8.8 suggested there might be some outliers in the finger-wrist tapping and IQ scores. Outliers can have an important impact on the conclusions of a study. It is important to definitely identify outliers and either exclude them outright or at least perform alternative analyses with and without the outliers present. Therefore, in this section we study some decision rules for outlier detection.

We refer to Figures 8.11 and 8.12, which provide stem-and-leaf and box plots from SAS of the finger-wrist tapping scores and the full-scale IQ scores for the control group and the exposed group, respectively. According to the box plots in Figure 8.11, there are potential outlying finger-wrist tapping scores (denoted by zeros in the plot) of 13, 23, 26, and 84 taps per 10 seconds for the control group and 13, 14, and 83 taps per 10 seconds for the exposed group. According to the box plots in Figure 8.12, there are potential outlying full-scale IQ scores of 50, 56, 125, 128, and 141 for the control group and 46 for the exposed group. All the potentially outlying

values are far from the mean in absolute value. Therefore, a useful way to quantify an extreme value is by the number of standard deviations that a value is from the mean. This statistic applied to the most extreme value in a sample is called the Extreme Studentized Deviate and is defined as follows.

**Definition 8.7** The Extreme Studentized Deviate (or ESD statistic) =  $\max_{i=1,\dots,n} |x_i - \bar{x}|/s$ .

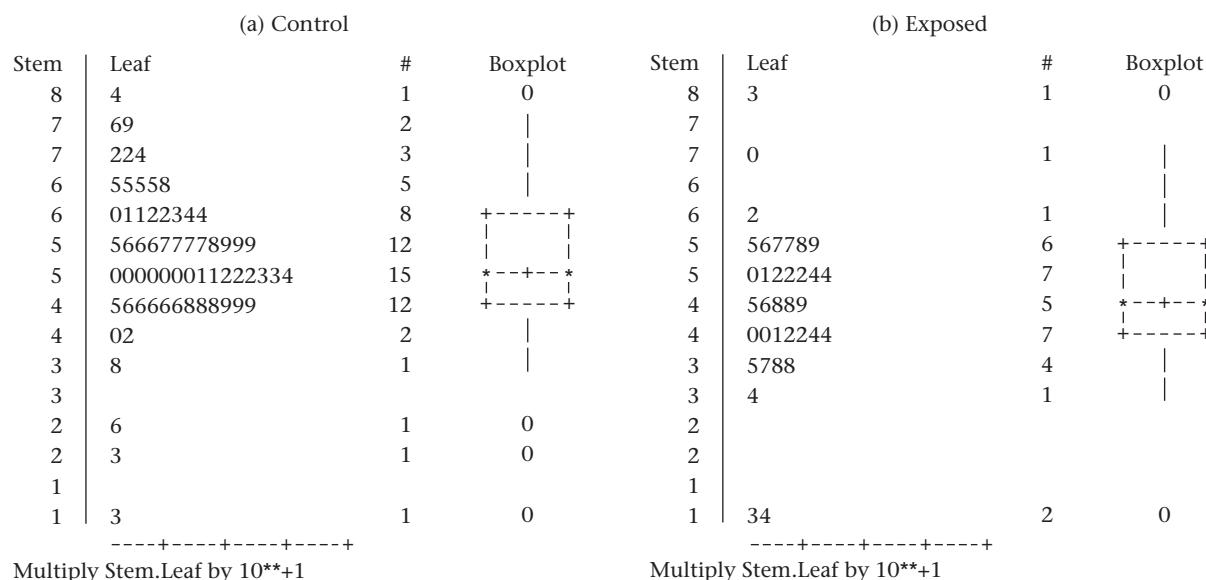
**Example 8.21** Compute the ESD statistic for the finger-wrist tapping scores for the control group.

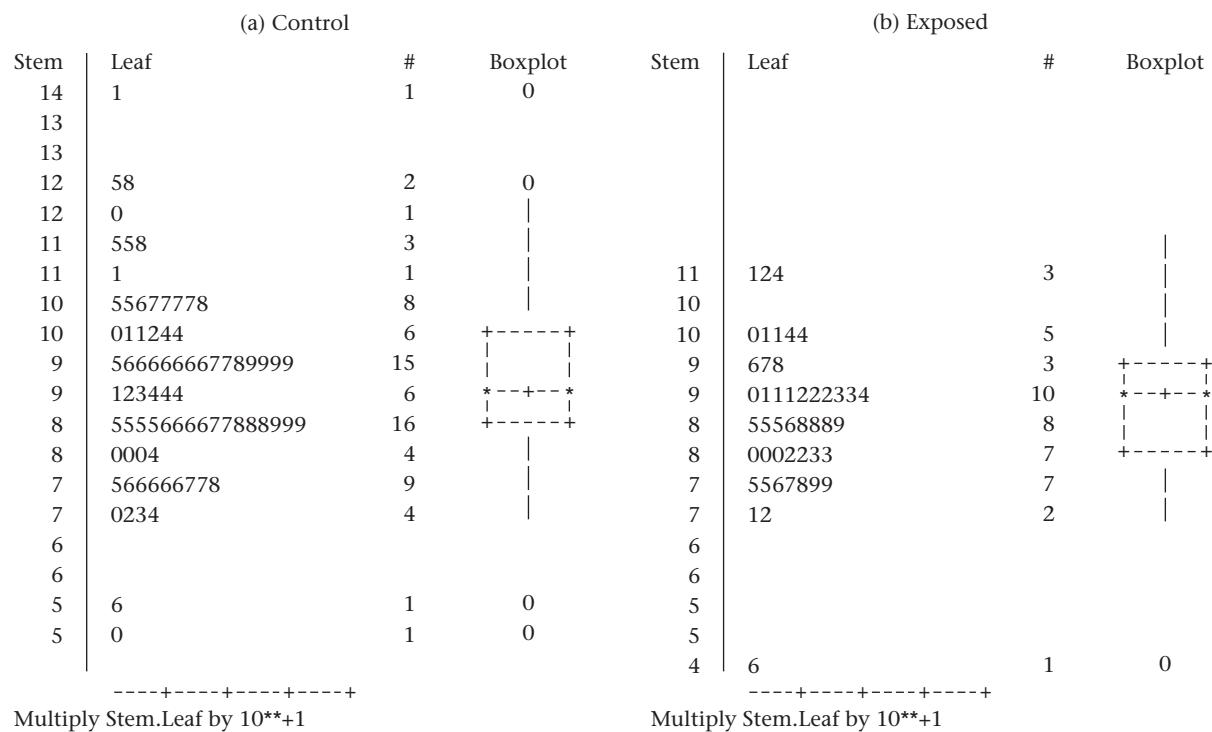
**Solution**

From Table 8.10, we see that  $\bar{x} = 54.4$ ,  $s = 12.1$ . From Figure 8.11a we note that the distance from the mean for the smallest and largest values are  $|13 - 54.4| = 41.4$  and  $|84 - 54.4| = 29.6$ , respectively. Therefore, because  $41.4 > 29.6$ , it follows that  $ESD = 41.4/12.1 = 3.44$ .

How large must the ESD statistic be for us to conclude that the most extreme value is an outlier? Remember that in a sample of size  $n$  without outliers, we would expect the largest value to correspond approximately to the  $100\% \times \left(\frac{n}{n+1}\right)$ th percentile. Thus, for a sample of size 64 from a normal distribution this would correspond to approximately the  $100 \times 64/65$ th percentile  $\approx 98.5$ th percentile = 2.17. If an outlier is present, then the ESD statistic will be larger than 2.17. The appropriate critical values depend on the sampling distribution of the ESD statistic for samples of size  $n$  from a normal distribution. Critical values from Rosner [3] based on an approximation provided by Quesenberry and David [4] are presented in Table 10 in the Appendix. The critical values depend on the sample size  $n$  and the percentile  $p$ . The  $p$ th percentile for the ESD statistic based on a sample of size  $n$  is denoted by  $ESD_{n,p}$ .

**Figure 8.11 Stem-and-leaf and box plots of finger-wrist tapping score by group, El Paso Lead Study**



**Figure 8.12 Stem-and-leaf and box plots of full-scale IQ by group, El Paso Lead Study**

**Example 8.22** Find the upper 5th percentile for the ESD statistic based on a sample of size 50.

**Solution** The appropriate percentile =  $\text{ESD}_{50,.95}$  is found by referring to the 50 row and the .95 column and is 3.13.

For values of  $n$  that are not in the table, we can sometimes assess significance by using the principle that for a given level of significance, the critical values increase as the sample size increases. This leads to the following procedure for the detection of a single outlier in normally distributed samples.

### Equation 8.24

#### ESD Single-Outlier Procedure

Suppose we have a sample  $x_1, \dots, x_n \sim N(\mu, \sigma^2)$  but feel that there may be some outliers present. To test the hypothesis,  $H_0$ : that no outliers are present vs.  $H_1$ : that a single outlier is present, with a type I error of  $\alpha$ ,

- (1) We compute the  $\text{ESD} = \max_{i=1,\dots,n} \frac{|x_i - \bar{x}|}{s}$ . The sample value  $x_i$ , such that  $\text{ESD} = \frac{|x_i - \bar{x}|}{s}$  is referred to as  $x^{(n)}$ .
- (2) We refer to Table 10 in the Appendix to obtain the critical value =  $\text{ESD}_{n,1-\alpha'}$ .
- (3) If  $\text{ESD} > \text{ESD}_{n,1-\alpha'}$ , then we reject  $H_0$  and declare that  $x^{(n)}$  is an outlier. If  $\text{ESD} \leq \text{ESD}_{n,1-\alpha'}$ , then we declare that no outliers are present.

### Example 8.23

Evaluate whether outliers are present for the finger-wrist tapping scores in the control group.

**Solution**

Following Equation 8.24, we compute the ESD test statistic. From Example 8.21, we have  $ESD = 3.44$  with 13 being the most extreme value. To assess statistical significance with  $\alpha = .05$ , we refer to Appendix Table 10. From Table 10,  $ESD_{70,.95} = 3.26$ . Because  $ESD = 3.44 > ESD_{70,.95} = 3.26 > ESD_{64,.95}$ , it follows that  $p < .05$ . Therefore, we infer that the finger-wrist tapping score of 13 taps per 10 seconds is an outlier.

In some instances, when multiple outliers are present, it is difficult to identify specific data points as outliers using the single-outlier detection procedure. This is because the standard deviation can get inflated in the presence of multiple outliers, reducing the magnitude of the ESD test statistic in Equation 8.24.

**Example 8.24**

Evaluate whether any outliers are present for the finger-wrist tapping scores in the exposed group.

**Solution**

Referring to Table 8.10, we see that  $\bar{x} = 47.4$ ,  $s = 13.2$ , and  $n = 35$  in the exposed group. Furthermore, the minimum and maximum values are 13 and 83, respectively. Because  $|83 - 47.4| = 35.6 > |13 - 47.4| = 34.4$ , it follows that the ESD statistic is  $35.6/13.2 = 2.70$ . From Appendix Table 10, we see that  $ESD_{35,.95} = 2.98 > ESD = 2.70$ . Therefore  $p > .05$ , and we accept the null hypothesis that no outliers are present.

The solution to Example 8.24 is unsettling because it is inconsistent with Figure 8.11b. It appears that the values 13, 14, and 83 are outliers, yet no outliers are identified by the single-outlier procedure in Equation 8.24. The problem is that the multiple outliers have artificially inflated the standard deviation. This is called the *masking problem*, because multiple outliers have made it difficult to identify the single most extreme sample point as an outlier. This is particularly true if the multiple outliers are roughly equidistant from the sample mean, as in Figure 8.11b. To overcome this problem, we must employ a flexible procedure that can accurately identify either single or multiple outliers and is less susceptible to the masking problem. For this purpose, we first must determine a reasonable upper bound for the number of outliers in a data set. In my experience, a reasonable upper bound for the number of possible outliers is  $\min([n/10], 5)$ , where  $[n/10]$  is the largest integer  $\leq n/10$ . If there are more than five outliers in a data set, then we most likely have an underlying nonnormal distribution, unless the sample size is very large. The following multiple-outlier procedure [3] achieves this goal.

**Equation 8.25****ESD Many-Outlier Procedure**

Suppose  $x_1, \dots, x_n \sim N(\mu, \sigma^2)$  for a large majority of the sample points, but we suspect that we may have as many as  $k$  outliers, where  $k = \min([n/10], 5)$ , where  $[n/10]$  is the largest integer  $\leq n/10$ . We wish to have a type I error of  $\alpha$  to test the hypothesis  $H_0$ : there are no outliers vs.  $H_1$ : there are between 1 and  $k$  outliers, and we would like to use a decision rule that can specifically identify the outliers. For this purpose,

1. We compute the ESD statistic based on the full sample =  $\max_{i=1,\dots,n} |x_i - \bar{x}|/s$ . We denote this statistic by  $ESD^{(n)}$  and the most outlying data point by  $x^{(n)}$ .
2. We remove  $x^{(n)}$  from the sample and compute the mean, standard deviation, and ESD statistic from the remaining  $n - 1$  data points. We denote the ESD statistic from the reduced sample by  $ESD^{(n-1)}$ .
3. We continue to remove the most outlying sample points and recompute the ESD statistic until we have computed  $k$  ESD statistics denoted by  $ESD^{(n)}, ESD^{(n-1)}, \dots, ESD^{(n-k+1)}$  based on the original sample size of  $n$ , and successively

reduced samples of size  $n - 1, \dots, n - k + 1$ . The most outlying values identified at each of the  $k$  steps are denoted by  $x^{(n)}, x^{(n-1)}, \dots, x^{(n-k+1)}$ .

4. The critical values corresponding to the ESD statistics are  $\text{ESD}_{n,1-\alpha}, \text{ESD}_{n-1,1-\alpha}, \dots, \text{ESD}_{n-k+1,1-\alpha}$ .
5. We then use the following decision rule for outlier detection:  
 If  $\text{ESD}^{(n-k+1)} > \text{ESD}_{n-k+1,1-\alpha}$ , then we declare the  $k$  values  $x^{(n)}, \dots, x^{(n-k+1)}$  as outliers  
 else If  $\text{ESD}^{(n-k+2)} > \text{ESD}_{n-k+2,1-\alpha}$ , then we declare the  $k - 1$  values  $x^{(n)}, \dots, x^{(n-k+2)}$  as outliers  
 :  
 else If  $\text{ESD}^{(n)} > \text{ESD}_{n,1-\alpha}$ , then we declare one outlier,  $x^{(n)}$   
 else If  $\text{ESD}^{(n)} \leq \text{ESD}_{n,1-\alpha}$ , then we declare no outliers present  
 Thus we can declare either 0, 1, ..., or  $k$  sample points as outliers.
6. We should use Table 10 from the Appendix to implement this procedure only if  $n \geq 20$ .

Note that we must compute all  $k$  outlier test statistics  $\text{ESD}^{(n)}, \text{ESD}^{(n-1)}, \dots, \text{ESD}^{(n-k+1)}$  regardless of whether any specific test statistic (e.g.,  $\text{ESD}^{(n)}$ ) is significant. This procedure has good power either to declare no outliers or to detect from 1 to  $k$  outliers with little susceptibility to masking effects unless the true number of outliers is larger than  $k$ .

### Example 8.25

Reanalyze the finger-wrist tapping scores for the exposed group in Figure 8.11b using the multiple-outlier procedure in Equation 8.25.

#### Solution

We will set the maximum number of outliers to be detected to be  $[35/10] = 3$ . From Example 8.24, we see that  $\text{ESD}^{(35)} = 2.70$  and the most outlying value =  $x^{(35)} = 83$ . We remove 83 from the sample and recompute the sample mean (46.4) and standard deviation (11.8) from the reduced sample of size 34. Because  $|13 - 46.4| = 33.4 > |70 - 46.4| = 23.6$ , 13 is the most extreme value and  $\text{ESD}^{(34)} = 33.4/11.8 = 2.83$ . We then remove 13 from the sample and recompute the sample mean (47.4) and standard deviation (10.4) from the reduced sample of size 33. Because  $|14 - 47.4| = 33.4 > |70 - 47.4| = 22.6$ , it follows that  $\text{ESD}^{(33)} = 33.4/10.4 = 3.22$ .

To assess statistical significance, we first compare 3.22 with the critical value  $\text{ESD}_{33,.95}$ . From Table 10 in the Appendix, we see that  $\text{ESD}^{(33)} = 3.22 > \text{ESD}_{35,.95} = 2.98 > \text{ESD}_{33,.95}$ . Therefore,  $p < .05$ , and we declare the three most extreme values (83, 13, and 14) as outliers. Note that although significance was achieved by an analysis of the third most extreme value (14), once it is identified as an outlier, then the more extreme points (13, 83) are also designated as outliers. Also, note that the results are consistent with Figure 8.11b and are different from the results of the single-outlier procedure, in which no outliers were declared.

### Example 8.26

Assess whether any outliers are present for the finger-wrist tapping scores for controls.

#### Solution

Because  $n = 64$ ,  $\min([64/10], 5) = \min(6, 5) = 5$ . Therefore, we set the maximum number of outliers to be detected to 5 and organize the appropriate test statistics and critical values in a table (Table 8.12).

**Table 8.12** Test statistics and critical values for Example 8.26

<i>n</i>	$\bar{x}$	<i>s</i>	$x^{(n)}$	$ESD^{(n)}$	$ESD_{n,95}$	<i>p</i> -value
64	54.4	12.1	13	3.44	$ESD_{64,95}^a$	<.05
63	55.1	10.9	23	2.94	$ESD_{63,95}^b$	NS
62	55.6	10.2	26	2.90	$ESD_{62,95}^b$	NS
61	56.1	9.6	84	2.92	$ESD_{61,95}^b$	NS
60	55.6	8.9	79	2.62	3.20	NS

<sup>a</sup> $ESD_{64,95} < ESD_{70,95} = 3.26$ <sup>b</sup> $ESD_{63,95}, \dots, ESD_{61,95}$  are all  $> ESD_{60,95} = 3.20$ 

From Table 8.12 we see that 79, 84, 26, and 23 are *not* identified as outliers, whereas 13 *is* identified as an outlier. Thus we declare one outlier present. This decision is consistent with the single-outlier test in Example 8.23.

In general, use the multiple-outlier test in Equation 8.25 rather than the single-outlier test in Equation 8.24 unless you are very confident there is at most one outlier.

The issue remains: What should we do now that we have identified one outlier among the controls and three outliers among the exposed? We have chosen to reanalyze the data, using a two-sample *t* test, after deleting the outlying observations.

**Example 8.27** Reanalyze the finger-wrist tapping score data in Table 8.10 after excluding the outliers identified in Examples 8.25 and 8.26.

**Solution** The *t* test results after excluding the outliers are given in Table 8.13.

**Table 8.13** Comparison of mean finger-wrist tapping scores for the exposed vs. control groups after excluding outliers, using the SAS *t* test procedure

The TTEST Procedure																
Statistics																
Variable	group	Lower CL			Upper CL			Lower CL			Upper CL					
		N	Mean	Mean	Mean	Std Dev	Std Dev	Std Dev	Std Err	Minimum	Maximum					
maxfwt	1	63	52.341	55.095	57.849	9.3033	10.935	13.266	1.3777	23	84					
maxfwt	2	32	45.343	48.438	51.532	6.8813	8.5833	11.411	1.5173	34	70					
maxfwt	Diff		2.2559	6.6577	11.06	8.9312	10.211	11.923	2.2167							
	(1-2)															
T-Tests																
Variable	Method	Variances			DF	t Value			Pr >  t							
maxfwt	Pooled	Equal			93	3.00			0.0034							
maxfwt	Satterthwaite	Unequal			77	3.25			0.0017							
Equality of Variances																
Variable	Method	Num DF	Den DF	F Value			Pr > F									
maxfwt	Folded F	62	31	1.62			0.1424									

We see that a significant difference remains between the mean finger-wrist tapping scores for the exposed and control groups ( $p = .003$ ). Indeed, the results are more significant than previously because the standard deviations are lower after exclusion of outliers, particularly for the exposed group.

We can take several approaches to the treatment of outliers in performing data analyses. One approach is to use efficient methods of outlier detection and either exclude outliers from further data analyses or perform data analyses with and without outliers present and compare results. Another possibility is not to exclude the outliers but to use a method of analysis that minimizes their effect on the overall results. One method for accomplishing this is to convert continuous variables such as finger-wrist tapping score to categorical variables (for example, high = above the median vs. low = below the median) and analyze the data using categorical-data methods. We discuss this approach in Chapter 10. Another possibility is to use nonparametric methods to analyze the data. These methods make much weaker assumptions about the underlying distributions than do the normal-theory methods such as the  $t$  test. We discuss this approach in Chapter 9. Another approach is to use “robust” estimators of important population parameters (such as  $\mu$ ). These estimators give less weight to extreme values in the sample but do not entirely exclude them. The subject of robust estimation is beyond the scope of this book. Using each of these methods may result in a loss of power relative to using ordinary  $t$  tests if no outliers exist but offer the advantage of a gain in power if some outliers are present. In general, there is no one correct way to analyze data; the conclusions from a study are strengthened if they are consistently found by using more than one analytic technique.

Software to implement the ESD Many-Outlier Procedure in Equation 8.25 in SAS is available at <http://www.biostat.harvard.edu/~carey/outlier.html>.

## 8.10 Estimation of Sample Size and Power for Comparing Two Means

### Estimation of Sample Size

Methods of sample-size estimation for the one-sample  $z$  test for the mean of a normal distribution with known variance were presented in Section 7.6. This section covers estimates of sample size that are useful in planning studies in which *two* samples are to be compared.

#### Example 8.28

**Hypertension** Consider the blood-pressure data for OC and non-OC users in Example 8.9 (p. 276) as a pilot study conducted to obtain parameter estimates to plan for a larger study. Suppose we assume the true blood-pressure distribution of 35- to 39-year-old OC users is normal with mean  $\mu_1$  and variance  $\sigma_1^2$ . Similarly, for non-OC users we assume the distribution is normal with mean  $\mu_2$  and variance  $\sigma_2^2$ . We wish to test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ . How can we estimate the sample size needed for the larger study?

Suppose we assume  $\sigma_1^2$  and  $\sigma_2^2$  are known and we anticipate equal sample sizes in the two groups. To conduct a two-sided test with significance level  $\alpha$  and power of  $1 - \beta$ , the appropriate sample size for *each* group is as follows:

**Equation 8.26****Sample Size Needed for Comparing the Means of Two Normally Distributed Samples of Equal Size Using a Two-Sided Test with Significance Level  $\alpha$  and Power  $1 - \beta$** 

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} = \text{sample size for each group}$$

where  $\Delta = |\mu_2 - \mu_1|$ . The means and variances of the two respective groups are  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ .

In words,  $n$  is the appropriate sample size in each group to have a probability of  $1 - \beta$  of finding a significant difference based on a two-sided test with significance level  $\alpha$ , if the absolute value of the true difference in means between the two groups is  $\Delta = |\mu_2 - \mu_1|$ , and a two-sided type I error of  $\alpha$  is used.

**Example 8.29**

**Hypertension** Determine the appropriate sample size for the study proposed in Example 8.28 using a two-sided test with a significance level of .05 and a power of .80.

**Solution**

In the small study,  $\bar{x}_1 = 132.86$ ,  $s_1 = 15.34$ ,  $\bar{x}_2 = 127.44$ , and  $s_2 = 18.23$ .

If the sample data  $(\bar{x}_1, s_1^2, \bar{x}_2, s_2^2)$  are used as estimates of the population parameters  $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ , then ensuring an 80% chance of finding a significant difference using a two-sided significance test with  $\alpha = .05$  would require a sample size of

$$n = (15.34^2 + 18.23^2)(1.96 + 0.84)^2 / (132.86 - 127.44)^2 = 151.5$$

or 152 people in *each* group. Not surprisingly, no significant difference was found in Example 8.10 with sample sizes of 8 and 21 in the two groups, respectively.

In many instances an imbalance between the groups can be anticipated and it can be predicted in advance that the number of people in one group will be  $k$  times the number in the other group for some number  $k \neq 1$ . In this case, where  $n_2 = kn_1$ , the appropriate sample size in the two groups for achieving a power of  $1 - \beta$  using a two-sided level  $\alpha$  significance test is given by the following formulas:

**Equation 8.27****Sample Size Needed for Comparing the Means of Two Normally Distributed Samples of Unequal Size Using a Two-Sided Test with Significance Level  $\alpha$  and Power  $1 - \beta$** 

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2/k)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} = \text{sample size of first group}$$

$$n_2 = \frac{(k\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} = \text{sample size of second group}$$

where  $\Delta = |\mu_2 - \mu_1|$ ;  $(\mu_1, \sigma_1^2)$ ,  $(\mu_2, \sigma_2^2)$ , are the means and variances of the two respective groups and  $k = n_2/n_1$  = the projected ratio of the two sample sizes.

Note that if  $k = 1$ , then the sample-size estimates given in Equation 8.27 are the same as those in Equation 8.26.

**Example 8.30**

**Hypertension** Suppose we anticipate twice as many non-OC users as OC users entering the study proposed in Example 8.28. Project the required sample size if a two-sided test is used with a 5% significance level and an 80% power is desired.

**Solution**

If Equation 8.27 is used with  $\mu_1 = 132.86$ ,  $\sigma_1 = 15.34$ ,  $\mu_2 = 127.44$ ,  $\sigma_2 = 18.23$ ,  $k = 2$ ,  $\alpha = .05$ , and  $1 - \beta = .8$ , then in order to achieve an 80% power in the study using a two-sided significance test with  $\alpha = .05$  we need to enroll

$$n_1 = \frac{(15.34^2 + 18.23^2 / 2)(1.96 + 0.84)^2}{(132.86 - 127.44)^2} = 107.1, \text{ or } 108 \text{ OC users}$$

and  $n_2 = 2(108) = 216$  non-OC users

If the variances in the two groups are the same, then for a given  $\alpha$ ,  $\beta$ , the smallest total sample size needed is achieved by the *equal-sample-size allocation rule* in Equation 8.26. Thus in the case of equal variances, the sample sizes in the two groups should be as nearly equal as possible.

Finally, to perform a one-sided rather than a two-sided test, we substitute  $\alpha$  for  $\alpha/2$  in Equations 8.26 and 8.27.

## Estimation of Power

In many situations, a predetermined sample size is available for study and how much power the study will have for detecting specific alternatives needs to be determined.

**Example 8.31**

**Hypertension** Suppose 100 OC users and 100 non-OC users are available for study and a true difference in mean SBP of 5 mm Hg is anticipated, with OC users having the higher mean SBP. How much power would such a study have assuming that the variance estimates in the pilot study in Example 8.9 are correct?

Assuming  $\sigma_1^2$  and  $\sigma_2^2$  are known, the power using a two-sided test with significance level  $\alpha$  is given by Equation 8.28.

**Equation 8.28**

### Power for Comparing the Means of Two Normally Distributed Samples Using a Significance Level $\alpha$

To test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$  for the specific alternative  $|\mu_1 - \mu_2| = \Delta$ , with significance level  $\alpha$ ,

$$\text{Power} = \Phi\left(-z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right)$$

where  $(\mu_1, \sigma_1^2)$ ,  $(\mu_2, \sigma_2^2)$  are the means and variances of the two respective groups and  $n_1, n_2$  are the sample sizes of the two groups.

**Example 8.32**

**Hypertension** Estimate the power available for the study proposed in Example 8.31 using a two-sided test with significance level = .05.

**Solution**

From Example 8.31,  $n_1 = n_2 = 100$ ,  $\Delta = 5$ ,  $\sigma_1 = 15.34$ ,  $\sigma_2 = 18.23$ , and  $\alpha = .05$ . Therefore, from Equation 8.28,

$$\begin{aligned}\text{Power} &= \Phi\left(-z_{.975} + \frac{5}{\sqrt{15.34^2/100 + 18.23^2/100}}\right) = \Phi\left(-1.96 + \frac{5}{2.383}\right) \\ &= \Phi(-1.96 + 2.099) = \Phi(0.139) = .555\end{aligned}$$

Thus there is a 55.5% chance of detecting a significant difference using a two-sided test with significance level = .05.

To calculate power for a one-sided rather than a two-sided test, simply substitute  $\alpha$  for  $\alpha/2$  in Equation 8.28.

## 8.11 Sample-Size Estimation for Longitudinal Studies

Longitudinal studies often involve comparing mean change scores between two groups. If there have been previous longitudinal studies, then the standard deviation of change scores can be obtained from these studies and the methods of power and sample size estimation in Section 8.10 can be used. However, often there are no previous longitudinal studies. Instead, there may be small reproducibility studies in which the correlation between repeated measures on the same subject over time is known.

### Example 8.33

**Hypertension** Suppose we are planning a longitudinal study to compare the mean change in SBP between a treated and a control group. It is projected, based on previous data, that the standard deviation of SBP at both baseline and follow-up is 15 mm Hg and that the correlation coefficient between repeated SBP values 1 year apart is approximately .70. How many participants do we need to study to have 80% power to detect a significant difference between groups using a two-sided test with  $\alpha = .05$  if the true mean decline in SBP over 1 year is 8 mm Hg for the treated group and 3 mm Hg for the control group?

To answer the question posed in Example 8.33, we would like to apply the sample-size formula given in Equation 8.26. However, using Equation 8.26 requires knowledge of the variances of change in SBP for each of the treated and control groups. Considering the control group first, let

$x_{1i}$  = SBP for the  $i$ th subject in the control group at baseline

$x_{2i}$  = SBP for the  $i$ th subject in the control group at 1 year

Therefore,

$d_i = x_{2i} - x_{1i}$  = change in SBP for the  $i$ th subject in the control group over 1 year

If  $x_{1i}$  and  $x_{2i}$  were independent, then from Equation 5.9 it would follow that

$Var(d_i) = \sigma_2^2 + \sigma_1^2$ , where

$\sigma_1^2$  = variance of baseline SBP in the control group

$\sigma_2^2$  = variance of 1-year SBP in the control group

However, repeated SBP measures on the same person are usually not independent. The correlation between them will, in general, depend on the time interval between the baseline and follow-up measures. Let us assume that the correlation coefficient

between measures 1 year apart is  $\rho$ . We have defined a correlation coefficient in Chapter 5, and in Chapter 11 we will discuss how to estimate correlation coefficients from sample data. Then from Equation 5.11, we have

**Equation 8.29**

$$\text{Var}(x_{2i} - x_{1i}) = \sigma_2^2 + \sigma_1^2 - 2\rho\sigma_1\sigma_2 = \sigma_d^2$$

where  $\sigma_d^2$  = variance of change in SBP.

For simplicity, we will assume that  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $\rho$ , and  $\sigma_d^2$  are the same in the treated and control groups. We wish to test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ , where

$\mu_1$  = true mean change in the control group

$\mu_2$  = true mean change in the active group

Based on Equations 8.26 and 8.29, we obtain the following sample-size estimate.

**Equation 8.30****Sample Size Needed for Longitudinal Studies Comparing Mean Change in Two Normally Distributed Samples with Two Time Points**

Suppose we are planning a longitudinal study with an equal number of subjects ( $n$ ) in each of two groups. We wish to test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ , where

$\mu_1$  = underlying mean change over time  $t$  in group 1

$\mu_2$  = underlying mean change over time  $t$  in group 2

We will conduct a two-sided test at level  $\alpha$  and wish to have a power of  $1 - \beta$  of detecting a significant difference if  $|\mu_1 - \mu_2| = \delta$  under  $H_1$ . The required sample size per group is

$$n = \frac{2\sigma_d^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

where

$$\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

$\sigma_1^2$  = variance of baseline values within a treatment group

$\sigma_2^2$  = variance of follow-up values within a treatment group

$\rho$  = correlation coefficient between baseline and follow-up values within a treatment group

**Solution to Example 8.33**

We have  $\sigma_1^2 = \sigma_2^2 = 15^2 = 225$ ,  $\rho = .70$ , and  $\delta = -8 - (-3) = -5$  mm Hg. Therefore,

$$\sigma_d^2 = 225 + 225 - 2(.70)(15)(15) = 135$$

Also,  $z_{1-\alpha/2} = z_{.975} = 1.96$ ,  $z_{1-\beta} = z_{.80} = 0.84$ . Thus

$$\begin{aligned} n &= \frac{2(135)(1.96 + 0.84)^2}{(-5)^2} \\ &= \frac{2116.8}{25} = 84.7, \text{ or } 85 \text{ subjects in each group} \end{aligned}$$

Similar to Equation 8.30, we can also consider the power of a longitudinal study given  $\alpha$ ,  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $\rho$ , and a specified sample size per group ( $n$ ).

**Equation 8.31**
**Power of a Longitudinal Study Comparing Mean Change Between Two Normally Distributed Samples with Two Time Points**

To test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ , where

$\mu_1$  = underlying mean change over time  $t$  in treatment group 1

$\mu_2$  = underlying mean change over time  $t$  in treatment group 2

for the specific alternative  $|\mu_1 - \mu_2| = \delta$ , with two-sided significance level  $\alpha$ , and sample of size  $n$  in each group, the power is given by

$$\text{Power} = \Phi\left(-Z_{1-\alpha/2} + \frac{\sqrt{n} \delta}{\sigma_d \sqrt{2}}\right)$$

where

$$\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

$\sigma_1^2$  = variance of baseline values within a treatment group

$\sigma_2^2$  = variance of follow-up values within a treatment group

$\rho$  = correlation between baseline and follow-up values over time  $t$  within a treatment group

**Example 8.34**

**Hypertension** Suppose that 75 participants per group are recruited for the study described in Example 8.33. How much power will the study have under the same assumptions as in Example 8.33?

**Solution**

We have  $n = 75$ ,  $\alpha = .05$ ,  $\delta = -5$  mm Hg,  $\sigma_d^2 = 135$  (from the solution to Example 8.33). Thus

$$\begin{aligned}\text{Power} &= \Phi\left(-Z_{.975} + \frac{\sqrt{75}(5)}{\sqrt{135(2)}}\right) \\ &= \Phi\left(-1.96 + \frac{43.30}{16.43}\right) \\ &= \Phi(0.675) = .750\end{aligned}$$

Thus the study will have 75% power to detect this difference.

Note that based on Equations 8.30 and 8.31, as the correlation coefficient between repeated measures decreases, the variance of change scores ( $\sigma_d^2$ ) increases, resulting in an increase in the required sample size for a given level of power (Equation 8.30) and a decrease in power for a given sample size (Equation 8.31). Thus measures that are less reproducible over time require a larger sample size for hypothesis-testing purposes.

Also, as the length of follow-up ( $t$ ) increases, the correlation between repeated measures usually decreases. Therefore, studies with longer follow-up (say, 2 years)

require a larger sample size than studies with a shorter follow-up (say, 1 year) to detect a difference of the same magnitude ( $\delta$ ). However, in some instances the expected difference between groups ( $\delta$ ) may increase as  $t$  increases. Thus the overall impact of length of follow-up on sample size is uncertain.

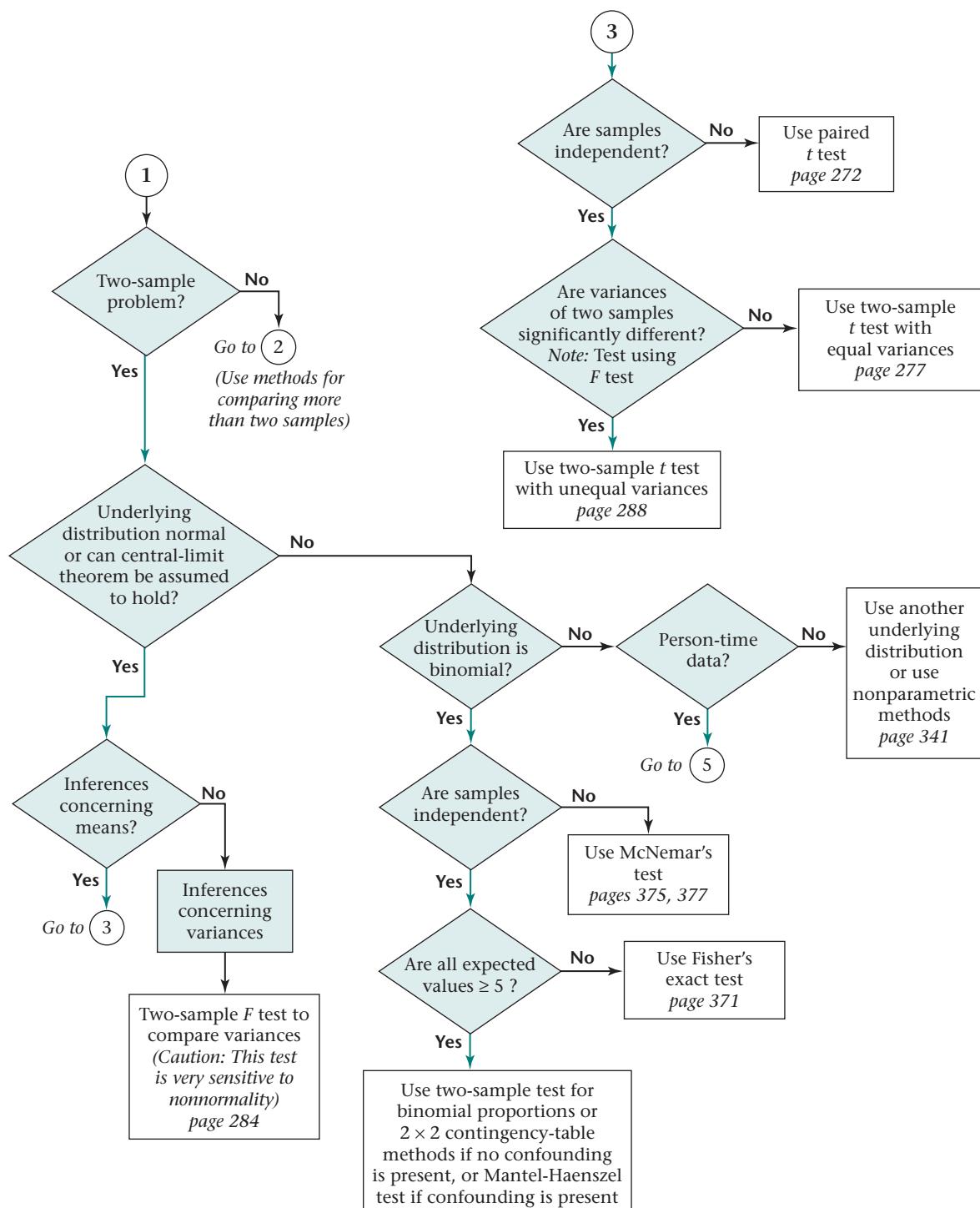
Finally, if data already exist on change scores over the time period  $t$  (either from a pilot study or from the literature), then the variance of change ( $\sigma_d^2$ ) can be estimated directly from the sample variance of change (in the pilot study or the literature), and it is unnecessary to use Equation 8.29 to compute  $\sigma_d^2$ . However, a common mistake is to run a pilot study based on repeated measures a short time apart (e.g., 1 week) and base the estimate of  $\sigma_d^2$  on difference scores from this pilot study, even when the length of the main investigation is much longer (e.g., 1 year). This usually results in an underestimate of  $\sigma_d^2$  (or, correspondingly, an overestimate of  $\rho$ ) and results in an underestimate (sometimes sizable) of required sample size for a given level of power or an overestimate of power for a given sample size [5].

The methods described in this section are for studies with a single follow-up visit. For studies with more follow-up visits, more complicated methods of sample size and power estimation are needed [5].

## 8.12 Summary

In this chapter, we studied methods of hypothesis testing for comparing the means and variances of two samples that are assumed to be normally distributed. The basic strategy is outlined in the flowchart in Figure 8.13, which is an extract from the larger flowchart in the back of this book (pp. 841–846). Referring to 1 in the upper left, first note that we are dealing with the case of a two-sample problem in which either the underlying distributions are normal or the central-limit theorem can be assumed to hold. If we are interested in comparing the means of the two samples, then we refer to box 3. If our two samples are paired—that is, if each person is used as his or her own control or if the samples consist of different people who are matched on a one-to-one basis—then the paired  $t$  test is appropriate. If the samples are independent, then the  $F$  test for the equality of two variances is used to decide whether the variances are significantly different. If the variances are not significantly different, then the two-sample  $t$  test with equal variances is used; if the variances are significantly different, then the two-sample  $t$  test with unequal variances is used. If we are only comparing the variances of the two samples, then only the  $F$  test for comparing variances is used, as indicated in the lower left of Figure 8.13.

The chapter concluded by providing methods for the detection of outliers and presenting the appropriate sample size and power formulas for planning investigations in which the goal is to compare the means from two independent samples. We considered sample size and power formulas for both cross-sectional and longitudinal studies. In Chapter 9, we extend our work on the comparison of two samples to the case in which there are two groups to be compared but the assumption of normality is questionable. We will introduce nonparametric methods to solve this problem to complement the parametric methods discussed in Chapters 7 and 8.

**Figure 8.13** Flowchart summarizing two-sample statistical inference—normal-theory methods

## PROBLEMS

**8.1** Find the lower 2.5th percentile of an  $F$  distribution with 14 and 7  $df$ . What symbol is used to denote this?

### Nutrition

The mean  $\pm 1\text{ }sd$  of  $\ln$  [calcium intake (mg)] among 25 females, 12 to 14 years of age, below the poverty level is  $6.56 \pm 0.64$ . Similarly, the mean  $\pm 1\text{ }sd$  of  $\ln$  [calcium intake (mg)] among 40 females, 12 to 14 years of age, above the poverty level is  $6.80 \pm 0.76$ .

**8.2** Test for a significant difference between the variances of the two groups.

**8.3** What is the appropriate procedure to test for a significant difference in means between the two groups?

**8.4** Implement the procedure in Problem 8.3 using the critical-value method.

**8.5** What is the  $p$ -value corresponding to your answer to Problem 8.4?

**8.6** Compute a 95% CI for the difference in means between the two groups.

Refer to the data in Table 2.11.

**8.7** Test for a significant difference in the variances of the initial white blood cell count between patients who did and patients who did not receive a bacterial culture.

**8.8** What is the appropriate test procedure to test for significant differences in mean white blood cell count between people who do and people who do not receive a bacterial culture?

**8.9** Perform the procedure in Problem 8.8 using the critical-value method.

**8.10** What is the  $p$ -value corresponding to your answer to Problem 8.9?

**8.11** Compute a 95% CI for the true difference in mean white blood cell count between the two groups.

Refer to Problem 8.2.

**\*8.12** Suppose an equal number of 12- to 14-year-old girls below and above the poverty level are recruited to study differences in calcium intake. How many girls should be recruited to have an 80% chance of detecting a significant difference using a two-sided test with  $\alpha = .05$ ?

**\*8.13** Answer Problem 8.12 if a one-sided rather than a two-sided test is used.

**\*8.14** Using a two-sided test with  $\alpha = .05$ , answer Problem 8.12, anticipating that two girls above the poverty level will be recruited for every one girl below the poverty level who is recruited.

**\*8.15** Suppose 50 girls above the poverty level and 50 girls below the poverty level are recruited for the study. How much power will the study have of finding a significant difference using a two-sided test with  $\alpha = .05$ , assuming

that the population parameters are the same as the sample estimates in Problem 8.2?

**\*8.16** Answer Problem 8.15 if a one-sided rather than a two-sided test is used.

**\*8.17** Suppose 50 girls above the poverty level and 25 girls below the poverty level are recruited for the study. How much power will the study have if a two-sided test is used with  $\alpha = .05$ ?

**\*8.18** Answer Problem 8.17 if a one-sided test is used with  $\alpha = .05$ .

### Ophthalmology

The drug diflunisal is used to treat mild to moderate pain, osteoarthritis (OA), and rheumatoid arthritis (RA). The ocular effects of diflunisal had not been considered until a study was conducted on its effect on intraocular pressure in glaucoma patients who were already receiving maximum therapy for glaucoma [6].

**\*8.19** Suppose the change (mean  $\pm sd$ ) in intraocular pressure after administration of diflunisal (follow-up – baseline) among 10 patients whose standard therapy was methazolamide and topical glaucoma medications was  $-1.6 \pm 1.5$  mm Hg. Assess the statistical significance of the results.

**\*8.20** The change in intraocular pressure after administration of diflunisal among 30 patients whose standard therapy was topical drugs only was  $-0.7 \pm 2.1$  mm Hg. Assess the statistical significance of these results.

**\*8.21** Compute 95% CIs for the mean change in pressure in each of the two groups identified in Problems 8.19 and 8.20.

**\*8.22** Compare the mean change in intraocular pressure in the two groups identified in Problems 8.19 and 8.20 using hypothesis-testing methods.

### Cardiovascular Disease, Pediatrics

A study in Pittsburgh measured various cardiovascular risk factors in children at birth and during their first 5 years of life [7]. In particular, heart rate was assessed at birth, 5 months, 15 months, 24 months, and annually thereafter until 5 years of age. Heart rate was related to age, sex, race, and socioeconomic status. The data in Table 8.14 were presented relating heart rate to race among newborns.

**Table 8.14 Relationship of heart rate to race among newborns**

Race	Mean heart rate (beats per minute)	sd	n
White	125	11	218
Black	133	12	156

Source: Reprinted with permission of the *American Journal of Epidemiology*, 119(4), 554–563.

**8.23** Test for a significant difference in mean heart rate between Caucasian and African-American newborns.

**8.24** Report a *p*-value for the test performed in Problem 8.23.

### Pharmacology

One method for assessing the bioavailability of a drug is to note its concentration in blood and/or urine samples at certain periods of time after the drug is given. Suppose we want to compare the concentrations of two types of aspirin (types A and B) in urine specimens taken from the same person 1 hour after he or she has taken the drug. Hence, a specific dosage of either type A or type B aspirin is given at one time and the 1-hour urine concentration is measured. One week later, after the first aspirin has presumably been cleared from the system, the same dosage of the other aspirin is given to the same person and the 1-hour urine concentration is noted. Because the order of giving the drugs may affect the results, a table of random numbers is used to decide which of the two types of aspirin to give first. This experiment is performed on 10 people; the results are given in Table 8.15.

**Table 8.15 Concentration of aspirin in urine samples**

Person	Aspirin A 1-hour concentration (mg%)	Aspirin B 1-hour concentration (mg%)
1	15	13
2	26	20
3	13	10
4	28	21
5	17	17
6	20	22
7	7	5
8	36	30
9	12	7
10	18	11
Mean	19.20	15.60
<i>sd</i>	8.63	7.78

Suppose we want to test the hypothesis that the mean concentrations of the two drugs are the same in urine specimens.

**\*8.25** What are the appropriate hypotheses?

**\*8.26** What are the appropriate procedures to test these hypotheses?

**\*8.27** Conduct the tests mentioned in Problem 8.26.

**\*8.28** What is the best point estimate of the mean difference in concentrations between the two drugs?

**\*8.29** What is a 95% CI for the mean difference?

**8.30** Suppose an  $\alpha$  level of .05 is used for the test in Problem 8.27. What is the relationship between the decision reached with the test procedure in Problem 8.27 and the nature of the CI in Problem 8.29?

### Pulmonary Disease

A 1980 study was conducted whose purpose was to compare the indoor air quality in offices where smoking was permitted with that in offices where smoking was not permitted [8]. Measurements were made of carbon monoxide (CO) at 1:20 p.m. in 40 work areas where smoking was permitted and in 40 work areas where smoking was not permitted. Where smoking was permitted, the mean CO level was 11.6 parts per million (ppm) and the standard deviation CO was 7.3 ppm. Where smoking was not permitted, the mean CO was 6.9 ppm and the standard deviation CO was 2.7 ppm.

**8.31** Test for whether the standard deviation of CO is significantly different in the two types of working environments.

**8.32** Test for whether or not the mean CO is significantly different in the two types of working environments.

**8.33** Provide a 95% CI for the difference in mean CO between the smoking and nonsmoking working environments.

### Ophthalmology

A camera has been developed to detect the presence of cataract more accurately. Using this camera, the gray level of each point (or pixel) in the lens of a human eye can be characterized into 256 gradations, where a gray level of 1 represents black and a gray level of 256 represents white. To test the camera, photographs were taken of 6 randomly selected normal eyes and 6 randomly selected cataractous eyes (the two groups consist of different people). The median gray level of each eye was computed over the 10,000+ pixels in the lens. The data are given in Table 8.16.

**Table 8.16 Median gray level for cataractous and normal eyes**

Patient number	Cataractous median gray level	Normal median gray level
1	161	158
2	140	182
3	136	185
4	171	145
5	106	167
6	149	177
$\bar{x}$	143.8	169.0
<i>s</i>	22.7	15.4

**8.34** What statistical procedure can be used to test whether there is a significant difference in the median gray levels between cataractous and normal eyes?

**8.35** Carry out the test procedure mentioned in Problem 8.34, and report a *p*-value.

**8.36** Provide a 99% CI for the mean difference in median gray levels between cataractous and normal eyes.

### Obstetrics

A clinical trial is conducted at the gynecology unit of a major hospital to determine the effectiveness of drug A in preventing premature birth. In the trial, 30 pregnant women are to be studied, 15 in a treatment group to receive drug A and 15 in a control group to receive a placebo. The patients are to take a fixed dose of each drug on a one-time-only basis between the 24th and 28th weeks of pregnancy. The patients are assigned to groups based on computer-generated random numbers, where for every two patients eligible for the study, one is assigned randomly to the treatment group and the other to the control group.

**8.37** Suppose you are conducting the study. What would be a reasonable way of allocating women to the treatment and control groups?

Suppose the weights of the babies are those given in Table 8.17.

**Table 8.17 Birthweights in a clinical trial to test a drug for preventing low-birthweight deliveries**

Patient number	Baby weight (lb)	
	Treatment group	Control group
1	6.9	6.4
2	7.6	6.7
3	7.3	5.4
4	7.6	8.2
5	6.8	5.3
6	7.2	6.6
7	8.0	5.8
8	5.5	5.7
9	5.8	6.2
10	7.3	7.1
11	8.2	7.0
12	6.9	6.9
13	6.8	5.6
14	5.7	4.2
15	8.6	6.8

**8.38** How would you assess the effects of drug A in light of your answer to Problem 8.37? Specifically, would you use a paired or an unpaired analysis, and of what type?

**8.39** Perform both a paired and an unpaired analysis of the data. Does the type of analysis affect the assessment of the results? Explain.

**8.40** Suppose patient 3 in the control group moves to another city before giving birth and her child's weight is unknown. Does this event affect the analyses in Problem 8.39? If so, how?

### Pulmonary Disease

A possible important environmental determinant of lung function in children is amount of cigarette smoking in the home. Suppose this question is studied by selecting two groups: Group 1 consists of 23 nonsmoking children 5–9 years of age, *both* of whose parents smoke, who have a mean forced expiratory volume (FEV) of 2.1 L and a standard deviation of 0.7 L; group 2 consists of 20 nonsmoking children of comparable age, *neither* of whose parents smoke, who have a mean FEV of 2.3 L and a standard deviation of 0.4 L.

**\*8.41** What are the appropriate null and alternative hypotheses in this situation?

**\*8.42** What is the appropriate test procedure for the hypotheses in Problem 8.41?

**\*8.43** Carry out the test in Problem 8.42 using the critical-value method.

**\*8.44** Provide a 95% CI for the true mean difference in FEV between 5- to 9-year-old children whose parents smoke and comparable children whose parents do not smoke.

**\*8.45** If this is regarded as a pilot study, how many children are needed in each group (assuming equal numbers in each group) to have a 95% chance of detecting a significant difference using a two-sided test with  $\alpha = .05$ ?

**\*8.46** Answer the question in Problem 8.45 if the investigators use a one-sided rather than a two-sided test.

Suppose 40 children, *both* of whose parents smoke, and 50 children, *neither* of whose parents smoke, are recruited for the study.

**\*8.47** How much power would such a study have using a two-sided test with significance level = .05, assuming that the estimates of the population parameters in the pilot study are correct?

**\*8.48** Answer Problem 8.47 if a one-sided rather than a two-sided test is used.

### Infectious Disease

The degree of clinical agreement among physicians on the presence or absence of generalized lymphadenopathy was assessed in 32 randomly selected participants from a prospective study of male sexual contacts of men with acquired immunodeficiency syndrome (AIDS) or an

AIDS-related condition (ARC) [9]. The total number of palpable lymph nodes was assessed by each of three physicians. Results from two of the three physicians are presented in Table 8.18.

**8.49** What is the appropriate test procedure to determine whether there is a systematic difference between the assessments of Doctor A vs. Doctor B?

**Table 8.18 Reproducibility of assessment of number of palpable lymph nodes among sexual contacts of AIDS or ARC patients**

Patient	Number of palpable lymph nodes		
	Doctor A	Doctor B	Difference
1	4	1	3
2	17	9	8
3	3	2	1
4	11	13	-2
5	12	9	3
6	5	2	3
7	5	6	-1
8	6	3	3
9	3	0	3
10	5	0	5
11	9	6	3
12	1	1	0
13	5	4	1
14	8	4	4
15	7	7	0
16	8	6	2
17	4	1	3
18	12	9	3
19	10	7	3
20	9	11	-2
21	5	0	5
22	3	0	3
23	12	12	0
24	5	1	4
25	13	9	4
26	12	6	6
27	6	9	-3
28	19	9	10
29	8	4	4
30	15	9	6
31	6	1	5
32	5	4	1
Mean	7.91	5.16	2.75
sd	4.35	3.93	2.83
n	32	32	32

**8.50** Should a one-sided or a two-sided test be performed? Why?

**8.51** Perform the test in Problem 8.49, and report a *p*-value.

**8.52** Compute a 95% CI for the true mean difference between observers. How does it relate to your answer to Problem 8.51?

**8.53** Suppose the results of Problem 8.51 show no significant difference. Does this mean this type of assessment is highly reproducible? Why or why not?

### Renal Disease

Ten patients with advanced diabetic nephropathy (kidney complications of diabetes) were treated with captopril over an 8-week period [10]. Urinary protein was measured before and after drug therapy, with results listed in Table 8.19 in both the raw and ln scale.

**Table 8.19 Changes in urinary protein after treatment with captopril**

Patient	Raw scale urinary protein (g/24 hr)		In Scale urinary protein ln (g/24 hr)	
	Before	After	Before	After
1	25.6	10.1	3.24	2.31
2	17.0	5.7	2.83	1.74
3	16.0	5.6	2.77	1.72
4	10.4	3.4	2.34	1.22
5	8.2	6.5	2.10	1.87
6	7.9	0.7	2.07	-0.36
7	5.8	6.1	1.76	1.81
8	5.4	4.7	1.69	1.55
9	5.1	2.0	1.63	0.69
10	4.7	2.9	1.55	1.06

**\*8.54** What is the appropriate statistical procedure to test whether mean urinary protein has changed over the 8-week period?

**\*8.55** Perform the test in Problem 8.54 using both the raw and ln scale, and report a *p*-value. Are there any advantages to using the raw or the ln scale?

**\*8.56** What is your best estimate of the percent change in urinary protein based on the data in Table 8.19?

**\*8.57** Provide a 95% CI associated with your estimate in Problem 8.56.

### Nutrition

An important hypothesis in hypertension research is that sodium restriction may lower blood pressure. However, it is difficult to achieve sodium restriction over the long term,

and dietary counseling in a group setting is sometimes used to achieve this goal. The data on urinary sodium in Table 8.20 were obtained on 8 individuals enrolled in a sodium-restricted group. Data were collected at baseline and after 1 week of dietary counseling.

**Table 8.20 Overnight sodium excretion (mEq/8hr) before and after dietary counseling**

Person	Week 0 (baseline)	Week 1	Difference
1	7.85	9.59	-1.74
2	12.03	34.50	-22.47
3	21.84	4.55	17.29
4	13.94	20.78	-6.84
5	16.68	11.69	4.99
6	41.78	32.51	9.27
7	14.97	5.46	9.51
8	12.07	12.95	-0.88
$\bar{x}$	17.65	16.50	1.14
s	10.56	11.63	12.22

**8.58** What are appropriate hypotheses to test whether dietary counseling is effective in reducing sodium intake over a 1-week period (as measured by overnight urinary sodium excretion)?

**8.59** Conduct the test mentioned in Problem 8.58, and report a p-value.

**8.60** Provide a 95% CI for the true mean change in overnight sodium excretion over a 1-week period.

**8.61** How many participants would be needed to have a 90% chance of detecting a significant change in mean urinary sodium excretion if a one-sided test is used with  $\alpha = .05$  and the estimates from the data in Table 8.20 are used as the true population parameters?

Refer to Data Set NIFED.DAT on the Companion Website. See p. 140 for a complete description of the data set.

**8.62** Assess whether there is any difference between the nifedipine and propranolol groups regarding their effects on blood pressure and heart rate. Refer to the indices of change defined in Problems 6.70–6.74.

## Genetics

A study was conducted of genetic and environmental influences on cholesterol levels. The data set used for the study were obtained from a twin registry in Sweden [11]. Specifically, four populations of adult twins were studied: (1) monozygotic (MZ) twins reared apart, (2) MZ twins reared together, (3) dizygotic (DZ) twins reared apart, and (4) DZ twins reared together. One issue is whether it is necessary to correct for sex before performing more complex genetic analyses. The data in Table 8.21 were presented for total cholesterol levels for MZ twins reared apart, by sex.

**Table 8.21 Comparison of mean total cholesterol for MZ twins reared apart, by sex**

	Men	Women
Mean	253.3	271.0
sd	44.1	44.1
n	44	48

Note: n = number of people (e.g., for males, 22 pairs of twins = 44 people)

**\*8.63** If we assume (a) serum cholesterol is normally distributed, (b) the samples are independent, and (c) the standard deviations for men and women are the same, then what is the name of the statistical procedure that can be used to compare the two groups?

**\*8.64** Suppose we want to use the procedure in Problem 8.63 using a two-sided test. State the hypotheses being tested, and implement the method. Report a p-value.

**\*8.65** Suppose we want to use the procedure in Problem 8.63 using a one-sided test in which the alternative hypothesis is that men have higher cholesterol levels than women. State the hypotheses being tested and implement the method in Problem 8.63. Report a p-value.

**\*8.66** Are the assumptions in Problem 8.63 likely to hold for these samples? Why or why not?

## Pulmonary Disease

A study was performed looking at the effect of mean ozone exposure on change in pulmonary function. Fifty hikers were recruited into the study; 25 study participants hiked on days with low-ozone exposure, and 25 hiked on days with high-ozone exposure. The change in pulmonary function after a 4-hour hike was recorded for each participant. The results are given in Table 8.22.

**Table 8.22 Comparison of change in FEV on high-ozone vs. low-ozone days**

Ozone level	Mean change in FEV <sup>a</sup>	sd	n
High	0.101	0.253	25
Low	0.042	0.106	25

<sup>a</sup>Change in FEV, forced expiratory volume, in 1 second (L) (baseline – follow-up)

**8.67** What test can be used to determine whether the mean change in FEV differs between the high-ozone and low-ozone days?

**8.68** Implement the test in Problem 8.67, and report a p-value (two-tailed).

**8.69** Suppose we determine a 95% CI for the true mean change in pulmonary function on high-ozone days. Is this CI

narrower, wider, or the same width as a 90% CI? (Do not actually compute the CI.)

### Rheumatology

A study was conducted [12] comparing muscle function between patients with rheumatoid arthritis (RA) and osteoarthritis (OA). A 10-point scale was used to assess balance and coordination in which a high score indicates better coordination. The results were as shown in Table 8.23 for 36 RA patients and 30 OA patients.

**Table 8.23 Comparison of balance scores for patients with RA vs. OA**

	Mean balance score	sd	n
RA	3.4	3.0	36
OA	2.5	2.8	30

\*8.70 What test can be used to determine whether the mean balance score is the same for RA and OA patients? What are some assumptions of this test?

\*8.71 Perform the test mentioned in Problem 8.70, and report a *p*-value.

\*8.72 What is your best estimate of the proportion of RA and OA patients with impaired balance, where impaired balance is defined as a balance score  $\leq 2$  and normality is assumed?

\*8.73 Suppose a larger study is planned. How many participants are needed to detect a difference of 1 unit in mean balance score with 80% power if the number of participants in each group is intended to be the same and a two-sided test is used with  $\alpha = .05$ ?

### Cardiology

A clinical trial compared percutaneous transluminal coronary angioplasty (PTCA) with medical therapy in treating single-vessel coronary-artery disease [13]. Researchers randomly assigned 107 patients to medical therapy and 105 to PTCA. Patients were given exercise tests at baseline and after 6 months of follow-up. Exercise tests were performed up to maximal effort until clinical signs (such as angina) were present. The results shown in Table 8.24

**Table 8.24 Change in total duration of exercise for patients with coronary-artery disease randomized to medical therapy vs. PTCA**

	Mean change (min)	sd	n
Medical therapy	0.5	2.2	100
PTCA	2.1	3.1	99

were obtained for change in total duration of exercise (min) (6 months – baseline).

\*8.74 What test can be performed to test for change in mean total duration of exercise for a *specific* treatment group?

\*8.75 Perform the test in Problem 8.74 for the medical therapy group, and report a *p*-value.

\*8.76 What test can be performed to compare the mean change in duration of exercise *between* the two treatment groups?

\*8.77 Perform the test mentioned in Problem 8.76, and report a *p*-value.

### Hypertension

A case-control study was performed among participants in the Kaiser Permanente Health Plan to compare body-fat distribution among “hypertensive” people who were normotensive at entry into the plan and became hypertensive over time as compared with “normotensives” who remained normotensive throughout their participation in the plan. A match was found for each “hypertensive” person of the same sex, race, year of birth, and year of entry into the plan; 609 matched pairs were created. The data in Table 8.25 represent body mass index (BMI) at baseline in the two groups [14].

8.78 What are the advantages of using a matched design for this study?

8.79 What test procedure can be used to test for differences in mean BMI between the two groups?

8.80 Implement the test procedure in Problem 8.79, and report a *p*-value.

8.81 Compute a 90% CI for the mean difference in BMI between the two groups.

### Hepatic Disease

An experiment was conducted to examine the influence of avian pancreatic polypeptide (aPP), cholecystokinin (CCK), vasoactive intestinal peptide (VIP), and secretin on pancreatic and biliary secretions in laying hens. In particular, researchers were concerned with the extent to which these hormones increase or decrease biliary and pancreatic flows and their pH values.

White leghorn hens, 14–29 weeks of age, were surgically fitted with cannulas for collecting pancreatic and biliary secretions and a jugular cannula for continuous infusion of aPP, CCK, VIP, or secretin. One trial per day was conducted on a hen, as long as her implanted cannulas remained functional. Thus there were varying numbers of trials per hen.

Each trial began with infusion of physiologic saline for 20 minutes. At the end of this period, pancreatic and biliary secretions were collected and the cannulas were attached to new vials. The biliary and pancreatic flow rates (in microliters per

**Table 8.25 Comparison of BMI for people who developed hypertension vs. people who remained normotensive**

	Cases		Controls		Mean difference	Test statistic
	Mean	sd	Mean	sd		
BMI ( $\text{kg}/\text{m}^2$ )	25.39	3.75	24.10	3.42	1.29	6.66

minute) and pH values (if possible) were measured. Infusion of a hormone was then begun and continued for 40 minutes. Measurements were then repeated.

Data Set HORMONE.DAT (on the Companion Website) contains data for the four hormones and saline, where saline indicates trials in which physiologic saline was infused in place of an active hormone during the second period. Each trial is one record in the file. There are 11 variables associated with each trial, as shown in Table 8.26.

**8.82** Assess whether there are significant changes in secretion rates or pH levels with any of the hormones or with saline.

**8.83** Compare the changes in secretion rate or pH levels for each active hormone vs. the placebo (saline) group. Use methods of hypothesis testing and/or CIs to express these comparisons statistically.

**8.84** For each active-hormone group, categorize dosage by high dose (above the median) vs. low dose (at or below the median) and assess whether there is any dose-response relationship (any differences in mean changes in secretion rates or pH between the high- and low-dose groups).

Refer to Data Set FEV.DAT, on the Companion Website.

**8.85** Compare the level of mean FEV between males and females separately in three distinct age groups (5–9, 10–14, and 15–19 years).

**8.86** Compare the level of mean FEV between smokers and nonsmokers separately for 10- to 14-year-old boys, 10- to 14-year-old girls, 15- to 19-year-old boys, and 15- to 19-year-old girls.

### Hypertension, Pediatrics

Refer to Data Set INFANTBP.DAT and INFANTBP.DOC, both on the Companion Website.

Consider again the salt-taste indices and sugar-taste indices constructed in Problems 6.56–6.57.

**8.87** Obtain a frequency distribution, and subdivide infants as high or low according to whether they are above or below the median value for the indices. Use hypothesis-testing and CI methodology to compare mean blood-pressure levels between children in the high and low groups.

**8.88** Answer Problem 8.87 in a different way by subdividing the salt- and sugar-taste indices more finely (such as quintiles or deciles). Compare mean blood-pressure level for children at the extremes (i.e., those at the highest quintile vs. the lowest quintile). Do you get the impression that the indices are related to blood-pressure level? Why or why

**Table 8.26 Format of HORMONE.DAT**

Column	Record number	Format of HORMONE.DAT	Code
1–8	1	Unique identification number for each chicken	xx.x
10–17	1	Biliary secretion rate (pre)	xx.x
19–26	1	Biliary pH (pre)	xx.x
28–35	1	Pancreatic secretion rate (pre)	xx.x
37–44	1	Pancreatic pH (pre)	x.x
46–53	1	Dosage of hormone	xx.x
55–62	1	Biliary secretion rate (post)	xx.x
64–71	1	Biliary pH (post)	x.x
1–8	2	Pancreatic secretion rate (post)	xx.x
10–17	2	Pancreatic pH (post)	x.x
19–26	2	Hormone (1 = saline; 2 = aPP, 3 = CCK; 4 = secretin; 5 = VIP)	xx.x
		Zero values for pH indicate missing values. The units for dosages are nanograms per mL of plasma for aPP, and $\mu\text{g}$ per kg per hour for CCK, VIP, and secretin.	

not? We discuss this from a different point of view in our work on regression analysis in Chapter 11 and the analysis of variance in Chapter 12.

### Sports Medicine

Tennis elbow is a painful condition that afflicts many tennis players at some time. A number of different treatments are used for this condition, including rest, heat, and anti-inflammatory medications. A clinical trial was conducted among 87 participants, comparing the effectiveness of Motrin (generic name, ibuprofen), a widely used anti-inflammatory agent, vs. placebo. Participants received both drug and placebo, but the order of administration of the two was determined by randomization. Specifically, approximately half the participants (group A) received an initial 3-week course of Motrin, while the other participants (group B) received an initial 3-week course of placebo. After the 3-week period, participants were given a 2-week **washout period** during which they received no study medication. The purpose of the washout period was to eliminate any residual biological effect of the first-course medication. After the washout period, a second period of active drug administration began, with group A participants receiving 3 weeks of placebo, and group B participants receiving 3 weeks of Motrin. At the end of each active drug period as well as at the end of the washout period, participants were asked to rate their degree of pain compared with baseline (before the beginning of the first active drug period). The goal of the study was to compare the degree of pain while on Motrin vs. the degree of pain while on a placebo. This type of study is called a **cross-over design**, which we discuss in more detail in Chapter 13.

Degree of pain vs. baseline was measured on a 1–6 scale, with 1 being “worse than baseline” and 6 being “completely improved.” The comparison was made in four different ways: (1) during maximum activity, (2) 12 hours following maximum activity, (3) during the average day, and (4) by overall impression of drug efficacy. The data are given in Data Set TENNIS2.DAT with documentation in TENNIS2.DOC, both on the Companion Website.

**8.89** Compare degree of pain while on Motrin with degree of pain on placebo during maximal activity.

**8.90** Answer Problem 8.89 for degree of pain 12 hours following maximum activity.

**8.91** Answer Problem 8.89 for degree of pain during the average day.

**8.92** Answer Problem 8.89 for the overall impression of drug efficacy.

### Environmental Health, Pediatrics

Refer to Figure 8.12 and Table 8.11.

**8.93** Assess whether there are any outliers for full-scale IQ in the control group.

**8.94** Assess whether there are any outliers for full-scale IQ in the exposed group.

**8.95** Based on your answers to Problems 8.93 and 8.94, compare mean full-scale IQ between the exposed and the control groups, after the exclusion of outliers.

### Pulmonary Disease

**8.96** Refer to Data Set FEV.DAT on the Companion Website. Assess whether there are any outliers in FEV for the following groups: 5- to 9-year-old boys, 5- to 9-year-old girls, 10- to 14-year-old boys, 10- to 14-year-old girls, 15- to 19-year-old boys, and 15- to 19-year-old girls.

### Ophthalmology

A study compared mean electroretinogram (ERG) amplitude of patients with different genetic types of retinitis pigmentosa (RP), a genetic eye disease that often results in blindness. The results shown in Table 8.27 were obtained for  $\ln(\text{ERG amplitude})$  among patients 18–29 years of age.

**Table 8.27 Comparison of mean  $\ln(\text{ERG amplitude})$  by genetic type among patients with RP**

Genetic type	Mean $\pm$ $sd$	$n$
Dominant	$0.85 \pm 0.18$	62
Recessive	$0.38 \pm 0.21$	35
X-linked	$-0.09 \pm 0.21$	28

**8.97** What is the standard error of  $\ln(\text{ERG amplitude})$  among patients with dominant RP? How does it differ from the standard deviation in the table?

**8.98** What test can be used to compare the variance of  $\ln(\text{ERG amplitude})$  between patients with dominant vs. recessive RP?

**8.99** Implement the test in Problem 8.98, and report a  $p$ -value (two-tailed). (*Hint:  $F_{34,61,975} = 1.778$ .*)

**8.100** What test can be used to compare the mean  $\ln(\text{ERG amplitude})$  between patients with dominant vs. recessive RP?

**8.101** Implement the test in Problem 8.100, and report a two-tailed  $p$ -value.

### Hypertension

A study was performed comparing different nonpharmacologic treatments for people with high-normal diastolic blood pressure (DBP) (80–89 mm Hg). One of the modes of treatment studied was stress management. People were randomly assigned to a stress management intervention (SMI) group or a control group. Participants randomized to SMI were given instructions in a group setting concerning different techniques for stress management and met periodically over a 1-month period. Participants randomized to the

control group were advised to pursue their normal lifestyles and were told that their blood pressure would be closely monitored and that their physician would be notified of any consistent elevation. The results for the SMI group ( $n = 242$ ) at the end of the study (18 months) were as follows:

Mean (change) =  $-5.53$  mm Hg (follow-up – baseline),  
 $sd$  (change) =  $6.48$  mm Hg.

**8.102** What test can be used to assess whether mean blood pressure has changed significantly in the SMI group?

**8.103** Implement the test in Problem 8.102, and report a  $p$ -value.

The results for the control group ( $n = 320$ ) at the end of the study were as follows:

Mean (change) =  $-4.77$  mm Hg,  
 $sd$  (change) =  $6.09$  mm Hg.

**8.104** What test can be used to compare mean blood-pressure change between the two groups? (Hint: For reference,  $F_{241,319,90} = 1.166$ ,  $F_{241,319,95} = 1.218$ ,  $F_{241,319,975} = 1.265$ .)

**8.105** Implement the test in Problem 8.104, and report a two-tailed  $p$ -value.

**8.106** How much power did the study have for detecting a significant difference between groups (using a two-sided test with a 5% level of significance) if the true effect of the SMI intervention is to reduce mean DBP by 2 mm Hg more than the control group and the standard deviation of change within a group is 6 mm Hg?

## Endocrinology

A study was performed to determine the effect of introducing a low-fat diet on hormone levels of 73 postmenopausal women not using exogenous hormones [15]. The data in Table 8.28 were presented for plasma estradiol in  $\log_{10}$  (picograms/milliliter).

**Table 8.28 Change in plasma estradiol after adopting a low-fat diet**

	Estradiol $\log_{10}$ (pg/mL) <sup>a</sup>
Preintervention	0.71 (0.26)
Postintervention	0.63 (0.26)
Difference (postintervention – preintervention)	$-0.08$ (0.20)

<sup>a</sup>Values are mean and  $sd$  (in parentheses) for  $\log_{10}$  of preintervention and postintervention measurements and for their difference.

**8.107** What test can be performed to assess the effects of adopting a low-fat diet on mean plasma-estradiol levels?

**8.108** Implement the test in Problem 8.107, and report a  $p$ -value.

**8.109** Provide a 95% CI for the change in mean  $\log_{10}$  (plasma estradiol). (Hint: The 95th percentile of a  $t$  distribution with  $72 df = 1.6663$ ; the 97.5th percentile of a  $t$  distribution with  $72 df = 1.9935$ .)

**8.110** Suppose a similar study is planned among women who use exogenous hormones. How many participants need to be enrolled if the mean change in  $\log_{10}$  (plasma estradiol) is  $-0.08$ , the standard deviation of change is  $0.20$ , and we want to conduct a two-sided test with an  $\alpha$  level of  $.05$  and a power of  $.80$ ?

## Cardiology

A study was performed concerning risk factors for carotid-artery stenosis (arterial narrowing) among 464 men born in 1914 and residing in the city of Malmö, Sweden [16]. The data reported for blood-glucose level are shown in Table 8.29.

**Table 8.29 Comparison of blood-glucose level between men with and without stenosis**

	No stenosis ( $n = 356$ )		Stenosis ( $n = 108$ )	
	Mean	$sd$	Mean	$sd$
Blood glucose (mmol/L)	5.3	1.4	5.1	0.8

**8.111** What test can be performed to assess whether there is a significant difference in mean blood-glucose level between men with and without stenosis? (Hint:  $F_{355,107,95} = 1.307$ ;  $F_{355,107,975} = 1.377$ .)

**8.112** Implement the test mentioned in Problem 8.111, and report a  $p$ -value (two-tailed).

## Ophthalmology

A study is being planned to assess whether a topical anti-allergic eye drop is effective in preventing the signs and symptoms of allergic conjunctivitis. In a pilot study, at an initial visit, participants are given an allergen challenge; that is, they are subjected to a substance that provokes allergy signs (e.g., cat dander) and their redness score is noted 10 minutes after the allergen challenge (visit 1 score). At a follow-up visit, the same procedure is followed, except that participants are given an active eye drop in one eye and the placebo in the fellow eye 3 hours before the challenge; a visit 2 score is obtained 10 minutes after the challenge. The data collected are shown in Table 8.30.

**8.113** Suppose we want to estimate the number of participants needed in the main study so that there is a 90% chance of finding a significant difference between active and placebo eyes using a two-sided test with a significance level of  $.05$ . We expect the active eyes to have a mean redness score 0.5 unit less than that of the placebo eyes. How many participants are needed in the main study?

**8.114** Suppose 60 participants are enrolled in the main study. How much power would the study have to detect a 0.5-unit mean difference if a two-sided test is used with a significance level of .05?

**8.115** In a substudy, participants will be subdivided into two equal groups according to the severity of previous allergy symptoms, and the effectiveness of the eye drop (vs. placebo) will be compared between the two groups. If 60 participants are enrolled in the main study (in the two groups combined), then how much power will the substudy have if there is a true mean difference in effectiveness of 0.25 [i.e., (mean change score active eye – mean change score placebo eye, subgroup 1) – (mean change score active eye – mean change score placebo eye, subgroup 2) = 0.25] between the two groups and a two-sided test is used with a significance level of .05?

**Table 8.30 Effect of an eye drop in reducing ocular redness among participants subjected to an allergen challenge**

	Active eye	Placebo eye	Change score in active eye – change score in placebo eye
	Mean $\pm$ sd	Mean $\pm$ sd	Mean $\pm$ sd
Change in average redness score <sup>a</sup> (visit 2 – visit 1 score)	$-0.61 \pm 0.70$	$-0.04 \pm 0.68$	$-0.57 \pm 0.86$

<sup>a</sup>The redness score ranges from 0 to 4 in increments of 0.5, where 0 is no redness at all and 4 is severe redness.

### Microbiology

A study sought to demonstrate that soy beans inoculated with nitrogen-fixing bacteria yield more and grow adequately without the use of expensive environmentally deleterious synthesized fertilizers. The trial was conducted under controlled conditions with uniform amounts of soil. The initial hypothesis was that inoculated plants would outperform their uninoculated counterparts. This assumption was based on the facts that plants need nitrogen to manufacture vital proteins and amino acids and that nitrogen-fixing bacteria would make more of this substance available to plants, increasing their size and yield. There were 8 inoculated plants (I) and 8 uninoculated plants (U). The plant yield as measured by pod weight for each plant is given in Table 8.31.

**8.116** Provide a 95% CI for the mean pod weight in each group.

**8.117** Suppose there is some overlap between the 95% CIs in Problem 8.116. Does this necessarily imply there is

**Table 8.31 Pod weight (g) from inoculated (I) and uninoculated (U) plants<sup>a</sup>**

	I	U
	1.76	0.49
	1.45	0.85
	1.03	1.00
	1.53	1.54
	2.34	1.01
	1.96	0.75
	1.79	2.11
	1.21	0.92
Mean	1.634	1.084
sd	0.420	0.510
n	8	8

<sup>a</sup>The data for this problem were supplied by David Rosner.

no significant difference between the mean pod weights for the two groups? Why or why not?

**8.118** What test can be used to compare the mean pod weight between the two groups?

**8.119** Perform the test in Problem 8.118, and report a p-value (two-tailed).

**8.120** Provide a 95% CI for the difference in mean pod weight between the two groups.

### Cardiovascular Disease

A study was performed to assess whether hyperinsulinemia is an independent risk factor for ischemic heart disease [17]. A group of 91 men who developed clinical manifestations of ischemic heart disease (IHD) over a 5-year period were compared with 105 control men (matched with regard to age, obesity [BMI = wt/ht<sup>2</sup> in units of kg/m<sup>2</sup>], cigarette smoking, and alcohol intake) who did not develop IHD over the period. The primary exposure variable of interest was level of fasting insulin at baseline. The data presented are shown in Table 8.32.

**Table 8.32 Mean fasting insulin ( $\pm$  sd) for men with IHD and control men**

	Controls (n = 105)	Cases (n = 91)
Fasting insulin (pmol/L)	$78.2 \pm 28.8$	$92.1 \pm 27.5$

**8.121** What test can be performed to compare the mean level of fasting insulin between case and control patients? (Hint:  $F_{104,90,975} = 1.498$ .)

**8.122** Implement the test in Problem 8.121, and report a p-value (two-tailed).

**8.123** Provide a 95% CI for the mean difference in fasting insulin between the two groups. (Note:  $t_{194, .975} = 1.972$ .)

**8.124** Suppose a 99% CI for the mean difference were also desired. Would this interval be of the same length, longer, or shorter than the 95% CI? (Do not actually compute the CI.)

### Renal Disease

The goal of the Swiss Analgesic Study was to assess the effect of taking phenacetin-containing analgesics on kidney function and other health parameters. A group of 624 women were identified from workplaces near Basel, Switzerland, with high intake of phenacetin-containing analgesics. This constituted the "study" group. In addition, a control group of 626 women were identified, from the same workplaces and with normal N-acetyl-P-aminophenyl (NAPAP) levels, who were presumed to have low or no phenacetin intake. The urine NAPAP level was used as a marker of phenacetin intake. The study group was then subdivided into high-NAPAP and low-NAPAP subgroups according to the absolute NAPAP level. However, both subgroups had higher NAPAP levels than the control group. The women were examined at baseline during 1967–1968 and also in 1969, 1970, 1971, 1972, 1975, and 1978, during which their kidney function was evaluated by several objective laboratory tests. Data Set SWISS.DAT on the Companion Website contains longitudinal data on serum-creatinine levels (an important index of kidney function) for 100 women in each of the high-NAPAP group, low-NAPAP group, and the control group. Documentation for this data set is given in SWISS.DOC on the Companion Website.

**8.125** One hypothesis is that analgesic abusers would have different serum-creatinine profiles at baseline. Using the data from the baseline visit, can you address this question?

**8.126** A major hypothesis of the study is that women with high phenacetin intake would show a greater change in serum-creatinine level compared with women with low phenacetin intake. Can you assess this issue using the longitudinal data in the data set? (Hint: A simple approach for accomplishing this is to look at the change in serum creatinine between the baseline visit and the last follow-up visit. More complex approaches using all the available data are considered in our discussion of regression analysis in Chapter 11.)

### Health Promotion

A study is planned on the effect of a new health-education program promoting smoking cessation among heavy-smoking teenagers ( $\geq 20$  cigarettes—equal to one pack—per day). A randomized study is planned whereby 50 heavy-smoking teenagers in two schools (A and B) will receive an active intervention with group meetings run by trained psychologists according to an American Cancer Society protocol; 50 other heavy-smoking teenagers in two different schools

(C and D) will receive pamphlets from the American Cancer Society promoting smoking cessation but will receive no active intervention by psychologists. Random numbers are used to select two of the four schools to receive the active intervention and the remaining two schools to receive the control intervention. The intervention is planned to last for 1 year, after which study participants in all schools will provide self-reports of the number of cigarettes smoked, which will be confirmed by biochemical tests of urinary cotinine levels. The main outcome variable is the change in the number of cigarettes smoked per day. A participant who completely stops smoking is scored as smoking 0 cigarettes per day.

It is hypothesized that the effect of the intervention will be to reduce the mean number of cigarettes smoked by 5 cigarettes per day over 1 year for the active-intervention group. It is also hypothesized that teenagers in the control group will increase their cigarette consumption by an average of 2 cigarettes per day over 1 year. Let us assume that the distribution of the number of cigarettes smoked per day at baseline in both groups is normal, with mean = 30 cigarettes per day and standard deviation = 5 cigarettes per day. Furthermore, it is expected, based on previous intervention studies, that the standard deviation of the number of cigarettes per day will increase to 7 cigarettes per day after 1 year. Finally, past data also suggest that the correlation coefficient between number of cigarettes smoked by the same person at baseline and 1 year will be .80.

**8.127** How much power will the proposed study have if a two-sided test is used with  $\alpha = .05$ ?

**8.128** Suppose the organizers of the study are reconsidering their sample-size estimate. How many participants should be enrolled in each of the active and control intervention groups to achieve 80% power if a two-sided test is used with  $\alpha = .05$ ?

### Hypertension

A study was recently reported comparing the effects of different dietary patterns on blood pressure within an 8-week follow-up period [18]. Subjects were randomized to three groups: A, a control diet group,  $N = 154$ ; B, a fruits-and-vegetables diet group,  $N = 154$ ; C, a combination-diet group consisting of a diet rich in fruits, vegetables, and low-fat dairy products and with reduced saturated and total fat,  $N = 151$ . The results reported for systolic blood pressure (SBP) are shown in Table 8.33.

**Table 8.33 Effects of dietary pattern on change in SBP**

Mean change in fruits-and-vegetables group	-2.8 mm Hg
Minus mean change in control group (97.5% CI)	(-4.7 to -0.9)

**8.129** Suppose we want to compute a two-sided  $p$ -value for this comparison. *Without doing any further calculation*, which statement(s) must be false?

- (1)  $p = .01$  (2)  $p = .04$  (3)  $p = .07$  (4)  $p = .20$

(Note: The actual  $p$ -value may differ from all these values.)

**8.130** Suppose we assume that the standard deviation of change in blood pressure is the same in each group and is known without error. Compute the exact  $p$ -value from the information provided.

**8.131** Suppose we want to compute a two-sided 95% CI for the true mean change in the fruits-and-vegetables group minus the true mean change in the control group, which we represent by  $(c_1, c_2)$ . *Without doing any further calculations*, which of the following statement(s) must be false?

- (1) The lower confidence limit  $(c_1) = -5.0$ .  
 (2) The upper confidence limit  $(c_2) = -1.0$ .  
 (3) The width of the CI  $(c_2 - c_1) = 3.0$ .

[Note: The actual values of  $c_1$ ,  $c_2$ , or  $(c_2 - c_1)$  may differ from those given above.]

**8.132** If we make the same assumption as in Problem 8.130, then compute the 95% CI from the information provided.

## Diabetes

The Diabetes Prevention Study was a randomized study conducted in Finland in which middle-aged participants (mean age, 55 years) with impaired glucose tolerance (IGT) were enrolled [19]. Study participants, who had high-normal glucose levels, were randomized to either an intervention group or a control group. People in the intervention group were encouraged to (a) reduce weight, (b) reduce fat intake, (c) increase fiber intake, and (d) increase hours per week of exercise, and they underwent intensive individual-level counseling to reduce risk-factor levels. People in the control group received pamphlets with general information concerning diet and exercise but did not receive individual counseling. Data regarding changes in weight after 1 year are shown in Table 8.34.

**Table 8.34 Mean weight change by treatment group among people with IGT in the Diabetes Prevention Study**

	Intervention group ( $n = 256$ )	Control group ( $n = 250$ )
	Mean $\pm$ $sd$	Mean $\pm$ $sd$
Change in weight (kg) over 1 year*	$-4.2 \pm 5.1$	$-0.8 \pm 3.7$

\*Follow-up weight – baseline weight.

For the purposes of this problem, for any degrees of freedom ( $d$ )  $\geq 200$  assume that  $t_d \cong N(0, 1)$  distribution.

**8.133** What test can be used to assess mean changes in weight in the intervention group?

**8.134** Perform the test in Problem 8.133, and report a two-tailed  $p$ -value.

**8.135** What test can be used to compare mean change in weight between the intervention and control groups? (Note:  $F_{255, 249, .975} = 1.281$ .)

**8.136** Perform the test in Problem 8.135, and report a two-tailed  $p$ -value.

## Health Promotion

A study looked at the influence of retirement on the level of physical activity among people ages 45–64 in the Atherosclerosis Risk in Communities (ARIC) Study [20]. For this purpose a sport score from 1 (low physical activity) to 5 (high physical activity) and a leisure score from 1 (low physical activity) to 5 (high physical activity) were constructed. The main outcome measure was the sum of the sport and leisure scores [range from 2 (low physical activity) to 10 (high physical activity)]. These scores were ascertained at baseline (year 0) and at follow-up (year 6). A comparison was made between people who were still working at year 6 vs. those who were retired at year 6. The data in Table 8.35 were presented for African-American women.

**Table 8.35 Change in combined sport and leisure score for African-American women in the ARIC Study (year 6 score – year 0 score)**

	Mean change	95% CI	$n$
Retired at year 6	0.29	(0.17, 0.42)	295
Working at year 6	0.15	(0.05, 0.25)	841

[Hint: Assume that for  $d > 200$ , a  $t_d$  distribution is the same as an  $N(0, 1)$  distribution.]

**8.137** What are the standard deviation and standard error of the mean for the change score for retired women?

**8.138** Construct a two-sided 90% CI for the mean change score for retired women. What does it mean?

**8.139** What test can be used to assess whether the underlying mean change score differs for retired women vs. working women?

**8.140** Implement the test in Problem 8.139, and report a two-tailed  $p$ -value.

## Health Promotion

Cigarette smoking has important health consequences and is positively associated with heart and lung diseases. Less well known are the consequences of quitting smoking. One study enrolled a group of 10 nurses, ages 50–54 years,

**Table 8.36 BMI change in 50- to 54-year-old women over a 6-year period**

Never-smoking women			Heavy-smoking women ( $\geq 1$ pk/day)		
ID	BMI at baseline	BMI at 6-year follow-up	ID	BMI at baseline	BMI 6 years after quitting smoking
1	26.5	29.3	11	25.6	31.1
2	33.8	32.9	12	24.4	27.6
3	27.6	25.5	13	31.0	36.6
4	24.4	28.3	14	20.4	20.8
5	21.6	23.3	15	22.3	23.2
6	32.3	37.1	16	22.2	23.8
7	31.9	35.4	17	20.8	26.1
8	23.0	24.8	18	23.5	31.0
9	31.2	30.4	19	26.6	29.2
10	36.3	37.1	20	23.0	24.0
Mean	28.9	30.4		24.0	27.3
sd	4.9	5.1		3.1	4.7
n	10	10		10	10

who had smoked at least 1 pack per day and quit for at least 6 years. The nurses reported their weight before and 6 years after quitting smoking. A commonly used measure of obesity that takes height and weight into account is  $BMI = \frac{wt}{ht^2}$  (in units of  $kg/m^2$ ). The BMI of the 10 women before and 6 years after quitting smoking are given in the last two columns of Table 8.36.

**8.141** What test can be used to assess whether the mean BMI changed among heavy-smoking women 6 years after quitting smoking?

**8.142** Implement the test in Problem 8.141, and report a two-tailed *p*-value.

One issue is that there has been a secular change in weight in society. For this purpose, a control group of 50- to 54-year-old never-smoking women were recruited and their BMI was reported at baseline (ages 50–54) and 6 years later at a follow-up visit. The results are given in the first two columns of Table 8.36.

**8.143** What test can be used to assess whether the mean change in BMI over 6 years is different between women who quit smoking and women who have never smoked?

**8.144** Implement the test in Problem 8.143, and report a two-tailed *p*-value.

**8.145** Suppose the true mean increase in BMI among heavy-smoking women 6 years after quitting is  $3.0\text{ kg}/m^2$  with a standard deviation of  $2.5\text{ kg}/m^2$ . The comparable true mean increase in BMI among never-smoking women over 6 years is  $1.5\text{ kg}/m^2$  with a standard deviation of  $2.0\text{ kg}/m^2$ . How much power does the study in Problem 8.144 have of finding a significant difference if a two-sided test is used with a 5% significance level?

### Cardiovascular Disease

An important emerging area in cardiovascular disease research is the study of subclinical markers of atherosclerosis, similar to the clogging of coronary arteries before any overt heart disease or stroke is apparent. One widely studied marker is the carotid-artery intima–media thickness (IMT), measured in mm, which can be measured noninvasively and is indicative of atherosclerosis in the carotid artery (in the neck). A study was performed to examine the relationship between IMT and childhood SES, which is measured by the number of years of schooling for the parent. The results in Table 8.37 were reported for black males.

**Table 8.37 Mean IMT by childhood SES (years of schooling of the parent)**

	SES (years of schooling)	
	0–8 (low)	>12 (high)
Mean IMT (mm)	0.749	0.738
N	327	57
95% CI	(0.732–0.766)	(0.698–0.779)

Note:  $t_{326,975} = 1.967$ ;  $t_{56,975} = 2.003$ .

**8.146** What is the standard error of the mean for the low- and high-SES groups, respectively?

**8.147** What test can be used to compare mean carotid-artery IMT between low- and high-SES black males? (Note:  $F_{325,56,975} = 1.542$ .)

**8.148** Perform the test in Problem 8.147, and report a two-tailed  $p$ -value.

### Cancer

Age at menarche (onset of menstrual periods) is an important risk factor for breast cancer and possibly ovarian cancer. In general, women who reach menarche at an earlier age have a higher incidence of breast cancer. The long-term trend in developed countries is that age at menarche has been declining over the past 50 years. One hypothesis is that women with higher childhood SES have an earlier age at menarche.

Suppose we identify 20 girls with low childhood SES (that is, head of household is a blue-collar worker) and find a mean age at menarche of 13.4 years with a standard deviation of 1.4 years. We identify an additional 30 girls with high childhood SES (head of household is a white-collar worker or executive) and find a mean age at menarche of 12.9 years with a standard deviation of 1.5 years. Assume that the underlying variance of age at menarche for girls with low childhood SES and girls with high childhood SES is the same.

**8.149** What test can be used to compare the mean of the two groups?

**8.150** Perform the test in Problem 8.149, and report a two-tailed  $p$ -value.

**8.151** How many participants should be enrolled in each group in a future study, if (a) the true mean difference in age at menarche between girls with low- and high-childhood-SES is 0.5 years, (b) standard deviation of age at menarche is 1.5 years within each group, (c) an equal number of girls are in each group, and (d) we would like to have a 90% chance of detecting a significant difference between the girls with high- and low-childhood SES?

### Diabetes

Type I diabetes is a common disease among children. It is widely accepted that maintaining glycemic control by regularly taking insulin shots is essential to avoid the long-term consequences of diabetes, which include neurologic, vision, and kidney problems and, eventually, premature heart disease or death.

What is less clear is whether maintaining diabetes control affects growth and development in childhood. For this purpose, a group of adolescent boys ages 9–15 were examined periodically (approximately every 3 months, but with wide variation). At each exam, the degree of diabetes control was assessed by measuring glycosylated hemoglobin (HgbA1c). The higher the HgbA1c, the poorer the diabetic control is (normals typically have HgbA1c <7.0). In addition, the age, height, and weight of each child were determined at each visit. Exact visit dates are available in the data set given in DIABETES.DAT on the Companion Website. Data are available for 94 boys over 910 visits.

The main question of interest here lies in the overall relationship between glycemic control and growth (weight mainly, but you might wish to consider other measures of growth as well) for the whole population, and not in this relationship for any particular boy.

**8.152** Do boys with better glycemic control have different growth patterns in weight than boys with poorer glycemic control?

One approach for looking at this is to calculate the average HgbA1c over all visits for each boy and to categorize boys as maintaining good control if their mean HgbA1c is below the median for the 94 boys and as in poor control otherwise. The simplest measure of growth is to calculate change in weight per year = (weight at last visit – weight at first visit)/(age in years at last visit – age in years at first visit). You can then use  $t$  test methods to compare the mean rate of growth between boys who are in good control and boys who are in poor control.

**8.153** Answer Problem 8.152 for height instead of weight.

**8.154** Answer Problem 8.152 for BMI ( $BMI = \text{wt}/\text{ht}^2$  in units of  $\text{kg}/\text{m}^2$ ).

In Chapter 11, we discuss regression methods that allow for more sophisticated analyses of these data.

### Pediatrics

A study was conducted to assess the association between climate conditions in infancy and adult blood pressure and anthropometric measures (e.g., height, weight) [21]. There were 3964 British women born between 1919 and 1940 who were divided into quartiles ( $n = 991$  per quartile) according to mean summer temperature ( $^{\circ}\text{C}$ ) in the first year of life. The data in Table 8.38 were presented.

**Table 8.38 Mean adult height by mean summer temperature in the first year of life**

Group	Quartile	Range of temperature ( $^{\circ}\text{C}$ )	Mean adult height (cm)	95% CI	n
Q1	1	10.8–13.7	159.1	(158.8, 159.4)	991
Q2	2	13.8–14.6	159.0	(158.7, 159.3)	991
Q3	3	14.7–15.7	158.4	(158.1, 158.7)	991
Q4	4	15.8–18.1	158.1	(157.8, 158.4)	991

We will assume that the distribution of adult height within a quartile is normally distributed and that the sample sizes are large enough that the  $t$  distribution can be approximated by a normal distribution.

**8.155** What is the standard deviation of adult height for the women in the first quartile group (group Q1)?

**8.156** What test can be performed to compare the mean adult height between the first (Q1) and the fourth (Q4)

quartiles? (Assume that the underlying variances of adult height in Q1 and Q4 are the same.)

**8.157** Perform the test in problem 8.156 and report a  $p$ -value (two-tailed). Assume that the underlying variances in Q1 and Q4 are the same.

**8.158** Provide a 95% CI for the difference in mean adult height between women in the first and fourth quartiles.

### Cardiovascular Disease

A study of genetic factors related to coronary heart disease (CHD) was performed as a substudy within the ARIC Study [22]. The data in Table 8.39 were provided by ethnic group for high-density lipoprotein (HDL) cholesterol (mg/dL).

**Table 8.39 Mean HDL cholesterol (mg/dL) by ethnic group**

	Caucasian	African-American
Mean	51.1	55.3
sd	16.8	17.2
n	9389	3167

**8.159** What is the standard error of the mean for each group?

**8.160** Provide a 95% CI for the difference between the means. [Hint: Assume that for  $d \geq 200$ ,  $t_d \approx N(0,1)$  distribution. Also assume that the underlying variances are the same in each group.]

**8.161** Suppose a new study is conducted in which there will be  $n_1$  Caucasian subjects and  $n_2$  African-American subjects, where  $n_2 = 2n_1$ . How large should  $n_1$  and  $n_2$  be in order for the new study to have 90% power to detect a mean difference of 5 mg/dL between the two groups, where a two-sided test is conducted with  $\alpha = .05$ ? (Hint: Assume that the standard deviations in the above table for each ethnic group are the true standard deviations.)

**8.162** Suppose that 100 Caucasian and 150 African-American subjects are actually enrolled in the study. How much power would the study have to detect a 5-mg/dL mean difference using a two-sided test with  $\alpha = .05$ ?

### Diabetes

The insulin pump is a device that delivers insulin to a diabetic patient at regular intervals. It presumably regulates insulin better than standard injections. However, data to establish this point are few, especially in children.

The following study was set up to assess the effect of use of the insulin pump on HgbA1c, which is a long-term marker

of compliance with insulin protocols. In general, a normal range for HgbA1c is <7%.

Data were collected on 256 diabetic patients for 1 year before and after using the insulin pump. A subset of the data for 10 diabetic patients is given in Table 8.40.

**Table 8.40 Mean HgbA1c 1 year before and 1 year after use of the insulin pump**

ID	Mean HgbA1c 1 year before (%)	Mean HgbA1c 1 year after (%)	Before – After (%)
1	6.7	7.0	-0.3
2	7.4	7.4	0
3	9.2	8.6	0.6
4	9.6	8.1	1.5
5	7.4	6.8	0.6
6	8.1	7.0	1.1
7	10.8	8.5	2.3
8	7.1	7.7	-0.6
9	7.9	9.7	-1.8
10	10.8	7.7	3.1
Mean	8.5	7.9	0.65
sd	1.5	0.9	1.44

**8.163** What test can be used to compare the mean HgbA1c 1 year before vs. mean HgbA1c 1 year after use of the insulin pump?

**8.164** Perform the test in Problem 8.163, and report a two-tailed  $p$ -value.

**8.165** Provide a 95% CI for the mean difference in HgbA1c before minus the mean HgbA1c after use of the insulin pump.

**8.166** Suppose we wanted a 99% CI. Would this interval be longer, shorter, or the same size as the 95% CI in Problem 8.165? (Do not actually compute the interval.)

### Health Promotion

An individual has been exercising at a local gym for about 10 years. He always begins with a 10- to 15-minute session on the treadmill at a speed of 3.7 mph. During a 12-day period in 2006, he recorded his heart rate before using the treadmill and after 5 minutes of use. The data are shown in Table 8.41.

**8.167** Provide a 95% CI for the mean change in heart rate after using the treadmill for 5 minutes.

The subject also recorded his heart rate at baseline and 5 minutes after starting treadmill exercise (at a speed of 2.5 mph) in 1996. The data are shown in Table 8.42.

**Table 8.41 Change in heart rate following treadmill use in 2006**

Day	Heart rate		
	Baseline	5 min	Change
1	85	103	18
2	77	92	15
3	81	97	16
4	81	96	15
5	74.5	93	18.5
6	83.75	96	12.25
7	76.5	93	16.5
8	77.75	94	16.25
9	79.25	90	10.75
10	84.25	99	14.75
11	76.5	94	17.5
12	84.75	97	12.25
Mean	80.1	95.3	15.2
sd	3.7	3.5	2.4

**Table 8.42 Change in heart rate following treadmill use in 1996**

Day	Heart rate		Day	Heart rate	
	Baseline	5 min		Baseline	5 min
1	84	87	6	86	92
2	87	92	7	88	93
3	90	93	8	84	90
4	94	98	9	86	92
5	98	100	10	98	104
Mean	89.5	94.1			
sd	5.4	5.1			

**8.168** Implement a test to compare the baseline heart rate initially (1996) and after 10 years of exercise (2006), and report a two-sided *p*-value.

**8.169** Interpret the results of the test in Problem 8.168.

**8.170** Provide a 95% CI for the difference in mean baseline heart rate between starting an exercise program (1996) and after 10 years of regular exercise (2006).

### Cardiovascular Disease

A recent multicenter observational study examined the relationship between aspirin use in the 48 hours following coronary artery bypass graft surgery (CABG) and death occurring in the hospital [23]. Before consideration of deaths among the 5022 patients in the study, the authors examined

differences in a variety of characteristics between those who did and did not receive aspirin.

**8.171** The mean age of the 2999 patients who received aspirin within 48 hours of CABG was 63.6 years (standard deviation = 9.7 years), whereas the mean age among the 2023 patients who did not receive aspirin during this time was 64.3 years (standard deviation = 9.8 years). Perform a test to determine whether the mean age for those not receiving aspirin was significantly different from those receiving aspirin. Assume that the underlying standard deviation of age is the same in the two groups and that if  $d > 200$ , then  $t_d \approx N(0,1)$ .

**8.172** Provide a 95% CI for the mean difference in age between the two groups.

**8.173** The investigators examined cause-specific mortality in this study and found that 4 patients who received aspirin after CABG died as a result of gastrointestinal (GI) complications. Suppose it is known from external evidence that 5 deaths per thousand due to GI complications are expected following CABG. What is the probability of observing 4 or fewer GI deaths, given this rate? (*Hint:* Assume that the Poisson approximation to the binomial is valid.)

**8.174** Provide a point estimate and a 95% CI for the GI death rate among the patients receiving aspirin. Interpret the results.

### Neurology

In a 5-year follow-up of bilateral stimulation of the subthalamic nucleus among 49 patients with advanced Parkinson's disease, investigators assessed changes in symptoms as measured by the Unified Parkinson's Disease Rating Scale (range = 0 to 108, with higher values denoting more symptoms). Assume this measure follows a normal distribution. The mean score at baseline was 55.7. The standard deviation of change from baseline to 5 years was 15.3.

**8.175** How much power does this study have to detect a mean difference of 5 units in the symptom scale if a two-sided test is used with  $\alpha = .05$ ?

**8.176** What minimum sample size would be needed to detect a mean change of 5 units with 80% power if a two-sided test is used with  $\alpha = .05$ ?

**8.177** The above study had no control group. Assuming that the same standard deviation of change would occur among controls and the mean change among controls = 0, how many participants would be necessary to detect a mean difference of 5 units of change between those receiving stimulation and a control group of equal size with 80% power if a two-sided test is used with  $\alpha = .05$ ?

### Ophthalmology

The following data are from a study on Botox injections. Patients received a high-dose injection in one eye (experimental

treatment = treatment *E*) and a low-dose injection in the other eye (control treatment = treatment *C*). Patients were asked to rate the level of pain in each eye on a 1–10 scale, with higher scores indicating more pain. Which eye received which treatment was randomized. The subjects came back over several visits. Data from the last visit are given in Table 8.43.

**Table 8.43 Effect of Botox injection on eye pain**

Subject	Pain score at the last visit	
	Pain in <i>E</i> eye	Pain in <i>C</i> eye
1	1.3	8.8
2	7.3	1.3
3	0	0.8
4	0	9.5
5	3	7.8
6	0	9.0
7	3.5	5.0
8	0	2.3
9	0	2.5
10	2.0	8.0
11	0	4.5
12	3.0	4.5
13	5.0	9.0
14	0.3	7.5
15	0	0.5
16	0.8	4.3

Suppose we wish to compare the pain score in the *E* eye vs. the pain score in the *C* eye.

**8.178** What test can be used to make this comparison?

**8.179** Perform this test, and report a *p*-value (two-tailed).

Another way to compare the *E*-treated eyes vs. the *C*-treated eyes is to look at the percentage of subjects who have less pain in the *E* eye vs. the *C* eye.

**8.180** If the *E* and *C* treatments are comparable, what test can we use to compare the percentage of subjects who have less pain with the *E* eye vs. the *C* eye?

**8.181** Perform the test in Problem 8.180, and report a *p*-value (two-tailed).

## Nutrition

The EPIC-Norfolk study, a study of diet and cancer in Great Britain, was performed to assess the relationship between dietary intake of vitamin C, plasma levels of vitamin C (in blood), and other predictors. One hypothesis is that smokers might have different vitamin C intake and vitamin C plasma levels than nonsmokers. Dietary intake of vitamin C was obtained using 7-day diet records in which a subject recorded what he or she ate in real time and a computer program was used to estimate nutrient intake based on the diet record data. The data in Table 8.44 were obtained for current smokers and nonsmokers.

**Table 8.44 Association between current smoking and diet record intake of vitamin C in the EPIC-Norfolk Study<sup>a</sup>**

Group	Mean vitamin C intake (mg/day)	sd (mg/day)	<i>N</i>
Nonsmokers	92.5	50.4	306
Smokers	57.0	26.3	17

<sup>a</sup>Diet record intake includes intake from foods but not from vitamin supplements.

**8.182** What test can be used to compare the standard deviation of diet record vitamin C intake between current smokers vs. nonsmokers?

**8.183** Perform the test in Problem 8.182, and identify whether there is a significant difference between the two variances (i.e., is *p* < .05 or *p* > .05).

**8.184** What test can be performed to compare the mean diet record vitamin C intake between the two groups?

**8.185** Perform the test in Problem 8.184, and report a *p*-value (two-tailed).

**8.186** Obtain a 95% CI for the mean difference in diet record vitamin C intake between the two groups.

## REFERENCES

- [1] Satterthwaite, E. W. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110–114.
- [2] Landrigan, P. J., Whitworth, R. H., Baloh, R. W., Staehling, N. W., Barthel, W. F., & Rosenblum, B. R (1975, March 29). Neuropsychological dysfunction in children with chronic low-level lead absorption. *Lancet*, 708–715.
- [3] Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2), 165–172.
- [4] Quesenberry, C. P., & David, H. A. (1961). Some tests for outliers. *Biometrika*, 48, 379–399.
- [5] Cook, N. R., & Rosner, B. A. (1997). Sample size estimation for clinical trials with longitudinal measures: Application to studies of blood pressure. *Journal of Epidemiology and Biostatistics*, 2, 65–74.
- [6] Yablonski, M. E., Maren, T. H., Hayashi, M., Naveh, N., Potash, S. D., & Pessah, N. (1988). Enhancement of the ocular hypertensive effect of acetazolamide by diflunisal. *American Journal of Ophthalmology*, 106, 332–336.

- [7] Schachter, J., Kuller, L. H., & Perfetti, C. (1984). Heart rate during the first five years of life: Relation to ethnic group (black or white) and to parental hypertension. *American Journal of Epidemiology*, 119(4), 554–563.
- [8] White, J. R., & Froeb, H. E. (1980). Small airway dysfunction in nonsmokers chronically exposed to tobacco smoke. *New England Journal of Medicine*, 302(13), 720–723.
- [9] Coates, R. A., Fanning, M. M., Johnson, J. K., & Calzavara, L. (1988). Assessment of generalized lymphadenopathy in AIDS research: The degree of clinical agreement. *Journal of Clinical Epidemiology*, 41(3), 267–273.
- [10] Taguma, Y., Kitamoto, Y., Futaki, G., Ueda, H., Monma, H., Ishizaki, M., Takahashi, H., Sekino, H., & Sasaki, Y. (1985). Effect of captopril on heavy proteinuria in azotemic diabetics. *New England Journal of Medicine*, 313(26), 1617–1620.
- [11] Heller, D. A., DeFaire, U., Pederson, N. L., Dahlen, G., & McClearn, G. E. (1993). Genetic and environmental influences on serum lipid levels in twins. *New England Journal of Medicine*, 328(16), 1150–1156.
- [12] Ekdahl, C., Andersson, S. I., & Svensson, B. (1989). Muscle function of the lower extremities in rheumatoid arthritis and osteoarthritis. A descriptive study of patients in a primary health care district. *Journal of Clinical Epidemiology*, 42(10), 947–954.
- [13] Parisi, A. F., Folland, E. D., & Hartigan, P. (1992). A comparison of angioplasty with medical therapy in the treatment of single-vessel coronary artery disease. *New England Journal of Medicine*, 326(1), 10–16.
- [14] Selby, J. V., Friedman, G. D., & Quesenberry, C. P., Jr. (1989). Precursors of essential hypertension: The role of body fat distribution pattern. *American Journal of Epidemiology*, 129(1), 43–53.
- [15] Prentice, R., Thompson, D., Clifford, C., Gorbach, S., Goldin, B., & Byar, D. (1990). Dietary fat reduction and plasma estradiol concentration in healthy postmenopausal women. The Women's Health Trial Study Group. *Journal of the National Cancer Institute*, 82, 129–134.
- [16] Jungquist, G., Hanson, B. S., Isacsson, S. O., Janzon, L., Steen, B., & Lindell, S. E. (1991). Risk factors for carotid artery stenosis: An epidemiological study of men aged 69 years. *Journal of Clinical Epidemiology*, 44(4/5), 347–353.
- [17] Despres, J. P., Lamarche, B., Mauriege, P., Cantin, B., Dagenais, G. R., Moosani, S., & Lupien, P. J. (1996). Hyperinsulinemia as an independent risk factor for ischemic heart disease. *New England Journal of Medicine*, 334, 952–957.
- [18] Appel, L. J., Moore, T. J., Oberzanek, E., Vollmer, W. M., Svetkey, L. P., et al. (1997). A clinical trial of the effects of dietary patterns on blood pressure. *New England Journal of Medicine*, 336, 1117–1124.
- [19] Tuomilehto, J., Lindstrom, J., Eriksson, J. G., Valle, T. T., Hamalainen, H., Ilanne-Parikka, P., Keinanen-Kiukaanniemi, S., Laakso, M., Louheranta, A., Rastas, M., Salminen, V., Aunola, S., Cepaitis, Z., Moltchanov, V., Hakumaki, M., Mannelin, M., Martikkala, V., Sundvall, J., Uusitupa, M., & the Finnish Diabetes Prevention Study Group. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *New England Journal of Medicine*, 344, 1343–1350.
- [20] Evenson, K. R., Rosamond, W. D., Cai, J., Diez-Roux, A. V., & Brancati, F. L. (2002). Influence of retirement on leisure-time physical activity: The Atherosclerosis Risk in Communities Study. *American Journal of Epidemiology*, 155, 692–699.
- [21] Lawlor, D. A., Smith, G. D., Mitchell, R., & Ebrahim, S. (2006). Adult blood pressure and climate conditions in infancy: A test of the hypothesis that dehydration in infancy is associated with higher adult blood pressure. *American Journal of Epidemiology*, 163(7), 608–614.
- [22] Bare, L. A., Morrison, A. C., Rowland, C. M., Shiffman, D., Luke, M. M., Iakoubova, O. A., Kane, J. P., Malloy, M. J., Ellis, S. G., Pankow, J. S., Willerson, J. T., Devlin, J. J., & Boerwinkle, E. (2007). Five common gene variants identify elevated genetic risk for coronary heart disease. *Genetics in Medicine*, 9(10), 682–689.
- [23] Mangano, D. T., for the Multicenter Study of Perioperative Ischemia Research Group. (2002). Aspirin and mortality from coronary bypass surgery. *New England Journal of Medicine* 347(17), 1309–1317.

## 9.1 Introduction

So far in this book, we've assumed that data come from some underlying distribution, such as the normal or binomial distribution, whose general form was assumed known. Methods of estimation and hypothesis testing have been based on these assumptions. These procedures are usually called **parametric statistical methods** because the parametric form of the distribution is assumed to be known. If these assumptions about the shape of the distribution are not made, and/or if the central-limit theorem also seems inapplicable because of small sample size, then **nonparametric statistical methods**, which make fewer assumptions about the distributional shape, must be used.

Another assumption so far in this text is that it is meaningful to measure the distance between possible data values. This assumption is characteristic of cardinal data.

---

**Definition 9.1**

**Cardinal data** are on a scale where it is meaningful to measure the distance between possible data values.

---

**Example 9.1**

Body weight is a cardinal variable because a difference of 6 lb is twice as large as a difference of 3 lb.

There are actually two types of cardinal data: interval-scale data and ratio-scale data.

---

**Definition 9.2**

For cardinal data, if the zero point is arbitrary, then the data are on an **interval scale**; if the zero point is fixed, then the data are on a **ratio scale**.

---

**Example 9.2**

Body temperature is on an interval scale because the zero point is arbitrary. For example, the zero point has a different meaning for temperatures measured in Fahrenheit vs. Celsius.

**Example 9.3**

Blood pressure and body weight are on ratio scales because the zero point is well defined in both instances.

It is meaningful to measure ratios between specific data values for data on a ratio scale (for example, person A's weight is 10% higher than person B's) but not for data

on an interval scale (for example, the ratio of specific temperatures is different in degrees F from what it is in degrees C). It is meaningful to use means and standard deviations for cardinal data of either type.

Another type of data that occurs frequently in medical and biological work but does not satisfy Definition 9.1 is ordinal data.

### Definition 9.3

**Ordinal data** can be ordered but do not have specific numeric values. Thus common arithmetic *cannot* be performed on ordinal data in a meaningful way.

### Example 9.4

**Ophthalmology** Visual acuity can be measured on an ordinal scale because we know 20–20 vision is better than 20–30, which is better than 20–40, and so on. However, a numeric value cannot easily be assigned to each level of visual acuity that all ophthalmologists would agree on.

### Example 9.5

In some clinical studies the major outcome variable is the change in a patient's condition after treatment. This variable is often measured on the following 5-point scale: 1 = much improved, 2 = slightly improved, 3 = stays the same, 4 = slightly worse, 5 = much worse. This variable is ordinal because the different outcomes, 1, 2, 3, 4, and 5, are ordered in the sense that condition 1 is better than condition 2, which is better than condition 3, and so on. However, we cannot say that the difference between categories 1 and 2 (2 minus 1) is the same as the difference between categories 2 and 3 (3 minus 2), and so on. If these categories were on a cardinal scale, then the variable would have this property.

Because ordinal variables cannot be given a numeric scale that makes sense, computing means and standard deviations for such data is not meaningful. Therefore, methods of estimation and hypothesis testing based on normal distributions, as discussed in Chapters 6 through 8, cannot be used. However, we are still interested in making comparisons between groups for variables such as visual acuity and outcome of treatment, and nonparametric methods can be used for this purpose.

Another type of data scale, which has even less structure than an ordinal scale concerning relationships between data values, is a nominal scale.

### Definition 9.4

Data are on a **nominal scale** if different data values can be classified into categories but the categories have no specific ordering.

### Example 9.6

**Renal Disease** In classifying cause of death among patients with documented analgesic abuse, the following categories were used: (1) cardiovascular disease, (2) cancer, (3) renal or urogenital disease, and (4) all other causes of death. Cause of death is a good example of a nominal scale because the values (the categories of death) have no specific order with respect to each other.

In this chapter the most commonly used nonparametric statistical tests are developed, assuming the data are on either a cardinal or an ordinal scale. If they are on a cardinal scale, then the methods are most useful if there is reason to question the normality of the underlying sampling distribution of the test statistic (for example, small sample size). For nominal (or categorical) data, discrete data methods, described in Chapter 10, are used.

## 9.2 The Sign Test

As discussed in Section 9.1, for ordinal data we can measure the relative ordering of different categories of a variable. In this section, we consider data with even more restrictive assumptions; namely, for any two people A, B we can identify whether the score for person A is greater than, less than, or equal to the score for person B, but not the relative magnitude of the differences.

### Example 9.7

**Dermatology** Suppose we want to compare the effectiveness of two ointments (A, B) in reducing excessive redness in people who cannot otherwise be exposed to sunlight. Ointment A is randomly applied to either the left or right arm, and ointment B is applied to the corresponding area on the other arm. The person is then exposed to 1 hour of sunlight, and the two arms are compared for degrees of redness. Suppose only the following qualitative assessments can be made:

- (1) Arm A is not as red as arm B.
- (2) Arm B is not as red as arm A.
- (3) Both arms are equally red.

Of 45 people tested with the condition, 22 are better off on arm A, 18 are better off on arm B, and 5 are equally well off on both arms. How can we decide whether this evidence is enough to conclude that ointment A is better than ointment B?

### Normal-Theory Method

In this section, we consider a large-sample method for addressing the question posed in Example 9.7.

Suppose that the degree of redness could be measured on a quantitative scale, with a higher number indicating more redness. Let  $x_i$  = degree of redness on arm A,  $y_i$  = degree of redness on arm B for the  $i$ th person. Let's focus on  $d_i = x_i - y_i$  = difference in redness between the A and B arms for the  $i$ th participant and test the hypothesis  $H_0$ :  $\Delta = 0$  vs.  $H_1$ :  $\Delta \neq 0$ , where  $\Delta$  = the population median of the  $d_i$  or the 50th percentile of the underlying distribution of the  $d_i$ .

- (1) If  $\Delta = 0$ , then the ointments are equally effective.
- (2) If  $\Delta < 0$ , then ointment A is better because arm A is less red than arm B.
- (3) If  $\Delta > 0$ , then ointment B is better because arm A is redder than arm B.

Notice that the actual  $d_i$  cannot be observed; we can only observe whether  $d_i > 0$ ,  $d_i < 0$ , or  $d_i = 0$ . The people for whom  $d_i = 0$  will be excluded because we cannot tell which ointment is better for them. The test will be based on the number of people C for whom  $d_i > 0$  out of the total of  $n$  people with nonzero  $d_i$ . This test makes sense because if C is large, then most people prefer treatment B over treatment A, whereas if C is small, they prefer treatment A over treatment B. We would expect under  $H_0$  that  $Pr(\text{nonzero } d_i > 0) = \frac{1}{2}$ . We will assume the normal approximation to the binomial is valid. This assumption will be true if

$$npq \geq 5 \quad \text{or} \quad n\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \geq 5$$

$$\text{Or} \quad \frac{n}{4} \geq 5 \quad \text{or} \quad n \geq 20$$

where  $n$  = the number of nonzero  $d_i$ 's.

The following test procedure for a two-sided level  $\alpha$  test, called the **sign test**, can then be used:

### Equation 9.1

#### The Sign Test

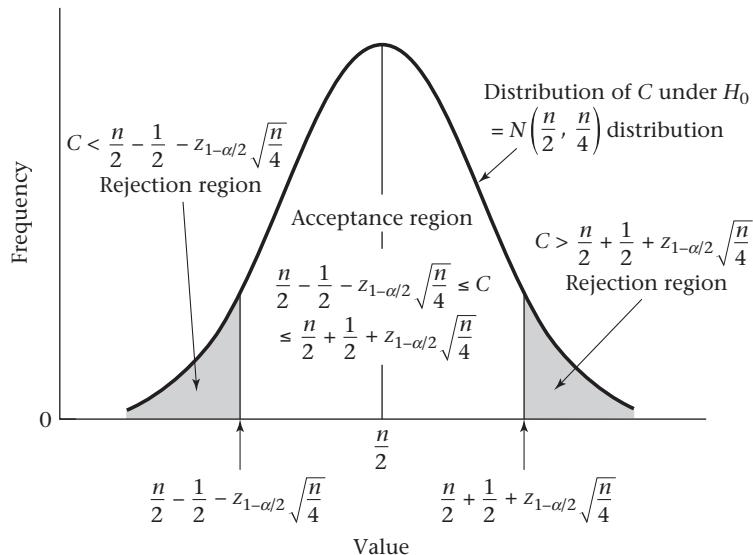
To test the hypothesis  $H_0: \Delta = 0$  vs.  $H_1: \Delta \neq 0$ , where the number of nonzero  $d_i$ 's  $= n \geq 20$  and  $C =$  the number of  $d_i$ 's where  $d_i > 0$ , if

$$C > c_2 = \frac{n}{2} + \frac{1}{2} + z_{1-\alpha/2} \sqrt{n/4} \quad \text{or} \quad C < c_1 = \frac{n}{2} - \frac{1}{2} - z_{1-\alpha/2} \sqrt{n/4}$$

then  $H_0$  is rejected. Otherwise,  $H_0$  is accepted.

The acceptance and rejection regions for this test are shown in Figure 9.1.

**Figure 9.1** Acceptance and rejection regions for the sign test



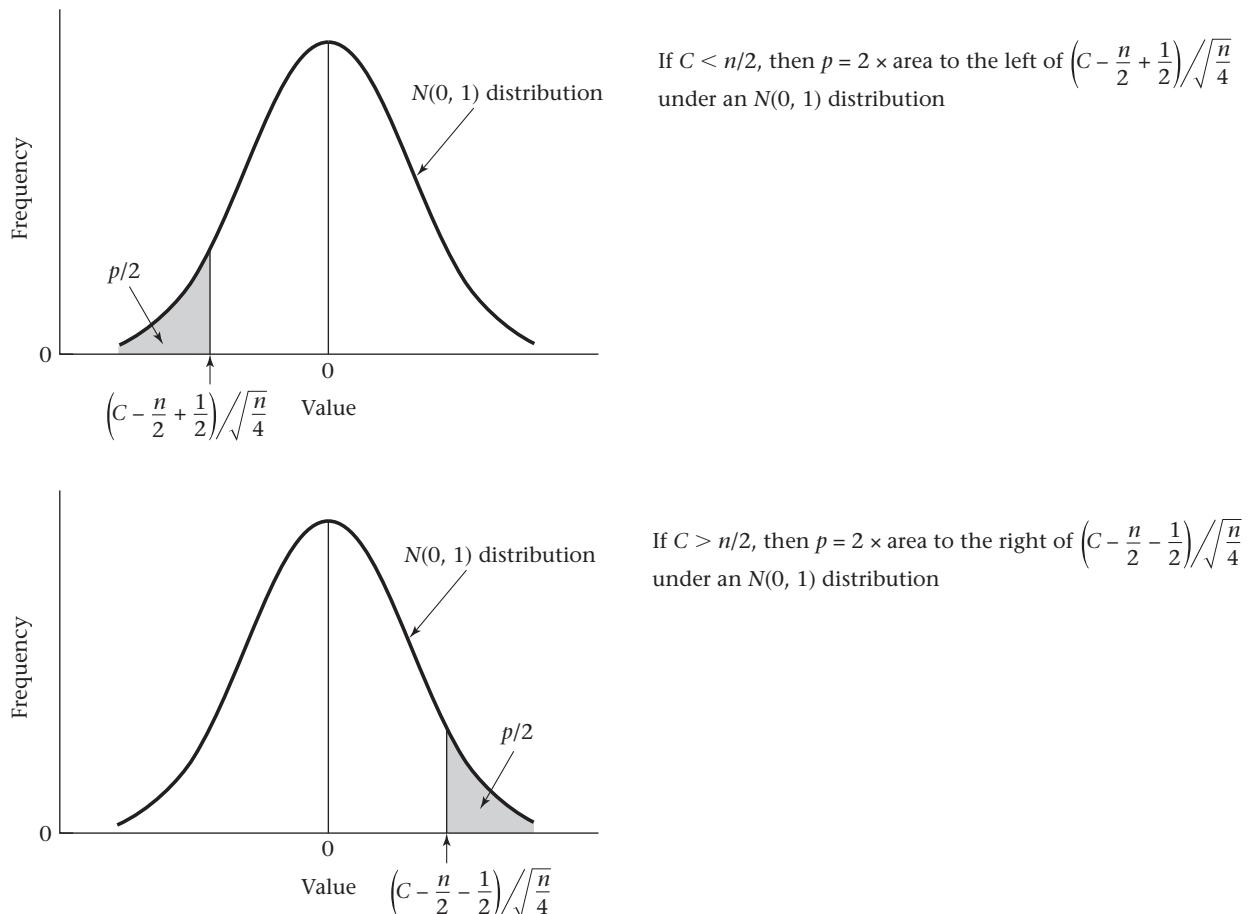
Similarly, the  $p$ -value for the procedure is computed using the following formula.

### Equation 9.2

#### Computation of the $p$ -Value for the Sign Test (Normal-Theory Method)

$$\begin{aligned} p &= 2 \times \left[ 1 - \Phi \left( \frac{C - \frac{n}{2} - .5}{\sqrt{n/4}} \right) \right] && \text{if } C > \frac{n}{2} \\ p &= 2 \times \Phi \left( \frac{C - \frac{n}{2} + .5}{\sqrt{n/4}} \right) && \text{if } C < \frac{n}{2} \\ p &= 1.0 && \text{if } C = \frac{n}{2} \end{aligned}$$

This computation is illustrated in Figure 9.2.

**Figure 9.2 Computation of the *p*-value for the sign test**

An alternative and equivalent formula for the *p*-value is given by

$$p = 2 \times \left[ 1 - \Phi\left(\frac{|C - D| - 1}{\sqrt{n}}\right) \right] \quad \text{if } C \neq D \quad \text{and} \quad p = 1.0 \quad \text{if } C = D$$

where  $C$  = the number of  $d_i > 0$  and  $D$  = the number of  $d_i < 0$ .

This test is called the *sign test* because it depends only on the sign of the differences and not on their actual magnitude.

The sign test is actually a special case of the one-sample binomial test in Section 7.10, where the hypothesis  $H_0: p = 1/2$  vs.  $H_1: p \neq 1/2$  was tested. In Equation 9.1 and Equation 9.2 a large-sample test is being used, and we are assuming the normal approximation to the binomial distribution is valid. Under  $H_0$ ,  $p = 1/2$  and  $E(C) = np = n/2$ ,  $Var(C) = npq = n/4$ , and  $C \sim N(n/2, n/4)$ . Furthermore, the .5 term in computing the critical region and *p*-value serves as a continuity correction and better approximates the binomial distribution by the normal distribution.

**Example 9.8**

**Dermatology** Assess the statistical significance of the skin-ointment data in Example 9.7.

**Solution**

In this case there are 40 untied pairs and  $C = 18 < n/2 = 20$ . From Equation 9.1, the critical values are given by

$$\begin{aligned}c_2 &= n/2 + 1/2 + z_{1-\alpha/2} \sqrt{n/4} \\&= 40/2 + 1/2 + z_{.975} \sqrt{40/4} = 20.5 + 1.96(3.162) = 26.7\\c_1 &= n/2 - 1/2 - z_{1-\alpha/2} \sqrt{n/4} = 19.5 - 1.96(3.162) = 13.3\end{aligned}$$

Because  $13.3 \leq C = 18 \leq 26.7$ ,  $H_0$  is accepted using a two-sided test with  $\alpha = .05$  and we conclude the ointments do not significantly differ in effectiveness. From Equation 9.2, because  $C = 18 < n/2 = 20$ , the exact  $p$ -value is given by

$$p = 2 \times \Phi \left[ \left( 18 - 20 + \frac{1}{2} \right) / \sqrt{40/4} \right] = 2 \times \Phi(-0.47) = 2 \times .3176 = .635$$

which is not statistically significant. Therefore, we accept  $H_0$ ; the ointments are equally effective.

Alternatively, we could compute the test statistic

$$z = \frac{|C - D| - 1}{\sqrt{n}}$$

where  $C = 18$ ,  $D = 22$ , and  $n = 40$ , yielding

$$z = \frac{|18 - 22| - 1}{\sqrt{40}} = \frac{3}{\sqrt{40}} = 0.47$$

and obtain the  $p$ -value from

$$p = 2 \times [1 - \Phi(0.47)] = .635$$

## Exact Method

If  $n < 20$ , then exact binomial probabilities rather than the normal approximation must be used to compute the  $p$ -value.  $H_0$  should still be rejected if  $C$  is very large or very small. The expressions for the  $p$ -value based on exact binomial probabilities are as follows:

**Equation 9.3**

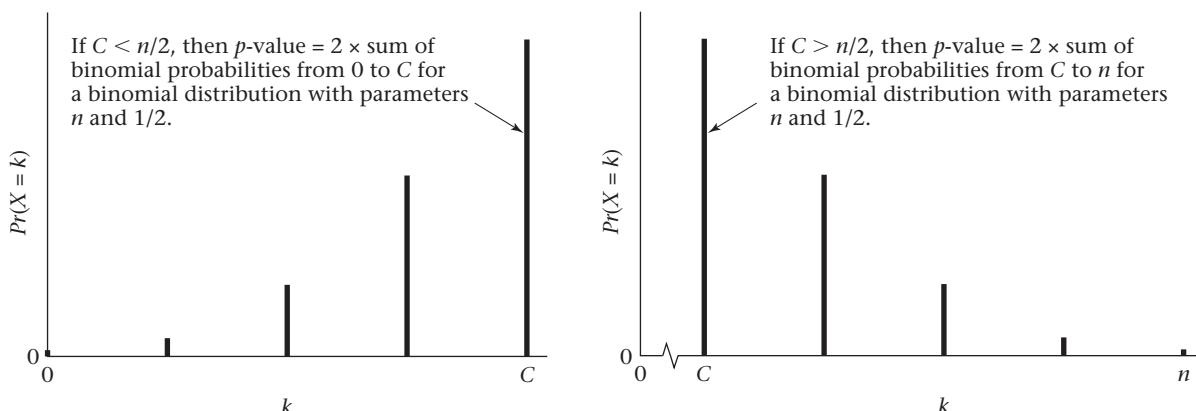
### Computation of the $p$ -Value for the Sign Test (Exact Test)

$$\text{If } C > n/2, \quad p = 2 \times \sum_{k=C}^n \binom{n}{k} \left( \frac{1}{2} \right)^n$$

$$\text{If } C < n/2, \quad p = 2 \times \sum_{k=0}^C \binom{n}{k} \left( \frac{1}{2} \right)^n$$

$$\text{If } C = n/2, \quad p = 1.0$$

This computation is depicted in Figure 9.3.

**Figure 9.3 Computation of the *p*-value for the sign test (exact test)**

This test is a special case of the small-sample, one-sample binomial test described in Equation 7.44, where the hypothesis  $H_0: p = \frac{1}{2}$  vs.  $H_1: p \neq \frac{1}{2}$  is tested.

#### **Example 9.9**

**Ophthalmology** Suppose we wish to compare two different types of eye drops (A, B) that are intended to prevent redness in people with hay fever. Drug A is randomly administered to one eye and drug B to the other eye. The redness is noted at baseline and after 10 minutes by an observer who is unaware of which drug has been administered to which eye. We find that for 15 people with an equal amount of redness in each eye at baseline, after 10 minutes the drug A eye is less red than the drug B eye for 2 people ( $d_i < 0$ ); the drug B eye is less red than the drug A eye for 8 people ( $d_i > 0$ ); and the eyes are equally red for 5 people ( $d_i = 0$ ). Assess the statistical significance of the results.

#### **Solution**

The test is based on the 10 people who had a differential response to the two types of eye drops. Because  $n = 10 < 20$ , the normal-theory method in Equation 9.2 cannot be used; the exact method in Equation 9.3 must be used instead. Because  $C = 8 > \frac{10}{2} = 5$ ,

$$p = 2 \times \sum_{k=8}^{10} \binom{10}{k} \left(\frac{1}{2}\right)^{10}$$

Refer to the binomial tables (Table 1 in the Appendix) using  $n = 10$ ,  $p = .5$ , and note that  $Pr(X = 8) = .0439$ ,  $Pr(X = 9) = .0098$ ,  $Pr(X = 10) = .0010$ . Thus  $p = 2 \times Pr(X \geq 8) = 2(.0439 + .0098 + .0010) = 2 \times .0547 = .109$ , which is not statistically significant. Thus we accept  $H_0$ , that the two types of eye drops are equally effective in reducing redness in people with hay fever.

## 9.3 The Wilcoxon Signed-Rank Test

#### **Example 9.10**

**Dermatology** Consider the data in Example 9.7 from a different perspective. We assumed that the only possible assessment was that the degree of sunburn with ointment A was either better or worse than that with ointment B. Suppose instead the degree of burn can be quantified on a 10-point scale, with 10 being the worst burn

and 1 being no burn at all. We can now compute  $d_i = x_i - y_i$ , where  $x_i$  = degree of burn for ointment A and  $y_i$  = degree of burn for ointment B. If  $d_i$  is positive, then ointment B is doing better than ointment A; if  $d_i$  is negative, then ointment A is doing better than ointment B. For example, if  $d_i = +5$ , then the degree of redness is 5 units greater on the ointment A arm than on the ointment B arm, whereas if  $d_i = -3$ , then the degree of redness is 3 units less on the ointment A arm than on the ointment B arm. How can this additional information be used to test whether the ointments are equally effective?

Suppose the sample data in Table 9.1 are obtained. The  $f_i$  values represent the frequency or the number of people with difference in redness  $d_i$  between the ointment A and ointment B arms.

Notice that there is only a slight excess of people with negative  $d_i$  (that is, who are better off with ointment A, 22) than with positive  $d_i$  (that is, who are better off with ointment B, 18). However, the extent to which the 22 people are better off appears far greater than that of the 18 people because the negative  $d_i$ 's generally have a much greater absolute value than the positive  $d_i$ 's. This point is illustrated in Figure 9.4.

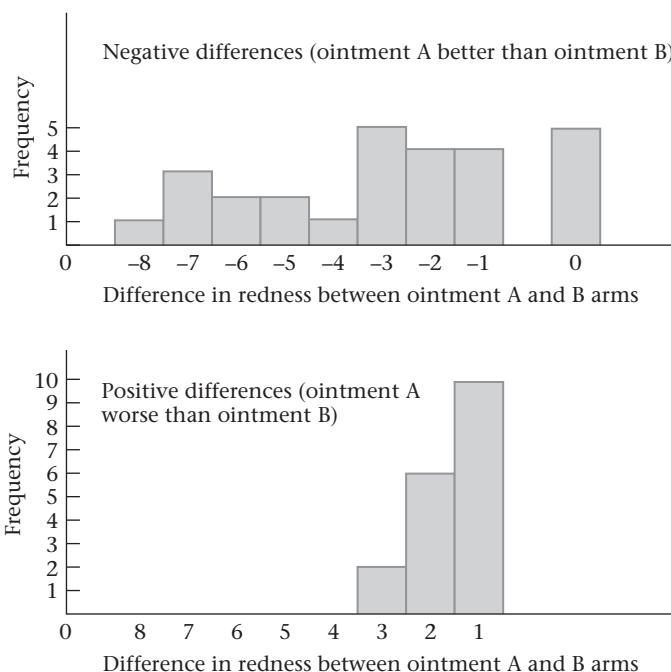
We wish to test the hypothesis  $H_0: \Delta = 0$  vs.  $H_1: \Delta \neq 0$ , where  $\Delta$  = the median score difference between the ointment A and ointment B arms. If  $\Delta < 0$ , then ointment A is better; if  $\Delta > 0$ , then ointment B is better. More generally, we can test the hypothesis  $H_{0r}$  that the distribution of  $d_i$  is symmetric about zero, vs.  $H_1$ , that the distribution of  $d_i$  is not symmetric about zero.

Let's assume that the  $d_i$ 's have an underlying continuous distribution. Based on Figure 9.4, a seemingly reasonable test of this hypothesis would be to take account of both the magnitude and the sign of the differences  $d_i$ . A paired  $t$  test might be used here, but the problem is that the rating scale is ordinal. The measurement  $d_i = -5$  does not mean that the difference in degree of burn is five times as great as  $d_i = -1$ , but rather it simply means there is a relative ranking of differences in degree of burn, with  $-8$  being most favorable to ointment A,  $-7$  the next most favorable, and so on. Thus a

**Table 9.1** Difference in degree of redness between ointment A and ointment B arms after 10 minutes of exposure to sunlight

$ d_i $	Negative		Positive		Number of people with same absolute value	Range of ranks	Average rank
	$d_i$	$f_i$	$d_i$	$f_i$			
10	-10	0	10	0		—	—
9	-9	0	9	0	0	—	—
8	-8	1	8	0	1	40	40.0
7	-7	3	7	0	3	37–39	38.0
6	-6	2	6	0	2	35–36	35.5
5	-5	2	5	0	2	33–34	33.5
4	-4	1	4	0	1	32	32.0
3	-3	5	3	2	7	25–31	28.0
2	-2	4	2	6	10	15–24	19.5
1	-1	4	1	10	14	1–14	7.5
		22		18			
0	0	5					

**Figure 9.4** Bar graph of the differences in redness between the ointment A and ointment B arms for the data in Example 9.10



nonparametric test that is analogous to the paired  $t$  test is needed here. The **Wilcoxon signed-rank test** is such a test. It is nonparametric, because it is based on the ranks of the observations rather than on their actual values, as is the paired  $t$  test.

The first step in this test is to compute ranks for each observation, as follows.

#### Equation 9.4

#### Ranking Procedure for the Wilcoxon Signed-Rank Test

- (1) Arrange the differences  $d_i$  in order of *absolute value* as in Table 9.1.
- (2) Count the number of differences with the same absolute value.
- (3) Ignore the observations where  $d_i = 0$ , and rank the remaining observations from 1 for the observation with the lowest absolute value, up to  $n$  for the observation with the highest absolute value.
- (4) If there is a group of several observations with the same absolute value, then find the lowest rank in the range  $= 1 + R$  and the highest rank in the range  $= G + R$ , where  $R$  = the highest rank used prior to considering this group and  $G$  = the number of differences in the *range of ranks* for the group. Assign the *average rank*  $= (\text{lowest rank in the range} + \text{highest rank in the range})/2$  as the rank for each difference in the group.

#### Example 9.11

**Dermatology** Compute the ranks for the skin-ointment data in Table 9.1.

#### Solution

First collect the differences with the same absolute value. Fourteen people have absolute value 1; this group has a rank range from 1 to 14 and an average rank of  $(1 + 14)/2 = 7.5$ . The group of 10 people with absolute value 2 has a rank range from  $(1 + 14)$  to  $(10 + 14) = 15$  to 24 and an average rank  $= (15 + 24)/2 = 19.5$ , . . . , and so on.

The test is based on the sum of the ranks, or **rank sum** ( $R_1$ ), for the group of people with positive  $d_i$ —that is, the rank sum for people for whom ointment A is not as good as ointment B. A large rank sum indicates that differences in burn degree in favor of treatment B tend to be larger than those for treatment A, whereas a small rank sum indicates that differences in burn degree in favor of treatment A tend to be larger than those for treatment B. If the null hypothesis is true, then the expected value and variance of the rank sum (when there are no ties) are given by

$$E(R_1) = n(n+1)/4, \quad Var(R_1) = n(n+1)(2n+1)/24$$

where  $n$  is the number of nonzero differences.

If the number of nonzero  $d_i$ 's is  $\geq 16$ , then a normal approximation can be used for the sampling distribution of  $R_1$ . This test procedure, the Wilcoxon signed-rank test, is given as follows.

### Equation 9.5

#### Wilcoxon Signed-Rank Test (Normal Approximation Method for Two-Sided Level $\alpha$ Test)

- (1) Rank the differences as shown in Equation 9.4.
- (2) Compute the rank sum  $R_1$  of the positive differences.
- (3) (a) If  $R_1 \neq \frac{n(n+1)}{4}$  and there *are no ties* (no groups of differences with the same absolute value), then

$$T = \left[ \left| R_1 - \frac{n(n+1)}{4} \right| - \frac{1}{2} \right] / \sqrt{n(n+1)(2n+1)/24}$$

- (b) If  $R_1 \neq \frac{n(n+1)}{4}$  and there *are ties*, where  $t_i$  refers to the number of differences with the same absolute value in the  $i$ th tied group and  $g$  is the number of tied groups, then

$$T = \left[ \left| R_1 - \frac{n(n+1)}{4} \right| - \frac{1}{2} \right] / \sqrt{n(n+1)(2n+1) / 24 - \sum_{i=1}^g (t_i^3 - t_i) / 48}$$

- (c) If  $R_1 = \frac{n(n+1)}{4}$ , then  $T = 0$ .
- (4) If

$$T > z_{1-\alpha/2}$$

then reject  $H_0$ . Otherwise, accept  $H_0$ .

- (5) The  $p$ -value for the test is given by

$$p = 2 \times [1 - \Phi(T)]$$

- (6) This test should be used only if the number of nonzero differences is  $\geq 16$  and if the difference scores have an underlying continuous symmetric distribution. The computation of the  $p$ -value is illustrated in Figure 9.5.

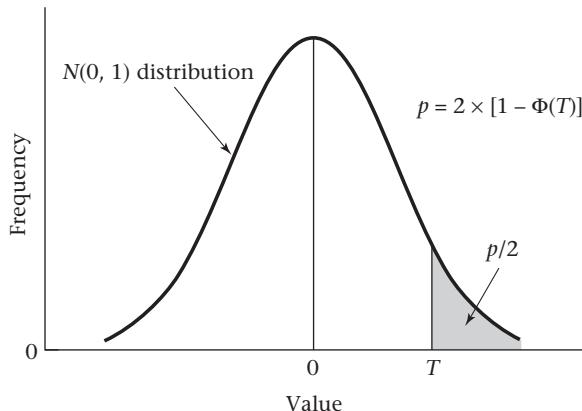
The rationale for the different test statistic in the absence (3a) or presence (3b) of tied values is that the variance of  $R_1$  is reduced in the presence of ties (sometimes substantially) [1].

An alternative variance formula for  $R_1$  is

$$\text{Var}(R_1) = \sum_{j=1}^n r_j^2 / 4$$

where  $r_j$  = rank of the absolute value of the  $j$ th observation and the sum is over all observations, whether positive or negative. This formula is valid both in the presence or absence of ties but is computationally easier than the variance formula in 3b if ties are present.

**Figure 9.5** Computation of the  $p$ -value for the Wilcoxon signed-rank test



The term  $1/2$  in the computation of  $T$  serves as a continuity correction in the same manner as for the sign test in Equations 9.1 and 9.2.

**Example 9.12**

**Dermatology** Perform the Wilcoxon signed-rank test for the data in Example 9.10.

**Solution**

Because the number of nonzero differences ( $22 + 18 = 40 \geq 16$ ), the normal approximation method in Equation 9.5 can be used. Compute the rank sum for the people with positive  $d_i$ —that is, where ointment B performs better than ointment A, as follows:

$$R_1 = 10(7.5) + 6(19.5) + 2(28.0) = 75 + 117 + 56 = 248$$

The expected rank sum is given by

$$E(R_1) = 40(41) / 4 = 410$$

The variance of the rank sum corrected for ties is given by

$$\begin{aligned} \text{Var}(R_1) &= 40(41)(81) / 24 - [(14^3 - 14) + (10^3 - 10) + (7^3 - 7) + (1^3 - 1) + (2^3 - 2) \\ &\quad + (2^3 - 2) + (3^3 - 3) + (1^3 - 1)] / 48 \\ &= 5535 - (2730 + 990 + 336 + 0 + 6 + 6 + 24 + 0) / 48 \\ &= 5535 - 4092 / 48 = 5449.75 \end{aligned}$$

If the alternative variance formula is used, then

$$\begin{aligned} \text{Var}(T) &= [14(7.5)^2 + 10(19.5)^2 + \dots + (40)^2] / 4 \\ &= 21,799 / 4 = 5449.75 \end{aligned}$$

Thus  $sd(R_1) = \sqrt{5449.75} = 73.82$ . Therefore, the test statistic  $T$  is given by

$$T = \left( |248 - 410| - \frac{1}{2} \right) / 73.82 = 161.5 / 73.82 = 2.19$$

The  $p$ -value of the test is given by

$$p = 2[1 - \Phi(2.19)] = 2 \times (1 - .9857) = .029$$

We therefore can conclude that there is a significant difference between ointments, with ointment A doing better than ointment B because the observed rank sum (248) is smaller than the expected rank sum (410). This conclusion differs from the conclusion based on the sign test in Example 9.8, where no significant difference between ointments was found. This result indicates that when the information is available, it is worthwhile to consider both magnitude and direction of the difference between treatments, as the signed-rank test does, rather than just the direction of the difference, as the sign test does.

In general, for a two-sided test, if the signed-rank test is based on negative differences rather than positive differences, the same test statistic and  $p$ -value will always result. Thus the rank sum can be arbitrarily computed based on either positive or negative differences.

### Example 9.13

**Dermatology** Perform the Wilcoxon signed-rank test for the data in Example 9.10 based on negative—rather than positive—difference scores.

#### Solution

$R_2$  = rank sum for negative differences

$$\begin{aligned} &= 4(7.5) + 4(19.5) + 5(28.0) + 1(32.0) + 2(33.5) + 2(35.5) + 3(38.0) + 1(40.0) \\ &= 572 \end{aligned}$$

Thus

$$\left| R_2 - \frac{n(n+1)}{4} \right| - .5 = |572 - 410| - .5 = 161.5 = \left| R_1 - \frac{n(n+1)}{4} \right| - .5$$

Because  $Var(R_1) = Var(R_2)$ , the same test statistic  $T = 2.19$  and  $p$ -value = .029 are obtained as when positive-difference scores are used.

If the number of pairs with nonzero  $d_i \leq 15$ , then the normal approximation is no longer valid and special small-sample tables giving significance levels for this test must be used. Such a table is Table 11 in the Appendix, which gives upper and lower critical values for  $R_1$  for a two-sided test with  $\alpha$  levels of .10, .05, .02, and .01, respectively. In general, the results are statistically significant at a particular  $\alpha$  level only if either  $R_1 \leq$  the lower critical value or  $R_1 \geq$  the upper critical value for that  $\alpha$  level.

### Example 9.14

Suppose there are 9 untied nonzero paired differences and a rank sum of 43. Evaluate the statistical significance of the results.

#### Solution

Because  $R_1 = 43 \geq 42$ , it follows that  $p < .02$ . Because  $R_1 = 43 < 44$ , it follows that  $p \geq .01$ . Thus  $.01 \leq p < .02$ , and the results are statistically significant.

If you have  $<16$  nonzero paired differences and there are ties among the paired differences, then Table 11 in the Appendix is no longer applicable and more complicated methods based on permutation tests should be used (see [1] for more details).

An example of the signed-rank test with ordinal data has been presented. This test and the other nonparametric tests can be applied to cardinal data as well, particularly if the sample size is small and the assumption of normality appears grossly violated. However, an assumption of the signed-rank test is that one has a continuous and symmetric, but not necessarily normal, distribution. If the actual distribution turns out to be normal, then the signed-rank test has less power than the paired *t* test, which is the penalty paid.

### REVIEW QUESTIONS 9A

- 1** What are the differences among cardinal data, ordinal data, and nominal data?
- 2** What is the difference between a parametric test and a nonparametric test?
- 3** What is the difference between the sign test and the signed-rank test?
- 4** Suppose researchers study an experimental surgical procedure for patients with retinitis pigmentosa among (RP) a group of 10 patients. The investigators find that 9 of the patients got worse after this procedure and 1 patient got better over the short term (3–6 months). Assuming that in the absence of treatment the patients would be expected to show no change over this time period, evaluate the results of the study.
- 5** The actual change scores on the electroretinogram (ERG), a measure of electrical activity in the retina, are presented for each patient in Table 9.2

**Table 9.2** ERG change scores following surgery for RP (Berson et al., 1996)

Patient	Score <sup>a</sup>	Patient	Score
1	-0.238	6	+0.090
2	-0.085	7	-0.736
3	-0.215	8	-0.365
4	-0.227	9	-0.179
5	-0.037	10	-0.048

<sup>a</sup>The change scores =  $\ln(\text{ERG amplitude})$  at follow-up –  $\ln(\text{ERG amplitude})$  at baseline. A negative score indicates decline.

Evaluate the significance of the results without assuming the change scores are normally distributed. What do the results mean?

## 9.4 The Wilcoxon Rank-Sum Test

In the previous section, a nonparametric analog to the paired *t* test—namely, the Wilcoxon signed-rank test—was presented. In this section, a nonparametric analog to the *t* test for two independent samples is described.

### Example 9.15

**Ophthalmology** Different genetic types of the disease retinitis pigmentosa (RP) are thought to have different rates of progression, with the dominant form of the disease progressing the slowest, the recessive form the next slowest, and

**Table 9.3 Comparison of visual acuity in people ages 10–19 with dominant and sex-linked RP**

Visual acuity	Dominant	Sex-linked	Combined sample	Range of ranks	Average rank
20–20	5	1	6	1–6	3.5
20–25	9	5	14	7–20	13.5
20–30	6	4	10	21–30	25.5
20–40	3	4	7	31–37	34.0
20–50	2	8	10	38–47	42.5
20–60	0	5	5	48–52	50.0
20–70	0	2	2	53–54	53.5
20–80	<u>0</u>	<u>1</u>	<u>1</u>	55	55.0
	25	30	55		

the sex-linked form the fastest. This hypothesis can be tested by comparing the visual acuity of people who have different genetic types of RP. Suppose there are 25 people ages 10–19 with dominant disease and 30 people with sex-linked disease. The best-corrected visual acuities (i.e., with appropriate glasses) in the better eye of these people are presented in Table 9.3. How can these data be used to test whether the distribution of visual acuity is different in the two groups?

We wish to test the hypothesis  $H_0: F_D = F_{SL}$  vs.  $H_1: F_D(x) = F_{SL}(x - \Delta)$ , where  $\Delta \neq 0$ .  $F_D$  = cumulative distribution function (c.d.f.) of visual acuity for the dominant group,  $F_{SL}$  = cumulative distribution function of visual acuity for the sex-linked group, and  $\Delta$  is a location shift of the c.d.f. for the sex-linked group relative to the dominant group. If  $\Delta > 0$ , then dominant patients tend to have better visual acuity than sex-linked patients; if  $\Delta < 0$ , then dominant patients tend to have worse visual acuity than sex-linked patients; if  $\Delta = 0$ , then dominant patients have the same acuity distribution as sex-linked patients. The two-sample  $t$  test for independent samples, discussed in Sections 8.4 and 8.7, would ordinarily be used for this type of problem. However, visual acuity cannot be given a specific numeric value that all ophthalmologists would agree on. Thus the  $t$  test is inapplicable, and a nonparametric analog must be used. The nonparametric analog to the independent-samples  $t$  test is the **Wilcoxon rank-sum test**. This test is nonparametric because it is based on the *ranks* of the individual observations rather than on their actual values, which would be used in the  $t$  test. The ranking procedure for this test is as follows.

**Equation 9.6****Ranking Procedure for the Wilcoxon Rank-Sum Test**

- (1) Combine the data from the two groups, and order the values from lowest to highest or, in the case of visual acuity, from best (20–20) to worst (20–80).
- (2) Assign ranks to the individual values, with the best visual acuity (20–20) having the lowest rank and the worst visual acuity (20–80) having the highest rank, or vice versa.
- (3) If a group of observations has the same value, then compute the *range of ranks* for the group, as was done for the signed-rank test in Equation 9.4, and assign the *average rank* for each observation in the group.

**Example 9.16****Solution**

Compute the ranks for the visual-acuity data in Table 9.3.

First collect all people with the same visual acuity over the two groups, as shown in Table 9.3. There are 6 people with visual acuity 20–20 who have a rank range of 1–6 and are assigned an average rank of  $(1 + 6)/2 = 3.5$ . There are 14 people for the two groups combined with visual acuity 20–25. The rank range for this group is from  $(1 + 6)$  to  $(14 + 6) = 7$  to 20. Thus all people in this group are assigned the average rank  $= (7 + 20)/2 = 13.5$ , and similarly for the other groups.

The test statistic for this test is the sum of the ranks in the first sample ( $R_1$ ). If this sum is large, then the dominant group has poorer visual acuity than the sex-linked group, whereas if it is small, the dominant group has better visual acuity. If the number of observations in the two groups are  $n_1$  and  $n_2$ , respectively, then the average rank in the combined sample is  $(1 + n_1 + n_2)/2$ . Thus, under  $H_0$  the expected rank sum in the first group  $\equiv E(R_1) = n_1 \times \text{average rank in the combined sample} = n_1(n_1 + n_2 + 1)/2$ . It can be shown that the variance of  $R_1$  under  $H_0$  if there are no tied values is given by  $\text{Var}(R_1) = n_1 n_2 (n_1 + n_2 + 1)/12$ . Furthermore, we will assume that the smaller of the two groups is of size at least 10 and that the variable under study has an underlying continuous distribution. Under these assumptions, the sampling distribution of the rank sum  $R_1$  is approximately normal. Thus the following test procedure is used.

**Equation 9.7****Wilcoxon Rank-Sum Test (Normal Approximation Method for Two-Sided Level  $\alpha$  Test)**

- (1) Rank the observations as shown in Equation 9.6.
- (2) Compute the rank sum  $R_1$  in the first sample (the choice of sample is arbitrary).
- (3) (a) If  $R_1 \neq n_1(n_1 + n_2 + 1)/2$  and there are no ties, then compute

$$T = \left[ \left| R_1 - \frac{n_1(n_1 + n_2 + 1)}{2} \right| - \frac{1}{2} \right] \sqrt{\left( \frac{n_1 n_2}{12} \right) (n_1 + n_2 + 1)}$$

- (b) If  $R_1 \neq n_1(n_1 + n_2 + 1)/2$  and there are ties, then compute

$$T = \left[ \left| R_1 - \frac{n_1(n_1 + n_2 + 1)}{2} \right| - \frac{1}{2} \right] \sqrt{\left( \frac{n_1 n_2}{12} \right) \left[ n_1 + n_2 + 1 - \frac{\sum_{i=1}^g t_i(t_i^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)} \right]}$$

where  $t_i$  refers to the number of observations with the same value in the  $i$ th tied group, and  $g$  is the number of tied groups.

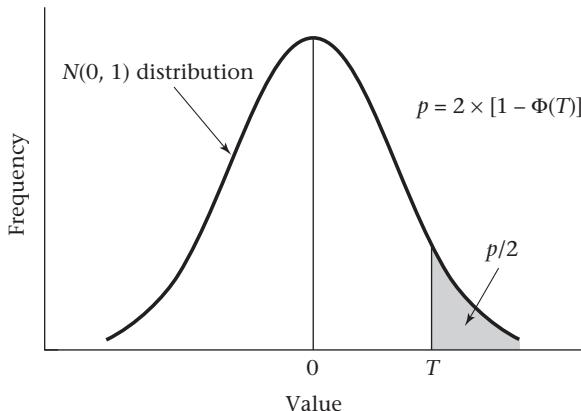
- (c) If  $R_1 = n_1(n_1 + n_2 + 1)/2$ , then  $T = 0$ .
- (4) If

$$T > z_{1-\alpha/2}$$

then reject  $H_0$ . Otherwise, accept  $H_0$ .

- (5) Compute the exact  $p$ -value by

$$p = 2 \times [1 - \Phi(T)]$$

**Figure 9.6** Computation of the *p*-value for the Wilcoxon rank-sum test

- (6) This test should be used only if both  $n_1$  and  $n_2$  are at least 10, and if there is an underlying continuous distribution.

The computation of the *p*-value is illustrated in Figure 9.6.

### Example 9.17

#### Solution

Perform the Wilcoxon rank-sum test for the data in Table 9.3.

Because the minimum sample size in the two samples is  $25 \geq 10$ , the normal approximation can be used. The rank sum in the dominant group is given by

$$\begin{aligned} R_1 &= 5(3.5) + 9(13.5) + 6(25.5) + 3(34) + 2(42.5) \\ &= 17.5 + 121.5 + 153 + 102 + 85 = 479 \end{aligned}$$

$$\text{Furthermore, } E(R_1) = \frac{25(56)}{2} = 700$$

and  $\text{Var}(R_1)$  corrected for ties is given by

$$\begin{aligned} &[25(30)/12]\{56 - [6(6^2 - 1) + 14(14^2 - 1) + 10(10^2 - 1) + 7(7^2 - 1) + 10(10^2 - 1) \\ &\quad + 5(5^2 - 1) + 2(2^2 - 1) + 1(1^2 - 1)]/[55(54)]\} \\ &= 62.5(56 - 5382/2970) = 3386.74 \end{aligned}$$

Thus the test statistic  $T$  is given by

$$T = \frac{(|479 - 700| - .5)}{\sqrt{3386.74}} = \frac{220.5}{58.2} = 3.79$$

which follows an  $N(0,1)$  distribution under  $H_0$ . The *p*-value of the test is

$$p = 2 \times [1 - \Phi(3.79)] < .001$$

We conclude that the visual acuities of the two groups are significantly different. Because the observed rank sum in the dominant group (479) is lower than the expected rank sum (700), the dominant group has better visual acuity than the sex-linked group.

If either sample size is less than 10, the normal approximation is not valid, and a small-sample table of exact significance levels must be used. Table 12 in the Appendix gives upper and lower critical values for the rank sum in the first of two samples ( $T$ ) for a two-sided test with  $\alpha$  levels of .10, .05, .02, and .01, respectively, under the

assumption that there are no ties. In general, the results are statistically significant at a particular  $\alpha$  level if either  $T \leq T_l$  = the lower critical value or  $T \geq T_r$  = the upper critical value for that  $\alpha$  level.

**Example 9.18**

Suppose there are two samples of size 8 and 15, with no ties, with a rank sum of 73 in the sample size of 8. Evaluate the statistical significance of the results.

**Solution**

Refer to  $n_1 = 8$ ,  $n_2 = 15$ ,  $\alpha = .05$  and find that  $T_l = 65$ ,  $T_r = 127$ . Because  $T = 73 > 65$  and  $T < 127$ , the results are not statistically significant using a two-sided test at the 5% level.

Appendix Table 12 was constructed under the assumption of no tied values. If there are ties, and  $\min(n_1, n_2) < 10$ , then an exact-permutation test must be used to assess statistical significance (see Lehmann [1] for more details).

The Wilcoxon rank-sum test is sometimes referred to in the literature as the **Mann-Whitney U test**. The test statistic for the Mann-Whitney  $U$  test is based on the number of pairs of observations  $(x_i, y_j)$ , one from each sample, such that  $x_i < y_j$ ; in addition, 0.5 is added to the test statistic for each  $(x_i, y_j)$  pair such that  $x_i = y_j$ . The Mann-Whitney  $U$  test and the Wilcoxon rank-sum test are completely equivalent because the same  $p$ -value is obtained by applying either test. Therefore, the choice of which test to use is a matter of convenience.

Because ranking all the observations in a large sample is tedious, a computer program is useful in performing this test. The Stata ranksum command was used on the data set in Table 9.3; the results are given in Table 9.4.

**Table 9.4** Stata ranksum command used on the data in Table 9.3

Two-sample Wilcoxon rank-sum (Mann-Whitney) test			
group	obs	rank sum	expected
1	25	479	700
2	30	1061	840
<b>combined  </b>	<b>55</b>	<b>1540</b>	<b>1540</b>
<b>unadjusted variance</b>		<b>3500.00</b>	
<b>adjustment for ties</b>		<b>-113.26</b>	
<b>adjusted variance</b>		<b>3386.74</b>	
<b>Ho: visual-y(group==1) = visual-y(group==2)</b>			
<b>z = -3.798</b>			
<b>Prob &gt;  z  = 0.0001</b>			

The median visual acuity in the dominant and sex-linked groups is listed first. The dominant group, with a lower median, tends to have better visual acuity than the sex-linked group. The Wilcoxon rank-sum (labeled  $W$ ) = 479 is given in the output. The two-tailed  $p$ -value after adjusting for ties is denoted by Prob  $> |z|$  and is 0.0001.

Finally, a necessary condition for the strict validity of the rank-sum test is that the underlying distributions being compared must be continuous. However, McNeil has investigated the use of this test in comparing discrete distributions and has found only small losses in power when applying this test to grouped data from normal distributions, compared with the actual ungrouped observations from such distributions [3]. He concludes that the rank-sum test is approximately valid in this case, with the appropriate provision for ties as given in Equation 9.7.

## 9.5 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children

In previous chapters, we have considered the effect of lead exposure on neurological and cognitive function in children as described in Data Set LEAD.DAT on the Companion Website. Other effects of lead as described in the literature are behavioral in nature. One such variable is hyperactivity. In this study, the children's parents were asked to rate the degree of hyperactivity in their children on a four-point scale, from normal (0) to very hyperactive (3). The scale is ordinal in nature. Thus to compare the degree of hyperactivity between the exposed and control groups, nonparametric methods are appropriate. Because this question was only asked for the younger children, data are available for only 49 control children and 35 exposed children. The raw data are given in Table 9.5. The columns of the table correspond to the groups (1 = control, 2 = exposed). The rows of the table correspond to the degree of hyperactivity. Within each group, the percentage of children with a specific hyperactivity score (the column percentages) is also given. The exposed children seem slightly more hyperactive than the control children.

We have used the Mann-Whitney  $U$  test to compare the hyperactivity distribution between the two groups. The results are given in Table 9.6. We see that the two-sided  $p$ -value (adjusted for ties) as given in the last row of the table is .46. Thus there is no significant difference in the distribution of hyperactivity level between the two groups.

### REVIEW QUESTIONS 9B

- 1 What is the difference between the Wilcoxon signed-rank test and the Wilcoxon rank-sum test?
- 2 A pilot study is planned to test the efficacy of vitamin E supplementation as a possible preventive agent for Alzheimer's disease. Twenty subjects age 65+ are randomized to either a supplement of vitamin E of 400 IU/day (group 1,  $n = 10$ ), or placebo (group 2). It is important to compare the total vitamin E intake (from food and supplements) of the two groups at baseline. The baseline intake of each group in IU/day is as follows:  
Group 1 ( $n = 10$ ): 7.5, 12.6, 3.8, 20.2, 6.8, 403.3, 2.9, 7.2, 10.5, 205.4  
Group 2 ( $n = 10$ ): 8.2, 13.3, 102.0, 12.7, 6.3, 4.8, 19.5, 8.3, 407.1, 10.2
  - (a) What test can be used to compare the baseline vitamin E intake between the two groups if we do not wish to assume normality?
  - (b) Implement the test in Review Question 9B.2a and report a two-tailed  $p$ -value.

## 9.6 Summary

This chapter presented some of the most widely used nonparametric statistical tests corresponding to the parametric procedures in Chapter 8. The main advantage of nonparametric methods is that the assumption of normality made in previous chapters can be relaxed when such assumptions are unreasonable. One drawback of nonparametric procedures is that some power is lost relative to using a parametric procedure (such as a  $t$  test) if the data truly follow a normal distribution or if the central-limit theorem is applicable. Also, the data typically must be expressed in terms of ranks, a scale some researchers find difficult to understand compared with maintaining the raw data in the original scale.

**Table 9.5** Hyperactivity data for case study

Tabulated statistics: Hyperact, Group			
	ROWS: HYPERACT	COLUMNS: GROUP	
	1	2	ALL
0	24 48.98	15 42.86	39 46.43
1	20 40.82	14 40.00	34 40.48
2	3 6.12	5 14.29	8 9.52
3	2 4.08	1 2.86	3 3.57
Missing	29 *	11 *	*
ALL	49 100.00	35 100.00	84 100.00
CELL CONTENTS --			
	COUNT		
	% OF COL		

**Table 9.6** Results of Mann-Whitney *U* test

Mann-Whitney Confidence Interval and Test		
	N	Median
hyper-1	49	1.0000
hyper-2	35	1.0000
Point estimate for ETA1-ETA2 is 0.0000		
95.1 Percent C.I. for ETA1-ETA2 is (-0.0000, 0.0001)		
W = 2008.5		
Test of ETA1 = ETA2 vs. ETA1 not = ETA2 is significant at 0.5049		
The test is significant at 0.4649 (adjusted for ties)		

The specific procedures covered for the comparison of two samples include the sign test, the Wilcoxon signed-rank test, and the Wilcoxon rank-sum test. Both the sign test and the signed-rank test are nonparametric analogs to the paired *t* test. For the sign test it is only necessary to determine whether one member of a matched pair has a higher or lower score than the other member of the pair. For the signed-rank test the magnitude of the absolute value of the difference score (which is then ranked), as well as its sign, is used in performing the significance test. The Wilcoxon rank-sum test (also known as the Mann-Whitney *U* test) is an analog to the two-sample *t* test for independent samples, in which the actual values are replaced by rank scores. Nonparametric procedures appropriate for regression, analysis of variance, and survival analysis are introduced in Chapters 11, 12, and 14.

The tests covered in this chapter are among the most basic of nonparametric tests. Hollander and Wolfe [4] and Lehmann [1] provide a more comprehensive treatment of nonparametric statistical methods.

## PROBLEMS

### Dentistry

In a study, 28 adults with mild periodontal disease are assessed before and 6 months after implementation of a dental-education program intended to promote better oral hygiene. After 6 months, periodontal status improved in 15 patients, declined in 8, and remained the same in 5.

\***9.1** Assess the impact of the program statistically (use a two-sided test).

Suppose patients are graded on the degree of change in periodontal status on a 7-point scale, with +3 indicating the greatest improvement, 0 indicating no change, and -3 indicating the greatest decline. The data are given in Table 9.7.

**9.2** What nonparametric test can be used to determine whether a significant change in periodontal status has occurred over time?

**9.3** Implement the procedure in Problem 9.2, and report a *p*-value.

**Table 9.7 Degree of change in periodontal status**

Change score	Number of patients
+3	4
+2	5
+1	6
0	5
-1	4
-2	2
-3	2

**9.4** Suppose there are two samples of size 6 and 7, with a rank sum of 58 in the sample of size 6. Using the Wilcoxon rank-sum test, evaluate the significance of the results, assuming there are no ties.

**9.5** Answer Problem 9.4 for two samples of size 7 and 10, with a rank sum of 47 in the sample of size 7. Assume there are no ties.

**9.6** Answer Problem 9.4 for two samples of size 12 and 15, with a rank sum of 220 in the sample of size 12. Assume there are no ties.

### Obstetrics

**9.7** Reanalyze the data in Table 8.17 using nonparametric methods. Assume the samples are unpaired.

**9.8** Would such methods be preferable to parametric methods in analyzing the data? Why or why not?

### Health Services Administration

Suppose we want to compare the length of hospital stay for patients with the same diagnosis at two different hospitals. The results are shown in Table 9.8.

**Table 9.8 Comparison of length of stay in 2 hospitals**

First	
hospital	21, 10, 32, 60, 8, 44, 29, 5, 13, 26, 33
Second	
hospital	86, 27, 10, 68, 87, 76, 125, 60, 35, 73, 96, 44, 238

\***9.9** Why might a *t* test not be very useful in this case?

\***9.10** Carry out a nonparametric procedure for testing the hypothesis that lengths of stay are comparable in the two hospitals.

### Infectious Disease

The distribution of white-blood-cell count is typically positively skewed, and assumptions of normality are usually not valid.

**9.11** To compare the distribution of white-blood-cell counts of patients on the medical and surgical services in Table 2.11 when normality is not assumed, what test can be used?

**9.12** Perform the test in Problem 9.11, and report a *p*-value.

### Sports Medicine

Refer to Data Set TENNIS2.DAT (on the Companion Website).

**9.13** What test can be used to compare degree of pain during maximal activity in the first period between people randomized to Motrin and a placebo?

**9.14** Perform the test in Problem 9.13, and report a *p*-value.

### Otolaryngology, Pediatrics

A common symptom of otitis media in young children is the prolonged presence of fluid in the middle ear, known as *middle-ear effusion*. The presence of fluid may result in temporary hearing loss and interfere with normal learning skills in the first 2 years of life. One hypothesis is that babies who are breastfed for at least 1 month build up some immunity against the effects of the infection and have less prolonged effusion than do bottle-fed babies. A small study of 24 pairs

of babies is set up, in which the babies are matched on a one-to-one basis according to age, sex, socioeconomic status, and type of medications taken. One member of the matched pair is a breastfed baby, and the other member is a bottle-fed baby. The outcome variable is the duration of middle-ear effusion after the first episode of otitis media. The results are given in Table 9.9.

**Table 9.9 Duration of middle-ear effusion in breast-fed and bottle-fed babies**

Pair number	Duration of effusion in breastfed baby (days)	Duration of effusion in bottle-fed baby (days)
1	20	18
2	11	35
3	3	7
4	24	182
5	7	6
6	28	33
7	58	223
8	7	7
9	39	57
10	17	76
11	17	186
12	12	29
13	52	39
14	14	15
15	12	21
16	30	28
17	7	8
18	15	27
19	65	77
20	10	12
21	7	8
22	19	16
23	34	28
24	25	20

\***9.15** What hypotheses are being tested here?

\***9.16** Why might a nonparametric test be useful in testing the hypotheses?

\***9.17** Which nonparametric test should be used here?

\***9.18** Test the hypothesis that the duration of effusion is different among breastfed babies than among bottle-fed babies using a nonparametric test.

## Hypertension

Polyunsaturated fatty acids in the diet favorably affect several risk factors for cardiovascular disease. The

principal dietary polyunsaturated fat is linoleic acid. To test the effects of dietary supplementation with linoleic acid on blood pressure, 17 adults consumed 23 g/day of safflower oil, high in linoleic acid, for 4 weeks. Systolic blood pressure (SBP) measurements were taken at baseline (before ingestion of oil) and 1 month later, with the mean values over several readings at each visit given in Table 9.10.

**Table 9.10 Effect of linoleic acid on SBP**

Subject	Baseline SBP	1-month SBP	Baseline – 1-month SBP
1	119.67	117.33	2.34
2	100.00	98.78	1.22
3	123.56	123.83	-0.27
4	109.89	107.67	2.22
5	96.22	95.67	0.55
6	133.33	128.89	4.44
7	115.78	113.22	2.56
8	126.39	121.56	4.83
9	122.78	126.33	-3.55
10	117.44	110.39	7.05
11	111.33	107.00	4.33
12	117.33	108.44	8.89
13	120.67	117.00	3.67
14	131.67	126.89	4.78
15	92.39	93.06	-0.67
16	134.44	126.67	7.77
17	108.67	108.67	0.00

**9.19** What parametric test could be used to test for the effect of linoleic acid on SBP?

**9.20** Perform the test in Problem 9.19, and report a *p*-value.

**9.21** What nonparametric test could be used to test for the effect of linoleic acid on SBP?

**9.22** Perform the test in Problem 9.21, and report a *p*-value.

**9.23** Compare your results in Problems 9.20 and 9.22, and discuss which method you feel is more appropriate here.

## Hypertension

An instrument that is in fairly common use in blood-pressure epidemiology is the random-zero device, in which the zero point of the machine is randomly set with each use and the observer is not aware of the actual level of blood pressure at the time of measurement. This instrument is intended to reduce observer bias. Before using such a machine, it is important to check that readings

are, on average, comparable to those of a standard cuff. For this purpose, two measurements were made on 20 children with both the standard cuff and the random-zero machine. The mean systolic blood pressure (SBP) for the two readings for each machine are given in Table 9.11. Suppose observers are reluctant to assume that the distribution of blood pressure is normal.

**Table 9.11 Comparison of mean SBP with the standard cuff vs. the random-zero machine (mm Hg)**

Person	Mean SBP (standard cuff)	Mean SBP (random-zero)
1	79	84
2	112	99
3	103	92
4	104	103
5	94	94
6	106	106
7	103	97
8	97	108
9	88	77
10	113	94
11	98	97
12	103	103
13	105	107
14	117	120
15	94	94
16	88	87
17	101	97
18	98	93
19	91	87
20	105	104

\***9.24** Which nonparametric test should be used to test the hypothesis that the mean SBPs for the two machines are comparable?

**9.25** Conduct the test recommended in Problem 9.24.

Another aspect of the same study is to compare the variability of blood pressure with each method. This comparison is achieved by computing  $|x_1 - x_2|$  for each participant and method (i.e., absolute difference between first and second readings) and comparing absolute differences between machines for individual participants. The data are given in Table 9.12. The observers are reluctant to assume that the distributions are normal.

\***9.26** Which nonparametric test should be used to test the hypothesis that variability of the two machines is comparable?

**9.27** Conduct the test recommended in Problem 9.26.

**Table 9.12 Comparison of variability of SBP with the standard cuff and the random-zero machine (mm Hg)**

Person	Absolute difference, standard cuff ( $a_s$ )	Absolute difference, random-zero ( $a_r$ )
1	2	12
2	4	6
3	6	0
4	4	2
5	8	4
6	4	4
7	2	6
8	2	8
9	4	2
10	2	4
11	0	6
12	2	6
13	6	6
14	2	4
15	8	8
16	0	2
17	6	6
18	4	6
19	2	14
20	2	4

### Health Promotion

Refer to Data Set SMOKE.DAT on the Companion Website.

**9.28** Use nonparametric methods to test whether there is a difference between the number of days abstinent from smoking for males vs. females.

**9.29** Divide the data set into age groups (above/below the median), and use nonparametric methods to test whether the number of days abstinent from smoking is related to age.

**9.30** Use the same approach as in Problem 9.29 to test whether the amount previously smoked is related to the number of days abstinent from smoking.

**9.31** Use the same approach as in Problem 9.29 to test whether the adjusted carbon monoxide (CO) level is related to the number of days abstinent from smoking.

**9.32** Why are nonparametric methods well suited to a study of risk factors for smoking cessation?

### Nutrition

Refer to the urinary-sodium data in Table 8.20.

**9.33** Use nonparametric methods to assess whether dietary counseling is effective in reducing sodium intake as judged by urinary-sodium excretion levels.

## Hepatic Disease

Refer to Data Set HORMONE.DAT on the Companion Website.

**9.34** Use nonparametric methods to answer Problem 8.82.

**9.35** Use nonparametric methods to answer Problem 8.83.

**9.36** Use nonparametric methods to answer Problem 8.84.

**9.37** Compare your results in Problems 9.34–9.36 with the corresponding results using parametric methods in Problems 8.82–8.84.

## Ophthalmology

Refer to the data set in Tables 7.7 and 7.8 (pp. 263–264).

**9.38** Answer the question in Problem 7.71 using nonparametric methods.

**9.39** Implement the test suggested in Problem 9.38, and report a two-sided *p*-value.

**9.40** Compare the results in Problem 9.39 with those obtained in Problem 7.72.

## Endocrinology

Refer to Data Set BONEDEN.DAT on the Companion Website.

**9.41** Answer the question in Problem 7.73 using nonparametric methods, and compare your results with those obtained using parametric methods.

**9.42** Answer the question in Problem 7.74 using nonparametric methods, and compare your results with those obtained using parametric methods.

## Infectious Disease

**9.43** Reanalyze the data in Table 8.18 using nonparametric methods, and compare your results with those obtained in Problem 8.51.

## Microbiology

Refer to the data in Table 8.31 (p. 318).

**9.44** What nonparametric test can be used to compare the distribution of pod weight for inoculated vs. noninoculated plants?

**9.45** Implement the test in Problem 9.44, and report a two-tailed *p*-value.

**9.46** Compare your results with those obtained in Problem 8.119.

## Diabetes

Growth during adolescence among girls with diabetes has been shown to relate to consistency in taking insulin injections. A similar hypothesis was tested in adolescent boys ages 13–17. Boys were seen at repeated visits approximately 90 days apart. Their weight and HgbA1c, a marker that

reflects consistency in taking insulin injections over the past 30 days, were measured at each visit. People with diabetes have a higher-than-normal HgbA1c; the goal of insulin treatment is to lower the HgbA1c level as much as possible. To look at the relationship between change in weight and change in HgbA1c, each of 23 boys was ascertained during one 90-day interval when HgbA1c change was minimal (i.e., change of <1%) (control period) and during another 90-day interval when HgbA1c increased by ≥1% (lack-of-consistency period); this is a fairly large increase, indicating lack of consistency in taking insulin injections. These data represent a subset of the data in DIABETES.DAT on the Companion Website. Weight change was compared between these intervals using the following measure:

$$\Delta = (\text{weight change, control period}) - (\text{weight change, lack-of-consistency period})$$

A frequency distribution of the data sorted in increasing order of  $\Delta$  is shown in Table 9.13.

Suppose we assume normality of the change scores in Table 9.13.

**Table 9.13 (Weight change, control period) – (weight change, lack-of-consistency period) among 23 adolescent diabetic boys**

<i>i</i>	$\Delta_i$	<i>i</i>	$\Delta_i$	<i>i</i>	$\Delta_i$
1	-12.6	9	+2.2	17	+11.5
2	-10.3	10	+3.5	18	+12.2
3	-5.9	11	+4.8	19	+13.9
4	-5.4	12	+5.4	20	+14.2
5	-4.5	13	+5.8	21	+18.0
6	-2.7	14	+6.0	22	+18.6
7	-1.8	15	+6.7	23	+21.7
8	+0.3	16	+9.6		
				Mean	4.83
				sd	9.33
				<i>n</i>	23

**9.47** What test can be used to compare weight change during the control period vs. weight change during the lack-of-consistency period?

**9.48** Implement the test in Problem 9.47, and report a two-tailed *p*-value.

**9.49** Answer the question in Problem 9.47 if we are not willing to make the assumption of normality.

**9.50** Implement the test in Problem 9.49, and report a two-tailed *p*-value.

## Cancer

Serum estradiol is an important risk factor for breast cancer in postmenopausal women. To better understand the etiology

of breast cancer, serum-estradiol samples were collected from 25 premenopausal women (at about the same time period of the menstrual cycle) of whom 10 were Caucasian and 15 were African-American. Data were collected on both serum estradiol as well as body mass index (BMI) = weight (kg)/height<sup>2</sup> (m<sup>2</sup>), which is an important measure of overall obesity. The data are shown in Table 9.14.

The distribution of serum estradiol is usually highly skewed (especially for premenopausal women), and we are reluctant to assume normality.

**9.51** What test can we use to compare the distribution of serum estradiol for Caucasian vs. African-American women?

**Table 9.14 Relationship of serum estradiol, BMI, and ethnicity in 25 premenopausal women**

ID	Serum estradiol (pg/mL)	Estradiol rank	BMI	Ethnic group*
1	94	25	18.9	0
2	54	20	19.7	1
3	31	9.5	20.7	0
4	21	5	23.4	1
5	46	18	23.3	1
6	56	21	25.1	0
7	18	3	35.6	1
8	19	4	26.2	1
9	12	1	22.3	1
10	14	2	20.4	0
11	25	7	21.7	0
12	35	12	20.0	1
13	22	6	21.0	1
14	71	23	21.8	0
15	43	16	32.7	1
16	35	12	23.6	1
17	42	15	24.7	1
18	50	19	23.2	1
19	44	17	33.9	1
20	41	14	20.6	0
21	28	8	24.7	0
22	65	22	26.3	0
23	31	9.5	20.1	0
24	35	12	22.5	1
25	91	24	29.1	1

\*1 = African-American, 0 = Caucasian.

**9.52** Implement the test in Problem 9.51, and report a two-tailed *p*-value.

Another important variable in the epidemiology of breast cancer is BMI, which has been found to be related to both serum estradiol and ethnicity in previous studies.

**9.53** Suppose we want to compare mean BMI between Caucasian and African-American premenopausal women based on the data in the table and are willing to assume the distribution of BMI is approximately normal. What test can we use to make this comparison?

(Note that for Caucasian women, mean BMI = 22.0, *sd* = 2.47, *n* = 10; for African-American women, mean BMI = 25.4, *sd* = 5.01, *n* = 15.)

**9.54** Implement the test in Problem 9.53, and report a two-tailed *p*-value.

## Ophthalmology

Refer to Data Set TEAR.DAT on the Companion Website. We want to compare tear break-up time (TBUT) immediately after eye-drop instillation vs. TBUT before instillation. For this purpose, we will compute the average TBUT over both eyes and over two replicates for each eye (that is, the summary score is an average of four values). Also, we will only use data with a blink period of 6 seconds.

**9.55** What test can we use to perform the analysis if we don't want to assume that TBUT is normally distributed?

**9.56** Implement the test in Problem 9.55, and report a *p*-value (two-tailed).

**9.57** Answer the question in Problem 9.56 comparing TBUT time 5 minutes after drop instillation vs. TBUT before instillation.

**9.58** Answer the question in Problem 9.56 comparing TBUT 10 minutes after drop instillation vs. TBUT before instillation.

**9.59** Answer the question in Problem 9.56 comparing TBUT 15 minutes after drop instillation vs. TBUT before instillation.

**9.60** Based on your results from Problems 9.55–9.59, do you think the response to the placebo eye drop is short-lasting or long-lasting?

## Endocrinology

A study to assess the effect of a low-fat diet on estrogen metabolism recruited 6 healthy women ages 21–32 [5]. The women were within 5% of their ideal body weight, were not participating in athletics, and were eating a typical American diet. For the first 4 weeks the women were fed a high-fat diet (40% of total calories from fat). They were then switched to a low-fat diet (25% of calories from fat) for 2 months. During the follicular phase of their menstrual cycle (days 5–7), each woman was given a sugar cube with [<sup>3</sup>H]E<sub>2</sub> (estradiol). This was done once during the high-fat period and again after the woman had been eating the low-fat diet for 2 months. The percentage of orally administered [<sup>3</sup>H]E<sub>2</sub> excreted in urine as 16 $\alpha$ -hydroxylated glucuronides is given in Table 9.15.

**Table 9.15 Percentage of orally administered [<sup>3</sup>H]E<sub>2</sub> excreted in urine as glucoronides of 16α-OHE<sub>1</sub>**

Subject	High-fat diet	Low-fat diet
1	2.55	1.27
2	2.92	1.60
3	1.71	0.53
4	4.00	1.02
5	0.77	0.74
6	1.03	0.67

**9.61** What parametric test can be used to compare the 16α-OHE<sub>1</sub> percentages on a high-fat diet vs. a low-fat diet?

**9.62** Perform the test in Problem 9.61 and report a *p*-value (two-tailed).

**9.63** What nonparametric test can be used to compare the 16α-OHE<sub>1</sub> percentages on a high-fat diet vs. a low-fat diet?

**9.64** Perform the test in Problem 9.63 and report a *p*-value (two-tailed).

## REFERENCES

- [1] Lehmann, E. L. (1998). *Nonparametrics: Statistical methods based on ranks*. New York: Springer.
- [2] Berson, E. L., Remulla, J. F. C., Rosner, B., Sandberg, M. A., & Weigel-DiFranco, C. (1996). Evaluation of patients with retinitis pigmentosa receiving electrical stimulation, ozonated blood, and ocular surgery in Cuba. *Archives of Ophthalmology*, 114, 560–563.
- [3] McNeil, D. R. (1967). Efficiency loss due to grouping in distribution free tests. *Journal of the American Statistical Association*, 62, 954–965.
- [4] Hollander, M., & Wolfe, D. (1999). *Nonparametric statistical methods*. 2nd ed. New York: Wiley.
- [5] Longcope, C., Gorbach, S., Goldin, B., Woods, M., Dwyer, J., Morrill, A., & Waram, J. (1987). The effect of a low fat diet on estrogen metabolism. *Journal of Clinical Endocrinology and Metabolism*, 64, 1246–1250.

# 10

## Hypothesis Testing: Categorical Data

### 10.1 Introduction

In Chapters 7 and 8, the basic methods of hypothesis testing for continuous data were presented. For each test, the data consisted of one or two samples, which were assumed to come from an underlying normal distribution(s); appropriate procedures were developed based on this assumption. In Chapter 9, the assumption of normality was relaxed and a class of nonparametric methods was introduced. Using these methods, we assumed that the variable under study can be ordered without assuming any underlying distribution.

If the variable under study is not continuous but is instead classified into categories, which may or may not be ordered, then different methods of inference should be used. Consider the problems in Examples 10.1 through 10.3.

#### Example 10.1

**Cancer** Suppose we are interested in the association between oral contraceptive (OC) use and the 5-year incidence of ovarian cancer from January 1, 2003 to January 1, 2008. Women who are disease-free on January 1, 2003 are classified into two OC-use categories as of that date: ever users and never users. We are interested in whether the 5-year incidence of ovarian cancer is different between ever users and never users. Hence, this is a two-sample problem comparing two binomial proportions, and the *t*-test methodology in Chapter 8 cannot be used because the outcome variable, the development of ovarian cancer, is a discrete variable with two categories (yes/no) rather than a continuous variable.

#### Example 10.2

**Cancer** Suppose the OC users in Example 10.1 are subdivided into “heavy” users, who have used the pill for 5 years or longer, and “light” users, who have used the pill for less than 5 years. We may be interested in comparing 5-year ovarian-cancer incidence rates among heavy users, light users, and nonusers. In this problem, *three* binomial proportions are being compared, and we need to consider methods comparing more than two binomial proportions.

#### Example 10.3

**Infectious Disease** The fitting of a probability model based on the Poisson distribution to the random variable defined by the annual number of deaths from polio in the United States during the period 1968–1977 has been discussed, as shown in Table 4.9. We want to develop a general procedure for testing the goodness of fit of this and other probability models on actual sample data.

In this chapter, methods of hypothesis testing for comparing two or more binomial proportions are developed. Methods for testing the goodness of fit of a previously specified probability model to actual data are also considered. We will also consider relationships between categorical and nonparametric approaches to data analysis.

## 10.2 Two-Sample Test for Binomial Proportions

### Example 10.4

**Cancer** A hypothesis has been proposed that breast cancer in women is caused in part by events that occur between the age at menarche (the age when menstruation begins) and the age at first childbirth. The hypothesis is that the risk of breast cancer increases as the length of this time interval increases. If this theory is correct, then an important risk factor for breast cancer is age at first birth. This theory would explain in part why the incidence of breast cancer seems higher for women in the upper socioeconomic groups, because they tend to have their children relatively late in reproductive life.

An international study was set up to test this hypothesis [1]. Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did *not* have breast cancer. All women were asked about their age at first birth.

The set of women with at least one birth was arbitrarily divided into two categories: (1) women whose age at first birth was  $\leq 29$  years and (2) women whose age at first birth was  $\geq 30$  years. The following results were found among women with at least one birth: 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth  $\geq 30$ . How can we assess whether this difference is significant or simply due to chance?

Let  $p_1$  = the probability that age at first birth is  $\geq 30$  in case women with at least one birth and  $p_2$  = the probability that age at first birth is  $\geq 30$  in control women with at least one birth. The question is whether the underlying probability of having an age at first birth of  $\geq 30$  is different in the two groups. This problem is equivalent to testing the hypothesis  $H_0: p_1 = p_2 = p$  vs.  $H_1: p_1 \neq p_2$  for some constant  $p$ .

Two approaches for testing the hypothesis are presented. The first approach uses normal-theory methods similar to those developed in Chapter 8 and discussed on page 354. The second approach uses contingency-table methods, which are discussed on page 362. These two approaches are *equivalent* in that they always yield the same  $p$ -values, so which one is used is a matter of convenience.

### Normal-Theory Method

It is reasonable to base the significance test on the difference between the sample proportions ( $\hat{p}_1 - \hat{p}_2$ ). If this difference is very different from 0 (either positive or negative), then  $H_0$  is rejected; otherwise,  $H_0$  is accepted. The samples will be assumed large enough so that the *normal approximation to the binomial distribution is valid*. Then, under  $H_0$ ,  $\hat{p}_1$  is normally distributed with mean  $p$  and variance  $pq/n_1$ , and  $\hat{p}_2$  is normally distributed with mean  $p$  and variance  $pq/n_2$ . Therefore, from Equation 5.10, because the samples are independent,  $\hat{p}_1 - \hat{p}_2$  is normally distributed with mean 0 and variance

$$\frac{pq}{n_1} + \frac{pq}{n_2} = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

If we divide  $\hat{p}_1 - \hat{p}_2$  by its standard error,

$$\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

then under  $H_0$ ,

**Equation 10.1** 
$$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{pq(1/n_1 + 1/n_2)} \sim N(0,1)$$

The problem is that  $p$  and  $q$  are unknown, and thus the denominator of  $z$  cannot be computed unless some estimate for  $p$  is found. The best estimator for  $p$  is based on a weighted average of the sample proportions  $\hat{p}_1, \hat{p}_2$ . This weighted average, referred to as  $\hat{p}$ , is given by

**Equation 10.2** 
$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

where  $x_1$  = the observed number of events in the first sample and  $x_2$  = the observed number of events in the second sample. This estimate makes intuitive sense because each of the sample proportions is weighted by the number of people in the sample. Thus we substitute the estimate  $\hat{p}$  in Equation 10.2 for  $p$  in Equation 10.1. Finally, to better accommodate the normal approximation to the binomial, a continuity correction is introduced in the numerator of Equation 10.1. If  $\hat{p}_1 \geq \hat{p}_2$ , then we subtract  $\left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)$ ; if  $\hat{p}_1 < \hat{p}_2$  then we add  $\left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)$ . Equivalently, we can rewrite the numerator in terms of  $|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)$  and reject  $H_0$  only for large positive values of  $z$ . This suggests the following test procedure.

**Equation 10.3**

#### Two-Sample Test for Binomial Proportions (Normal-Theory Test)

To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ , where the proportions are obtained from two independent samples, use the following procedure:

- (1) Compute the test statistic

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where  $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$ ,  $\hat{q} = 1 - \hat{p}$

and  $x_1, x_2$  are the number of events in the first and second samples, respectively.

(2) For a two-sided level  $\alpha$  test,

if  $z > z_{1-\alpha/2}$

then reject  $H_0$ ;

if  $z \leq z_{1-\alpha/2}$

then accept  $H_0$ .

(3) The approximate  $p$ -value for this test is given by

$$p = 2[1 - \Phi(z)]$$

(4) Use this test only when the normal approximation to the binomial distribution is valid for each of the two samples—that is, when  $n_1\hat{p}\hat{q} \geq 5$  and  $n_2\hat{p}\hat{q} \geq 5$ .

The acceptance and rejection regions for this test are shown in Figure 10.1. Computation of the exact  $p$ -value is illustrated in Figure 10.2.

### Example 10.5

**Cancer** Assess the statistical significance of the results from the international study in Example 10.4.

#### Solution

The sample proportion of case women whose age at first birth was  $\geq 30$  is  $683/3220 = .212 = \hat{p}_1$ , and the sample proportion of control women whose age at first birth was  $\geq 30$  is  $1498/10,245 = .146 = \hat{p}_2$ . To compute the test statistic  $z$  in Equation 10.3, the estimated common proportion  $\hat{p}$  must be computed, which is given by

$$\hat{p} = (683 + 1498)/(3220 + 10,245) = 2181/13,465 = .162$$

$$\hat{q} = 1 - .162 = .838$$

Note that

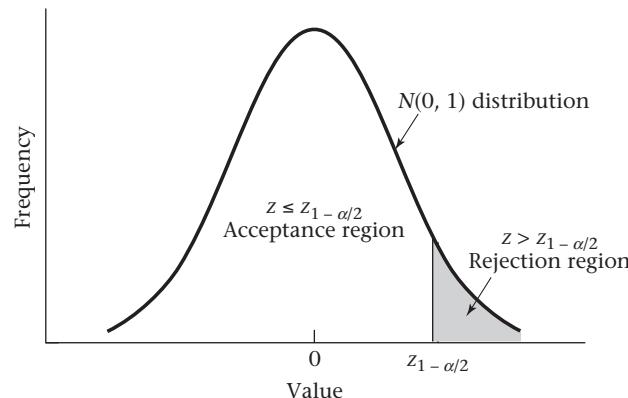
$$n_1\hat{p}\hat{q} = 3220(.162)(.838) = 437 \geq 5$$

$$\text{and } n_2\hat{p}\hat{q} = 10,245(.162)(.838) = 1391 \geq 5$$

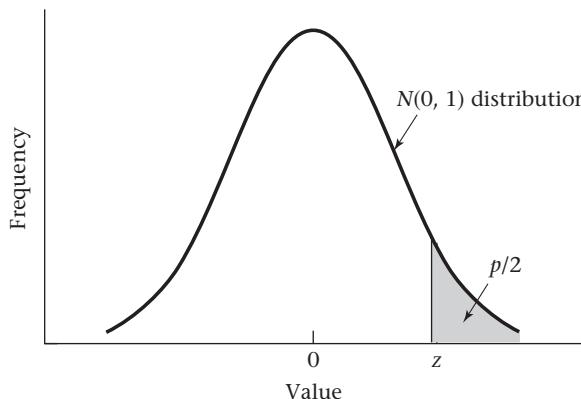
Thus the test in Equation 10.3 can be used.

**Figure 10.1**

**Acceptance and rejection regions for the two-sample test for binomial proportions (normal-theory test)**



**Figure 10.2** Computation of the exact *p*-value for the two-sample test for binomial proportions (normal-theory test)



The test statistic is given by

$$\begin{aligned} z &= \left\{ |.212 - .146| - \left[ \frac{1}{2(3220)} + \frac{1}{2(10,245)} \right] \right\} / \sqrt{.162(.838)\left(\frac{1}{3220} + \frac{1}{10,245}\right)} \\ &= .0657/.00744 \\ &= 8.8 \end{aligned}$$

The *p*-value =  $2 \times [1 - \Phi(8.8)] < .001$ , and the results are highly significant. Therefore, we can conclude that women with breast cancer are significantly more likely to have had their first child after age 30 than are comparable women without breast cancer.

### Example 10.6

**Cardiovascular Disease** A study looked at the effects of OC use on heart disease in women 40 to 44 years of age. The researchers found that among 5000 current OC users at baseline, 13 women developed a myocardial infarction (MI) over a 3-year period, whereas among 10,000 non-OC users, 7 developed an MI over a 3-year period. Assess the statistical significance of the results.

#### Solution

Note that  $n_1 = 5000$ ,  $\hat{p}_1 = 13/5000 = .0026$ ,  $n_2 = 10,000$ ,  $\hat{p}_2 = 7/10,000 = .0007$ . We want to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ . The best estimate of the common proportion  $p$  is given by

$$\hat{p} = \frac{13+7}{15,000} = \frac{20}{15,000} = .00133$$

Because  $n_1 \hat{p} \hat{q} = 5000(.00133)(.99867) = 6.7$ ,  $n_2 \hat{p} \hat{q} = 10,000(.00133)(.99867) = 13.3$ , the normal-theory test in Equation 10.3 can be used. The test statistic is given by

$$z = \frac{|.0026 - .0007| - \left[ \frac{1}{2(5000)} + \frac{1}{2(10,000)} \right]}{\sqrt{.00133(.99867)(1/5000 + 1/10,000)}} = \frac{.00175}{.00063} = 2.77$$

The *p*-value is given by  $2 \times [1 - \Phi(2.77)] = .006$ . Thus there is a highly significant difference between MI incidence rates for current OC users vs. non-OC users. In other words, OC use is significantly associated with MI incidence over a 3-year period.

## Contingency-Table Method

The same test posed in this section on page 353 is now approached from a different perspective.

### Example 10.7

**Cancer** Suppose all women with at least one birth in the international study in Example 10.4 are classified as either cases or controls and with age at first birth as either  $\leq 29$  or  $\geq 30$ . The four possible combinations are shown in Table 10.1.

**Table 10.1**

**Data for the international study in Example 10.4 comparing age at first birth in breast-cancer cases with comparable controls**

Status	Age at first birth		Total
	$\geq 30$	$\leq 29$	
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

Source: Reprinted with permission from *WHO Bulletin*, 43, 209–221, 1970.

Case-control status is displayed along the rows of the table, and age at first birth groups are presented in the columns of the table. Hence, each woman falls into one of the four boxes, or *cells*, of the table. In particular, there are 683 women with breast cancer whose age at first birth is  $\geq 30$ , 2537 women with breast cancer whose age at first birth is  $\leq 29$ , 1498 control women whose age at first birth is  $\geq 30$ , and 8747 control women whose age at first birth is  $\leq 29$ . Furthermore, the number of women in each row and column can be totaled and displayed in the margins of the table. Thus, there are 3220 case women ( $683 + 2537$ ), 10,245 control women ( $1498 + 8747$ ), 2181 women with age at first birth  $\geq 30$  ( $683 + 1498$ ), and 11,284 women with age at first birth  $\leq 29$  ( $2537 + 8747$ ). These sums are referred to as row margins and column margins, respectively. Finally, the total number of units = 13,465 is given in the lower right-hand corner of the table; this total can be obtained either by summing the four cells ( $683 + 2537 + 1498 + 8747$ ) or by summing the row margins ( $3220 + 10,245$ ) or the column margins ( $2181 + 11,284$ ). This sum is sometimes referred to as the *grand total*.

Table 10.1 is called a  $2 \times 2$  *contingency table* because it has two categories for case-control status and two categories for age-at-first-birth status.

### Definition 10.1

A  $2 \times 2$  *contingency table* is a table composed of two rows cross-classified by two columns. It is an appropriate way to display data that can be classified by two different variables, *each* of which has only two possible outcomes. One variable is arbitrarily assigned to the rows and the other to the columns. Each of the four cells represents the number of units (women, in the previous example), with a specific value for each of the two variables. The cells are sometimes referred to by number, with the (1, 1) cell being the cell in the first row and first column, the (1, 2) cell being the cell in the first row and second column, the (2, 1) cell being the cell in the second row and first column, and the (2, 2) cell being the cell in the second row and second column. The observed number of units in the four cells is likewise referred to as  $O_{11}$ ,  $O_{12}$ ,  $O_{21}$ , and  $O_{22}$ , respectively.

Furthermore, it is customary to total

- (1) The number of units in each row and display them in the right margins, which are called **row marginal totals** or **row margins**.
- (2) The number of units in each column and display them in the bottom margins, which are called **column marginal totals** or **column margins**.
- (3) The total number of units in the four cells, which is displayed in the lower right-hand corner of the table and is called the **grand total**.

**Example 10.8**

**Cardiovascular Disease** Display the MI data in Example 10.6 in the form of a  $2 \times 2$  contingency table.

**Solution**

Let the rows of the table represent the OC-use group, with the first row representing current OC users and the second row representing non-OC users. Let the columns of the table represent MI, with the first column representing Yes and the second column representing No. We studied 5000 current OC users, of whom 13 developed MI and 4987 did not. We studied 10,000 non-OC users, of whom 7 developed MI and 9993 did not. Thus the contingency table should look like Table 10.2.

**Table 10.2**  $2 \times 2$  contingency table for the OC–MI data in Example 10.6

OC-use group	MI status over 3 years		Total
	Yes	No	
OC users	13	4987	5000
Non-OC users	7	9993	10,000
Total	20	14,980	15,000

Two different sampling designs lend themselves to a contingency-table framework. The breast-cancer data in Example 10.4 have two independent samples (i.e., case women and control women), and we want to compare the proportion of women in each group who have a first birth at a late age. Similarly, in the OC–MI data in Example 10.6 there are two independent samples of women with different contraceptive-use patterns and we want to compare the proportion of women in each group who develop an MI. In both instances, we want to test whether the proportions are the same in the two independent samples. This test is called a **test for homogeneity of binomial proportions**. In this situation, one set of margins is fixed (e.g., the rows) and the number of successes in each row is a random variable. For example, in Example 10.4 the total number of breast-cancer cases and controls is fixed, and the number of women with age at first birth  $\geq 30$  is a binomial random variable conditional on the fixed-row margins (i.e., 3220 cases and 10,245 controls).

Another possible design from which contingency tables arise is in testing for the independence of two characteristics in the same sample when neither characteristic is particularly appropriate as a denominator. In this setting, both sets of margins are assumed to be fixed. The number of units in one particular cell of the table [e.g., the (1, 1) cell] is a random variable, and all other cells can be determined from the fixed margins and the (1, 1) cell. An example of this design is given in Example 10.9.

**Example 10.9**

**Nutrition** The food-frequency questionnaire is widely used to measure dietary intake. A person specifies the number of servings consumed per day of each of many different food items. The total nutrient composition is then calculated from the specific dietary components of each food item. One way to judge how well a questionnaire measures dietary intake is by its reproducibility. To assess reproducibility the questionnaire is administered at two different times to 50 people and the reported nutrient intakes from the two questionnaires are compared. Suppose dietary cholesterol is quantified on each questionnaire as high if it exceeds 300 mg/day and as normal otherwise. The contingency table in Table 10.3 is a natural way to compare the results of the two surveys. Notice that this example has no natural denominator. We simply want to test whether there is some association between the two reported measures of dietary cholesterol for the same person. More specifically, we want to assess how unlikely it is that 15 women will report high dietary cholesterol intake on both questionnaires, given that 20 of 50 women report high intake on the first questionnaire and 24 of 50 women report high intake on the second questionnaire. This test is called a **test of independence** or a **test of association** between the two characteristics.

**Table 10.3** A comparison of dietary cholesterol assessed by a food-frequency questionnaire at two different times

		Second food-frequency questionnaire		Total
		High	Normal	
First food-frequency questionnaire	High	15	5	20
	Normal	9	21	30
Total		24	26	50

Fortunately, the same test procedure is used whether a test of homogeneity or a test of independence is performed, so we will no longer distinguish between these two tests in this section.

### Significance Testing Using the Contingency-Table Approach

Table 10.1 is an **observed contingency table** or an **observed table**. To determine statistical significance, we need to develop an **expected table**, which is the contingency table that would be expected if there were no relationship between breast cancer and age at first birth—that is, if  $H_0: p_1 = p_2 = p$  were true. In this example  $p_1$  and  $p_2$  are the probabilities (among women with at least one birth) of a breast-cancer case and a control, respectively, having a first birth at an age  $\geq 30$ . For this purpose, a general observed table, if there were  $x_1$  exposed out of  $n_1$  women with breast cancer and  $x_2$  exposed out of  $n_2$  control women, is given in Table 10.4.

If  $H_0$  were true, then the best estimate of the common proportion  $p$  is  $\hat{p}$ , which is given in Equation 10.2 as

$$(n_1 \hat{p}_1 + n_2 \hat{p}_2) / (n_1 + n_2)$$

or as  $(x_1 + x_2) / (n_1 + n_2)$

**Table 10.4 General contingency table for the international-study data in Example 10.4 if (1) of  $n_1$  women in the case group,  $x_1$  are exposed and (2) of  $n_2$  women in the control group,  $x_2$  are exposed (that is, having an age at first birth  $\geq 30$ )**

Case-control status	Age at first birth		Total
	$\geq 30$	$\leq 29$	
Case	$x_1$	$n_1 - x_1$	$n_1$
Control	$x_2$	$n_2 - x_2$	$n_2$
Total	$x_1 + x_2$	$n_1 + n_2 - (x_1 + x_2)$	$n_1 + n_2$

where  $x_1$  and  $x_2$  are the number of exposed women in groups 1 and 2, respectively. Furthermore, under  $H_0$  the expected number of units in the (1, 1) cell equals the expected number of women with age at first birth  $\geq 30$  among women with breast cancer, which is given by

$$n_1 \hat{p} = n_1 (x_1 + x_2) / (n_1 + n_2)$$

However, in Table 10.4 this number is simply the product of the first row margin ( $n_1$ ) multiplied by the first column margin ( $x_1 + x_2$ ), divided by the grand total ( $n_1 + n_2$ ). Similarly, the expected number of units in the (2, 1) cell equals the expected number of control women with age at first birth  $\geq 30$ :

$$n_2 \hat{p} = n_2 (x_1 + x_2) / (n_1 + n_2)$$

which is equal to the product of the second row margin multiplied by the first column margin, divided by the grand total. In general, the following rule can be applied.

#### Equation 10.4

##### Computation of Expected Values for $2 \times 2$ Contingency Tables

The **expected number of units** in the  $(i, j)$  cell, which is usually denoted by  $E_{ij}$ , is the product of the **i**th row margin multiplied by the **j**th column margin, divided by the grand total.

#### Example 10.10

**Cancer** Compute the expected table for the breast-cancer data in Example 10.4.

#### Solution

Table 10.1 gives the observed table for these data. The row totals are 3220 and 10,245; the column totals are 2181 and 11,284; and the grand total is 13,465. Thus

$$\begin{aligned} E_{11} &= \text{expected number of units in the (1, 1) cell} \\ &= 3220(2181)/13,465 = 521.6 \end{aligned}$$

$$\begin{aligned} E_{12} &= \text{expected number of units in the (1, 2) cell} \\ &= 3220(11,284)/13,465 = 2698.4 \end{aligned}$$

$$\begin{aligned} E_{21} &= \text{expected number of units in the (2, 1) cell} \\ &= 10,245(2181)/13,465 = 1659.4 \end{aligned}$$

$$\begin{aligned} E_{22} &= \text{expected number of units in the (2, 2) cell} \\ &= 10,245(11,284)/13,465 = 8585.6 \end{aligned}$$

These expected values are shown in Table 10.5.

**Table 10.5** Expected table for the breast-cancer data in Example 10.4

Case-control status	Age at first birth		Total
	$\geq 30$	$\leq 29$	
Case	521.6	2698.4	3220
Control	1659.4	8585.6	10,245
Total	2181	11,284	13,465

**Example 10.11**

**Cardiovascular Disease** Compute the expected table for the OC–MI data in Example 10.6.

**Solution**

From Table 10.2, which gives the observed table for these data,

$$E_{11} = \frac{5000(20)}{15,000} = 6.7$$

$$E_{12} = \frac{5000(14,980)}{15,000} = 4993.3$$

$$E_{21} = \frac{10,000(20)}{15,000} = 13.3$$

$$E_{22} = \frac{10,000(14,980)}{15,000} = 9986.7$$

These expected values are displayed in Table 10.6.

**Table 10.6**

Expected table for the OC–MI data in Example 10.6

OC-use group	MI status over 3 years		Total
	Yes	No	
Current OC users	6.7	4993.3	5000
Non-OC users	13.3	9986.7	10,000
Total	20	14,980	15,000

We can show from Equation 10.4 that the *total* of the expected number of units in any row or column should be the same as the corresponding observed row or column total. This relationship provides a useful check that the expected values are computed correctly.

**Example 10.12**

Check that the expected values in Table 10.5 are computed correctly.

**Solution**

The following information is given:

- (1) The total of the expected values in the first row  $= E_{11} + E_{12} = 521.6 + 2698.4 = 3220$  = first row total in the observed table.
- (2) The total of the expected values in the second row  $= E_{21} + E_{22} = 1659.4 + 8585.6 = 10,245$  = second row total in the observed table.

- (3) The total of the expected values in the first column =  $E_{11} + E_{21} = 521.6 + 1659.4 = 2181$  = first column total in the observed table.
- (4) The total of the expected values in the second column =  $E_{12} + E_{22} = 2698.4 + 8585.6 = 11,284$  = second column total in the observed table.

We now want to compare the observed table in Table 10.1 with the expected table in Table 10.5. If the corresponding cells in these two tables are close, then  $H_0$  will be accepted; if they differ enough, then  $H_0$  will be rejected. How should we decide how different the cells should be for us to reject  $H_0$ ? It can be shown that the best way of comparing the cells in the two tables is to use the statistic  $(O - E)^2/E$ , where  $O$  and  $E$  are the observed and expected number of units, respectively, in a particular cell. In particular, under  $H_0$  it can be shown that the sum of  $(O - E)^2/E$  over the four cells in the table approximately follows a chi-square distribution with 1 degree of freedom ( $df$ ).  $H_0$  is rejected only if this sum is large and is accepted otherwise because small values of this sum correspond to good agreement between the two tables, whereas large values correspond to poor agreement. This test procedure will be used only when the normal approximation to the binomial distribution is valid. In this setting the normal approximation can be shown to be approximately true if no expected value in the table is less than 5.

Furthermore, under certain circumstances a version of this test statistic with a *continuity correction* yields more accurate *p*-values than does the uncorrected version when approximated by a chi-square distribution. For the continuity-corrected version, the statistic  $\left(\left|O - E\right| - \frac{1}{2}\right)^2/E$  rather than  $(O - E)^2/E$  is computed for each cell and the preceding expression is summed over the four cells. This test procedure is called the Yates-corrected chi-square and is summarized as follows.

### Equation 10.5

#### **Yates-Corrected Chi-Square Test for a $2 \times 2$ Contingency Table**

Suppose we wish to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$  using a contingency-table approach, where  $O_{ij}$  represents the observed number of units in the  $(i, j)$  cell and  $E_{ij}$  represents the expected number of units in the  $(i, j)$  cell.

- (1) Compute the test statistic

$$X^2 = \left(\left|O_{11} - E_{11}\right| - .5\right)^2/E_{11} + \left(\left|O_{12} - E_{12}\right| - .5\right)^2/E_{12} \\ + \left(\left|O_{21} - E_{21}\right| - .5\right)^2/E_{21} + \left(\left|O_{22} - E_{22}\right| - .5\right)^2/E_{22}$$

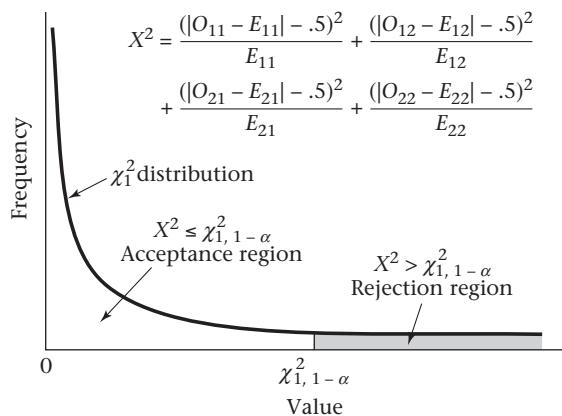
which under  $H_0$  approximately follows a  $\chi_1^2$  distribution.

- (2) For a level  $\alpha$  test, reject  $H_0$  if  $X^2 > \chi_{1,1-\alpha}^2$  and accept  $H_0$  if  $X^2 \leq \chi_{1,1-\alpha}^2$ .
- (3) The approximate *p*-value is given by the area to the right of  $X^2$  under a  $\chi_1^2$  distribution.
- (4) Use this test only if none of the four expected values is less than 5.

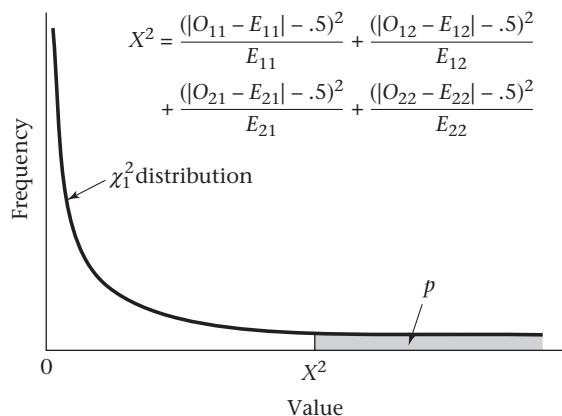
The acceptance and rejection regions for this test are shown in Figure 10.3. Computation of the *p*-value is illustrated in Figure 10.4.

The Yates-corrected chi-square test is a *two-sided* test even though the critical region, based on the chi-square distribution, is one-sided. The rationale is that large values of  $|O_{ij} - E_{ij}|$  and, correspondingly, of the test statistic  $X^2$  will be obtained under  $H_1$  regardless of whether  $p_1 < p_2$  or  $p_1 > p_2$ . Small values of  $X^2$  are evidence in favor of  $H_0$ .

**Figure 10.3** Acceptance and rejection regions for the Yates-corrected chi-square test for a  $2 \times 2$  contingency table



**Figure 10.4** Computation of the  $p$ -value for the Yates-corrected chi-square test for a  $2 \times 2$  contingency table



**Example 10.13** **Cancer** Assess the breast-cancer data in Example 10.4 for statistical significance, using a contingency-table approach.

**Solution**

First compute the observed and expected tables as given in Tables 10.1 and 10.5, respectively. Check that all expected values in Table 10.5 are at least 5, which is clearly the case. Thus, following Equation 10.5,

$$\begin{aligned} \chi^2 &= \frac{(|683 - 521.6| - .5)^2}{521.6} + \frac{(|2537 - 2698.4| - .5)^2}{2698.4} \\ &\quad + \frac{(|1498 - 1659.4| - .5)^2}{1659.4} + \frac{(|8747 - 8585.6| - .5)^2}{8585.6} \\ &= \frac{160.9^2}{521.6} + \frac{160.9^2}{2698.4} + \frac{160.9^2}{1659.4} + \frac{160.9^2}{8585.6} \\ &= 49.661 + 9.599 + 15.608 + 3.017 = 77.89 \sim \chi^2_1 \text{ under } H_0 \end{aligned}$$

Because  $\chi^2_{1,999} = 10.83 < 77.89 = X^2$ , we have  $p < 1 - .999 = .001$  and the results are extremely significant. Thus breast cancer incidence is significantly associated with having a first child after age 30.

**Example 10.14**

**Cardiovascular Disease** Assess the OC-MI data in Example 10.6 for statistical significance, using a contingency-table approach.

**Solution**

First compute the observed and expected tables as given in Tables 10.2 and 10.6, respectively. Note that the minimum expected value in Table 10.6 is 6.7, which is  $\geq 5$ . Thus the test procedure in Equation 10.5 can be used:

$$\begin{aligned} X^2 &= \frac{(|13 - 6.7| - .5)^2}{6.7} + \frac{(|4987 - 4993.3| - .5)^2}{4993.3} \\ &\quad + \frac{(|7 - 13.3| - .5)^2}{13.3} + \frac{(|9993 - 9986.7| - .5)^2}{9986.7} \\ &= \frac{5.8^2}{6.7} + \frac{5.8^2}{4993.3} + \frac{5.8^2}{13.3} + \frac{5.8^2}{9986.7} \\ &= 5.104 + 0.007 + 2.552 + 0.003 = 7.67 \sim \chi^2_1 \text{ under } H_0 \end{aligned}$$

Because  $\chi^2_{1,99} = 6.63$ ,  $\chi^2_{1,995} = 7.88$ , and  $6.63 < 7.67 < 7.88$  it follows that  $1 - .995 < p < 1 - .99$ , or  $.005 < p < .01$ , and the results are highly significant. Thus there is a significant difference between MI incidence rates for OC users and non-OC users among 40- to 44-year-old women, with OC users having higher rates.

The test procedures in Equation 10.3 and Equation 10.5 are equivalent in the sense that they always give the same  $p$ -values and always result in the same decisions about accepting or rejecting  $H_0$ . Which test procedure is used is a matter of convenience. Most researchers find the contingency-table approach more understandable, and results are more frequently reported in this format in the scientific literature.

At this time statisticians disagree widely about whether a continuity correction is needed for the contingency-table test in Equation 10.5. Generally,  $p$ -values obtained using the continuity correction are slightly larger. Thus results obtained are slightly less significant than comparable results obtained without using a continuity correction. However, the difference in results obtained using these two methods should be small for tables based on large sample sizes. The Yates-corrected test statistic is slightly more widely used in the applied literature, and therefore is used in this section. Another possible approach for performing hypothesis tests based on  $2 \times 2$  contingency tables is to use Fisher's exact test. This procedure is discussed in [Section 10.3](#).

### Short Computational Form for the Yates-Corrected Chi-Square Test for $2 \times 2$ Contingency Tables

The test statistic  $X^2$  in Equation 10.5 has another computational version that is more convenient to use with a hand calculator and does not require the computation of an expected table.

**Equation 10.6****Short Computational Form for the Yates-Corrected Chi-Square Test for  $2 \times 2$  Contingency Tables**

Suppose we have the  $2 \times 2$  contingency table in Table 10.7. The  $X^2$  test statistic in Equation 10.5 can be written

$$X^2 = n \left( |ad - bc| - \frac{n}{2} \right)^2 / [(a+b)(c+d)(a+c)(b+d)]$$

Thus the test statistic  $X^2$  depends only on (1) the grand total  $n$ , (2) the row and column margins  $a + b$ ,  $c + d$ ,  $a + c$ ,  $b + d$ , and (3) the magnitude of the quantity  $ad - bc$ . To compute  $X^2$ ,

- (1) Compute

$$\left( |ad - bc| - \frac{n}{2} \right)^2$$

Start with the first column margin, and proceed counterclockwise.

- (2) Divide by each of the two column margins.
- (3) Multiply by the grand total.
- (4) Divide by each of the two row margins.

This computation is particularly easy with a hand calculator because previous products and quotients can be maintained in the display and used for further calculations.

**Table 10.7** General contingency table

$a$	$b$	$a + b$
$c$	$d$	$c + d$
$a + c$	$b + d$	$n = a + b + c + d$

**Example 10.15**

**Nutrition** Compute the chi-square statistic for the nutrition data in Example 10.9 using the short computational form in Equation 10.6.

**Solution**

From Table 10.3,

$$a = 15 \quad b = 5 \quad c = 9 \quad d = 21 \quad n = 50$$

Furthermore, the smallest expected value =  $(24 \times 20)/50 = 9.6 \geq 5$ . Thus it is valid to use the chi-square test. Use the approach in Equation 10.6:

- (1) Compute

$$\begin{aligned} \left( |ad - bc| - \frac{n}{2} \right)^2 &= \left[ |15 \times 21 - 5 \times 9| - \frac{50}{2} \right]^2 \\ &= (270 - 25)^2 = 245^2 = 60,025 \end{aligned}$$

- (2) Divide the result in step 1 (60,025) by each of the two column margins (24 and 26), thus obtaining 96.194.

- (3) Multiply the result in step 2 (96.194) by the grand total (50), obtaining 4809.70.
- (4) Divide the result in step 3 (4809.70) by each of the two row margins (20 and 30), obtaining 8.02.

Because the critical value =  $\chi^2_{1,95} = 3.84$  and  $X^2 = 8.02 > 3.84$ , the results are statistically significant. To find a range for the  $p$ -value, note from the chi-square table that  $\chi^2_{1,995} = 7.88$ ,  $\chi^2_{1,999} = 10.83$ , and thus, because  $7.88 < 8.02 < 10.83$ ,  $.001 < p < .005$ .

These data have also been analyzed using the SPSS<sup>x</sup>/PC CROSSTABS program, as shown in Table 10.8. The program prints out the cell counts, the row and column totals and percentages, and the grand total. Furthermore, it prints out the minimum expected frequency (min E.F. = 9.6) and notes there are no expected frequencies  $< 5$ . Finally, the significance test is performed using both the Yates-corrected chi-square test (chi-square = 8.02,  $df = 1$ ,  $p$ -value = .0046) and the uncorrected chi-square test (chi-square = 9.74,  $df = 1$ ,  $p$ -value = .0018).

The results show a highly significant association between dietary-cholesterol intake reported by the same person at two different points in time. This is to be expected because, if the dietary instrument is at all meaningful, then there should be an association between the responses of the same person at two different points in time. We discuss measures of reproducibility for categorical data in more detail in Section 10.8.

In this section, we have discussed the two-sample test for binomial proportions. This is the analog to the two-sample  $t$  test for comparing means from two independent samples introduced in Chapter 8, except that here we are comparing proportions instead of means.

We refer to the master flowchart (Figure 10.16, p. 409). For all the methods in Chapter 10, we answer yes to (1) only one variable of interest? no to (2) one-sample problem?

**Table 10.8 Use of SPSS<sup>x</sup>/PC CROSSTABS program to analyze the nutrition data in Table 10.3**

SPSS <sup>x</sup> /PC Release 1.0					
Crosstabulation:		CHOL1 1ST FOOD FREQUENCY QUESTIONNAIRE			
		By CHOL2 2ND FOOD FREQUENCY QUESTIONNAIRE			
CHOL2→	Count	HIGH	NORMAL		
CHOL1		1.00	2.00	Row Total	
	1.00	15	5	20	40.0
HIGH		2.00	9	30	60.0
	Column Total	24	26	50	
	Total	48.0	52.0	100.0	
 Chi-Square D.F. Significance Min E.F. Cells with E.F.<5					
-----	----	-----	-----	-----	-----
8.01616	1	0.0046	9.600	None	
9.73558	1	0.0018	(Before Yates Correction)		
 Number of Missing Observations = 0					

yes to (3) two-sample problem? no to (4) underlying distribution normal or can central-limit theorem be assumed to hold? and yes to (5) underlying distribution binomial?

We now refer to the flowchart at the end of this chapter (p. 409). We answer yes to (1) are samples independent? (2) are all expected values  $\geq 5$ ? and (3)  $2 \times 2$  contingency table? This leads us to the box labeled “Use the two-sample test for binomial proportions or  $2 \times 2$  contingency-table methods if no confounding is present, or Mantel-Haenszel test if confounding is present.” In brief, a confounder is another variable that is potentially related to both the row and column classification variables, and it must be controlled for. We discuss methods for controlling for confounding in Chapter 13. In this chapter, we assume no confounding is present. Thus we use either the two-sample test for binomial proportions (Equation 10.3) or the equivalent chi-square test for  $2 \times 2$  contingency tables (Equation 10.5).

### REVIEW QUESTIONS 10A

- 1** What is a contingency table?
- 2** Suppose we have 50 ovarian-cancer cases and 100 controls, all of whom are age 50–54. Ten of the ovarian-cancer cases and 12 of the controls reached menarche (age when periods begin) at  $< 11$  years.
  - (a)** What test can be used to assess whether there is a significant association between early age at menarche and ovarian cancer?
  - (b)** Perform the test in Review Question 10A.2a, and report a two-tailed *p*-value.

## 10.3 Fisher's Exact Test

In Section 10.2, we discussed methods for comparing two binomial proportions using either normal-theory or contingency-table methods. Both methods yield identical *p*-values. However, they require that the normal approximation to the binomial distribution be valid, which is not always the case, especially for small samples.

### Example 10.16

**Cardiovascular Disease, Nutrition** Suppose we want to investigate the relationship between high salt intake and death from cardiovascular disease (CVD). Groups of high- and low-salt users could be identified and followed over a long time to compare relative frequency of death from CVD in the two groups. In contrast, a much less expensive study would involve looking at death records, separating CVD deaths from non-CVD deaths, asking a close relative (such as a spouse) about the dietary habits of the deceased, and then comparing salt intake between people who died of CVD vs. people who died of other causes.

The latter type of study, a retrospective study, may be impossible to perform for a number of reasons. But if it is possible, it is almost always less expensive than the former type, a prospective study.

### Example 10.17

**Cardiovascular Disease, Nutrition** Suppose a retrospective study is done among men ages 50–54 in a specific county who died over a 1-month period. The investigators try to include approximately an equal number of men who died from CVD (the cases) and men who died from other causes (the controls). Of 35 people who died from CVD, 5 were on a high-salt diet before they died, whereas of 25 people who died from other causes 2 were on such a diet. These

data, presented in Table 10.9, are in the form of a  $2 \times 2$  contingency table, so the methods of Section 10.2 may be applicable.

**Table 10.9**
**Data concerning the possible association between cause of death and high salt intake**

Cause of death	Type of diet		Total
	High salt	Low salt	
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

However, the expected values of this table are too small for such methods to be valid. Indeed,

$$E_{11} = 7(25)/60 = 2.92$$

$$E_{12} = 7(35)/60 = 4.08$$

thus two of the four cells have expected values less than 5. How should the possible association between cause of death and type of diet be assessed?

In this case, **Fisher's exact test** can be used. This procedure gives exact levels of significance for any  $2 \times 2$  table but is only necessary for tables with small expected values, tables in which the standard chi-square test as given in Equation 10.5 is not applicable. For tables in which use of the chi-square test is appropriate, the two tests give very similar results. Suppose the probability that a man was on a high-salt diet given that his cause of death was noncardiovascular (non-CVD) =  $p_1$  and the probability that a man was on a high-salt diet given that his cause of death was cardiovascular (CVD) =  $p_2$ . We wish to test the hypothesis  $H_0: p_1 = p_2 = p$  vs.  $H_1: p_1 \neq p_2$ . Table 10.10 gives the general layout of the data.

**Table 10.10**
**General layout of data for Fisher's exact test example**

Cause of death	Type of diet		Total
	High salt	Low salt	
Non-CVD	a	b	a + b
CVD	c	d	c + d
Total	a + c		n
	b + d		

For mathematical convenience, we assume the margins of this table are *fixed*; that is, the numbers of non-CVD deaths and CVD deaths are fixed at  $a + b$  and  $c + d$ , respectively, and the numbers of people on high- and low-salt diets are fixed at  $a + c$  and  $b + d$ , respectively. Indeed, it is difficult to compute exact probabilities unless one assumes fixed margins. The exact probability of observing the table with cells  $a, b, c, d$  is as follows.

**Equation 10.7****Exact Probability of Observing a Table with Cells  $a, b, c, d$** 

$$Pr(a, b, c, d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

The formula in Equation 10.7 is easy to remember because the numerator is the product of the factorials of each of the row and column margins, and the denominator is the product of the factorial of the grand total and the factorials of the individual cells.

**Example 10.18**

Suppose we have the  $2 \times 2$  table shown in Table 10.11. Compute the exact probability of obtaining this table assuming the margins are fixed.

**Solution**

$$Pr(2, 5, 3, 1) = \frac{7!4!5!6!}{11!2!5!3!1!} = \frac{5040(24)(120)(720)}{39,916,800(2)(120)(6)} = \frac{1.0450944 \times 10^{10}}{5.7480192 \times 10^{10}} = .182$$

**Table 10.11****Hypothetical  $2 \times 2$  contingency table in Example 10.18**

2	5	7
3	1	4
5	6	11

**The Hypergeometric Distribution**

Suppose we consider all possible tables with fixed row margins denoted by  $N_1$  and  $N_2$  and fixed column margins denoted by  $M_1$  and  $M_2$ . We assume the rows and columns have been rearranged so that  $M_1 \leq M_2$  and  $N_1 \leq N_2$ . We refer to each table by its  $(1, 1)$  cell because all other cells are then determined from the fixed row and column margins. Let the random variable  $X$  denote the cell count in the  $(1, 1)$  cell. The probability distribution of  $X$  is given by

**Equation 10.8**

$$Pr(X = a) = \frac{N_1!N_2!M_1!M_2!}{N!a!(N_1-a)!(M_1-a)!(M_2-N_1+a)!}, a = 0, \dots, \min(M_1, N_1)$$

and  $N = N_1 + N_2 = M_1 + M_2$ . This probability distribution is called the **hypergeometric distribution**.

It will be useful for our subsequent work on combining evidence from more than one  $2 \times 2$  table in Chapter 13 to refer to the expected value and variance of the hypergeometric distribution. These are as follows.

**Equation 10.9****Expected Value and Variance of the Hypergeometric Distribution**

Suppose we consider all possible tables with fixed row margins  $N_1, N_2$  and fixed column margins  $M_1, M_2$ , where  $N_1 \leq N_2$ ,  $M_1 \leq M_2$ , and  $N = N_1 + N_2 = M_1 + M_2$ . Let the random variable  $X$  denote the cell count in the  $(1, 1)$  cell. The expected value and variance of  $X$  are

$$E(X) = \frac{M_1 N_1}{N}$$

$$Var(X) = \frac{M_1 M_2 N_1 N_2}{N^2 (N - 1)}$$

Thus the exact probability of obtaining a table with cells  $a, b, c, d$  in Equation 10.7 is a special case of the hypergeometric distribution, where  $N_1 = a + b$ ,  $N_2 = c + d$ ,  $M_1 = a + c$ ,  $M_2 = b + d$ , and  $N = a + b + c + d$ . We can evaluate this probability by calculator using Equation 10.7, or we can use the HYPGEOMDIST function of Excel. In the latter case, to evaluate  $Pr(a, b, c, d)$ , we specify HYPGEOMDIST ( $a, a + b, a + c, N$ ). In words, the hypergeometric distribution evaluates the probability of obtaining  $a$  successes out of a sample of  $a + b$  observations, given that the total population (in this case, the two samples combined), is of size  $N$ , of which  $a + c$  observations are successes. Thus, to evaluate the exact probability in Table 10.11, we specify HYPGEOMDIST (2, 7, 5, 11) = .182, which is the probability of obtaining two successes in a sample of 7 observations given that the total population consists of 11 observations, of which 5 are successes. The hypergeometric distribution differs from the binomial distribution, because in the latter case, we simply evaluate the probability of obtaining  $a$  successes out of  $a + b$  observations, assuming that each outcome is independent. For the hypergeometric distribution, the outcomes are not independent because once a success occurs it is less likely that another observation will be a success, as the total number of successes is fixed (at  $a + c$ ). If  $N$  is large, the two distributions are very similar because there is only a slight deviation from independence for the hypergeometric.

The basic strategy in testing the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$  will be to enumerate all possible tables with the same margins as the observed table and to compute the exact probability for each such table based on the hypergeometric distribution. A method for accomplishing this is as follows.

#### Equation 10.10

##### Enumeration of All Possible Tables with the Same Margins as the Observed Table

- (1) Rearrange the rows and columns of the observed table so the smaller row total is in the first row and the smaller column total is in the first column.

Suppose that after the rearrangement, the cells in the observed table are  $a, b, c, d$ , as shown in Table 10.10.

- (2) Start with the table with 0 in the (1, 1) cell. The other cells in this table are then determined from the row and column margins. Indeed, to maintain the same row and column margins as the observed table, the (1, 2) element must be  $a + b$ , the (2, 1) cell must be  $a + c$ , and the (2, 2) element must be  $(c + d) - (a + c) = d - a$ .
- (3) Construct the next table by increasing the (1, 1) cell by 1 (i.e., from 0 to 1), decreasing the (1, 2) and (2, 1) cells by 1, and increasing the (2, 2) cell by 1.
- (4) Continue increasing and decreasing the cells by 1, as in step 3, until one of the cells is 0, at which point all possible tables with the given row and column margins have been enumerated. Each table in the sequence of tables is referred to by its (1, 1) element. Thus, the first table is the "0" table, the next table is the "1" table, and so on.

#### Example 10.19

**Cardiovascular Disease, Nutrition** Enumerate all possible tables with the same row and column margins as the observed data in Table 10.9.

**Solution**

The observed table has  $a = 2$ ,  $b = 23$ ,  $c = 5$ ,  $d = 30$ . The rows or columns do not need to be rearranged because the first row total is smaller than the second row total, and the first column total is smaller than the second column total. Start with the 0 table, which has 0 in the (1, 1) cell, 25 in the (1, 2) cell, 7 in the (2, 1) cell, and  $30 - 2$ , or 28, in the (2, 2) cell. The 1 table then has 1 in the (1, 1) cell,  $25 - 1 = 24$  in the (1, 2) cell,  $7 - 1 = 6$  in the (2, 1) cell, and  $28 + 1 = 29$  in the (2, 2) cell. Continue in this fashion until the 7 table is reached, which has 0 in the (2, 1) cell, at which point all possible tables with the given row and column margins have been enumerated. The set of hypergeometric probabilities in Table 10.12 can be easily evaluated using the recursive properties of Excel by (1) setting up a column with consecutive values from 0 to 7 (say from B1 to B8), (2) using the function HYPGEOMDIST to compute  $Pr(0) = \text{HYPGEOMDIST}(B1, 25, 7, 60)$  and placing it in C1, and then (3) dragging the cursor down column C to compute the remaining hypergeometric probabilities. See the Companion Website for more details on the use of the HYPGEOMDIST function. The collection of tables and their associated probabilities based on the hypergeometric distribution in Equation 10.8 are given in Table 10.12.

**Table 10.12** Enumeration of all possible tables with fixed margins and their associated probabilities, based on the hypergeometric distribution for Example 10.19

0	25	1	24	2	23	3	22
7	28	6	29	5	30	4	31
	.017		.105		.252		.312
4	21	5	20	6	19	7	18
3	32	2	33	1	34	0	35
	.214		.082		.016		.001

The question now is: What should be done with these probabilities to evaluate the significance of the results? The answer depends on whether a one-sided or a two-sided alternative is being used. In general, the following method can be used.

**Equation 10.11****Fisher's Exact Test: General Procedure and Computation of *p*-Value**

To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ , where the expected value of at least one cell is <5 when the data are analyzed in the form of a  $2 \times 2$  contingency table, use the following procedure:

- (1) Enumerate all possible tables with the same row and column margins as the observed table, as shown in Equation 10.10.
- (2) Compute the exact probability of each table enumerated in step 1, using either the computer or the formula in Equation 10.7.
- (3) Suppose the observed table is the  $a$  table and the last table enumerated is the  $k$  table.
  - (a) To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ , the *p*-value =  $2 \times \min[Pr(0) + Pr(1) + \dots + Pr(a), Pr(a) + Pr(a+1) + \dots + Pr(k), .5]$ .
  - (b) To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 < p_2$ , the *p*-value =  $Pr(0) + Pr(1) + \dots + Pr(a)$ .

- (c) To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 > p_2$ , the  $p$ -value =  $Pr(a) + Pr(a+1) + \dots + Pr(k)$ .

For each of these three alternative hypotheses, the  $p$ -value can be interpreted as the probability of obtaining a table as extreme as or more extreme than the observed table.

**Example 10.20**

**Cardiovascular Disease, Nutrition** Evaluate the statistical significance of the data in Example 10.17 using a two-sided alternative.

**Solution**

We want to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ . Our table is the 2 table whose probability is .252 in Table 10.12. Thus, to compute the  $p$ -value, the smaller of the tail probabilities corresponding to the 2 table is computed and doubled. This strategy corresponds to the procedures for the various normal-theory tests studied in Chapters 7 and 8. First compute the left-hand tail area,

$$Pr(0) + Pr(1) + Pr(2) = .017 + .105 + .252 = .375$$

and the right-hand tail area,

$$Pr(2) + Pr(3) + \dots + Pr(7) = .252 + .312 + .214 + .082 + .016 + .001 = .878$$

$$\text{Then } p = 2 \times \min(.375, .878, .5) = 2(.375) = .749$$

If a one-sided alternative of the form  $H_0: p_1 = p_2$  vs.  $H_1: p_1 < p_2$  is used, then the  $p$ -value equals

$$Pr(0) + Pr(1) + Pr(2) = .017 + .105 + .252 = .375$$

Thus the two proportions in this example are *not* significantly different with either a one-sided or two-sided test, and we *cannot* say, on the basis of this limited amount of data, that there is a significant association between salt intake and cause of death.

In most instances, computer programs are used to implement Fisher's exact test using statistical packages such as SAS. There are other possible approaches to significance testing in the two-sided case. For example, the approach used by SAS is to compute

$$p\text{-value (two-tailed)} = \sum_{\{i: Pr(i) \leq Pr(a)\}} Pr(i)$$

In other words, the two-tailed  $p$ -value using SAS is the sum of the probabilities of all tables whose probabilities are  $\leq$  the probability of the observed table. Using this approach, the two-tailed  $p$ -value would be

$$\begin{aligned} p\text{-value (two-tailed)} &= Pr(0) + Pr(1) + Pr(2) + Pr(4) + Pr(5) + Pr(6) + Pr(7) \\ &= .017 + .105 + .252 + .214 + .082 + .016 + .001 = .688 \end{aligned}$$

In this section, we learned about Fisher's exact test, which is used for comparing binomial proportions from two independent samples in  $2 \times 2$  tables with small expected counts ( $< 5$ ). This is the two-sample analog to the exact one-sample binomial test given in Equation 7.44. If we refer to the flowchart at the end of this chapter (Figure 10.16, p. 409), we answer yes to (1) are samples independent? and no to (2) are all expected values  $\geq 5$ ? This leads us to the box labeled "Use Fisher's exact test."

**REVIEW QUESTIONS 10B**

- 1** What is the difference between the chi-square test for  $2 \times 2$  tables and Fisher's exact test? Under what circumstances do we use each test?
- 2** Suppose we have a  $2 \times 2$  table with cell counts  $a, b, c$ , and  $d$  as in Table 10.10. For each of the following tables, identify which test in Review Question 10B.1 we should use:
  - (a)**  $a = 5, b = 10, c = 11, d = 7$
  - (b)**  $a = 2, b = 3, c = 7, d = 10$
  - (c)**  $a = 1, b = 99, c = 10, d = 90$
- 3** Suppose that 2 of 4000 men and 3 of 3500 women (all ages 40–44) develop lung cancer over the next year. All these people have smoked 1 pack of cigarettes per day from age 18 to their current age. Perform a significance test to compare the incidence of lung cancer between 40- to 44-year-old men and women. Report a two-tailed  $p$ -value.

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

**Example 10.21**

**Cancer** Suppose we want to compare two different chemotherapy regimens for breast cancer after mastectomy. The two treatment groups should be as comparable as possible on other prognostic factors. To accomplish this goal, a matched study is set up such that a random member of each matched pair gets treatment A (chemotherapy) perioperatively (within 1 week after mastectomy) and for an additional 6 months, whereas the other member gets treatment B (chemotherapy only perioperatively). The patients are assigned to pairs matched on age (within 5 years) and clinical condition. The patients are followed for 5 years, with survival as the outcome variable. The data are displayed in a  $2 \times 2$  table, as in Table 10.13. Notice the small difference in 5-year survival between the two treatment groups: treatment A group =  $526/621 = .847$  vs. treatment B group =  $515/621 = .829$ . Indeed, the Yates-corrected chi-square statistic as given in Equation 10.5 is 0.59 with 1  $df$ , which is not significant. However, *use of this test is valid only if the two samples are independent*. From the manner in which the samples were selected it is obvious they are *not* independent because the two members of each matched pair are similar in age and clinical condition. Thus the Yates-corrected chi-square test *cannot* be used in this situation because the  $p$ -value will not be correct. How then can the two treatments be compared using a hypothesis test?

**Table 10.13** A  $2 \times 2$  contingency table comparing treatments A and B for breast cancer based on 1242 patients

Treatment	Outcome		Total
	Survive for 5 years	Die within 5 years	
A	526	95	621
B	515	106	621
Total	1041	201	1242

Suppose a different kind of  $2 \times 2$  table is constructed to display these data. In Table 10.13 the *person* was the unit of analysis, and the sample size was 1242 people. In Table 10.14 the *matched pair* is the unit of analysis and *pairs* are classified according to whether the members of that pair survived for 5 years. Notice that Table 10.14 has 621 units rather than the 1242 units in Table 10.13. Furthermore, there are 90 pairs in which both patients died within 5 years, 510 pairs in which both patients survived for 5 years, 16 pairs in which the treatment A patient survived and the treatment B patient died, and 5 pairs in which the treatment B patient survived and the treatment A patient died. The dependence of the two samples can be illustrated by noting that the probability that the treatment B member of the pair survived given that the treatment A member of the pair survived =  $510/526 = .970$ , and the probability that the treatment B member of the pair survived given that the treatment A member of the pair died =  $5/95 = .053$ . If the samples were independent, then these two probabilities should be about the same. Thus we conclude that the samples are highly dependent and the chi-square test cannot be used.

**Table 10.14** A  $2 \times 2$  contingency table with the matched pair as the sampling unit based on 621 matched pairs

		Outcome of treatment B patient		Total
Outcome of treatment A patient	Survive for 5 years	Survive for 5 years	Die within 5 years	
Survive for 5 years	510	16		526
Die within 5 years	5	90		95
Total	515	106		621

In Table 10.14, for 600 pairs ( $90 + 510$ ), the outcomes of the two treatments are the same, whereas for 21 pairs ( $16 + 5$ ), the outcomes of the two treatments are different. The following special names are given to each of these types of pairs:

---

**Definition 10.2** A **concordant pair** is a matched pair in which the outcome is the same for each member of the pair.

---



---

**Definition 10.3** A **discordant pair** is a matched pair in which the outcomes differ for the members of the pair.

---

**Example 10.22** | There are 600 concordant pairs and 21 discordant pairs for the data in Table 10.14.

The concordant pairs provide no information about *differences between treatments* and are not used in the assessment. Instead, we focus on the discordant pairs, which can be divided into two types.

---

**Definition 10.4** A **type A discordant pair** is a discordant pair in which the treatment A member of the pair has the event and the treatment B member does not. Similarly, a **type B**

**discordant pair** is a discordant pair in which the treatment B member of the pair has the event and the treatment A member does not.

**Example 10.23** If we define having an event as dying within 5 years, there are 5 type A discordant pairs and 16 type B discordant pairs for the data in Table 10.14.

Let  $p$  = the probability that a discordant pair is of type A. If the treatments are equally effective, then about an equal number of type A and type B discordant pairs would be expected, and  $p$  should =  $\frac{1}{2}$ . If treatment A is more effective than treatment B, then fewer type A than type B discordant pairs would be expected, and  $p$  should be  $< \frac{1}{2}$ . Finally, if treatment B is more effective than treatment A, then more type A than type B discordant pairs would be expected, and  $p$  should be  $> \frac{1}{2}$ .

Thus we wish to test the hypothesis  $H_0: p = \frac{1}{2}$  vs.  $H_1: p \neq \frac{1}{2}$ .

### Normal-Theory Test

Suppose that of  $n_D$  discordant pairs,  $n_A$  are type A. Then given that the observed number of discordant pairs =  $n_D$ , under  $H_0$ ,  $E(n_A) = n_D/2$  and  $Var(n_A) = n_D/4$ , from the mean and variance of a binomial distribution, respectively. Let's assume that the normal approximation to the binomial distribution holds, but use a continuity correction for a better approximation. This approximation will be valid if  $npq = n_D/4 \geq 5$  or  $n_D \geq 20$ . The following test procedure, called McNemar's test, can then be used.

#### Equation 10.12

#### McNemar's Test for Correlated Proportions—Normal-Theory Test

- (1) Form a  $2 \times 2$  table of matched pairs, where the outcomes for the treatment A members of the matched pairs are listed along the rows and the outcomes for the treatment B members are listed along the columns.
- (2) Count the total number of discordant pairs ( $n_D$ ) and the number of type A discordant pairs ( $n_A$ ).
- (3) Compute the test statistic

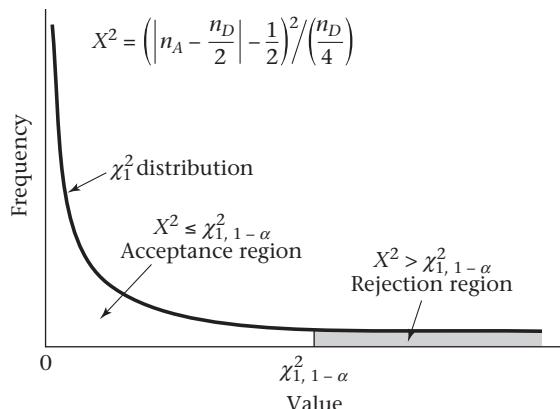
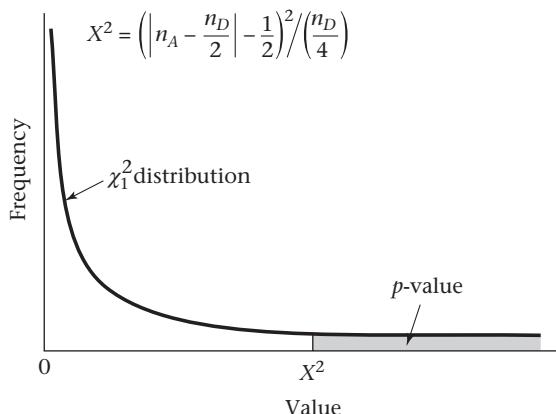
$$X^2 = \left( \left| n_A - \frac{n_D}{2} \right| - \frac{1}{2} \right)^2 / \left( \frac{n_D}{4} \right)$$

An equivalent version of the test statistic is also given by

$$X^2 = (|n_A - n_B| - 1)^2 / (n_A + n_B)$$

where  $n_B$  = number of type B discordant pairs.

- (4) For a two-sided level  $\alpha$  test,
  - if  $X^2 > \chi_{1,1-\alpha}^2$  then reject  $H_0$ ;
  - if  $X^2 \leq \chi_{1,1-\alpha}^2$  then accept  $H_0$ .
- (5) The exact  $p$ -value is given by  $p\text{-value} = Pr(\chi_1^2 \geq X^2)$ .
- (6) Use this test only if  $n_D \geq 20$ .

**Figure 10.5** Acceptance and rejection regions for McNemar's test—normal-theory method**Figure 10.6** Computation of the *p*-value for McNemar's test—normal-theory method

The acceptance and rejection regions for this test are shown in Figure 10.5. Computation of the *p*-value for McNemar's test is depicted in Figure 10.6.

This is a two-sided test despite the one-sided nature of the critical region in Figure 10.5. The rationale for this is that if either  $p < \frac{1}{2}$  or  $p > \frac{1}{2}$ ,  $|n_A - n_D/2|$  will be large and, correspondingly,  $X^2$  will be large. Thus, for alternatives on either side of the null hypothesis ( $p = \frac{1}{2}$ ),  $H_0$  is rejected if  $X^2$  is large and accepted if  $X^2$  is small.

**Example 10.24**

**Cancer** Assess the statistical significance of the data in Table 10.14.

**Solution**

Note that  $n_D = 21$ . Because  $n_D\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = 5.25 \geq 5$ , the normal approximation to the binomial distribution is valid and the test in Equation 10.12 can be used. We have

$$X^2 = \frac{\left(|5 - 10.5| - \frac{1}{2}\right)^2}{21/4} = \frac{\left(5.5 - \frac{1}{2}\right)^2}{5.25} = \frac{5^2}{5.25} = \frac{25}{5.25} = 4.76$$

Equivalently, we could also compute the test statistic from

$$X^2 = \frac{(|5 - 16| - 1)^2}{5 + 16} = \frac{10^2}{21} = 4.76$$

From Table 6 in the Appendix, note that

$$\begin{aligned}\chi^2_{1,95} &= 3.84 \\ \chi^2_{1,975} &= 5.02\end{aligned}$$

Thus, because  $3.84 < 4.76 < 5.02$ , it follows that  $.025 < p < .05$ , and the results are statistically significant. The exact  $p$ -value using Excel is  $p\text{-value} = \text{CHIDIST}(4.76, 1) = .029$ .

We conclude that *if the treatments give different results from each other* for the members of a matched pair, then the treatment A member of the pair is significantly more likely to survive for 5 years than the treatment B member. Thus, all other things being equal (such as toxicity, cost, etc.), treatment A would be the treatment of choice.

## Exact Test

If  $n_D/4 < 5$ —that is, if  $n_D < 20$ —then the normal approximation to the binomial distribution cannot be used, and a test based on exact binomial probabilities is required. The details of the test procedure are similar to the small sample one-sample binomial test in Equation 7.44 and are summarized as follows.

### Equation 10.13

#### McNemar's Test for Correlated Proportions—Exact Test

- (1) Follow the procedure in step 1 in Equation 10.12.
- (2) Follow the procedure in step 2 in Equation 10.12.
- (3) The exact  $p$ -value is given by

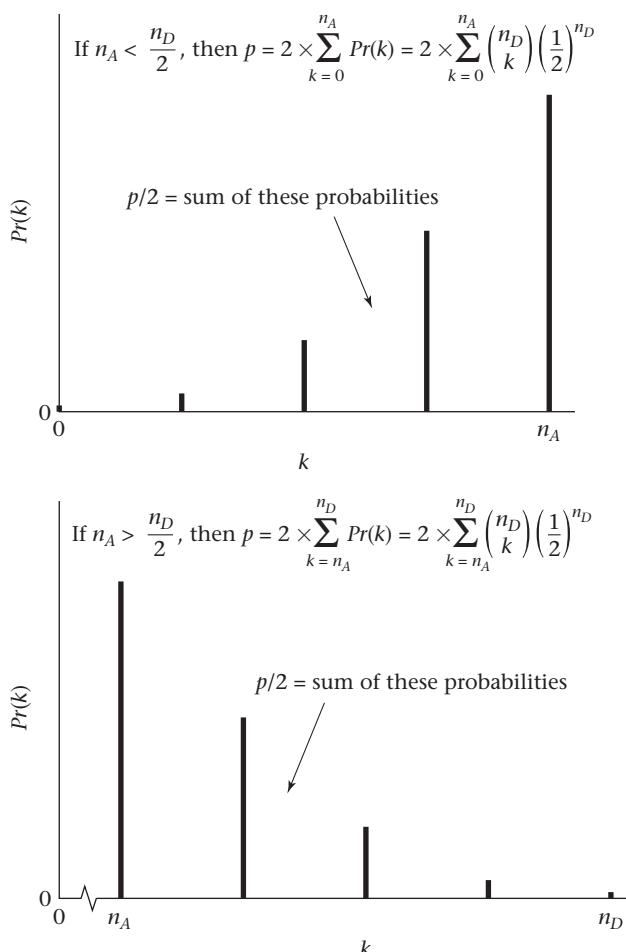
$$(a) \quad p = 2 \times \sum_{k=0}^{n_A} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D} \text{ if } n_A < n_D/2$$

$$(b) \quad p = 2 \times \sum_{k=n_A}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D} \text{ if } n_A > n_D/2$$

$$(c) \quad p = 1 \text{ if } n_A = n_D/2$$

- (4) This test is valid for any number of discordant pairs ( $n_D$ ) but is particularly useful for  $n_D < 20$ , when the normal-theory test in Equation 10.12 cannot be used.

The computation of the  $p$ -value for this test is shown in Figure 10.7.

**Figure 10.7 Computation of the  $p$ -value for McNemar's test—exact method****Example 10.25**

**Hypertension** A recent phenomenon in the recording of blood pressure is the development of the automated blood-pressure machine, where for a small fee a person can sit in a booth and have his or her blood pressure measured by a computer device. A study is conducted to compare the computer device with standard methods of blood pressure measurement. Twenty patients are recruited, and their hypertensive status is assessed by both the computer device and a trained observer. Hypertensive status is defined as either hypertensive (+) if systolic blood pressure is  $\geq 160$  mm Hg or higher or if diastolic blood pressure is  $\geq 95$  mm Hg or higher; the patient is considered normotensive (−) otherwise. The data are given in Table 10.15. Assess the statistical significance of these findings.

**Solution**

An ordinary Yates-corrected chi-square test cannot be used for these data because each person is being used as his or her own control and there are *not* two independent samples. Instead, a  $2 \times 2$  table of matched pairs is formed, as in Table 10.16. Note that 3 people are measured as hypertensive by both the computer device and the trained observer, 9 people are normotensive by both methods, 7 people are hypertensive by the computer device and normotensive by the trained

**Table 10.15** Hypertensive status of 20 patients as judged by a computer device and a trained observer

Person	Hypertensive status		Person	Hypertensive status	
	Computer device	Trained observer		Computer device	Trained observer
1	–	–	11	+	–
2	–	–	12	+	–
3	+	–	13	–	–
4	+	+	14	+	–
5	–	–	15	–	+
6	+	–	16	+	–
7	–	–	17	+	–
8	+	+	18	–	–
9	+	+	19	–	–
10	–	–	20	–	–

**Table 10.16** Comparison of hypertensive status as judged by a computer device and a trained observer

		Trained observer	
		+	–
Computer device	+	3	7
	–	1	9

observer, and 1 person is normotensive by the computer device and hypertensive by the trained observer. Therefore, there are 12 (9 + 3) concordant pairs and 8 (7 + 1) discordant pairs ( $n_D$ ). Because  $n_D < 20$ , the exact method must be used. We see that  $n_A = 7$ ,  $n_D = 8$ . Therefore, because  $n_A > n_D/2 = 4$ , it follows from Equation 10.13 that

$$p = 2 \times \sum_{k=7}^8 \binom{8}{k} \left(\frac{1}{2}\right)^8$$

This expression can be evaluated using Table 1 in the Appendix by referring to  $n = 8$ ,  $p = .5$  and noting that  $Pr(X \geq 7 | p = .5) = .0313 + .0039 = .0352$ . Thus, the two-tailed  $p$ -value =  $2 \times .0352 = .070$ .

Alternatively, a computer program could be used to perform the computations, as in Table 10.17. The first and second columns have been interchanged so the discordant pairs appear in the diagonal elements (and are easier to identify). In summary, the results are not statistically significant, and we cannot conclude that there is a significant difference between the two methods, although a *trend* can be detected toward the computer device identifying more hypertensives than the trained observer.

**Table 10.17** Use of SPSS<sup>X</sup>/PC McNemar's test program to evaluate the significance of the data in Table 10.16

SPSS <sup>X</sup> /PC Release 1.0					
-----McNemar Test					
		COMPUTER DEVICE			
with OBS		TRAINED OBSERVER			
		OBS			
		2.00      1.00		Cases	
COMP		1.00	7      3	(Binomial)	
		2.00	9      1	2-tailed P	
				0.0703	

Note that for a two-sided one-sample binomial test, the hypothesis  $H_0: p = p_0$  vs.  $H_1: p \neq p_0$  is tested. In the special case where  $p_0 = 1/2$ , the same test procedure as for McNemar's test is also followed.

Finally, if we interchange the designation of which of two outcomes is an event, then the *p*-values will be the same in Equations 10.12 and 10.13. For example, if we define an event as surviving for 5+ years, rather than dying within 5 years in Table 10.14, then  $n_A = 16$ ,  $n_B = 5$  (rather than  $n_A = 5$ ,  $n_B = 16$  in Example 10.23). However, the test statistic  $X^2$  and the *p*-value are the same because  $|n_A - n_B|$  remains the same in Equation 10.12. Similarly, the *p*-value remains the same in Equation 10.13 because of the symmetry of the binomial distribution when  $p = 1/2$  (under  $H_0$ ).

In this section, we have studied McNemar's test for correlated proportions, which is used to compare two binomial proportions from matched samples. We studied both a large-sample test when the normal approximation to the binomial distribution is valid (i.e., when the number of discordant pairs,  $n_D$ , is  $\geq 20$ ) and a small-sample test when  $n_D < 20$ . Referring to the flowchart at the end of this chapter (Figure 10.16, p. 409), we answer no to (1) are samples independent? which leads us to the box labeled "Use McNemar's test."

### REVIEW QUESTIONS 10C

- 1 **(a)** What is the difference between McNemar's test and the chi-square test for  $2 \times 2$  tables?
- 1 **(b)** When do we use each test?
- 2 What is a discordant pair? A concordant pair? Which type of pair is used in McNemar's test?
- 3 A twin design is used to study age-related macular degeneration (AMD), a common eye disease of the elderly that results in substantial losses in vision. Suppose we contact 66 twinships in which one twin has AMD and the other twin does not. The twins are given a dietary questionnaire to report their usual diet. We find that in 10 twinships the AMD twin takes multivitamin supplements and the normal twin does not. In 8 twinships the normal twin takes multivitamin supplements and the AMD twin does not. In 3 twinships both twins take multivitamin supplements, and in 45 twinships neither twin takes multivitamin supplements.
  - (a)** What test can be used to assess whether there is an association between AMD and taking multivitamin supplements?
  - (b)** Are AMD and taking multivitamin supplements significantly associated based on these data?

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

In Section 8.10, methods for estimating the sample size needed to compare means from two normally distributed populations were presented. In this section, similar methods for estimating the sample size required to compare two proportions are developed.

### Independent Samples

#### Example 10.26

**Cancer, Nutrition** Suppose we know from Connecticut tumor-registry data that the incidence rate of breast cancer over a 1-year period for initially disease-free women ages 45–49 is 150 cases per 100,000 [2]. We wish to study whether ingesting large doses of vitamin E in capsule form will prevent breast cancer. The study is set up with (1) a control group of 45- to 49-year-old women who are mailed placebo pills and are expected to have the same disease rate as indicated in the Connecticut tumor-registry data and (2) a study group of similar-age women who are mailed vitamin E pills and are expected to have a 20% reduction in risk. How large a sample is needed if a two-sided test with a significance level of .05 is used and a power of 80% is desired?

We want to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ . Suppose we want to conduct a test with significance level  $\alpha$  and power  $1 - \beta$  and we anticipate there will be  $k$  times as many people in group 2 as in group 1; that is,  $n_2 = kn_1$ . The sample size required in each of the two groups to achieve these objectives is as follows.

#### Equation 10.14

#### Sample Size Needed to Compare Two Binomial Proportions Using a Two-Sided Test with Significance Level $\alpha$ and Power $1 - \beta$ , Where One Sample ( $n_2$ ) Is $k$ Times as Large as the Other Sample ( $n_1$ ) (Independent-Sample Case)

To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$  for the specific alternative  $|p_1 - p_2| = \Delta$ , with a significance level  $\alpha$  and power  $1 - \beta$ , the following sample size is required

$$n_1 = \left[ \sqrt{\bar{p} \bar{q} \left( 1 + \frac{1}{k} \right)} z_{1-\alpha/2} + \sqrt{p_1 q_1 + \frac{p_2 q_2}{k}} z_{1-\beta} \right]^2 / \Delta^2$$

$$n_2 = kn_1$$

where  $p_1, p_2$  = projected true probabilities of success in the two groups

$$q_1, q_2 = 1 - p_1, 1 - p_2$$

$$\Delta = |p_2 - p_1|$$

$$\bar{p} = \frac{p_1 + kp_2}{1 + k}$$

$$\bar{q} = 1 - \bar{p}$$

#### Example 10.27

**Cancer, Nutrition** Estimate the sample size required for the study proposed in Example 10.26 if an equal sample size is anticipated in each group.

#### Solution

$$p_1 = 150 \text{ per } 100,000 \text{ or } 150/10^5 = .00150$$

$$q_1 = 1 - .00150 = .99850$$

If we want to detect a 20% reduction in risk, then  $p_2 = 0.8p_1$  or

$$p_2 = (150 \times .8)/10^5 = 120/10^5 = .00120$$

$$q_2 = 1 - .00120 = .99880$$

$$\alpha = .05$$

$$1 - \beta = .8$$

$$k = 1 (\text{because } n_1 = n_2)$$

$$\bar{p} = \frac{.00150 + .00120}{2} = .00135$$

$$\bar{q} = 1 - .00135 = .99865$$

$$z_{1-\alpha/2} = z_{.975} = 1.96$$

$$z_{1-\beta} = z_{.80} = 0.84$$

Thus, referring to Equation 10.14,

$$n_1 = \frac{\left[ \sqrt{.00135(.99865)(1+1)}(1.96) + \sqrt{.00150(.99850) + .00120(.99880)}(0.84) \right]^2}{(.00150 - .00120)^2}$$

$$= \frac{[.05193(1.96) + .05193(0.84)]^2}{.00030^2} = \frac{.14539^2}{.00030^2} = 234,881 = n_2$$

or about 235,000 women in each group.

To perform a one-tailed rather than a two-tailed test, simply substitute  $\alpha$  for  $\alpha/2$  in the sample-size formula in Equation 10.14.

Clearly, from the results in Example 10.27, we could not conduct such a large study over a 1-year period. The sample size needed would be reduced considerably if the period of study was lengthened beyond 1 year because the expected number of events would increase in a multiyear study.

In many instances, the sample size available for investigation is fixed by practical constraints, and what is desired is an estimate of statistical power with the anticipated available sample size. In other instances, after a study is completed, we want to calculate the power using the sample sizes that were actually used in the study. For these purposes the following estimate of power is provided to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$  with significance level  $\alpha$  and sample sizes of  $n_1$  and  $n_2$  in the two groups.

### Equation 10.15

#### Power Achieved in Comparing Two Binomial Proportions Using a Two-Sided Test with Significance Level $\alpha$ and Samples of Size $n_1$ and $n_2$ (Independent-Sample Case)

To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$  for the specific alternative  $|p_1 - p_2| = \Delta$ , compute

$$\text{Power} = \Phi \left[ \frac{\Delta}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}} - z_{1-\alpha/2} \frac{\sqrt{\bar{p} \bar{q} (1/n_1 + 1/n_2)}}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}} \right]$$

where

$p_1, p_2$  = projected true probabilities of success in groups 1 and 2, respectively

$$\begin{aligned}q_1, q_2 &= 1 - p_1, 1 - p_2 \\ \Delta &= |p_2 - p_1| \\ \bar{p} &= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \\ \bar{q} &= 1 - \bar{p}\end{aligned}$$

**Example 10.28**

**Otolaryngology** Suppose a study comparing a medical and a surgical treatment for children who have an excessive number of episodes of otitis media (OTM) during the first 3 years of life is planned. Success rates of 50% and 70% are assumed in the medical and surgical groups, respectively, and the recruitment of 100 patients for each group is realistically anticipated. Success is defined as  $\leq 1$  episode of OTM in the first 12 months after treatment. How much power does such a study have of detecting a significant difference if a two-sided test with an  $\alpha$  level of .05 is used?

**Solution**

Note that  $p_1 = .5$ ,  $p_2 = .7$ ,  $q_1 = .5$ ,  $q_2 = .3$ ,  $n_1 = n_2 = 100$ ,  $\Delta = .2$ ,  $\bar{p} = (.5 + .7)/2 = .6$ ,  $\bar{q} = .4$ ,  $\alpha = .05$ ,  $z_{1-\alpha/2} = z_{.975} = 1.96$ . Thus from Equation 10.15 the power can be computed as follows:

$$\begin{aligned}\text{Power} &= \Phi\left[\frac{.2}{\sqrt{[.5(.5) + .7(.3)]/100}} - \frac{1.96\sqrt{.6(.4)(1/100 + 1/100)}}{\sqrt{[.5(.5) + .7(.3)]/100}}\right] \\ &= \Phi\left[\frac{.2}{.0678} - 1.96 \frac{(.0693)}{.0678}\right] = \Phi(2.949 - 2.002) = \Phi(0.947) = .83\end{aligned}$$

Thus there is an 83% chance of finding a significant difference using the anticipated sample sizes.

If a one-sided test is used, then Equation 10.15 can be used after replacing  $z_{1-\alpha/2}$  by  $z_{1-\alpha}$ .

**Paired Samples**

In Section 10.4, McNemar's test for comparing binomial proportions in paired samples was introduced. As noted there, this test is a special case of the one-sample binomial test. Therefore, to estimate sample size and power, the more general formulas for the one-sample binomial test given in Section 7.10 can be used. Specifically, referring to Equation 7.46 to test the hypothesis  $H_0: p = p_0$  vs.  $H_1: p \neq p_0$  using a two-sided test with significance level  $\alpha$  and power  $1 - \beta$  for the specific alternative  $p = p_1$ , a sample size of

$$n = \frac{p_0 q_0 [z_{1-\alpha/2} + z_{1-\beta} \sqrt{p_1 q_1 / (p_0 q_0)}]^2}{(p_1 - p_0)^2}$$

is needed. To use this formula in the case of McNemar's test, set  $p_0 = q_0 = 1/2$ ,  $p_1 = p_A$  = the proportion of discordant pairs that are of type A, and  $n = n_D$  = the number of discordant pairs. On substitution,

$$n_D = \frac{(z_{1-\alpha/2} + 2z_{1-\beta} \sqrt{p_A q_A})^2}{4(p_A - .5)^2}$$

However, the number of discordant pairs ( $n_D$ ) = the total number of pairs ( $n$ )  $\times$  the probability that a matched pair is discordant. If the latter probability is denoted by  $p_D$ , then  $n_D = np_D$ , or  $n = n_D/p_D$ . Thus the following sample-size formula can be used.

**Equation 10.16****Sample Size Needed to Compare Two Binomial Proportions Using a Two-Sided Test with Significance Level  $\alpha$  and Power  $1 - \beta$  (Paired-Sample Case)**

If McNemar's test for correlated proportions is used to test the hypothesis  $H_0: p = \frac{1}{2}$  vs.  $H_1: p \neq \frac{1}{2}$ , for the specific alternative  $p = p_A$ , where  $p$  = the probability that a discordant pair is of type A, with a significance level of  $\alpha$  and power  $1 - \beta$ , then use

$$n = \frac{(z_{1-\alpha/2} + 2z_{1-\beta}\sqrt{p_A q_A})^2}{4(p_A - .5)^2 p_D} \text{ matched pairs}$$

$$\text{or } 2n = \frac{(z_{1-\alpha/2} + 2z_{1-\beta}\sqrt{p_A q_A})^2}{2(p_A - .5)^2 p_D} \text{ individuals}$$

where  $p_D$  = projected proportion of discordant pairs among all pairs  
 $p_A$  = projected proportion of discordant pairs of type A among discordant pairs

**Example 10.29**

**Cancer** Suppose we want to compare two different regimens of chemotherapy (A, B) for treatment of breast cancer where the outcome measure is recurrence of breast cancer or death over a 5-year period. A matched-pair design is used, in which patients are matched on age and clinical stage of disease, with one patient in a matched pair assigned to treatment A and the other to treatment B. Based on previous work, it is estimated that patients in a matched pair will respond similarly to the treatments in 85% of matched pairs (i.e., both will either die or have a recurrence or both will be alive and not have a recurrence over 5 years). Furthermore, for matched pairs in which there is a difference in response, it is estimated that in two-thirds of the pairs the treatment A patient will either die or have a recurrence, and the treatment B patient will not; in one-third of the pairs the treatment B patient will die or have a recurrence, and the treatment A patient will not. How many participants (or matched pairs) need to be enrolled in the study to have a 90% chance of finding a significant difference using a two-sided test with type I error = .05?

**Solution**

We have  $\alpha = .05$ ,  $\beta = .10$ ,  $p_D = 1 - .85 = .15$ ,  $p_A = \frac{2}{3}$ ,  $q_A = \frac{1}{3}$ . Therefore, from Equation 10.16,

$$\begin{aligned} n (\text{pairs}) &= \frac{\left[ z_{.975} + 2z_{.90} \sqrt{(2/3)(1/3)} \right]^2}{4(2/3 - 1/2)^2 (.15)} \\ &= \frac{\left[ 1.96 + 2(1.28)(.4714) \right]^2}{4(1/6)^2 (.15)} = \frac{3.1668^2}{.0167} = 602 \text{ matched pairs} \end{aligned}$$

$$2n = 2 \times 602 = 1204 \text{ individuals}$$

Therefore, 1204 women in 602 matched pairs must be enrolled. This will yield approximately  $.15 \times 602 = 90$  discordant pairs.

In some instances, sample size is fixed and we want to determine what power a study has (or had) to detect specific alternatives. For a two-sided one-sample binomial test with significance level  $\alpha$ , to test the hypothesis  $H_0: p = p_0$  vs.  $H_1: p \neq p_0$  for the specific alternative  $p = p_1$ , the power is given by (see Equation 7.45)

$$\text{Power} = \Phi \left[ \sqrt{p_0 q_0 / (p_1 q_1)} \left( z_{\alpha/2} + \frac{|p_1 - p_0| \sqrt{n}}{\sqrt{p_0 q_0}} \right) \right]$$

For McNemar's test, set  $p_0 = q_0 = \frac{1}{2}$ ,  $p_1 = p_A$ , and  $n = n_D$ , yielding

$$\text{Power} = \Phi \left[ \frac{1}{2\sqrt{p_A q_A}} \left( z_{\alpha/2} + 2|p_A - .5| \sqrt{n_D} \right) \right]$$

On substituting  $n_D = np_D$ , the following power formula is obtained.

### Equation 10.17

#### Power Achieved in Comparing Two Binomial Proportions Using a Two-Sided Test with Significance Level $\alpha$ (Paired-Sample Case)

If McNemar's test for correlated proportions is used to test the hypothesis  $H_0: p = 1/2$  vs.  $H_1: p \neq 1/2$ , for the specific alternative  $p = p_A$ , where  $p$  = the probability that a discordant pair is of type A,

$$\text{Power} = \Phi \left[ \frac{1}{2\sqrt{p_A q_A}} \left( z_{\alpha/2} + 2|p_A - .5| \sqrt{np_D} \right) \right]$$

where

$n$  = number of matched pairs

$p_D$  = projected proportion of discordant pairs among all pairs

$p_A$  = projected proportion of discordant pairs of type A among discordant pairs

### Example 10.30

**Cancer** Consider the study in Example 10.29. If 400 matched pairs are enrolled, how much power would such a study have?

#### Solution

We have  $\alpha = .05$ ,  $p_D = .15$ ,  $p_A = \frac{2}{3}$ ,  $n = 400$ . Therefore, from Equation 10.17,

$$\begin{aligned} \text{Power} &= \Phi \left[ \frac{1}{2\sqrt{(2/3)(1/3)}} \left[ z_{.025} + 2|2/3 - .5| \sqrt{400(.15)} \right] \right] \\ &= \Phi \{1.0607[-1.96 + 2(1/6)(7.7460)]\} \\ &= \Phi[1.0607(0.6220)] = \Phi(0.660) = .745 \end{aligned}$$

Therefore, the study would have 74.5% power, or a 74.5% chance of detecting a statistically significant difference.

To compute sample size and power for a one-sided alternative, substitute  $\alpha$  for  $\alpha/2$  in the formulas in Equations 10.16 and 10.17.

Note that a crucial element in calculating sample size and power for matched-pair studies based on binomial proportions using Equations 10.16 and 10.17 is knowledge of the probability of discordance between outcome for members of a matched pair ( $p_D$ ). This probability depends on the strictness of the matching criteria and on how strongly related the matching criteria are to the outcome variable.

Also, the methods in the paired sample case are for matched studies with 1:1 matching (i.e., in Example 10.29, each treatment A patient was matched to a single treatment B patient). Dupont [3] discusses more advanced methods of power calculation for matched studies with  $m:1$  matching (i.e.,  $m$  controls per case).

## Sample Size and Power in a Clinical Trial Setting

In Examples 10.27 and 10.28, we have estimated sample size and power in proposed clinical trials assuming that compliance with (ability to follow) treatment regimens is perfect. To be more realistic, we should examine how these estimates will change if compliance is not perfect.

Suppose we are planning a clinical trial comparing an active treatment vs. placebo. There are potentially two types of noncompliance to consider.

---

### Definition 10.5

The **dropout rate** is defined as the proportion of participants in the active-treatment group who fail to actually receive the active treatment.

---



---

### Definition 10.6

The **drop-in rate** is defined as the proportion of participants in the placebo group who actually receive the active treatment outside the study protocol.

---

### Example 10.31

**Cardiovascular Disease** The Physicians' Health Study was a randomized clinical trial, one goal of which was to assess the effect of aspirin in preventing myocardial infarction (MI). Participants were 22,000 male physicians ages 40–84 and free of cardiovascular disease in 1982. The physicians were randomized to either active aspirin (one white pill containing 325 mg of aspirin taken every other day) or aspirin placebo (one white placebo pill taken every other day). As the study progressed, it was estimated from self-report that 10% of the participants in the aspirin group were not complying (that is, were not taking their study [aspirin] capsules). Thus the dropout rate was 10%. Also, it was estimated from self-report that 5% of the participants in the placebo group were taking aspirin regularly on their own outside the study protocol. Thus the drop-in rate was 5%. The issue is: How does this lack of compliance affect the sample size and power estimates for the study?

Let  $\lambda_1$  = dropout rate,  $\lambda_2$  = drop-in rate,  $p_1$  = incidence of MI over a 5-year period among physicians who actually take aspirin, and  $p_2$  = incidence of MI over a 5-year period among physicians who don't take aspirin under the assumption of perfect compliance. Finally, let  $p_1^*$ ,  $p_2^*$  = observed rate of MI over a 5-year period in the aspirin and placebo groups, respectively (i.e., assuming that compliance is not perfect). We can estimate  $p_1^*$ ,  $p_2^*$  using the total-probability rule. Specifically,

### Equation 10.18

$$\begin{aligned} p_1^* &= \Pr(\text{MI} | \text{assigned to aspirin group}) \\ &= \Pr(\text{MI} | \text{aspirin-group complier}) \times \Pr(\text{compliance in the aspirin group}) \end{aligned}$$

$$\begin{aligned}
 & + Pr(\text{MI}|\text{aspirin-group noncomplier}) \times Pr(\text{noncompliance in the aspirin group}) \\
 & = p_1(1 - \lambda_1) + p_2\lambda_1
 \end{aligned}$$

Here we have assumed that the observed incidence rate for a noncompliant participant in the aspirin group =  $p_2$ .

**Equation 10.19**

$$\begin{aligned}
 \text{Similarly, } p_2^* &= Pr(\text{MI}|\text{assigned to placebo group}) \\
 &= Pr(\text{MI}|\text{placebo-group complier}) \times Pr(\text{compliance in the placebo group}) \\
 & + Pr(\text{MI}|\text{placebo-group noncomplier}) \times Pr(\text{noncompliance in the placebo group}) \\
 & = p_2(1 - \lambda_2) + p_1\lambda_2
 \end{aligned}$$

Here we have assumed that noncompliance in the placebo group means that the participant takes aspirin on his own and that such a participant has incidence rate =  $p_1$  = rate for aspirin-group compliers. Placebo-group participants who don't take their study capsules and refrain from taking aspirin outside the study are considered compliers from the viewpoint of the preceding discussion; that is, their incidence rate is the same as that for placebo-group compliers =  $p_2$ .

If we subtract  $p_2^*$  from  $p_1^*$ , we obtain

**Equation 10.20**

$$\begin{aligned}
 p_1^* - p_2^* &= p_1(1 - \lambda_1 - \lambda_2) - p_2(1 - \lambda_1 - \lambda_2) = (p_1 - p_2)(1 - \lambda_1 - \lambda_2) \\
 &= \text{compliance-adjusted rate difference}
 \end{aligned}$$

In the presence of noncompliance, sample size and power estimates should be based on the compliance-adjusted rates ( $p_1^*, p_2^*$ ) rather than on the perfect compliance rates ( $p_1, p_2$ ). These results are summarized as follows.

**Equation 10.21****Sample-Size Estimation to Compare Two Binomial Proportions in a Clinical Trial Setting (Independent-Sample Case)**

Suppose we want to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$  for the specific alternative  $|p_1 - p_2| = \Delta$  with a significance level  $\alpha$  and a power  $1 - \beta$  in a randomized clinical trial in which group 1 receives active treatment, group 2 receives placebo, and an equal number of subjects are allocated to each group. We assume that  $p_1, p_2$  are the rates of disease in treatment groups 1 and 2 under the assumption of perfect compliance. We also assume that

$\lambda_1$  = dropout rate = proportion of participants in the active-treatment group who fail to comply

$\lambda_2$  = drop-in rate = proportion of participants in the placebo group who receive the active treatment outside the study protocol

- (1) The appropriate sample size in each group is

$$n_1 = n_2 = \left( \sqrt{2\bar{p}^*\bar{q}^*} z_{1-\alpha/2} + \sqrt{p_1^*q_1^* + p_2^*q_2^*} z_{1-\beta} \right)^2 / \Delta^{*2}$$

where

$$\begin{aligned} p_1^* &= (1 - \lambda_1)p_1 + \lambda_1 p_2 \\ p_2^* &= (1 - \lambda_2)p_2 + \lambda_2 p_1 \\ \bar{p}^* &= (p_1^* + p_2^*)/2, \bar{q}^* = 1 - \bar{p}^*, \Delta^* = |p_1^* - p_2^*| = (1 - \lambda_1 - \lambda_2)|p_1 - p_2| \\ &= (1 - \lambda_1 - \lambda_2)\Delta \end{aligned}$$

- (2) If noncompliance rates are low ( $\lambda_1, \lambda_2$  each  $\leq .10$ ), then an approximate sample-size estimate is given by

$$\begin{aligned} n_{1,\text{approx}} = n_{2,\text{approx}} &= \frac{\left( \sqrt{2\bar{p}\bar{q}} z_{1-\alpha/2} + \sqrt{p_1 q_1 + p_2 q_2} z_{1-\beta} \right)^2}{\Delta^2} \times \frac{1}{(1 - \lambda_1 - \lambda_2)^2} \\ &= n_{\text{perfect compliance}} / (1 - \lambda_1 - \lambda_2)^2 \end{aligned}$$

where  $n_{\text{perfect compliance}}$  is the sample size in each group under the assumption of perfect compliance, as computed in Equation 10.14 with  $p_1^* = p_1$ ,  $p_2^* = p_2$ , and  $k = n_2/n_1 = 1$ .

### Example 10.32

**Cardiovascular Disease** Refer to Example 10.31. Suppose we assume that the incidence of MI is .005 per year among participants who actually take placebo and that aspirin prevents 20% of MIs (i.e., relative risk =  $p_1/p_2 = 0.8$ ). We also assume that the duration of the study is 5 years and that the dropout rate in the aspirin group = 10% and the drop-in rate in the placebo group = 5%. How many participants need to be enrolled in each group to achieve 80% power using a two-sided test with significance level = .05?

### Solution

This is a 5-year study, so the 5-year incidence of MI among participants who actually take placebo  $\approx 5(.005) = .025 = p_2$ . Because the risk ratio = 0.8, we have  $p_1/p_2 = 0.8$  or  $p_1 = .020 = 5\text{-year incidence of MI among participants who actually take aspirin}$ . To estimate the true incidence rates to be expected in the study, we must factor in the expected rates of noncompliance. Based on Equation 10.21, the compliance-adjusted rates  $p_1^*$  and  $p_2^*$  are given by

$$\begin{aligned} p_1^* &= (1 - \lambda_1)p_1 + \lambda_1 p_2 \\ &=.9(.020) + .1(.025) = .0205 \end{aligned}$$

$$\begin{aligned} p_2^* &= (1 - \lambda_2)p_2 + \lambda_2 p_1 \\ &=.95(.025) + .05(.020) = .02475 \end{aligned}$$

$$\text{Also, } \Delta^* = |p_1^* - p_2^*| = .00425$$

$$\bar{p}^* = \frac{p_1^* + p_2^*}{2} = \frac{.0205 + .02475}{2} = .02263, \bar{q}^* = 1 - \bar{p}^* = .97737$$

Finally,  $z_{1-\beta} = z_{.80} = 0.84$ ,  $z_{1-\alpha/2} = z_{.975} = 1.96$ . Therefore, from Equation 10.21, the required sample size in each group is

$$\begin{aligned} n_1 = n_2 &= \frac{\left[ \sqrt{2(.02263)(.97737)}(1.96) + \sqrt{.0205(.9795) + .02475(.97525)}(0.84) \right]^2}{.00425^2} \\ &= \left[ \frac{.2103(1.96) + .2103(0.84)}{.00425} \right]^2 = 19,196 \text{ per group} \end{aligned}$$

The total sample size needed = 38,392.

If we don't factor compliance into our sample-size estimates, then based on Equation 10.14, we would need

$$\begin{aligned} n_1 = n_2 &= \frac{\left( \sqrt{2\bar{p}\bar{q}}z_{1-\alpha/2} + \sqrt{p_1q_1 + p_2q_2}z_{1-\beta} \right)^2}{\Delta^2} \\ &= \frac{\left[ \sqrt{2(.0225)(.9775)}1.96 + \sqrt{.02(.98) + .025(.975)}0.84 \right]^2}{|.02 - .025|^2} \\ &= 13,794 \text{ per group} \end{aligned}$$

or a total sample size =  $2(13,794) = 27,588$ .

The approximate sample-size formula in step 2 of Equation 10.21 would yield

$$\begin{aligned} n_{1,\text{approx}} = n_{2,\text{approx}} &= \frac{n_{\text{perfect compliance}}}{(1 - .10 - .05)^2} \\ &= \frac{13,794}{.85^2} = 19,093 \end{aligned}$$

or a total sample size of  $2(19,093) = 38,186$  participants.

Thus the effect of noncompliance is to narrow the observed difference in risk between the aspirin and placebo groups and as a result to increase the required sample size by approximately  $100\% \times (1/.85^2 - 1) = 38\%$  or more exactly  $100\% \times (38,392 - 27,588)/27,588 = 39\%$ .

The Physicians' Health Study actually enrolled 22,000 participants, thus implying that the power of the study with 5 years of follow-up would be somewhat lower than 80%. In addition, the physicians were much healthier than expected and the risk of MI in the placebo group was much lower than expected. However, aspirin proved much more effective than anticipated, preventing 40% of MIs (relative risk = 0.6) rather than the 20% anticipated. This led to a highly significant treatment benefit for aspirin after 5 years of follow-up and an eventual change in the FDA-approved indications for aspirin to include labeling as an agent to prevent cardiovascular disease for men over age 50.

The power formula for the comparison of binomial proportions in Equation 10.15 also assumes perfect compliance. To correct these estimates for noncompliance in a clinical trial setting, replace  $p_1$ ,  $p_2$ ,  $\Delta$ ,  $\bar{p}$ , and  $\bar{q}$  in Equation 10.15 with  $p_1^*$ ,  $p_2^*$ ,  $\Delta^*$ ,  $\bar{p}^*$ ,  $\bar{q}^*$  as given in Equation 10.21. The resulting power is a compliance-adjusted power estimate.

**REVIEW QUESTIONS 10D**

- 1 Suppose we are planning a randomized trial of dietary interventions affecting weight gain in women. We want to compare women randomized to a high-fiber diet vs. women randomized to a low-fiber diet, with the outcome being 10+ lb weight gain after 5 years. We anticipate that after 5 years 20% of the women in the low-fiber group and 10% of the women in the high-fiber group will have gained 10+ lb.
  - (a) How many women need to be randomized in each group to achieve 80% power if a two-sided test will be used with a 5% significance level?
  - (b) Suppose we recruit 250 women in each group. How much power will the study have?
- 2 Consider the study in Review Question 10D.1. Suppose 20% of the women randomized to the high-fiber diet don't follow the dietary instructions (and instead eat a standard Western diet, which we will assume is a low-fiber diet).
  - (a) How many women would be needed for the study under the conditions in Review Question 10D.1a?
  - (b) How much power will the study have if 250 women are recruited for each group?
- 3 Suppose we plan a comparative study of two eye drops (A, B) to reduce intraocular pressure (IOP) among patients with glaucoma. A *contralateral design* is used, in which drop A is assigned to a random eye and drop B is assigned to the fellow eye. The patients take the eye drops for 1 month, after which their IOP is measured again. The outcome is a decrease in IOP of 5+ mm Hg in an eye. We expect the following:
  - (i) that both eyes will be failures (i.e., not show a decrease of 5+ mm Hg) in 50% of patients;
  - (ii) that both eyes will be successes (i.e., will show a decrease of 5+ mm Hg) in 30% of patients;
  - (iii) that in 15% of patients the drop A eye will result in a decrease in IOP of 5+ mm Hg but the drop B eye will not; and
  - (iv) that in 5% of patients the drop B eye will show a decrease in IOP of 5+ mm Hg but the drop A eye will not.
  - (a) What method of analysis can be used to compare the efficacy of drop A vs. drop B?
  - (b) How many patients do we need to randomize to achieve 80% power if we have a two-sided test with  $\alpha = .05$ , assuming that all patients take their drops?

## 10.6 *R* × *C* Contingency Tables

### Tests for Association for *R* × *C* Contingency Tables

In the previous sections of this chapter, methods of analyzing data that can be organized in the form of a  $2 \times 2$  contingency table—that is, where each variable under study has only two categories—were studied. Frequently, one or both variables under study have more than two categories.

**Definition 10.7**

An ***R* × *C* contingency table** is a table with *R* rows and *C* columns. It displays the relationship between two variables, where the variable in the rows has *R* categories and the variable in the columns has *C* categories.

**Example 10.33**

**Cancer** Suppose we want to study further the relationship between age at first birth and development of breast cancer, as in Example 10.4. In particular, we would like

to know whether the effect of age at first birth follows a consistent trend, that is, (1) more protection for women whose age at first birth is <20 than for women whose age at first birth is 25–29 and (2) higher risk for women whose age at first birth is ≥35 than for women whose age at first birth is 30–34. The data are presented in Table 10.18, where case-control status is indicated along the rows and age at first birth categories are indicated along the columns. The data are arranged in the form of a  $2 \times 5$  contingency table because case-control status has two categories and age at first birth has five categories. We want to test for a relationship between age at first birth and case-control status. How should this be done?

**Table 10.18** Data from the international study in Example 10.4 investigating the possible association between age at first birth and case-control status

		Age at first birth					Total
Case-control status		<20	20–24	25–29	30–34	≥35	
	Case	320	1206	1011	463	220	3220
	Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465	
% cases	.184	.214	.259	.298	.351	.239	

Source: Reprinted with permission by WHO Bulletin, 43, 209–221, 1970.

Generalizing our experience from the  $2 \times 2$  situation, the expected table for an  $R \times C$  table can be formed in the same way as for a  $2 \times 2$  table.

### Equation 10.22

#### Computation of the Expected Table for an $R \times C$ Contingency Table

The expected number of units in the  $(i, j)$  cell =  $E_{ij}$  = the product of the number of units in the  $i$ th row multiplied by the number of units in the  $j$ th column, divided by the total number of units in the table.

### Example 10.34

**Cancer** Compute the expected table for the data in Table 10.18.

#### Solution

$$\text{Expected value of the (1,1) cell} = \frac{\text{first row total} \times \text{first column total}}{\text{grand total}} = \frac{3220(1742)}{13,465} = 416.6$$

$$\text{Expected value of the (1,2) cell} = \frac{\text{first row total} \times \text{second column total}}{\text{grand total}} = \frac{3220(5638)}{13,465} = 1348.3$$

⋮

$$\text{Expected value of the (2,5) cell} = \frac{\text{second row total} \times \text{fifth column total}}{\text{grand total}} = \frac{10,245(626)}{13,465} = 476.3$$

All 10 expected values are given in Table 10.19.

The sum of the expected values across any row or column must equal the corresponding row or column total, as was the case for  $2 \times 2$  tables. This fact provides a good check that the expected values are computed correctly. The expected values in Table 10.19 fulfill this criterion except for roundoff error.

**Table 10.19** Expected table for the international study data in Table 10.18

		Age at first birth					Total
Case-control status		<20	20–24	25–29	30–34	≥35	
	Case	416.6	1348.3	933.6	371.9	149.7	3220
Control	1325.4	4289.7	2970.4	1183.1	476.3		10,245
Total	1742	5638	3904	1555	626		13,465

We again want to compare the observed table with the expected table. The more similar these tables are, the more willing we will be to accept the null hypothesis that there is no association between the two variables. The more different the tables are, the more willing we will be to reject  $H_0$ . Again the criterion  $(O - E)^2/E$  is used to compare the observed and expected counts for a particular cell. Furthermore,  $(O - E)^2/E$  is summed over all the cells in the table to get an overall measure of agreement for the observed and expected tables. Under  $H_0$ , for an  $R \times C$  contingency table, the sum of  $(O - E)^2/E$  over the  $RC$  cells in the table will approximately follow a chi-square distribution with  $(R - 1) \times (C - 1)$  df.  $H_0$  will be rejected for large values of this sum and will be accepted for small values.

Generally speaking, the continuity correction is not used for contingency tables larger than  $2 \times 2$  because statisticians have found empirically that the correction does not help in the approximation of the test statistic by the chi-square distribution. As for  $2 \times 2$  tables, this test should not be used if the expected values of the cells are too small. Cochran [4] has studied the validity of the approximation in this case and recommends its use if

- (1) No more than 1/5 of the cells have expected values <5.  
and
- (2) No cell has an expected value <1.

The test procedure can be summarized as follows.

### Equation 10.23

#### Chi-Square Test for an $R \times C$ Contingency Table

To test for the relationship between two discrete variables, where one variable has  $R$  categories and the other has  $C$  categories, use the following procedure:

- (1) Analyze the data in the form of an  $R \times C$  contingency table, where  $O_{ij}$  represents the observed number of units in the  $(i, j)$  cell.
- (2) Compute the expected table as shown in Equation 10.22, where  $E_{ij}$  represents the expected number of units in the  $(i, j)$  cell.
- (3) Compute the test statistic

$$X^2 = (O_{11} - E_{11})^2/E_{11} + (O_{12} - E_{12})^2/E_{12} + \dots + (O_{RC} - E_{RC})^2/E_{RC}$$

which under  $H_0$  approximately follows a chi-square distribution with  $(R - 1) \times (C - 1)$  df.

- (4) For a level  $\alpha$  test,

if  $X^2 > \chi^2_{(R-1) \times (C-1), 1-\alpha}$ , then reject  $H_0$ .

If  $X^2 \leq \chi^2_{(R-1) \times (C-1), 1-\alpha}$ , then accept  $H_0$ .

- (5) The approximate  $p$ -value is given by the area to the right of  $X^2$  under a  $\chi_{(R-1) \times (C-1)}^2$  distribution.
- (6) Use this test only if both of the following two conditions are satisfied:
- No more than 1/5 of the cells have expected values <5.
  - No cell has an expected value <1.

The acceptance and rejection regions for this test are shown in Figure 10.8. Computation of the  $p$ -value for this test is illustrated in Figure 10.9.

### Example 10.35

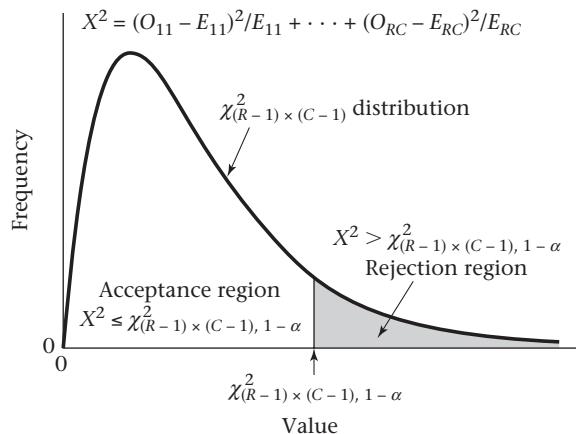
#### Solution

**Cancer** Assess the statistical significance of the data in Example 10.33.

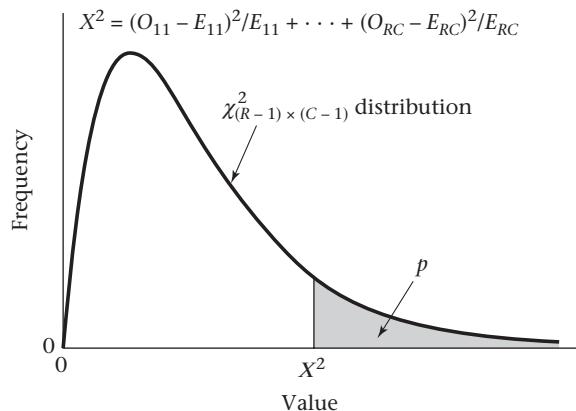
From Table 10.19 we see that all expected values are  $\geq 5$ , so the test procedure in Equation 10.23 can be used. From Tables 10.18 and 10.19,

$$X^2 = \frac{(300 - 416.6)^2}{416.6} + \frac{(1206 - 1348.3)^2}{1348.3} + \dots + \frac{(406 - 476.3)^2}{476.3} = 130.3$$

**Figure 10.8** Acceptance and rejection regions for the chi-square test for an  $R \times C$  contingency table



**Figure 10.9** Computation of the  $p$ -value for the chi-square test for an  $R \times C$  contingency table



Under  $H_0$ ,  $X^2$  follows a chi-square distribution with  $(2 - 1) \times (5 - 1)$ , or 4,  $df$ . Because

$$\chi^2_{4,.999} = 18.47 < 130.3 = X^2$$

it follows that  $p < 1 - .999 = .001$

Therefore, the results are very highly significant, and we can conclude there is a significant relationship between age at first birth and prevalence of breast cancer.

### Chi-Square Test for Trend in Binomial Proportions

Refer again to the international study data in Table 10.18. In Example 10.35 the test procedure in Equation 10.23 was used to analyze the data. For the special case of a  $2 \times k$  table, this test procedure enables us to test the hypothesis  $H_0: p_1 = p_2 = \dots = p_k$  vs.  $H_1$ : at least two of the  $p_i$ 's are unequal, where  $p_i$  = probability of success for the  $i$ th group = probability that an observation from the  $i$ th column falls in the first row. When this test procedure was employed in Example 10.35, a chi-square statistic of 130.3 with 4  $df$  was found, which was highly significant ( $p < .001$ ). As a result,  $H_0$  was rejected and we concluded the prevalence of breast-cancer cases in at least 2 of the 5 age-at-first-birth groups were different. However, although this result shows some relationship between breast cancer and age at first birth, it does not tell us specifically about the nature of the relationship. In particular, from Table 10.18 we notice an increasing *trend* in the proportion of women with breast cancer in each succeeding column. We would like to employ a specific test to detect such trends. For this purpose a **score variable**  $S_i$  is introduced to correspond to the  $i$ th group. The score variable can represent some particular numeric attribute of the group. In other instances, for simplicity, 1 is assigned to the first group, 2 to the second group, ...,  $k$  to the  $k$ th (last) group.

#### Example 10.36

**Cancer** Construct a score variable for the international-study data in Table 10.18.

#### Solution

It is natural to use the average age at first birth within a group as the score variable for that group. This rule presents no problem for the second, third, and fourth groups, in which the average age is estimated as 22.5  $[(20 + 25)/2]$ , 27.5, and 32.5 years, respectively. However, a similar calculation cannot be performed for the first and fifth groups because they are defined as  $<20$  and  $\geq 35$ , respectively. By symmetry, a score of 17.5 years could be assigned to the first group and 37.5 years to the fifth group. However, because the scores are equally spaced our purposes are equally well served by assigning scores of 1, 2, 3, 4, and 5 to the five groups. For simplicity, this scoring method will be adopted.

We want to relate the proportion of breast-cancer cases in a group to the score variable for that group. In other words, we wish to test whether the proportion of breast-cancer cases increases or decreases as age at first birth increases. For this purpose the following test procedure is introduced.

#### Equation 10.24

##### Chi-Square Test for Trend in Binomial Proportions (Two-Sided Test)

Suppose there are  $k$  groups and we want to test whether there is an increasing (or decreasing) trend in the proportion of "successes"  $p_i$  (the proportion of units in the first row of the  $i$ th group) as  $i$  increases.

- (1) Set up the data in the form of a  $2 \times k$  contingency table, where success or failure is listed along the rows and the  $k$  groups are listed along the columns.

- (2) Denote the number of successes in the  $i$ th group by  $x_i$ , the total number of units in the  $i$ th group by  $n_i$ , and the proportion of successes in the  $i$ th group by  $\hat{p}_i = x_i / n_i$ . Denote the total number of successes over all groups by  $x$ , the total number of units over all groups by  $n$ , the overall proportion of successes by  $\bar{p} = x / n$ , and the overall proportion of failures by  $\bar{q} = 1 - \bar{p}$ .
- (3) Construct a score variable  $S_i$  to correspond to the  $i$ th group. This variable will usually either be 1, 2, ...,  $k$  for the  $k$  groups or be defined to correspond to some other numeric attribute of the group.
- (4) More specifically, we wish to test the hypothesis  $H_0$ : There is no trend among the  $p_i$ 's vs.  $H_1$ : The  $p_i$  are an increasing or decreasing function of the  $S_i$ , expressed in the form  $p_i = \alpha + \beta S_i$  for some constants  $\alpha, \beta$ . To relate  $p_i$  and  $S_i$ , compute the test statistic  $X_1^2 = A^2 / B$ , where

$$\begin{aligned} A &= \sum_{i=1}^k n_i (\hat{p}_i - \bar{p})(S_i - \bar{S}) \\ &= \left( \sum_{i=1}^k x_i S_i \right) - x \bar{S} = \left( \sum_{i=1}^k x_i S_i \right) - x \left( \sum_{i=1}^k n_i S_i \right) / n \\ B &= \bar{p} \bar{q} \left[ \left( \sum_{i=1}^k n_i S_i^2 \right) - \left( \sum_{i=1}^k n_i S_i \right)^2 / n \right] \end{aligned}$$

which under  $H_0$  approximately follows a chi-square distribution with 1 df.

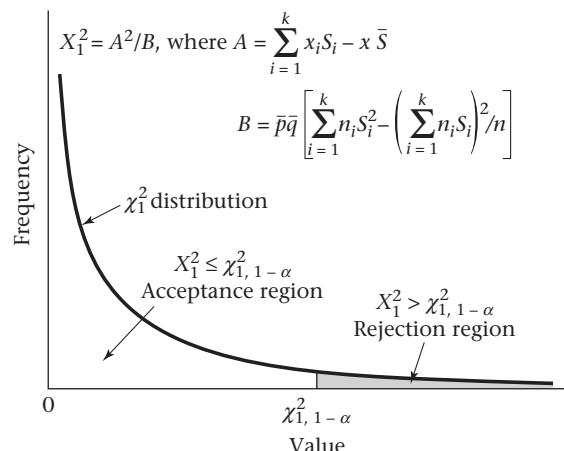
- (5) For a two-sided level  $\alpha$  test,

if  $X_1^2 > \chi_{1,1-\alpha}^2$ , then reject  $H_0$ .

If  $X_1^2 \leq \chi_{1,1-\alpha}^2$ , then accept  $H_0$ .

- (6) The approximate  $p$ -value is given by the area to the right of  $X_1^2$  under a  $\chi_1^2$  distribution.
- (7) The direction of the trend in proportions is indicated by the sign of  $A$ . If  $A > 0$ , then the proportions increase with increasing score, if  $A < 0$ , then the proportions decrease with increasing score.

**Figure 10.10** Acceptance and rejection regions for the chi-square test for trend in binomial proportions

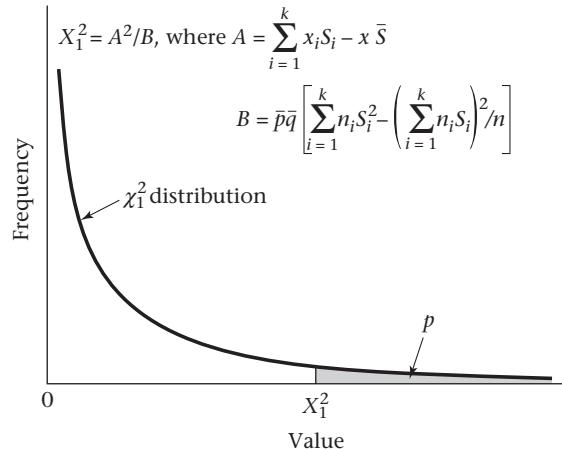


(8) Use this test only if  $n\bar{p}\bar{q} \geq 5.0$ .

The acceptance and rejection regions for this test are shown in Figure 10.10. Computation of the  $p$ -value is illustrated in Figure 10.11.

The test statistic in Equation 10.24 is reasonable, because if  $\hat{p}_i$  (or  $\hat{p}_i - \bar{p}$ ) increases as  $S_i$  increases, then  $A > 0$ , whereas if  $\hat{p}_i$  decreases as  $S_i$  increases, then  $A < 0$ . In either case  $A^2$  and the test statistic  $X_1^2$  will be large. However, if  $\hat{p}_i$  shows no particular trend regarding  $S_i$ , then  $A$  will be close to 0 and the test statistic  $X_1^2$  will be small. This test can be used even if some of the groups have small sample size because the test is based on the overall trend in the proportions. This property is in contrast to the overall chi-square test in Equation 10.23, which tests for heterogeneity among proportions and requires that the expected number of units in individual cells not be too small.

**Figure 10.11** Computation of the  $p$ -value for the chi-square test for trend in binomial proportions



### Example 10.37

**Cancer** Using the international study data in Table 10.18, assess whether there is an increasing trend in the proportion of breast-cancer cases as age at first birth increases.

### Solution

Note that  $S_i = 1, 2, 3, 4, 5$  in the five groups, respectively. Furthermore, from Table 10.18,  $x_i = 320, 1206, 1011, 463, 220$ , and  $n_i = 1742, 5638, 3904, 1555, 626$  in the five respective groups, whereas  $x = 3220, n = 13,465, \bar{p} = x/n = .239, \bar{q} = 1 - \bar{p} = .761$ . From Equation 10.24 it follows that

$$\begin{aligned} A &= 320(1) + 1206(2) + \dots + 220(5) \\ &\quad - (3220)[1742(1) + 5638(2) + \dots + 626(5)]/13,465 \\ &= 8717 - (3220)(34,080)/13,465 = 8717 - 8149.84 = 567.16 \\ B &= (.239)(.761)\{1742(1^2) + 5638(2^2) + \dots + 626(5^2) \\ &\quad - [1742(1) + (5638)(2) + \dots + 626(5)]^2/13,465\} \\ &= .239(.761)(99,960 - 34,080^2/13,465) \\ &= .239(.761)(99,960 - 86,256.70) = 2493.33 \end{aligned}$$

Thus  $X_1^2 = A^2/B = \frac{567.16^2}{2493.33} = 129.01 \sim \chi_1^2$  under  $H_0$

Because  $\chi_{1,999}^2 = 10.83 < 129.01 = X_1^2$ ,  $H_0$  can be rejected with  $p < .001$  and we can conclude there is a significant trend in the proportion of breast-cancer cases among age-at-first-birth groups. Because  $A > 0$ , it follows that as age at first birth increases, the proportion of breast-cancer cases rises.

With a  $2 \times k$  table, the chi-square test for trend in Equation 10.24 is often more relevant to the hypotheses of interest than the chi-square test for heterogeneity in Equation 10.23. This is because the former procedure tests for specific trends in the proportions, whereas the latter tests for any differences in the proportions, where the proportions may follow any pattern. Other, more advanced methods for assessing  $R \times C$  contingency tables are given in Maxwell's *Analyzing Qualitative Data* [5].

In this section, we have discussed tests for association between two categorical variables with  $R$  and  $C$  categories, respectively, where either  $R > 2$  and/or  $C > 2$ . If both  $R$  and  $C$  are  $> 2$ , then the chi-square test for  $R \times C$  contingency tables is used. Referring to the flowchart at the end of this chapter (Figure 10.16, p. 409), we answer no to (1)  $2 \times 2$  contingency table? and (2)  $2 \times k$  contingency table? which leads to (3)  $R \times C$  contingency table with  $R > 2$  and  $C > 2$  and then to the box labeled "Use chi-square test for  $R \times C$  tables." If either  $R$  or  $C = 2$ , then assume we have rearranged the row and column variables so the row variable has two categories. Let's designate the number of column categories by  $k$  (rather than  $C$ ). If we are interested in assessing trend over the  $k$  binomial proportions formed by the proportions of units in the first row of each of the  $k$  columns, then we use the chi-square test for trend in binomial proportions. Referring to the flowchart in Figure 10.16 at the end of this chapter (p. 409), we answer no to (1)  $2 \times 2$  contingency table? yes to (2)  $2 \times k$  contingency table? and yes to (3) interested in trend over  $k$  binomial proportions? This leads us to the box labeled "Use chi-square test for trend if no confounding is present, or the Mantel Extension test if confounding is present."

## Relationship Between the Wilcoxon Rank-Sum Test and the Chi-Square Test for Trend

The Wilcoxon rank-sum test given in Equation 9.7 is actually a special case of the chi-square test for trend.

**Equation 10.25**

### Relationship Between the Wilcoxon Rank-Sum Test and the Chi-Square Test for Trend

Suppose we have a  $2 \times k$  table as shown in Table 10.20.

**Table 10.20**

**A hypothetical  $2 \times k$  table relating a dichotomous disease-variable  $D$  to a categorical exposure-variable  $E$  with  $k$  ordered categories**

		$E$			
		1	2	$k$	
		$x_1$	$x_2$	$\dots$	$x_k$
$D$	+				
	-	$n_1 - x_1$	$n_2 - x_2$	$\dots$	$n_k - x_k$
Score		$n_1$	$n_2$	$\dots$	$n_k$
		$S_1$	$S_2$	$\dots$	$S_k$
					$n$

The  $i$ th exposure category is assumed to have an associated score  $S_i$ ,  $i = 1, \dots, k$ . Let  $p_i$  = probability of disease in the  $i$ th exposure group. If  $p_i = \alpha + \beta S_i$  and we want to test the hypothesis  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ , then

- (1) We can use the chi-square test for trend, where we can write the test statistic in the form

$$X^2 = \frac{(|O - E| - 0.5)^2}{V} \sim \chi^2_1 \text{ under } H_0$$

where

$$O = \text{observed total score among subjects with disease} = \sum_{i=1}^k x_i S_i$$

$$E = \text{expected total score among subjects with disease under } H_0 = \frac{x}{n} \sum_{i=1}^k n_i S_i$$

$$V = \frac{x(n-x)}{n(n-1)} \left[ \sum_{i=1}^k n_i S_i^2 - \left( \sum_{i=1}^k n_i S_i \right)^2 / n \right]$$

and we reject  $H_0$  if  $X^2 > \chi^2_{1,1-\alpha}$

and accept  $H_0$  otherwise.

- (2) We can use the Wilcoxon rank-sum test as given in Equation 9.7, where we have the test statistic

$$T = \frac{\left| R_1 - \frac{x(n+1)}{2} \right| - \frac{1}{2}}{\sqrt{\left[ \frac{x(n-x)}{12} \right] \left[ n+1 - \frac{\sum_{i=1}^k n_i (n_i^2 - 1)}{n(n-1)} \right]}}$$

and reject  $H_0$  if  $T > z_{1-\alpha/2}$

and accept  $H_0$  if  $T \leq z_{1-\alpha/2}$

where  $z_{1-\alpha/2}$  = upper  $\alpha/2$  percentile of an  $N(0, 1)$  distribution.

- (3) If the scores  $S_i$  are set equal to the midrank for the  $i$ th group as defined in Equation 9.6, where the midrank for the  $i$ th exposure category = number of observations in the first  $i - 1$  groups +  $\left( \frac{1+n_i}{2} \right)$

$$= \sum_{j=1}^{i-1} n_j + \frac{(1+n_i)}{2} \text{ if } i > 1$$

$$= \frac{1+n_1}{2} \text{ if } i = 1$$

then the test procedures in steps (1) and (2) yield the same  $p$ -values and are equivalent. In particular,

$$O = R_1 = \text{Rank sum in the first row}, E = \frac{x(n+1)}{2}$$

$$V = \left[ \frac{x(n-x)}{12} \right] \left[ n+1 - \sum_{i=1}^k \frac{n_i (n_i^2 - 1)}{n(n-1)} \right], \text{ and } X^2_1 = T^2$$

**Example 10.38**

**Ophthalmology** Test the hypothesis that the average visual acuity is different for dominant and sex-linked people in Table 9.3 or, equivalently, that the proportion of dominant subjects changes in a consistent manner as visual acuity declines, using the chi-square test for trend.

**Solution**

We have the following 2 × 8 table:

		Visual acuity								
		20–20	20–25	20–30	20–40	20–50	20–60	20–70	20–80	
Dominant	5	9	6	3	2	0	0	0	25	
	1	5	4	4	8	5	2	1	30	
	6	14	10	7	10	5	2	1	55	
Score	3.5	13.5	25.5	34.0	42.5	50.0	53.5	55.0		

If the scores are set equal to the average ranks given in Table 9.3, then we have

$$\begin{aligned}
 O &= 5(3.5) + 9(13.5) + 6(25.5) + 3(34.0) + 2(42.5) = 479 \\
 E &= \frac{25}{55} [6(3.5) + 14(13.5) + 10(25.5) + 7(34.0) + 10(42.5) + 10(4.5) + 5(50.0) \\
 &\quad + 2(53.5) + 1(55.0)] \\
 &= \frac{25}{55}(1540) = 700 \\
 V &= \frac{25(30)}{55(54)} \left\{ [6(3.5)^2 + 14(13.5)^2 + 10(25.5)^2 + 7(34.0)^2 + 10(42.5)^2 + 5(50.0)^2 \right. \\
 &\quad \left. + 2(53.5)^2 + 1(55.0)^2] - \frac{1540^2}{55} \right\} \\
 &= \frac{25(30)}{55(54)} (56,531.5 - 43,120) \\
 &= \frac{25(30)}{55(54)} (13,411.5) = 3386.74 \\
 X^2 &= \frac{(|479 - 700| - 0.5)^2}{3386.74} = 14.36 \sim \chi^2_1 \text{ under } H_0
 \end{aligned}$$

The  $p$ -value =  $\Pr(\chi^2_1 > 14.36) < .001$ . Also, referring to Example 9.17, we see that  $O = R_1 = 479$ ,  $E = E(R_1) = 700$ ,  $V = V(R_1)$  corrected for ties = 3386.74 and

$$X^2 = 14.36 = T^2 = 3.79^2$$

Thus the two test procedures are equivalent. However, if we had chosen different scores (e.g., 1, . . . , 8) for the 8 visual-acuity groups, then the test procedures would *not* be the same. The choice of scores is somewhat arbitrary. If each column corresponds to a specific quantitative exposure category, then it is reasonable to use the average exposure within the category as the score. If the exposure level is not easily quantified, then either midranks or consecutive integers are reasonable choices for scores. If the number of subjects in each exposure category is the same, then these two methods of scoring will yield identical test statistics and  $p$ -values using the chi-square test for trend.

The estimate of variance ( $V$ ) given in Equation 10.25 is derived from the hypergeometric distribution and differs slightly from the variance estimate for the chi-square test for trend in Equation 10.24 given by

$$V = \frac{x(n-x)}{n^2} \left[ \sum_{i=1}^k n_i S_i^2 - \left( \sum_{i=1}^k n_i S_i \right)^2 / n \right]$$

which is based on the binomial distribution. The hypergeometric distribution is more appropriate, although the difference is usually slight, particularly for large  $n$ . Also, a continuity correction of 0.5 is used in the numerator of  $X^2$  in Equation 10.25, but not in  $A$  in the numerator of  $X_1^2$  in Equation 10.24. This difference is also usually slight.

### REVIEW QUESTIONS 10E

- 1** **(a)** What is the difference between the chi-square test for trend and the chi-square test for heterogeneity?  
**(b)** When do we use each test?
- 2** Suppose we are given the following  $2 \times 5$  table with two disease categories and five exposure categories, as in Table 10.21.

**Table 10.21** Hypothetical table illustrating the association between exposure and disease

		Exposure category				
		1	2	3	4	5
Disease category	+	2	3	4	6	3
	-	6	5	5	4	2

- (a)** If exposure is treated as a nominal categorical variable, is it valid to use the chi-square test for heterogeneity on these data? Why or why not?  
**(b)** If exposure is treated as an ordinal categorical variable, is it valid to use the chi-square test for trend on these data? Why or why not?
- 3** We are interested in studying the relationship between the prevalence of hypertension in adolescents and ethnic group, where hypertension is defined as being above the 90th percentile for a child's age, sex, and height, based on national norms.  
**(a)** Suppose that 8 of 100 Caucasian adolescent girls, 12 out of 95 African-American adolescent girls, and 10 of 90 Hispanic adolescent girls are above the 90th percentile for blood pressure. What test can be used to assess whether there is an association between adolescent hypertension and ethnic group?  
**(b)** Implement the test in Review Question 10E.3a, and report a two-tailed  $p$ -value.
- 4** We are also interested in the relationship between adolescent hypertension and obesity. For this purpose, we choose 30 normal-weight adolescent boys (i.e., body-mass index [BMI] =  $\text{kg}/\text{m}^2 < 25$ ), 30 overweight adolescent boys ( $25 \leq \text{BMI} < 30$ ), and 35 obese adolescent boys ( $\text{BMI} \geq 30$ ). We find that 2 of the normal-weight boys, 5 of the overweight boys, and 10 of the obese boys are hypertensive.

- (a) What test can be used to assess whether there is an association between adolescent hypertension and BMI?  
 (b) Implement the test in Review Question 10E.4a, and report a two-tailed  $p$ -value.

## 10.7 Chi-Square Goodness-of-Fit Test

In our previous work on estimation and hypothesis testing, we usually assumed the data came from a specific underlying probability model and then proceeded either to estimate the parameters of the model or test hypotheses concerning different possible values of the parameters. This section presents a general method of testing for the *goodness-of-fit of a probability model*. Consider the problem in Example 10.39.

### Example 10.39

**Hypertension** Diastolic blood-pressure measurements were collected at home in a community-wide screening program of 14,736 adults ages 30–69 in East Boston, Massachusetts, as part of a nationwide study to detect and treat hypertensive people [6]. The people in the study were each screened in the home, with two measurements taken during one visit. A frequency distribution of the mean diastolic blood pressure is given in Table 10.22 in 10-mm Hg intervals.

We would like to assume these measurements came from an underlying normal distribution because standard methods of statistical inference could then be applied on these data as presented in this text. How can the validity of this assumption be tested?

**Table 10.22** Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	77.9	≥80, <90	4604	4478.5
≥50, <60	330	547.1	≥90, <100	2119	2431.1
≥60, <70	2132	2126.7	≥100, <110	659	684.1
≥70, <80	4584	4283.3	≥110	251	107.2
			Total	14,736	14,736

This assumption can be tested by first computing what the expected frequencies would be in each group if the data did come from an underlying normal distribution and by then comparing these expected frequencies with the corresponding observed frequencies.

### Example 10.40

**Hypertension** Compute the expected frequencies for the data in Table 10.22, assuming an underlying normal distribution.

### Solution

Assume the mean and standard deviation of this hypothetical normal distribution are given by the sample mean and standard deviation, respectively ( $\bar{x} = 80.68$ ,  $s = 12.00$ ). The expected frequency within a group interval from  $a$  to  $b$  would then be given by

$$14,736 \{ \Phi[(b - \mu)/\sigma] - \Phi[(a - \mu)/\sigma] \}$$

Thus the expected frequency within the ( $\geq 50, < 60$ ) group would be

$$\begin{aligned} 14,736 \times [\Phi[(60 - 80.68)/12] - \Phi[(50 - 80.68)/12]] \\ = 14,736 \times [\Phi(-1.723) - \Phi(-2.557)] \\ = 14,736 \times (.0424 - .0053) = 14,736(.0371) = 547.1 \end{aligned}$$

Also, the expected frequency less than  $a$  would be  $\Phi[(a - \mu)/\sigma]$ , and the expected frequency greater than or equal to  $b$  would be  $1 - \Phi[(b - \mu)/\sigma]$ . The expected frequencies for all the groups are given in Table 10.22.

We use the same measure of agreement between the observed and expected frequencies in a group that we used in our work on contingency tables, namely,  $(O - E)^2/E$ . Furthermore, the agreement between observed and expected frequencies can be summarized over the whole table by summing  $(O - E)^2/E$  over all the groups. If we have the correct underlying model, then this sum will approximately follow a chi-square distribution with  $g - 1 - k$  df, where  $g$  = the number of groups and  $k$  = the number of parameters estimated from the data to compute the expected frequencies. Again, this approximation will be valid only if the expected values in the groups are not too small. In particular, the requirement is that no expected value can be  $< 1$  and not more than  $1/5$  of the expected values can be  $< 5$ . If there are too many groups with small expected frequencies, then some groups should be combined with other adjacent groups so the preceding rule is not violated. The test procedure can be summarized as follows.

#### Equation 10.26

##### Chi-Square Goodness-of-Fit Test

To test for the goodness of fit of a probability model, use the following procedure:

- (1) Divide the raw data into groups. The considerations for grouping data are similar to those in Section 2.7. In particular, the groups must not be too small, so step 7 is not violated.
- (2) Estimate the  $k$  parameters of the probability model from the data using the methods described in Chapter 6.
- (3) Use the estimates in step 2 to compute the probability  $\hat{p}$  of obtaining a value within a particular group and the corresponding expected frequency within that group ( $n\hat{p}$ ), where  $n$  is the total number of data points.
- (4) If  $O_i$  and  $E_i$  are, respectively, the observed and expected number of units within the  $i$ th group, then compute

$$X^2 = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2 + \dots + (O_g - E_g)^2/E_g$$

where  $g$  = the number of groups.

- (5) For a test with significance level  $\alpha$ , if

$$X^2 > \chi_{g-k-1, 1-\alpha}^2$$

then reject  $H_0'$ ; if

$$X^2 \leq \chi_{g-k-1, 1-\alpha}^2$$

then accept  $H_0$ .

- (6) The approximate  $p$ -value for this test is given by

$$Pr(\chi_{g-k-1}^2 > X^2)$$

- (7) Use this test only if
- No more than 1/5 of the expected values are <5.
  - No expected value is <1.

The acceptance and rejection regions for this test are shown in Figure 10.12. Computation of the  $p$ -value for this test is illustrated in Figure 10.13.

- (8) Note: If the parameters of the probability model were specified a priori, without using the present sample data, then  $k = 0$  and  $X^2 \sim \chi_{g-1}^2$ . We call such a model an *externally specified model*, as opposed to the internally specified model described in the preceding steps 1 to 7.

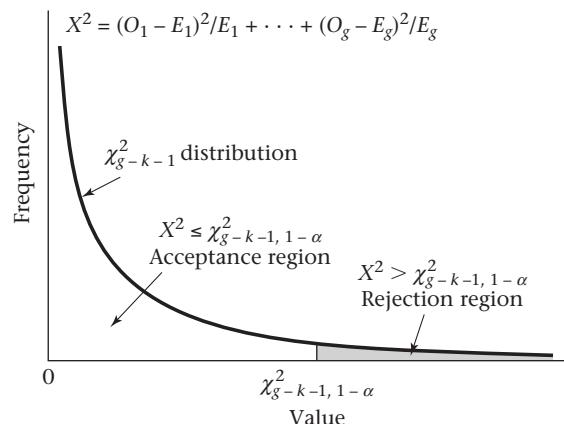
### Example 10.41

**Hypertension** Test for goodness of fit of the normal-probability model using the data in Table 10.22.

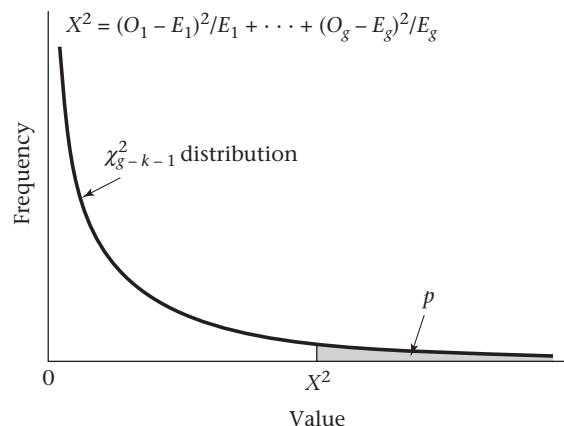
### Solution

Two parameters have been estimated from the data  $(\mu, \sigma^2)$ , and there are 8 groups. Therefore,  $k = 2$ ,  $g = 8$ . Under  $H_0$ ,  $X^2$  follows a chi-square distribution with  $8 - 2 - 1 = 5$  df.

**Figure 10.12** Acceptance and rejection regions for the chi-square goodness-of-fit test



**Figure 10.13** Computation of the  $p$ -value for the chi-square goodness-of-fit test



$$\begin{aligned} X^2 &= (O_1 - E_1)^2 / E_1 + \dots + (O_8 - E_8)^2 / E_8 \\ &= (57 - 77.9)^2 / 77.9 + \dots + (251 - 107.2)^2 / 107.2 = 350.2 \sim \chi^2_5 \text{ under } H_0 \end{aligned}$$

Because  $\chi^2_{5,999} = 20.52 < 350.2 = X^2$ , the  $p$ -value  $< 1 - .999 = .001$  and the results are very highly significant.

Thus the normal model does not provide an adequate fit to the data. The normal model appears to fit fairly well in the middle of the distribution (between 60 and 110 mm Hg) but fails badly in the tails, predicting too many blood pressures below 60 mm Hg and too few over 110 mm Hg.

The test procedure in Equation 10.26 can be used to assess the goodness of fit of any probability model, not just the normal model. The expected frequencies would be computed from the probability distribution of the proposed model and then the same goodness-of-fit test statistic as given in Equation 10.26 would be used. Also, the test procedure can be used to test for the goodness of fit of both a model in which the parameters are estimated from the data set used for testing the model as described in steps 1 to 7 and a model in which the parameters are specified a priori as in step 8.

### REVIEW QUESTIONS 10F

- 1 What is the difference between the chi-square goodness-of-fit test and the chi-square test for  $2 \times 2$  tables? When do we use each test?
- 2 The data in Table 4.12 refers to the monthly number of cases of Guillain-Barré syndrome in Finland from April 1984 to October 1985.
  - (a) Use the chi-square goodness-of-fit test to assess the adequacy of the Poisson model to these data. (Note: Exclude the month March 1985 from this analysis because this month appears to be an outlier.)
  - (b) What are your overall conclusions?

## 10.8 The Kappa Statistic

Most of our previous work has been concerned with *tests of association* between two categorical variables (usually a disease and an exposure variable). In some instances, some association is expected between the variables and the issue is quantifying the *degree of association*. This is particularly true in **reliability studies**, where the researcher wants to quantify the reproducibility of the same variable (e.g., dietary intake of a particular food) measured more than once.

### Example 10.42

**Nutrition** A diet questionnaire was mailed to 537 female American nurses on two separate occasions several months apart. The questions asked included the quantities eaten of more than 100 separate food items. The data from the two surveys for the amount of beef consumption are presented in Table 10.23. Notice that the responses on the two surveys are the same only for  $136 + 240 = 376$  out of 537 (70.0%) women. How can reproducibility of response for the beef-consumption data be quantified?

A chi-square test for association between the survey 1 and survey 2 responses could be performed. However, this test would not give a quantitative measure of

reproducibility between the responses at the two surveys. Instead, we focus on the percentage of women with concordant responses in the two surveys. We noted in Example 10.42 that 70.0% of the women gave concordant responses. We want to compare the observed concordance rate ( $p_o$ ) with the expected concordance rate ( $p_e$ ) if the responses of the women in the two surveys were statistically independent. The motivation behind this definition is that the questionnaire would be virtually worthless if the frequency of consumption reported at one survey had no relationship to the frequency of consumption reported at a second survey. Suppose there are  $c$  response categories and the probability of response in the  $i$ th category is  $a_i$  for the first survey and  $b_i$  for the second survey. These probabilities can be estimated from the row and column margins of the contingency table (Table 10.23). The expected concordance rate ( $p_e$ ) if the survey responses are independent is  $\sum a_i b_i$ .

**Table 10.23** Amount of beef consumption reported by 537 female American nurses at two different surveys

		Survey 2		Total
Survey 1		≤1 serving/week	>1 serving/week	
	≤ 1 serving/week	136	92	228
	>1 serving/week	69	240	309
Total		205	332	537

### Example 10.43

**Nutrition** Compute the expected concordance rate using the beef-consumption data in Table 10.23.

#### Solution

From Table 10.23

$$a_1 = \frac{228}{537} = .425$$

$$a_2 = \frac{309}{537} = .575$$

$$b_1 = \frac{205}{537} = .382$$

$$b_2 = \frac{332}{537} = .618$$

Thus  $p_e = (.425 \times .382) + (.575 \times .618) = .518$

Therefore, 51.8% concordance would be expected if the participants were responding independently regarding beef consumption at the two surveys.

We could use  $p_o - p_e$  as the measure of reproducibility. However, it is intuitively preferable to use a measure that equals +1.0 in the case of perfect agreement and 0.0 if the responses on the two surveys are completely independent. Indeed, the maximum possible value for  $p_o - p_e$  is  $1 - p_e$ , which is achieved with  $p_o = 1$ . Therefore, the Kappa statistic, defined as  $(p_o - p_e)/(1 - p_e)$ , is used as the measure of reproducibility:

**Equation 10.27****The Kappa Statistic**

- (1) If a categorical variable is reported at two surveys by each of  $n$  subjects, then the Kappa statistic ( $\kappa$ ) is used to measure reproducibility between surveys, where

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

and  $p_o$  = observed probability of concordance between the two surveys

$$\begin{aligned} p_o &= \text{expected probability of concordance between the two surveys} \\ &= \sum a_i b_i \end{aligned}$$

where  $a_i$ ,  $b_i$  are the marginal probabilities for the  $i$ th category in the  $c \times c$  contingency table relating response at the two surveys.

- (2) Furthermore,

$$se(\kappa) = \sqrt{\frac{1}{n(1-p_e)^2} \times \left\{ p_e + p_e^2 - \sum_{i=1}^c [a_i b_i (a_i + b_i)] \right\}}$$

To test the one-sided hypothesis  $H_0: \kappa = 0$  vs.  $H_1: \kappa > 0$ , use the test statistic

$$z = \frac{\kappa}{se(\kappa)}$$

which follows an  $N(0, 1)$  distribution under  $H_0$ .

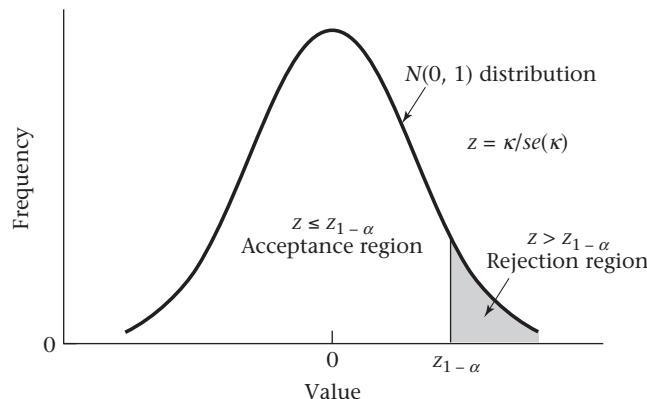
- (3) Reject  $H_0$  at level  $\alpha$  if  $z > z_{1-\alpha}$  and accept  $H_0$  otherwise.  
 (4) The exact  $p$ -value is given by  $p = 1 - \Phi(z)$ .

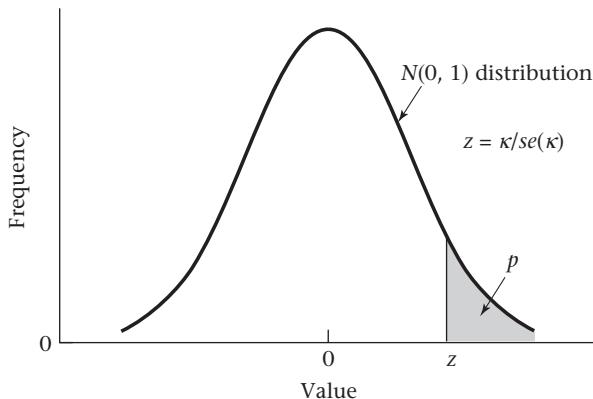
The acceptance and rejection regions for this test are depicted in Figure 10.14. Computation of the  $p$ -value is shown in Figure 10.15.

Note that we are customarily interested in one-tailed tests in Equation 10.27 because negative values for Kappa usually have no biological significance.

**Example 10.44** **Nutrition** Compute the Kappa statistic, and assess its statistical significance using the beef-consumption data in Table 10.23.

**Figure 10.14** Acceptance and rejection regions for the significance test for Kappa



**Figure 10.15 Computation of the  $p$ -value for the significance test for Kappa**

**Solution** From Examples 10.42 and 10.43,

$$p_o = .700$$

$$p_e = .518$$

Therefore, the Kappa statistic is given by

$$\kappa = \frac{.700 - .518}{1 - .518} = \frac{.182}{.482} = .378$$

Furthermore, from Equation 10.27 and the results of Example 10.43, the standard error of  $\kappa$  is given by

$$se(\kappa) = \sqrt{\frac{1}{537(1 - .518)^2} \times \left\{ .518 + .518^2 - \sum [a_i b_i (a_i + b_i)] \right\}}$$

where

$$\begin{aligned} \sum [a_i b_i (a_i + b_i)] &= .425 \times .382 \times (.425 + .382) + .575 \times .618 \times (.575 + .618) \\ &= .555 \end{aligned}$$

Thus

$$se(\kappa) = \sqrt{\frac{1}{537 \times .232} \times (.518 + .268 - .555)} = \sqrt{\frac{1}{124.8} \times .231} = .0430$$

The test statistic is given by

$$z = \frac{.378}{.043} = 8.8 \sim N(0,1) \text{ under } H_0$$

The  $p$ -value is  $p = 1 - \Phi(8.8) < .001$

Thus the Kappa statistic indicates highly significant reproducibility between the first and second surveys for beef consumption.

Although the Kappa statistic was significant in Example 10.44, it still shows that reproducibility was far from perfect. Indeed, Landis and Koch [7] provide the following guidelines for the evaluation of Kappa.

**Equation 10.28****Guidelines for Evaluating Kappa**

$\kappa > .75$  denotes *excellent* reproducibility.

$.4 \leq \kappa \leq .75$  denotes *good* reproducibility.

$0 \leq \kappa < .4$  denotes *marginal* reproducibility.

In general, reproducibility is not good for many items in dietary surveys, indicating the need for multiple dietary assessments to reduce variability. See Fleiss [8] for further information about the Kappa statistic, including assessments of reproducibility for more than two surveys.

Kappa is usually used as a measure of reproducibility between repeated assessments of the same variable. If we are interested in the concordance between responses on two different variables, where one variable can be considered to be a gold standard, then sensitivity and specificity, which are measures of validity of a screening test (see Chapter 3), are more appropriate indices than Kappa, which is a measure of reliability.

## 10.9 Summary

This chapter discussed the most widely used techniques for analyzing qualitative (or categorical) data. First, the problem of how to compare binomial proportions from two independent samples was studied. For the large-sample case, this problem was solved in two different (but equivalent) ways: using either the two-sample test for binomial proportions or the chi-square test for  $2 \times 2$  contingency tables. The former method is similar to the *t* test methodology introduced in Chapter 8, whereas the contingency-table approach can be easily generalized to more complex problems involving qualitative data. For the small-sample case, Fisher's exact test is used to compare binomial proportions in two independent samples. To compare binomial proportions in paired samples, such as when a person is used as his or her own control, McNemar's test for correlated proportions should be used.

The  $2 \times 2$  contingency-table problem was extended to the investigation of the relationship between two qualitative variables, in which one or both variables have more than two possible categories of response. A chi-square test for  $R \times C$  contingency tables was developed, which is a direct generalization of the  $2 \times 2$  contingency-table test. In addition, if one of the variables is a disease variable with a binary outcome and the other variable is an exposure variable with  $k$  ordered categories, then the chi-square test for trend was introduced to assess whether the disease rate follows a consistent increasing or decreasing trend as the level of exposure increases. Also, we studied how to assess the goodness-of-fit of probability models proposed in earlier chapters. The chi-square goodness-of-fit test was used to address this problem. Furthermore, the Kappa statistic was introduced as an index of reproducibility for categorical data.

Finally, formulas to compute sample size and power for comparing two binomial proportions were provided in the independent-sample and paired-sample case. Special considerations in computing sample size and power were discussed in a clinical trial setting. The methods in this chapter are illustrated in the flowchart in Figure 10.16 (p. 409).

In Chapters 8, 9, and 10, we considered the comparison between two groups for variables measured on a continuous, ordinal, and categorical scale, respectively. In Chapter 11, we discuss methods for studying the relationship between a continuous response variable and one or more predictor variables, which can be either continuous or categorical.

## PROBLEMS

### Cardiovascular Disease

Consider the Physicians' Health Study data presented in Example 10.32.

**10.1** How many participants need to be enrolled in each group to have a 90% chance of detecting a significant difference using a two-sided test with  $\alpha = .05$  if compliance is perfect?

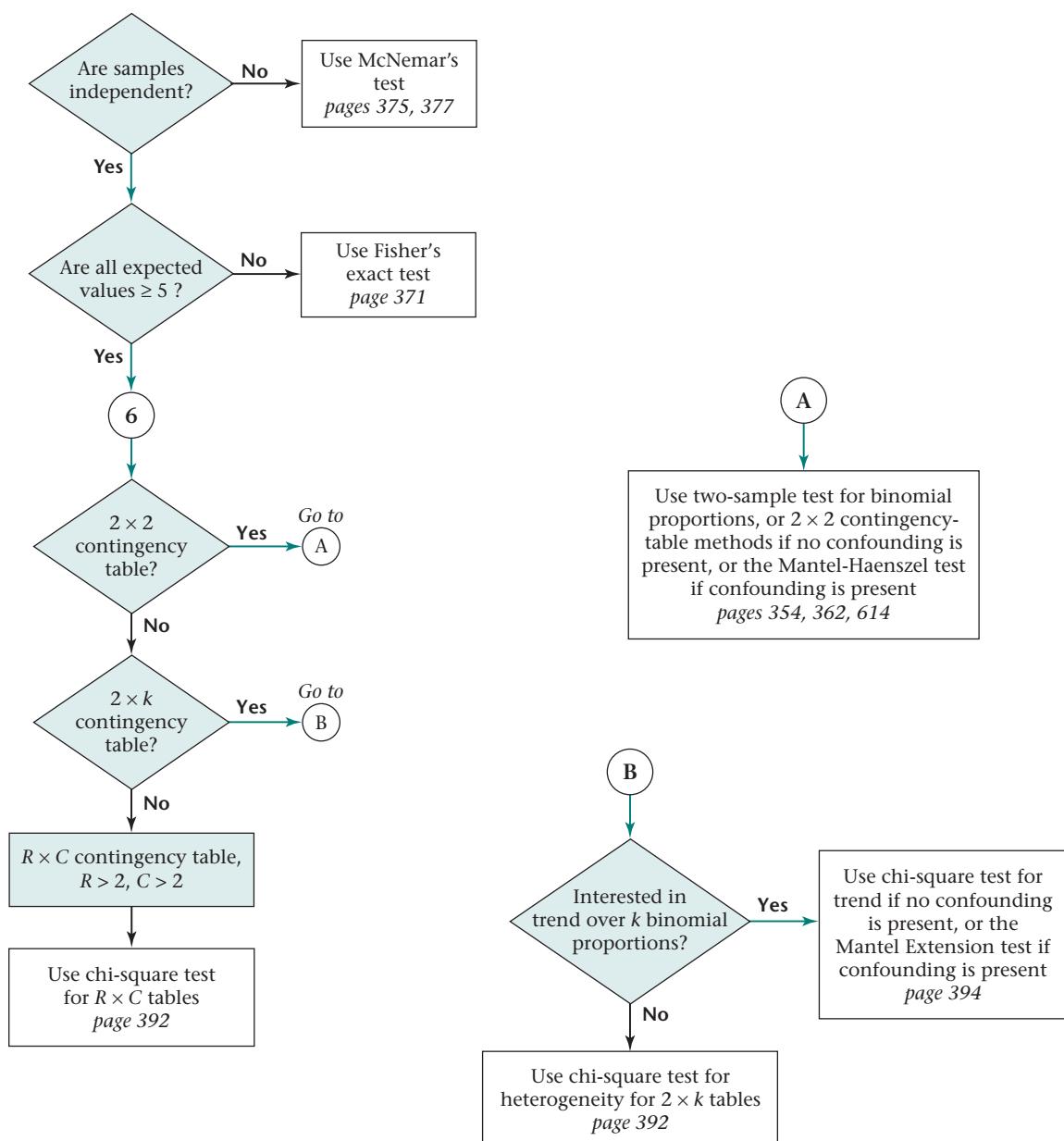
**10.2** Answer Problem 10.1 if compliance is as given in Example 10.32.

**10.3** Answer Problem 10.1 if a one-sided test with power = .8 is used and compliance is perfect.

**10.4** Suppose 11,000 men are actually enrolled in each treatment group. What would be the power of such a study if a two-sided test with  $\alpha = .05$  were used and compliance is perfect?

**10.5** Answer Problem 10.4 if compliance is as given in Example 10.32.

**Figure 10.16 Flowchart for appropriate methods of statistical inference for categorical data**



Refer to Table 2.11.

**10.6** What significance test can be used to assess whether there is a relationship between receiving an antibiotic and receiving a bacterial culture while in the hospital?

**10.7** Perform the test in Problem 10.6, and report a  $p$ -value.

### Gastroenterology

Two drugs (A, B) are compared for the medical treatment of duodenal ulcer. For this purpose, patients are carefully matched with regard to age, sex, and clinical condition. The treatment results based on 200 matched pairs show that for 89 matched pairs both treatments are effective; for 90 matched pairs both treatments are ineffective; for 5 matched pairs drug A is effective, whereas drug B is ineffective; and for 16 matched pairs drug B is effective, whereas drug A is ineffective.

\***10.8** What test procedure can be used to assess the results?

\***10.9** Perform the test in Problem 10.8, and report a  $p$ -value.

In the same study, if the focus is on the 100 matched pairs consisting of male patients, then the following results are obtained: for 52 matched pairs both drugs are effective; for 35 matched pairs both drugs are ineffective; for 4 matched pairs drug A is effective, whereas drug B is ineffective; and for 9 matched pairs drug B is effective, whereas drug A is ineffective.

\***10.10** How many concordant pairs are there among the male matched pairs?

\***10.11** How many discordant pairs are there among the male matched pairs?

\***10.12** Perform a significance test to assess any differences in effectiveness between the drugs among males. Report a  $p$ -value.

### Sexually Transmitted Disease

Suppose researchers do an epidemiologic investigation of people entering a sexually transmitted disease clinic. They find that 160 of 200 patients who are diagnosed as having gonorrhea and 50 of 105 patients who are diagnosed as having nongonococcal urethritis have had previous episodes of urethritis.

\***10.13** Are the present diagnosis and prior episodes of urethritis associated?

### Cancer

**10.14** A 1980 study investigated the relationship between the use of OCs and the development of endometrial cancer [9]. The researchers found that of 117 endometrial-cancer patients, 6 had used the OC Oracan at some time in their lives, whereas 8 of the 395 controls had used this agent. Test for an association between the use of Oracan and the incidence of endometrial cancer, using a two-tailed test.

### Ophthalmology

Retinitis pigmentosa is a disease that manifests itself via different genetic modes of inheritance. Cases have been documented with a dominant, recessive, and sex-linked mode of inheritance. It has been conjectured that mode of inheritance is related to the ethnic origin of the individual. Cases of the disease have been surveyed in an English and a Swiss population with the following results: Of 125 English cases, 46 had sex-linked disease, 25 had recessive disease, and 54 had dominant disease. Of 110 Swiss cases, 1 had sex-linked disease, 99 had recessive disease, and 10 had dominant disease.

\***10.15** Do these data show a significant association between ethnic origin and genetic type?

### Pulmonary Disease

One important aspect of medical diagnosis is its reproducibility. Suppose that two different doctors examine 100 patients for dyspnea in a respiratory-disease clinic and that doctor A diagnosed 15 patients as having dyspnea, doctor B diagnosed 10 patients as having dyspnea, and both doctor A and doctor B diagnosed 7 patients as having dyspnea.

**10.16** Compute the Kappa statistic and its standard error regarding reproducibility of the diagnosis of dyspnea in this clinic.

### Infectious Disease

Suppose a computerized database contains all charts of patients at nine hospitals in Cleveland, Ohio. One concern of the group conducting the study is the possibility that the attending physician underreports or overreports various diagnoses that seem consistent with a patient's chart. An investigator notes that 50 of the 10,000 people in the database are reported as having a particular viral infection by their attending physician. A computer using an automated method of diagnosis claims that 68 of the 10,000 people have the infection, 48 of them from the attending physician's 50 positives and 20 from the attending physician's 9950 negatives.

**10.17** Test the hypothesis that the probability of detecting this viral infection are the same for the computer and the attending physician.

### Cardiovascular Disease

An investigator wants to study the effect of cigarette smoking on the development of myocardial infarction (MI) in women. In particular, some question arises in the literature as to the relationship of timing of cigarette smoking to development of disease. One school of thought says that current smokers are at much higher risk than ex-smokers. Another school of thought that says a considerable latent period of nonsmoking is needed before the risk of ex-smokers becomes less

than that of current smokers. A third school of thought is that ex-smokers may actually have a higher incidence of MI than current smokers because they may include more women with some prior cardiac symptoms (e.g., angina) than current smokers. To test this hypothesis, 2000 disease-free currently smoking women and 1000 disease-free ex-smoking women, ages 50–59, are identified in 1996, and the incidence of MI between 1996 and 1998 is noted at follow-up visits 2 years later. Investigators find that 40 currently smoking women and 10 ex-smoking women have developed the disease.

\***10.18** Is a one-sample or a two-sample test needed here?

\***10.19** Is a one-sided or a two-sided test needed here?

**10.20** Which of the following test procedures should be used to test this hypothesis? (More than one may be necessary.) (*Hint:* Use the flowchart in Figure 10.16.)

(1)  $\chi^2$  test for  $2 \times 2$  contingency tables

(2) Fisher's exact test

(3) McNemar's test

(4) One-sample binomial test

(5) One-sample  $t$  test

(6) Two-sample  $t$  test with equal variances

\***10.21** Carry out the test procedure(s) mentioned in Problem 10.20, and report a  $p$ -value.

## Cancer

The following data are provided from the SEER Cancer Registries of the National Cancer Institute based on 17 cancer registries in the United States. Data are available for stage of disease at diagnosis for women with breast cancer by age and race as shown in Table 10.24 [10].

**10.22** Test whether the distribution of stage at disease is significantly different between Caucasian and African-American women with breast cancer who are younger than 50 years of age. Please provide a  $p$ -value (two-tailed). Ignore the unstaged cases in your analysis.

The 5-year survival rates by stage of disease, age at diagnosis, and race are provided in Table 10.25.

**10.23** Test whether the 5-year survival rate for breast cancer is significantly different between African-American and Caucasian women who are younger than 50 years of age and have localized disease. Provide a  $p$ -value (two-tailed).

## Sexually Transmitted Disease

Suppose a study examines the relative efficacy of penicillin and spectinomycin in treating gonorrhea. Three treatments are considered: (1) penicillin, (2) spectinomycin, low dose, and (3) spectinomycin, high dose. Three possible responses

**Table 10.24 Stage of breast cancer at diagnosis by age and race, SEER Cancer data, 1999–2005**

	Caucasian females		African-American females	
	<50	50+	<50	50+
Stage	( <i>n</i> = 53,060)	( <i>n</i> = 174,080)	( <i>n</i> = 8063)	( <i>n</i> = 16,300)
Localized	54*	64	46	53
Regional	41	29	46	35
Distant	3	5	7	9
Unstaged	2	2	2	3

\*percent.

**Table 10.25 Five-year survival rates for breast cancer by stage at diagnosis, age at diagnosis, and race, SEER Cancer data, 1999–2005**

	Caucasian females		African-American females	
	<50	50+	<50	50+
Stage	( <i>n</i> = 53,060)	( <i>n</i> = 174,080)	( <i>n</i> = 8063)	( <i>n</i> = 16,300)
Localized	96.5*	99.6	91.6	94.9
Regional	84.6	85.0	71.3	72.6
Distant	33.2	22.5	15.0	16.4
Unstaged	76.7	53.5	49.7	42.2

\*percent.

are recorded: (1) positive smear, (2) negative smear, positive culture, (3) negative smear, negative culture. The data in Table 10.26 are obtained.

**Table 10.26 Efficacy of different treatments for gonorrhea**

Treatment	Response				Total
	+ Smear	– Smear + Culture	– Smear – Culture		
Penicillin	40	30	130	200	
Spectinomycin (low dose)	10	20	70	100	
Spectinomycin (high dose)	15	40	45	100	
Total	65	90	245	400	

**10.24** Is there any relationship between the type of treatment and the response? What form does the relationship take?

**10.25** Suppose either a positive smear or a positive culture is regarded as a positive response and distinguished from the negative smear, negative culture response. Is there an association between the type of treatment and this measure of response?

### Diabetes

Improving control of blood-glucose levels is an important motivation for the use of insulin pumps by diabetic patients. However, certain side effects have been reported with pump therapy. Table 10.27 provides data on the occurrence of diabetic ketoacidosis (DKA) in patients before and after start of pump therapy [11].

Text not available due to copyright restrictions

\***10.26** What is the appropriate procedure to test whether the rate of DKA is different before and after start of pump therapy?

\***10.27** Perform the significance test in Problem 10.26, and report a *p*-value.

### Renal Disease

A study group of 576 working women 30–49 years of age who took phenacetin-containing analgesics and a control group of 533 comparably aged women without such intake

were identified in 1968 and followed for mortality and morbidity outcomes. One hypothesis to be tested was that phenacetin intake may influence renal (kidney) function and hence have an effect on specific indices of renal morbidity and mortality. The mortality status of these women was determined from 1968 to 1987. The researchers found that 16 of the women in the study group and 1 of the women in the control group died, where at least one cause of death was considered renal [12].

**10.28** To test for differences in renal mortality between the two groups in either direction, what statistical test should be used?

**10.29** Implement the test in Problem 10.28, and report a *p*-value.

The cohort was also followed for total mortality. The researchers found that 74 women in the study group died, compared with 27 in the control group.

**10.30** What statistical test should be used to compare the total mortality experience of the study group with that of the control group?

**10.31** Implement the test in Problem 10.30, and report a *p*-value.

### Mental Health

A study was performed in Lebanon looking at the effect of widowhood on mortality [13]. Each of 151 widowers and 544 widows were matched to a person married at the time of widowhood and of the same age ( $\pm 2$  years) and sex. The people in the matched pairs were followed until one member of the matched pair died. The results in Table 10.28 were obtained for those matched pairs in which at least one member had died by 1980.

\***10.32** Suppose all the matched pairs in Table 10.28 are considered. What method of analysis can be used to test whether there is an association between widowhood and mortality?

**Table 10.28 Effect of widowhood on mortality**

Age (years)	Males		Females	
	<i>n</i> <sub>1</sub> <sup>a</sup>	<i>n</i> <sub>2</sub> <sup>b</sup>	<i>n</i> <sub>1</sub>	<i>n</i> <sub>2</sub>
36–45	4	8	3	2
46–55	20	17	17	10
56–65	42	26	16	15
66–75	21	10	18	11
Unknown	0	2	3	2
Total	87	63	57	40

<sup>a</sup>*n*<sub>1</sub> = number of pairs in which the widowed subject is deceased and the married subject is alive.

<sup>b</sup>*n*<sub>2</sub> = number of pairs in which the widowed subject is alive and the married subject is deceased.

Source: Reprinted with the permission of the *American Journal of Epidemiology*, 125(1), 127–132, 1987.

\***10.33** Implement the test in Problem 10.32, and report a *p*-value.

**10.34** Answer the same question as Problem 10.32 considering 36- to 45-year-old males only.

**10.35** Implement the test in Problem 10.34, and report a *p*-value.

**10.36** Based on all matched pairs, how much power did the study just mentioned have vs. the alternative hypothesis that a widower is twice as likely to die before a married person of the same age and sex, assuming that all age groups are considered?

### Hepatic Disease

Refer to Data Set HORMONE.DAT, on the Companion Website. (See p. 315 for a description of this Data Set.)

**10.37** What test procedure can be used to compare the percentage of hens whose pancreatic secretions increased (post-pre) among the five treatment regimens?

**10.38** Implement the test procedure in Problem 10.37, and report a *p*-value.

**10.39** Answer Problem 10.38 for biliary secretions.

**10.40** For all hormone groups except saline, different doses of hormones were administered to different groups of hens. Is there a dose-response relationship between the proportion of hens with increasing pancreatic secretions and the hormone dose? This should be assessed separately for each specific active hormone.

**10.41** Answer Problem 10.40 for biliary secretions.

### Cardiovascular Disease

A secondary prevention trial of lipid lowering is planned in patients with previous myocardial infarction (MI). Patients are to be randomized to either a treatment group receiving diet therapy and cholesterol-lowering drugs or a control group receiving diet therapy and placebo pills. The study endpoint is to be a combined endpoint consisting of either definite fatal coronary heart disease or nonfatal MI (i.e., a new nonfatal MI distinct from previous events). Suppose it is projected that the incidence of combined events among controls is 7% per year.

\***10.42** What proportion of controls will have events over 5 years?

Suppose the treatment benefit is projected to be a reduction in the 5-year event rate by 30%.

**10.43** What is the expected event rate in the treated group?

\***10.44** If the rates in Problems 10.42 and 10.43 are the true rates, how many participants will be needed in

each group if a one-sided test with  $\alpha = .05$  is to be used and an 80% chance of finding a significant difference is desired?

Investigators expect that not all participants will comply. In particular, they project that 5% of the treatment group will not comply with drug therapy and that 10% of the control group will start taking cholesterol-lowering drugs outside the study.

**10.45** What will be the expected rates in Problems 10.42 and 10.43 if this level of lack of compliance is realized?

**10.46** What will be the revised sample size estimate in Problem 10.44 if the lack of compliance is taken into account?

### Pediatrics, Endocrinology

A study was performed among 40 boys in a school in Edinburgh to look at the presence of spermatozoa in urine samples according to age [14]. The boys entered the study at 8–11 years of age and left the study at 12–18 years of age. A 24-hour urine sample was supplied every 3 months by each boy. Table 10.29 gives the presence or absence of sperm cells in the urine samples for each boy together with the ages at entrance and exit of the study and the age at the first sperm-positive urine sample.

For all parts of this question, we exclude boys who exited this study without 1 sperm-positive urine sample (i.e., boys 8, 9, 14, 25, 28, 29, 30).

**10.47** Provide a stem-and-leaf plot of the age at first sperm-positive urine specimen.

**10.48** If we assume that all boys have no sperm cells at age 11 (11.0 years) and all have sperm cells at age 18, then estimate the probability of first developing sperm cells at ages 12 (i.e., between 12.0 and 12.9 years), 13, 14, 15, 16, and 17.

**10.49** Suppose mean age at spermatogenesis = 13.67 years, with standard deviation = 0.89 years and we assume that the age at spermatogenesis follows a normal distribution. The pediatrician would like to know what is the earliest age (in months) before which 95% of boys experience spermatogenesis because he or she would like to refer boys who haven't experienced spermatogenesis by this age to a specialist for further follow-up. Can you estimate this age from the information provided in this part of the problem?

**10.50** Suppose we are uncertain whether a normal distribution provides a good fit to the distribution of age at spermatogenesis. Answer this question using the results from Problems 10.47–10.49. (Assume that the large-sample method discussed in this chapter is applicable to these data.)

**Table 10.29** Presence (+) or absence (-) of spermatazoa in consecutive urine samples for all boys; age at first collected urine sample, at first positive, and at last sample

Boy	Age at			Observations
	Entrance	First positive	Exit	
1	10.3	13.4	16.7	-----+-----++++-
2	10.0	12.1	17.0	-----+---++-+---+-+-----++
3	9.8	12.1	16.4	-----+---++-+---++-+---++-
4	10.6	13.5	17.7	-----+---++-+---+----
5	9.3	12.5	16.3	-----+---+---+-----+-----
6	9.2	13.9	16.2	-----+-----+-----
7	9.6	15.1	16.7	-----+-----+---++-
8	9.2	—	12.2	-----
9	9.7	—	12.1	-----
10	9.6	12.7	16.4	-----+---++-+---++-+---++
11	9.6	12.5	16.7	-----+---+---+---++-
12	9.3	15.7	16.0	-----+-----+-----++
14	9.6	—	12.0	-----
16	9.4	12.6	13.1	-----+---++-
17	10.5	12.6	17.5	-----+---++-+---++-+---++
18	10.5	13.5	14.1	-----+--
19	9.9	14.3	16.8	-----+-----+---++-
20	9.3	15.3	16.2	-----+-----+-----++
21	10.4	13.5	17.3	-----+---++-+---+---++
22	9.8	12.9	16.7	-----+---++-+---++-+---++
23	10.8	14.2	17.3	-----+-----+---++-+---++
24	10.9	13.3	17.8	-----+---++-+---++-+---++
25	10.6	—	13.8	-----
26	10.6	14.3	16.3	-----+---++-+---++
27	10.5	12.9	17.4	-----+---++-+---++-+---++
28	11.0	—	12.4	-----
29	8.7	—	12.3	-----
30	10.9	—	14.5	-----
31	11.0	14.6	17.5	-----+---++-+---++-+---++
32	10.8	14.1	17.6	-----+---+---+-----
33	11.3	14.4	18.2	-----+---+---+-----
34	11.4	13.8	18.3	-----+---+---+---+---++-
35	11.3	13.7	17.8	-----+---+---+---+---++
36	11.2	13.5	15.7	-----+-----
37	11.3	14.5	16.3	-----+---+---+---++
38	11.2	14.3	17.2	-----+---+---+---++-+---++
39	11.6	13.9	14.7	-----+---
40	11.8	14.1	17.9	-----+---+---+---+---++
41	11.4	13.3	18.2	-----+---+---+---+---++
42	11.5	14.0	17.9	-----+---+---+---+---++

**Table 10.30 Reproducibility of types of malpractice**

		(a) Adverse events		(b) Negligence	
		Review process B		Review process B	
		+	-	+	-
Review process A	+	35	13	+	4
	-	21	249	-	12 293

### Health Services Administration

In the Harvard Medical Practice Study [15], a sample of 31,429 medical records of hospital patients were reviewed to assess the frequency of medical malpractice. Two types of malpractice were identified:

- (1) An *adverse event* was defined as an injury caused by medical management (rather than by the underlying disease).
- (2) *Negligence* was defined as care that fell below the standard expected of physicians in the community.

An approximate 1% sample of records was reviewed on two different occasions by different review teams. The data in Table 10.30 were obtained.

**10.51** Are frequencies of reporting of adverse events or negligence comparable in review process A (the original review) and review process B (the re-review)?

**10.52** Can you assess the reproducibility of adverse events and negligence designations? Which seems to be more reproducible?

### Infectious Disease

A study looked at risk factors for human immunodeficiency virus (HIV) infection among intravenous drug users enrolled in a methadone program [16]. The data in Table 10.31 were presented regarding the presence of HIV antibody among 120 non-Hispanic white subjects by total family income, as assessed by blood testing at one point in time.

**Table 10.31 Presence of HIV antibody among 120 non-Hispanic whites, by income level**

Group	Total family income (\$)	Number of HIV-positive	Number tested
A	< 10,000	13	72
B	10,000–20,000	5	26
C	>20,000	2	22

**10.53** What test can be used to assess whether total family income is related in a consistent manner to percentage of HIV-positive?

**10.54** Implement the test, and report a *p*-value.

### Sports Medicine

Data from a study concerning tennis elbow (a painful condition experienced by many tennis players) are given in TENNIS1.DAT, and the format is given in TENNIS1.DOC (both on the Companion Website).

Members of several tennis clubs in the greater Boston area were surveyed. Each participant was asked whether he or she had ever had tennis elbow and, if so, how many episodes. An attempt was made to enroll roughly an equal number of participants with at least one episode of tennis elbow (the cases) and participants with no episodes of tennis elbow (the controls). The interviewer asked the participants about other possibly related factors, including demographic factors (e.g., age, sex) and characteristics of their tennis racquet (e.g., string type of racquet used, material(s) of racquet). This type of study is a case-control study and also can be considered as an **observational study**. It is distinctly different from a clinical trial, in which treatments are assigned at random. Because of randomization, participants receiving different treatments in a clinical trial will, on average, tend to be comparable. In an observational study, we are interested in relating risk factors to disease outcomes. However, it is difficult to make causal inferences (e.g., “wood racquets cause tennis elbow”) because participants are not assigned to a type of racquet at random. Indeed, if we find differences in the frequency of tennis elbow by type of racquet, there may be some other variable(s) that are related to both tennis elbow and to the type of racquet that are more direct “causes” of tennis elbow. Nevertheless, observational studies are useful in obtaining important clues as to disease etiology. One interesting aspect of observational studies is that there are often no prior leads as to which risk factors are even associated with disease. Therefore, investigators tend to ask many questions about possible risk factors without having a firm idea as to which risk factors are really important.

**10.55** In this problem, act like a detective and look at each risk factor in the data set separately and relate this risk factor to tennis elbow. How you define tennis elbow is somewhat arbitrary. You may want to compare participants with 1+ episodes of tennis elbow vs. participants with no episodes. Or you may want to focus specifically on participants with multiple episodes of tennis elbow, or perhaps create a graded scale according to the number of episodes of tennis elbow (e.g., 0 episodes, 1 episode, 2+ episodes), etc. In this exercise, consider each risk factor separately. In Chapter 13, we will discuss logistic regression methods, where we will be able to study the effects of more than one risk factor simultaneously on disease.

### Otolaryngology

Consider Data Set EAR.DAT, on the Companion Website. (See p. 64.)

**10.56** For children with one affected ear at baseline, compare the efficacy of the two study medications.

**10.57** In children with two affected ears at baseline, compare the efficacy of the two study medications, treating the response at 14 days as a graded scale (two cleared ears, one cleared ear, no cleared ears).

**10.58** For children with two affected ears, test the hypothesis that response to the study medications for the first and second ears is independent.

### Hospital Epidemiology

Death of a patient in the hospital is a high-priority medical outcome. Some hospital deaths may be due to inadequate care and are potentially preventable. An *adverse event* during a hospital stay is defined as a problem of any nature and seriousness experienced by a patient during his or her stay in the hospital that is potentially attributable to clinical or administrative management rather than the underlying disease. A study in a hospital in Granada, Spain assessed whether there was a relationship between adverse events and deaths during hospital stay [17]. In this study, 524 cases (i.e., people who died in the hospital) were identified between January 1, 1990 and January 1, 1991. For each case, a control patient was matched on admission diagnosis and admission date. A retrospective chart review determined occurrence of adverse events among all cases and controls. There were 299 adverse events occurring among the cases and 225 among the controls. Among the 299 cases in which an adverse event occurred, 126 of their corresponding matched controls also had an adverse event.

**10.59** What method of analysis can be used to compare the proportion of adverse events between cases and controls?

**10.60** Implement the method suggested in Problem 10.59, and report a two-tailed *p*-value.

### Cancer

A topic of current interest is whether abortion is a risk factor for breast cancer. One issue is whether women who have had abortions are comparable to women who have not had abortions in terms of other breast-cancer risk factors. One of the best-known breast-cancer risk factors is parity (i.e., number of children), with parous women with many children having about a 30% lower risk of breast cancer than nulliparous women (i.e., women with no children). Hence it is important to assess whether the parity distribution of women with and without previous abortions is comparable. The data in Table 10.32 were obtained from the Nurses' Health Study on this issue.

**Table 10.32 Parity distribution of women with abortions and women without abortions**

Parity	Induced abortion	
	Yes (n = 16,353)	No (n = 77,220)
0	34%	29%
1	23%	18%
2	30%	34%
3	10%	15%
4+	3%	4%

**10.61** What test can be performed to compare the parity distribution of women with and without induced abortions?

**10.62** Implement the test in Problem 10.61, and report a two-tailed *p*-value.

Suppose that with each additional child, breast-cancer risk is reduced by 10% (i.e., women with 1 child have a risk of breast cancer that is 90% of that of a nulliparous woman of the same age; women with 2 children have a risk that is .9<sup>2</sup> or 81% of that of a nulliparous woman, etc.). (For the purposes of this problem, consider women with 4+ births as having exactly 4 births.)

**10.63** Suppose there is no causal effect of induced abortion on breast cancer. Based on the parity distribution in the two groups, would women with induced abortion be expected to have the same, higher, or lower risk of breast cancer? If higher or lower, by how much? (Assume that the age distributions are the same between women who have or have not had previous abortions.)

### Ophthalmology

A 5-year study among 601 participants with retinitis pigmentosa assessed the effects of high-dose vitamin A (15,000 IU per day) and vitamin E (400 IU per day) on the course of their disease. One issue is to what extent supplementation with vitamin A affected their serum-retinol levels.

The serum-retinol data in Table 10.33 were obtained over 3 years of follow-up among 73 males taking 15,000 IU/day of vitamin A (vitamin A group) and among 57 males taking 75 IU/day of vitamin A (the trace group; this is a negligible amount compared with usual dietary intake of 3000 IU/day).

**Table 10.33 Effect of vitamin A supplementation on serum-retinol levels**

Retinol (mmol/L)	<i>N</i>	Year 0	Year 3
		Mean $\pm$ <i>sd</i>	Mean $\pm$ <i>sd</i>
Vitamin A group	73	1.89 $\pm$ 0.36	2.06 $\pm$ 0.53
Trace group	57	1.83 $\pm$ 0.31	1.78 $\pm$ 0.30

**10.64** What test can be used to assess whether mean serum retinol has increased over 3 years among subjects in the vitamin A group?

**10.65** Can the test be implemented based on the data just presented? Why or why not? If yes, implement the test and report a two-tailed *p*-value.

**10.66** One assumption of the test in Problem 10.64 is that the distribution of serum retinol is approximately normal. To verify this assumption, the investigators obtained a frequency distribution of serum retinol at year 0 among males in the vitamin A group, with data as shown in Table 10.34.

**Table 10.34 Distribution of serum retinol in a retinitis-pigmentosa population**

Serum-retinol group ( $\mu\text{mol/L}$ )	<i>n</i>
$\leq 1.40$	6
1.41–1.75	22
1.76–2.10	22
2.11–2.45	20
$\geq 2.46$	3
	73

Perform a statistical test to check on the normality assumption. Given your results, do you feel the assumption of normality is warranted? Why or why not?

## Ophthalmology

One interesting aspect of the study described in Problem 10.64 is to assess changes in other parameters as a result of supplementation with vitamin A. One quantity of interest is the level of serum triglycerides. Researchers found that among 133 participants in the vitamin A group (males and females combined) who were in the normal range at baseline ( $<2.13 \mu\text{mol/L}$ ), 15 were above the upper limit of normal at each of their last 2 consecutive study visits. Similarly, among 138

participants in the trace group who were in the normal range at baseline ( $<2.13 \mu\text{mol/L}$ ), 2 were above the upper limit of normal at each of their last two consecutive study visits [18].

**10.67** What proportion of people in each group developed abnormal triglyceride levels over the course of the study? Are these proportions measures of prevalence, incidence, or neither?

**10.68** What test can be performed to compare the percentage of participants who developed abnormal triglyceride levels between the vitamin A group and the trace group?

**10.69** Implement the test in Problem 10.68, and report a two-tailed *p*-value.

## Zoology

A study was performed to look at the preference of different species of birds for different types of sunflower seeds. Two bird feeders were set up with different types of sunflower seeds, one with a black oil seed and one with a striped seed. The bird feeders were observed for a 1-hour period for each of 12 days over a 1-month period. The number of birds of different species who ate seeds from a specific bird feeder was counted for each bird feeder for each of a number of species of birds. (The data for this problem were supplied by David Rosner.)

On the first day of testing, 1 titmouse ate the black oil seeds and 4 titmice ate the striped seeds. Of the goldfinches, 19 ate the black oil seeds and 5 ate the striped seeds.

**10.70** What test can be performed to assess whether the feeding preferences of titmice and goldfinches are comparable on the first day of testing?

**10.71** Implement the test in Problem 10.70, and report a *p*-value.

One assumption in the entire experiment is that the feeding preferences of the same species of bird remain the same over time. To test this assumption, the data for goldfinches were separated by the 6 different days on which they were observed (they were not observed at all for the other 6 days). For 2 of the 6 days small numbers of goldfinches were observed (2 on one day and 1 on another day). Thus data from these two days were also not included. The results for the remaining 4 days are shown in Table 10.35.

**Table 10.35 Feeding preferences of goldfinches on different days**

Type of seed	Day				
	1	2	3	4	Total
Black oil	19	14	9	45	87
Striped	5	10	6	39	60

**10.72** What test can be used to assess whether the feeding preference of goldfinches are the same on different days?

**10.73** Implement the test in Problem 10.72, and report a  $p$ -value.

## Cancer

The Physicians' Health Study was a randomized double-blind placebo-controlled trial of beta-carotene (50 mg every other day). In 1982, the study enrolled 22,071 male physicians ages 40–84. The participants were followed until December 31, 1995 for the development of new cancers (malignant neoplasms). The results reported [19] are shown in Table 10.36.

**Table 10.36 Comparison of cancer incidence rates between the beta-carotene and placebo groups**

	Beta-carotene ( $n = 11,036$ )	Placebo ( $n = 11,035$ )
Malignant neoplasms	1273	1293

**10.74** What test can be used to compare cancer incidence rates between the two treatment groups?

**10.75** Implement the test in Problem 10.74, and report a two-tailed  $p$ -value.

**10.76** The expectation before the study started was that beta-carotene might prevent 10% of incident cancers relative to placebo. How much power did the study have to detect an effect of this magnitude if a two-sided test is used with  $\alpha = .05$  and we assume that the true incidence rate in the placebo group is the same as the observed incidence rate and that compliance with study medications is perfect?

## Demography

A common assumption is that the gender of successive offspring is independent. To test this assumption, birth records were collected from the first 5 births in 51,868 families. For families with exactly 5 children, Table 10.37 shows a frequency distribution of the number of male offspring (from Data Set SEXRAT.DAT; see p. 103).

**Table 10.37 Frequency distribution of number of male offspring in families of size 5**

Number of male offspring	$n$
0	518
1	2245
2	4621
3	4753
4	2476
5	549
Total	15,162

Suppose the investigators doubt the probability of a male birth is exactly 50% but are willing to assume the gender distributions of successive offspring are independent.

**10.77** What is the best estimate of the probability of a male birth based on the observed data?

**10.78** What is the probability of 0, 1, 2, 3, 4, and 5 male births out of 5 births based on the estimate in Problem 10.77?

**10.79** Test the hypothesis that the gender distributions of successive offspring are independent based on the model in Problem 10.78. What are your conclusions concerning the hypothesis?

## Pediatrics, Urology

Nighttime bladder control is an important developmental milestone, with failure dependent on age. Continence is usually achieved between 4 and 6 years of age, but an important minority of children experience delays in success. A longitudinal study was conducted in Britain in which nighttime bedwetting was assessed at ages 4, 6, 8, 9, 11, and 15 years among 3272 children in the Medical Research Council's 1946 National Survey of Health [20].

The following data were presented. There were 1362 boys and 1313 girls who reported no bedwetting at any of the six ages just listed. Consider this as the control group. There were 6 boys and 2 girls who reported no bedwetting at ages 4, 6, and 8 but reported some bedwetting at both ages 9 and 11. Consider this as the case group. Ignore children with any other pattern of bedwetting over the six ages.

**10.80** What test can be used to assess whether the percentage of cases among boys is significantly different from the percentage of cases among girls?

**10.81** Implement the test in problem 10.80, and report a two-tailed  $p$ -value. (*Hint:* Table 10.38 may be helpful in answering this question.)

**Table 10.38 Hypergeometric probabilities that are useful in analyzing the bedwetting data set**

x1	n1	x	n	HYPGEOMDIST(x1,n1,x,n)
0	1368	8	2683	.003
1	1368	8	2683	.028
2	1368	8	2683	.101
3	1368	8	2683	.210
4	1368	8	2683	.273
5	1368	8	2683	.227
6	1368	8	2683	.118
7	1368	8	2683	.035
8	1368	8	2683	.005
Total				1

## Health Promotion

Obesity is an important risk factor for many diseases. However, in studying the effects of obesity it is important to be aware of other risk factors that may be potentially related to obesity. One commonly used measure of obesity is body-mass index (BMI) ( $\text{kg}/\text{m}^2$ ), which is often categorized as follows: normal = BMI < 25, overweight = BMI 25.0–29.9, and obese = BMI  $\geq 30.0$ . The data in Table 10.39 were presented in a study relating education to BMI category.

**10.82** What test can be used to compare the percentage of obese and normal individuals with at least a high-school education?

**10.83** Implement the test in Problem 10.82, and report a two-tailed  $p$ -value.

**Table 10.39 Relationship between BMI category and education level ( $n = 261$ )**

BMI category	$n$	% $\geq$ high-school education
Normal	77	91
Overweight	120	87.5
Obese	64	83

**10.84** What test can be used to compare the percentage of individuals with at least a high-school education between the three BMI groups in Table 10.39?

**10.85** Implement the test in Problem 10.84, and report a two-tailed  $p$ -value.

## Otolaryngology

Acute OTM early in infancy may be an important predictor of subsequent morbidity, including psychological and educational difficulties. A study was performed among high-risk infants who had already experienced either a single episode of acute OTM prior to the age of 6 months or two or more episodes before 12 months of life [21].

Children were randomized to one of three treatment groups, (a) amoxicillin (AMX), (b) sulfisoxazole (SUL), or (c) placebo (PLA), and their parents were told to administer the drug daily for a 6-month period (even in the absence of symptoms). If children had an acute OTM episode during the study period, they received the best antibiotic care, regardless of their study-drug group. The results were as shown in Table 10.40.

**Table 10.40 Experience with acute OTM in 6 months after entry into the study**

Drug group	% Acute OTM-free	$n$
AMX	70	40
SUL	47	36
PLA	32	41

**10.86** What test can be used to compare the percentage of children who were acute OTM-free between the AMX group and the PLA group?

**10.87** Perform the test in Problem 10.86, and report a two-tailed  $p$ -value.

The children were followed for an additional 6 months after the study-drug period (first 6 months) was over. The results reported concerning acute OTM experience over the entire 12-month period are shown in Table 10.41.

**Table 10.41 Experience with acute OTM in 12 months after entry into the study**

Drug group	% Acute OTM-free	$n$
AMX	38	40
SUL	28	36
PLA	22	41

**10.88** Perform a test to compare the percentage acute OTM-free in the AMX group vs. the PLA group over the second 6 months of the study among children who were acute OTM-free after the first 6 months. Report a two-tailed  $p$ -value.

## Infectious Disease

Smallpox vaccine is highly effective in immunizing against smallpox when given as late as 2 to 3 days after exposure. Smallpox vaccine is in short supply and is currently available in different dosages. A randomized double-blind study was performed to compare the efficacy of undiluted vaccine vs. different doses of diluted vaccine [22].

One definition of clinical success was the formation of a vesicle at the inoculation site 7 to 9 days after vaccination. The results from the trial are shown in Table 10.42.

**Table 10.42 Clinical success rate of various types of smallpox vaccine**

Type of vaccine	Success/total
Undiluted vaccine dose = $10^{7.8}$ pfu/mL	19/20
1:10 diluted vaccine dose = $10^{6.5}$ pfu/mL	14/20
1:100 diluted vaccine dose = $10^{5.0}$ pfu/mL	3/20

**10.89** Perform a test to compare the success rate of undiluted vaccine vs. 1:100 diluted vaccine. Report a two-tailed  $p$ -value.

**10.90** Perform a test to assess whether the success rate is a function of the  $\log_{10}$ (dilution ratio). Report a two-tailed

*p*-value. [Note: For the 1:10 group,  $\log_{10}(\text{dilution ratio}) = \log_{10}(1/10) = -1, \dots, \text{etc.}$ ]

A more quantitative measure of efficacy was the cytotoxic T-cell response. The results are shown in Table 10.43.

**Table 10.43 Cytotoxic T-cell response (lytic units per  $10^6$  cells) to various doses of smallpox vaccine<sup>a</sup>**

Type of vaccine	T-cell response		
	0	1–99	100 +
Undiluted	1	8	10
1:10 dilution	4	9	4
1:100 dilution	15	3	1

<sup>a</sup>Some assays were missing, so not all sample sizes add up to 20.

**10.91** What test can be used to compare the cytotoxic T-cell response of the undiluted group vs. the 1:100 diluted group?

**10.92** Perform the test mentioned in Problem 10.91, and report a two-tailed *p*-value. What is your conclusion regarding the effectiveness of the undiluted in contrast to the 1:100 diluted vaccine?

### Obstetrics, Infectious Disease

A study was performed comparing two surveillance methods to detect patients with infections after caesarean section. Five hospitals were involved. Let's call the two methods the *standard Centers for Disease Control and Prevention (CDC) method* and the *enhanced method*.

In hospital 1, 197 patients were assessed; 3 of the patients were identified as having an infection by the CDC method and 6 were identified as having an infection by the enhanced method. Two of the patients were identified as having an infection by both methods.

**10.93** What test can be used to identify differences between the methods for identifying patients with infection in hospital 1?

**10.94** Implement the method in Problem 10.93, and report a two-tailed *p*-value.

In hospital 2, 870 patients were assessed. Eleven of the patients were identified as having an infection by the CDC method, and 32 were identified as having an infection by the enhanced method. Five of the patients were identified as having an infection by both methods.

**10.95** What test can be used to identify difference between the methods for identifying patients with infection in hospital 2?

**10.96** Implement the test in Problem 10.95, and report a two-tailed *p*-value.

**10.97** Is there a similarity between the types of patients identified as having an infection by the two methods based on the data in hospital 2? Provide a two-tailed *p*-value to support your conclusion.

### Health Promotion

It is fairly well known that perception of weight by adolescents does not always agree with actual weight. What is less clear is whether perception of weight differs by gender. For this purpose, a study was performed among students in a local high school, where students provided their actual height and weight by self-report. The following data were obtained from 286 students (143 boys and 143 girls). (The data for this problem were provided by Laura Rosner.) The students were classified as underweight if their body-mass index (BMI) ( $\text{kg}/\text{m}^2$ ) was less than  $18.0 \text{ kg}/\text{m}^2$ , as normal if their BMI was  $\geq 18.0$  and  $< 25.0$ , and overweight if their BMI was  $\geq 25.0$ .

Based on these criteria, 17 of the girls were underweight, 113 were of normal weight, and 13 were overweight. For the boys, 7 were underweight, 115 were of normal weight, and 21 were overweight.

**10.98** What test procedure can be used to assess whether the weight status of boys significantly differs from girls?

**10.99** Perform the test procedure in Problem 10.98, and provide a two-tailed *p*-value.

One issue in comparing BMI between groups is the underlying distribution of BMI. The mean BMI of the 143 boys was 21.8 with  $sd = 3.4$ . The distribution of BMI was as shown in Table 10.44.

**Table 10.44 Distribution of BMI among 143 high school boys**

BMI	Frequency
$\leq 19.9$	41
20–22.9	65
23–25.9	20
26–28.9	9
29+	8
Total	143

**10.100** What test procedure can be used to assess whether the distribution of BMI among the boys is normal?

**10.101** Perform the test procedure in Problem 10.100. Provide a *p*-value. (Assume that BMI is measured exactly with no measurement error; thus the range 20–22.9 includes all values  $\geq 20$  and  $< 23$ , etc.)

## Cancer, Genetics

A case-control study was performed of renal-cell carcinoma (RCC) (kidney cancer) [23]. The purpose of the study was to look at environmental risk factors for RCC and to assess whether genetic factors modify the role of environmental risk factors. There were a total of 113 cases and 256 controls studied.

The participants were subdivided into “slow acetylators” and “rapid acetylators” according to the NAT2 genotype. The hypothesis was that slow acetylators might metabolize potentially toxic substances more slowly than rapid acetylators and show different relationships with environmental risk factors such as smoking. Table 10.45 presents data for slow acetylators according to the number of cigarettes smoked per day.

**10.102** Test for the association between the number of cigarettes smoked per day and RCC among slow acetylators. (Report a two-tailed  $p$ -value.)

**Table 10.45 Relationship between number of cigarettes smoked per day and RCC among slow acetylators**

No. of cigarettes smoked per day <sup>a</sup>	Cases	Controls
0	19	69
1–20	19	27
>20	31	27

<sup>a</sup>1 pack contains 20 cigarettes.

Similar data were presented for rapid acetylators, as shown in Table 10.46.

**Table 10.46 Relationship between number of cigarettes smoked per day and RCC among rapid acetylators**

No. of cigarettes smoked per day <sup>a</sup>	Cases	Controls
0	18	70
1–20	11	37
>20	15	26

<sup>a</sup>1 pack contains 20 cigarettes.

**10.103** Answer the question in Problem 10.102 for rapid acetylators. Report a two-tailed  $p$ -value.

**10.104** Do you think that genetic factors influence the relationship between smoking and RCC based on the preceding data? Why or why not?

## Cancer

The effect of using postmenopausal hormones (PMH) on health outcomes is controversial. Most previous data collected have been from observational studies, and users of PMH may selectively differ from nonusers in ways that are difficult to quantify (e.g., more health conscious, more physician visits in which disease outcomes can be identified). A clinical trial was planned to randomize postmenopausal women to either PMH use or no PMH use and follow them for disease outcomes over a 10-year period. One outcome of special interest was breast cancer.

Suppose the incidence rate of breast cancer among postmenopausal 50-year-old women who do not use PMH is 200 per  $10^5$  women per year. Suppose also it is hypothesized that PMH increases the incidence of breast cancer by 20%.

**10.105** How many women need to be studied in each group (equal sample size per group) to have an 80% probability of detecting a significant difference if a two-sided test is used with a type I error of 5%?

**10.106** Suppose that 20,000 women are recruited in each group in the actual study. How much power would the study have under the same assumptions just given?

**10.107** One problem with the design is that about 20% of the women who are randomized to PMH will not comply (i.e., will go off PMH during the trial). In addition, 10% of the participants in the control group will go on PMH on their own during the study. How much power will the study have under these revised assumptions under the simplifying assumption that this lack of compliance occurs at the beginning of the study and 20,000 women are recruited for each treatment group?

## Mental Health

Researchers collected the following data concerning comparability of diagnoses of schizophrenia obtained from primary-care physician report as compared with proxy report (from spouses). Data were collected concerning 953 people (referred to as *index subjects*). The researchers found that schizophrenia was identified as present on 115 physician reports and 124 proxy reports. Both physician and proxy informants identified 34 people as positive and they are included among the 115 and 124 individuals described.

**10.108** If the physician report is considered the gold standard, what is the sensitivity and specificity of proxy reports of schizophrenia?

Suppose neither the physician report nor the proxy report is considered the gold standard.

**10.109** Compare the percentage of subjects identified as schizophrenic by physician report with those so identified by proxy report. Perform a hypothesis test, and report a two-tailed  $p$ -value.

**10.110** Suppose there is no difference in the percentage of subjects identified as schizophrenic by physician and by proxy informants. Does this mean the two sources of information are the same for each individual? Why or why not?

**10.111** In a reproducibility study, researchers contacted the 953 spouses a second time 1 year later and asked them again whether the index subject was schizophrenic. One hundred twelve positive reports of schizophrenia were obtained, of which 89 were positive on both first and second report. Compute an index of reproducibility for proxy report of schizophrenia based on these data, and provide an interpretation of what it means.

### Health Promotion

Refer to Data Set ESTRADL.DAT on the Companion Website. Suppose we classify women as overweight if their body-mass index (BMI) exceeds  $25 \text{ kg/m}^2$ .

**10.112** Compare the percentage of Caucasian and African-American women who are overweight. Report a two-tailed  $p$ -value.

A more exact classification of BMI is as follows:  $<25$  is normal;  $\geq 25$ ,  $<30$  is overweight;  $\geq 30$  is obese.

**10.113** Compare the distribution of BMI between Caucasian and African-American women using this finer classification. Report a two-tailed  $p$ -value.

### SIMULATION

Suppose we are planning a clinical trial and expect a 20% success rate in the active group and a 10% success rate in the placebo group. We expect to enroll 100 participants in each group and are interested in the power of the study.

**10.114** Perform a simulation study, and generate 100 participants from a binomial distribution with  $p = .2$  and 100 participants from a binomial distribution with  $p = .1$ . Test to find out whether the observed sample proportion of successes is significantly different, using a two-sided test with  $\alpha = .05$ .

**10.115** Repeat the Problem 10.114 simulation 100 times, and compute the proportion of simulations in which a significant difference is found (i.e., an estimate of the power of the study).

**10.116** What should the theoretical power be for the exercise in Problem 10.115? How do the estimated power and theoretical power compare?

### Infectious Disease

Aminoglycosides are powerful broad-spectrum antibiotics used for gram-negative infections often in seriously ill patients. For example, the drugs are often prescribed for drug-resistant tuberculosis as recommended by the World

Health Organization. However, these drugs have serious side effects, including irreversible hearing loss referred to as ototoxicity. The most commonly prescribed aminoglycoside is gentamicin.

A clinical trial was set up in China to assess whether the addition of aspirin to a standard regimen of gentamicin would have an effect on the incidence of ototoxicity [24]. One hundred ninety-five patients were enrolled in a prospective, randomized, double-blind clinical trial. Of these, 106 patients were randomized to a twice daily regimen of 80–160 mg of gentamicin plus placebo and 89 patients were randomized to receive the same regimen of gentamicin + 3 g of aspirin (ASA) daily.

**10.117** (i) What is a prospective study? What are its advantages?

(ii) What is a randomized study? What are its advantages?

(iii) What is a double-blind study? Are there advantages vs. other approaches?

After 2 weeks of treatment, 14 of the placebo patients and 3 of the ASA patients developed clinically significant hearing loss.

**10.118** Implement a test to assess whether the incidence of hearing loss is different between the groups. State clearly which test you are using, and report a two-sided  $p$ -value.

**10.119** Interpret the results of your test in Problem 10.118.

Suppose another research group wishes to replicate the findings of this study. The investigators conservatively estimate that the incidence of hearing loss will be 10% in a gentamicin + placebo group and 5% in a gentamicin + ASA group and plan to enroll an equal number of subjects in each group.

**10.120** How many subjects need to be enrolled in total in the study if (1) the investigators want to have a 90% chance of detecting a significant difference using a two-sided test with  $\alpha = .05$  and (2) it is anticipated that 5% of the enrolled subjects will not complete the study?

### Obstetrics

The standard screening test for Down's syndrome is based on a combination of maternal age and the level of serum alpha-fetoprotein. Using this test 80% of Down's syndrome cases can be identified, while 5% of normals are detected as positive.

**10.121** What is the sensitivity and specificity of the test?

Suppose that 1 out of 500 infants are born with Down's syndrome.

**10.122** What percentage of infants who test positive using the test will actually have Down's syndrome?

A new test is proposed that may be better or worse than the standard test. To assess their relative efficacy, both tests are used on the same subjects and compared with the true

diagnosis. Let + = correct assessment, – = incorrect assessment. The results are given in Table 10.47.

**Table 10.47 Comparison of two screening tests for Down's syndrome**

Standard test	New test	N
+	+	82
+	–	5
–	+	10
–	–	3
		100

Thus for 82 infants, both tests make the correct classification, for 3 infants both tests make the wrong classification, for 5 infants the standard test makes the correct diagnosis while the new test does not, and for 10 infants the new test makes the correct diagnosis while the standard test does not.

**10.123** Perform a hypothesis test to compare the accuracy of the two tests, and report a two-sided *p*-value.

### Ophthalmology

Dry eye is the most prevalent form of ocular discomfort and irritation, with approximately 20 million people in the United States having mild to moderate dry eye. A small clinical trial was performed to compare the effectiveness of an active drug vs. placebo for relieving symptoms of dry eye. Specifically, patients were randomized to either active drug or placebo for 2 weeks. They then came for a clinic visit where they were exposed to a chamber with a controlled adverse environment (CAE) for 90 minutes (with low humidity intended to exacerbate symptoms of dry eye). The patients were then asked to report their degree of discomfort while in the CAE using the following scale: (0 = none, 1 = intermittent awareness, 2 = constant awareness, 3 = intermittent discomfort, 4 = constant discomfort). The results by treatment group are shown in Table 10.48.

**Table 10.48 Comparison of active drug vs. placebo for the prevention of dry eye symptoms**

	Ocular discomfort			
	2	3	4	Total
Active drug	6	17	37	60
Placebo	2	13	44	59

**10.124** What is the difference between a nominal and ordinal categorical variable? What type of variable is ocular discomfort?

**10.125** Treating ocular discomfort as a *nominal* scale, assess whether significant differences in ocular discomfort exist between active drug and placebo patients. Report a *p*-value (two-tailed). (*Hint:* Assume that large sample methods are appropriate for the data.) Interpret the results.

**10.126** Treating ocular discomfort as an *ordinal* scale, assess whether significant differences in ocular discomfort exist between active and placebo patients. Report a *p*-value (two-tailed). (*Hint:* Assume that large sample methods are appropriate for these data.) Interpret the results.

### Environmental Health, Mental Health

A study was performed in China relating selenium level as assessed by nail samples and cognitive function [25]. A comparison of selenium levels by study site is shown in Table 10.49.

**Table 10.49 Comparison of selenium levels among four locations in China**

Selenium group	Site 1 (Qionglui) (%)	Site 2 (Gaomi) (%)	Site 3 (Jiange) (%)	Site 4 (Zichuan) (%)
4 (5.5*)	15.2	5.8	0.2	59.8
3 (4.9*)	30.4	23.6	0.4	26.8
2 (4.0*)	24.2	40.8	1.2	11.8
1 (3.0*)	30.2	29.8	98.2	1.6
<i>N</i>	500	500	500	500

\*Median selenium intake in mg/100 g.

**10.127** What test can be used to compare the median selenium intake between site 1 and site 2?

**10.128** State the hypotheses to be tested under the null and alternative hypotheses.

**10.129** Perform the test in problem 10.127 and report a *p*-value (two-tailed).

**10.130** Estimate the relative risk for being in selenium group 4 vs. selenium group 1 for site 1 compared with site 2.

### Cardiology

Antithrombotic drugs are used after coronary stenting to prevent stent thrombosis. A study was performed to compare the efficacy and safety of three antithrombotic drug regimens after coronary stenting: aspirin alone, aspirin + warfarin, and aspirin + ticlopidine [26]. Patients were randomly assigned to one of the three regimens. The primary endpoint was any of the following outcomes: (a) death, (b) revascularization of the target lesion, (c) angiographic evidence of thrombosis or MI within 30 days.

**10.131** What is the principal benefit(s) of assigning patients randomly to the treatment regimens?

The results for the primary endpoint are given in Table 10.50.

**Table 10.50 Comparison of vascular outcomes by treatment group in a clinical trial of three antithrombotic agents**

Group number	Group name	Number of events	Total
1	Aspirin alone	20	557
2	Aspirin + warfarin	15	550
3	Aspirin + ticlopidine	3	546

**10.132** What test can be used to compare event rates among the three groups?

**10.133** State the hypotheses to be tested under the null and alternative hypotheses.

**10.134** Implement the test in Problem 10.132, and report a  $p$ -value (two-tailed).

### Ophthalmology

A case-control study was performed among 145 subjects with macular degeneration and 34 controls, all of whom were 70- to 79-year-old women. A genetic risk score was developed to help differentiate the cases from the controls. The risk score was categorized into six groups (1, 2, 3, 4, 5, 6). The data in Table 10.51 were obtained relating the risk score to case/control status.

**Table 10.51 Association between a genetic risk score and macular degeneration**

Risk score	Cases	Controls
1	3	11
2	7	3
3	6	6
4	10	8
5	11	2
6	108	4
Total	145	34

**10.135** What test can be performed to study the association between case/control status and risk score? Specifically, we are interested in testing whether cases tend to have consistently higher risk scores or consistently lower risk scores than controls.

**10.136** Perform the test in Problem 10.135, and report a  $p$ -value (two-tailed).

Another use for the risk score is to compute a prevalence estimate of AMD for women with different risk scores. In the general population, the prevalence of AMD among 70- to 79-year-old women is .025.

**10.137** What is the estimated prevalence of AMD among women with a risk score in group 1? In group 6? Hint: Use Bayes' Theorem.

**10.138** What is the relative risk for AMD among women in group 6 compared with women in group 1? What does the relative risk mean in words?

### Pulmonary Disease

Asthma is an important health problem for inner-city children, frequently resulting in hospital admission if symptoms become exacerbated. It is well known that compliance of children with asthma medication is often poor. Also, many household allergens (e.g., roaches) worsen asthma symptoms. A study is proposed in which children will be randomized to either an active intervention where a community health worker comes to the home and educates the children and parents as to approaches to reduce the risk of asthma symptoms or a control intervention where households will receive the same information in print but no home visits will be performed. It is expected that 30% of the children in the active group vs. 10% of the children in the control group will have an improvement in asthma symptoms.

**10.139** How many subjects should be recruited in each group (same number per group) to have a 90% chance of detecting a significant difference using a two-sided test with  $\alpha = .05$ ?

**10.140** Suppose that 50 households are randomized per group. How much power would the study have under the above assumptions?

The results of the study were as follows: 14 of the 50 active intervention children had an improvement in symptoms compared with 6 of the 50 control intervention children.

**10.141** What test can be used to compare the results in the active and control groups?

**10.142** Perform the test in Problem 10.141, and report a  $p$ -value (two-tailed).

### Ophthalmology

The Sorbinil Retinopathy Trial was conducted among 497 type I (insulin-dependent) diabetic patients who had little or no evidence of retinopathy at baseline [27]. Retinopathies are abnormalities of the retina that sometimes occur among diabetic patients and can result in advanced stages in substantial losses of vision. Patients were randomized to either Sorbinil, an aldose-reductase inhibitor, or placebo and were seen at 1 year and then every 9 months up to 48 months after randomization. In addition, all subjects had a final visit at the end of the trial (max = 56 months). Sixteen

**Table 10.52 Outcome data for the Sorbinil Retinopathy Trial ( $n = 481$ ) change in retinopathy level**

Group	Better					Worse				Total
	2+ Levels	1 Level	No change	1 Level	2 Levels	3 Levels	4 Levels	5+ Levels		
Placebo	5	17	84	59	37	18	9	14	243	
Sorbinil	4	21	97	50	22	16	14	14	238	

of the patients provided no follow-up. The primary endpoint of the trial was based on change in retinopathy severity level from baseline to maximum follow-up (i.e., severity level at maximum follow-up – severity level at baseline). An ordinal grading scale was used to evaluate change: 2+ levels better, 1 level better, no change, 1 level worse, ..., 5+ levels worse. The outcome data by treatment group are given in Table 10.52.

The primary outcome for the study was worsening by 2 or more levels.

**10.143** What test can be used to compare the two treatment groups on the primary endpoint?

**10.144** Implement the test in Problem 10.143, and provide a  $p$ -value (two-tailed).

A more efficient method of analysis would leave the change in retinopathy level in its raw form without grouping the data but would take into account the ordinal nature of the change scores.

**10.145** What test can be used to compare the two groups if this more efficient method is used?

**10.146** Implement the test in Problem 10.145, and provide a  $p$ -value (two-tailed).

## REFERENCES

- [1] MacMahon, B., Cole, P., Lin, T. M., Lowe, C. R., Mirra, A. P., Ravnhar, B., Salber, E. J., Valaoras, V. G., & Yuasa, S. (1970). Age at first birth and breast cancer risk. *Bulletin of the World Health Organization*, 43, 209–221.
- [2] Doll, R., Muir, C., & Waterhouse, J. (Eds.). (1970). *Cancer in five continents* (Vol. 2). Berlin: Springer.
- [3] Dupont, W. D. (1988). Power calculations for matched case-control studies. *Biometrics*, 44, 1157–1168.
- [4] Cochran, W. G. (1954). Some methods for strengthening the common  $\chi^2$  test. *Biometrics*, 10, 417–451.
- [5] Maxwell, A. E. (1961). *Analyzing qualitative data*. London: Methuen.
- [6] Hypertension Detection and Follow-up Program Cooperative Group. (1977). Blood pressure studies in 14 communities—A two-stage screen for hypertension. *JAMA*, 237(22), 2385–2391.
- [7] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- [8] Fleiss, J. (1981). *Statistical methods for rates and proportions*. New York: Wiley.
- [9] Weiss, N. S., & Sayetz, T. A. (1980). Incidence of endometrial cancer in relation to the use of oral contraceptives. *New England Journal of Medicine*, 302(10), 551–554.
- [10] Horner, M. J., Ries, L. A. G., Krapcho, M., Neyman, N., Aminou, R., Howlader, N., Altekruse, S. F., Feuer, E. J., Huang, L., Mariotto, A., Miller, B. A., Lewis, D. R., Eisner, M. P., Stinchcomb, D. G., Edwards, B. K. (Eds.). *SEER cancer statistics review, 1975–2006*. Bethesda, MD: National Cancer Institute; [http://seer.cancer.gov/csr/1975\\_2006/](http://seer.cancer.gov/csr/1975_2006/), based on November 2008 SEER data submission, posted to the SEER website, 2009.
- [11] Mecklenburg, R. S., Benson, E. A., Benson, J. W., Fredlung, P. N., Quinn, T., Metz, R. J., Nielsen, R. L., & Sananar, C. A. (1984). Acute complications associated with insulin pump therapy: Report of experience with 161 patients. *JAMA*, 252(23), 3265–3269.
- [12] Dubach, U. C., Rosner, B., & Stürmer, T. (1991). An epidemiological study of abuse of analgesic drugs: Effects of phenacetin and salicylate on mortality and cardiovascular morbidity (1968–1987). *New England Journal of Medicine*, 324, 155–160.
- [13] Armenian, H., Saadeh, F. M., & Armenian, S. L. (1987). Widowhood and mortality in an Armenian church parish in Lebanon. *American Journal of Epidemiology*, 125(1), 127–132.
- [14] Jorgensen, M., Keiding, N., & Skakkebaek, N. E. (1991). Estimation of spermatache from longitudinal spermaturia data. *Biometrics*, 47, 177–193.
- [15] Brennan, T. A., Leake, L. L., Laird, N. M., Hebert, L., Localio, A. S., Lawthers, A. G., Newhouse, J. P., Weiler, P. G., & Hiatt, H. H. (1991). Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *New England Journal of Medicine*, 324(6), 370–376.
- [16] Schoenbaum, E. E., Hartel, D., Selwyn, P. A., Klein, R. S., Davenny, K., Rogers, M., Feiner, C., & Friedland, G. (1989). Risk factors for human immunodeficiency virus infection in intravenous drug users. *New England Journal of Medicine*, 321(13), 874–879.
- [17] García-Martín, M., Lardelli-Claret, P., Bueno-Cavanillas, A., Luna-del-Castillo, J. D., Espigares-García, M.,

- & Galvez-Vargas, R. (1997). Proportion of hospital deaths associated with adverse events. *Journal of Clinical Epidemiology*, 50(12), 1319–1326.
- [18] Sibulesky, L., Hayes, K. C., Pronczuk, A., Weigel-DiFranco, C., Rosner, B., & Berson, E. L. (1999). Safety of <7500 RE (<25000 IU) vitamin A daily in adults with retinitis pigmentosa. *American Journal of Clinical Nutrition*, 69(4), 656–663.
- [19] Hennekens, C. H., Buring, J. E., Manson, J. E., Stampfer, M., Rosner, B., Cook, N. R., Belanger, C., LaMotte, F., Gaziano, J. M., Ridker, P. M., Willett, W., & Peto, R. (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *New England Journal of Medicine*, 334(18), 1145–1149.
- [20] Croudace, T. J., Jarvelin, M.-R., Wadsworth, M. E. J., & Jones, P. E. (2003). Developmental typology of trajectories to nighttime bladder control: Epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology*, 157, 834–842.
- [21] Teele, D. W., Klein, J. O., Word, B. M., Rosner, B. A., Starobin, S., Earle II, R., Ertel, C. S., Fisch, G., Michaels, R., Heppen, R., Strause, N. P., & the Greater Boston Otitis Media Study Group. (2000). Antimicrobial prophylaxis for infants at risk for recurrent otitis media. *Vaccine*, 19 (Suppl. 1), S140–S143.
- [22] Frey, S. E., Newman, F. K., Cruz, J., Shelton, W. B., Tenant, J. M., Polach, T., Rothman, A. L., Kennedy, J. S., Wolff, M., Belshe, R. B., & Ennis, F. A. (2002). Dose-related effects of smallpox vaccine. *New England Journal of Medicine*, 346, 1275–1280.
- [23] Semenza, J. C., Ziogas, A., Largent, J., Peel, D., & Hoda, A.-C. (2001). Gene-environment interactions in renal cell carcinoma. *American Journal of Epidemiology*, 153, 851–859.
- [24] Sha, S.-H., Qiu, J.-H., & Schacht, J. (2006). Aspirin to prevent gentamicin-induced hearing loss. *New England Journal of Medicine*, 354(17), 1856–1857.
- [25] Gao, S., Jin, Y., Hall, K. S., Liang, C., Unverzagt, F. W., Ji, R., Murrell, J. R., Cao, J., Shen, J., Ma, F., Matesan, J., Ying, B., Cheng, Y., Bian, J., Li, P., & Hendrie, H. C. (2007). Selenium level and cognitive function in rural elderly Chinese. *American Journal of Epidemiology* 165(8), 955–965.
- [26] Leon, M. B., Baim, D. S., Popma, J. J., Gordon, P. C., Cutlip, D. E., Ho, K. K. L., Giambartolomei, A., Diver, D. J., Lasorda, D. M., Williams, D. O., Pocock, S. J., & Kuntz, R. E., for the Stent Anticoagulation Restenosis Study Investigators. (1998). A clinical trial comparing three antithrombotic-drug regimens after coronary-artery stenting. *New England Journal of Medicine*, 339(23), 1665–1671.
- [27] Sorbinil Retinopathy Trial Research Group. (1990). A randomized trial of Sorbinil, an aldose reductase inhibitor, in diabetic retinopathy. *Archives of Ophthalmology*, 108, 1234–1244.

## 11.1 Introduction

In Chapter 8, statistical methods for comparing the means of a normally distributed outcome variable between two populations were presented based on  $t$  tests. Suppose we call the outcome variable  $y$  and the group classification (or class) variable  $x$ . For  $t$  test applications,  $x$  takes on two values. Another way of looking at the methods in Chapter 8 is as techniques for assessing the possible association between a normally distributed variable  $y$  and a categorical variable  $x$ . We will see that these techniques are special cases of **linear-regression methods**. In linear regression, we will study how to relate a normally distributed outcome variable  $y$  to one or more predictor variables  $x_1, \dots, x_k$  where the  $x$ 's may be either continuous or categorical variables.

**Example 11.1**

**Obstetrics** Obstetricians sometimes order tests to measure estriol levels from 24-hour urine specimens taken from pregnant women who are near term because level of estriol has been found to be related to infant birthweight. The test can provide indirect evidence of an abnormally small fetus. The relationship between estriol level and birthweight can be quantified by fitting a *regression line* that relates the two variables.

In Chapter 10, we also studied the Kappa statistic, which is a measure of association between two categorical variables. This index is useful when we are interested in how strong the association is between two categorical variables rather than in predicting one variable as a function of the other variable. To quantify the association between two continuous variables, we can use the correlation coefficient introduced in Section 5.6. In this chapter we consider hypothesis-testing methods for correlation coefficients and extend the concept of a correlation coefficient to describe association among several continuous variables.

**Example 11.2**

**Hypertension** Much discussion has taken place in the literature concerning the familial aggregation of blood pressure. In general, children whose parents have high blood pressure tend to have higher blood pressure than their peers. One way of expressing this relationship is by computing a *correlation coefficient* relating the blood pressure of parents and children over a large collection of families.

In this chapter, we discuss methods of regression and correlation analysis in which *two* different variables in the same sample are related. The extension of these

methods to the case of multiple-regression analysis, where the relationship between more than two variables at a time is considered, is also discussed.

## 11.2 General Concepts

### Example 11.3

**Obstetrics** Greene and Touchstone conducted a study to relate birthweight and estriol level in pregnant women [1]. Figure 11.1 is a plot of the data from the study, and the actual data points are listed in Table 11.1. As can be seen from the figure, there appears to be a relationship between estriol level and birthweight, although this relationship is not consistent and considerable scatter exists throughout the plot. How can this relationship be quantified?

If  $x$  = estriol level and  $y$  = birthweight, then we can postulate a linear relationship between  $y$  and  $x$  that is of the following form:

### Equation 11.1

$$E(y|x) = \alpha + \beta x$$

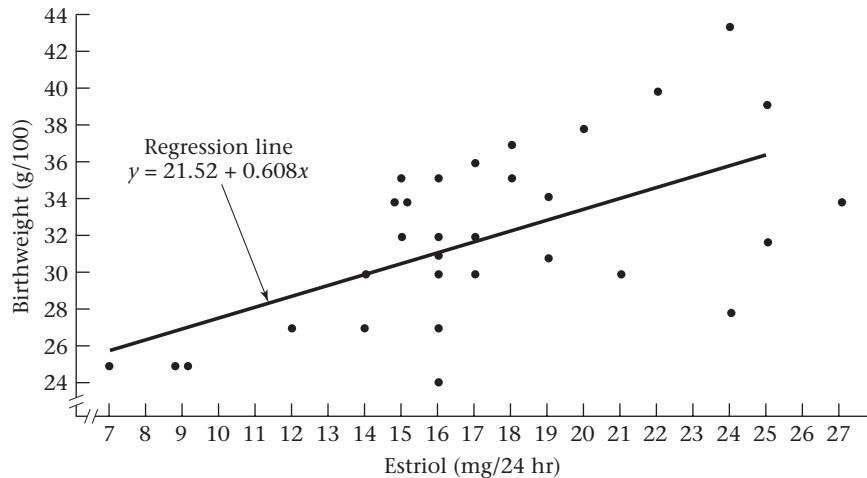
That is, for a given estriol-level  $x$ , the average birthweight  $E(y|x) = \alpha + \beta x$ .

### Definition 11.1

The line  $y = \alpha + \beta x$  is the **regression line**, where  $\alpha$  is the intercept and  $\beta$  is the slope of the line.

The relationship  $y = \alpha + \beta x$  is not expected to hold exactly for every woman. For example, not all women with a given estriol level have babies with identical birthweights. Thus an error term  $e$ , which represents the variance of birthweight among all babies of women with a given estriol level  $x$ , is introduced into the model. Let's

**Figure 11.1** Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term



Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology*, 85(1), 1–9, 1963.

**Table 11.1** Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

<i>i</i>	Estriol (mg/24 hr) <i>x<sub>i</sub></i>	Birthweight (g/100) <i>y<sub>i</sub></i>	<i>i</i>	Estriol (mg/24 hr) <i>x<sub>i</sub></i>	Birthweight (g/100) <i>y<sub>i</sub></i>
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology*, 85(1), 1–9, 1963.

assume  $e$  follows a normal distribution, with mean 0 and variance  $\sigma^2$ . The full linear-regression model then takes the following form:

**Equation 11.2**

$$y = \alpha + \beta x + e$$

where  $e$  is normally distributed with mean 0 and variance  $\sigma^2$ .

**Definition 11.2**

For any linear-regression equation of the form  $y = \alpha + \beta x + e$ ,  $y$  is called the **dependent variable** and  $x$  is called the **independent variable** because we are trying to predict  $y$  as a function of  $x$ .

**Example 11.4**

**Obstetrics** Birthweight is the dependent variable and estriol is the independent variable for the problem posed in Example 11.3 because estriol levels are being used to try to predict birthweight.

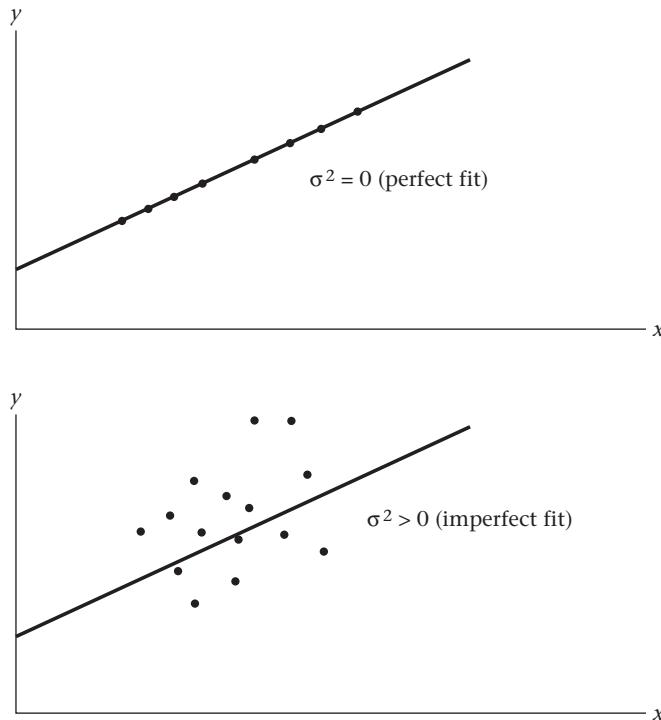
One interpretation of the regression line is that for a woman with estriol level  $x$ , the corresponding birthweight will be normally distributed with mean  $\alpha + \beta x$  and variance  $\sigma^2$ . If  $\sigma^2$  were 0, then every point would fall exactly on the regression line, whereas the larger  $\sigma^2$  is, the more scatter occurs about the regression line. This relationship is illustrated in Figure 11.2.

How can  $\beta$  be interpreted? If  $\beta$  is greater than 0, then as  $x$  increases, the expected value of  $y = \alpha + \beta x$  will increase.

**Example 11.5**

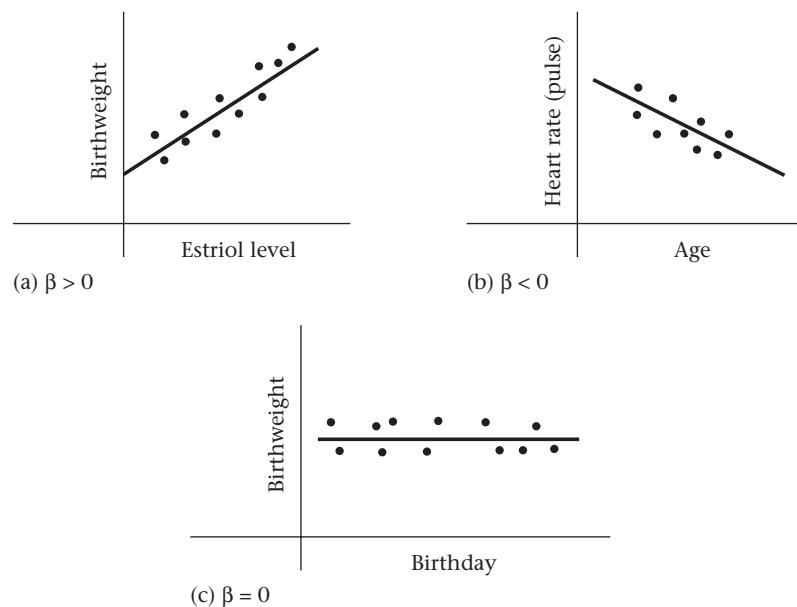
**Obstetrics** This situation appears to be the case in Figure 11.3a for birthweight ( $y$ ) and estriol ( $x$ ) because as estriol increases, the average birthweight correspondingly increases.

If  $\beta$  is less than 0, then as  $x$  increases, the expected value of  $y$  will decrease.

**Figure 11.2** The effect of  $\sigma^2$  on the goodness of fit of a regression line

**Example 11.6** **Pediatrics** This situation might occur in a plot of pulse rate ( $y$ ) vs. age ( $x$ ), as illustrated in Figure 11.3b, because infants are born with rapid pulse rates that gradually slow with age.

If  $\beta$  is equal to 0, then there is no linear relationship between  $x$  and  $y$ .

**Figure 11.3** Interpretation of the regression line for different values of  $\beta$ 

**Example 11.7** This situation might occur in a plot of birthweight vs. birthday, as shown in Figure 11.3c, because there is no relationship between birthweight and birthday.

### 11.3 Fitting Regression Lines—The Method of Least Squares

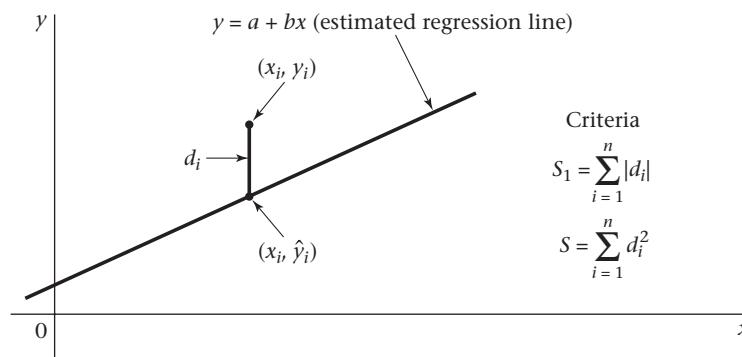
The question remains as to how to fit a regression line (or, equivalently, to obtain estimates of  $\alpha$  and  $\beta$ , denoted by  $a$  and  $b$ , respectively) when data appear in the form of Figure 11.1. We could eyeball the data and draw a line that is not too distant from any of the points, but in practice this approach is difficult and can be quite imprecise, with either a large number of points or a lot of scatter. A better method is to set up a specific criterion that defines the closeness of a line to a set of points and to find the line closest to the sample data according to this criterion.

Consider the data in Figure 11.4 and the estimated regression line  $y = a + bx$ . The distance  $d_i$  of a typical sample point  $(x_i, y_i)$  from the line could be measured along a direction parallel to the  $y$ -axis. If we let  $(x_i, \hat{y}_i) = (x_i, a + bx_i)$  be the point on the estimated regression line at  $x_i$ , then this distance is given by  $d_i = y_i - \hat{y}_i = y_i - a - bx_i$ . A good-fitting line would make these distances as small as possible. Because the  $d_i$  cannot all be 0, the criterion  $S_1 = \text{sum of the absolute deviations of the sample points from the line} = \sum_{i=1}^n |d_i|$  can be used and the line that minimizes  $S_1$  can be found. Instead, for both theoretical reasons and ease of derivation, the following least-squares criterion is commonly used:

$S = \text{sum of the squared distances of the points from the line}$

$$= \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

**Figure 11.4** Possible criteria for judging the fit of a regression line



**Definition 11.3** The **least-squares line**, or **estimated regression line**, is the line  $y = a + bx$  that minimizes the sum of squared distances of the sample points from the line given by

$$S = \sum_{i=1}^n d_i^2$$

This method of estimating the parameters of a regression line is known as the **method of least squares**.

The following notation is needed to define the slope and intercept of a regression line.

---

**Definition 11.4** The **raw sum of squares for  $x$**  is defined by

$$\sum_{i=1}^n x_i^2$$

The **corrected sum of squares for  $x$**  is denoted by  $L_{xx}$  and defined by

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

It represents the sum of squares of the deviations of the  $x_i$  from the mean. Similarly, the **raw sum of squares for  $y$**  is defined by

$$\sum_{i=1}^n y_i^2$$

The **corrected sum of squares for  $y$**  is denoted by  $L_{yy}$  and defined by

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2 / n$$


---

Notice that  $L_{xx}$  and  $L_{yy}$  are simply the numerators of the expressions for the sample variances of  $x$  (i.e.,  $s_x^2$ ) and  $y$  (i.e.,  $s_y^2$ ), respectively, because

$$s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1) \text{ and } s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)$$

---

**Definition 11.5** The **raw sum of cross products** is defined by

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

The **corrected sum of cross products** is defined by

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

which is denoted by  $L_{xy}$ .

It can be shown that a short form for the corrected sum of cross products is given by

$$\sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right) / n$$


---

What does the corrected sum of cross products mean? Suppose  $\beta > 0$ . From Figure 11.3a, we see that if  $\beta > 0$ , then as  $x$  increases,  $y$  will tend to increase as well. Another way of expressing this relationship is that if  $(x_i - \bar{x})$  is greater than 0 (which will be true for large values of  $x_i$ ), then  $y_i$  will tend to be large or  $y_i - \bar{y}$  will be greater than 0 and  $(x_i - \bar{x})(y_i - \bar{y})$  will be the product of two positive numbers and thus will be positive. Similarly, if  $x_i - \bar{x}$  is  $< 0$ , then  $y_i - \bar{y}$  will also tend to be  $< 0$  and  $(x_i - \bar{x})(y_i - \bar{y})$  will be the product of two negative numbers and thus will be positive. Thus if  $\beta > 0$ , the sum of cross products will tend to be positive. Suppose that  $\beta < 0$ . From Figure 11.3b, when  $x$  is small,  $y$  will tend to be large and when  $x$  is large,  $y$  will

tend to be small. In both cases,  $(x_i - \bar{x})(y_i - \bar{y})$  will often be the product of 1 positive and 1 negative number and will be negative. Thus if  $\beta < 0$ , the sum of cross products will tend to be negative. Finally, if  $\beta = 0$ , then  $x$  and  $y$  bear no linear relation to each other and the sum of cross products will be close to 0.

It can be shown that the estimate  $b$  of the underlying slope  $\beta$ , which minimizes  $S$ , is given by  $b = L_{xy}/L_{xx}$ . Thus we refer to  $b$  as the least-squares slope. Because  $L_{xx}$  is always positive (except in the degenerate case when all  $x$ 's in the sample are the same), the sign of  $b$  is the same as the sign of the sum of cross products  $L_{xy}$ . This makes good intuitive sense based on the preceding discussion. Furthermore, for a given estimate of the slope  $b$ , it can be shown that the value of the intercept for the line that satisfies the least-squares criterion (i.e., that minimizes  $S$ ) is given by  $a = \bar{y} - b\bar{x}$ . We summarize these results as follows.

### Equation 11.3

#### Estimation of the Least-Squares Line

The coefficients of the least-squares line  $y = a + bx$  are given by

$$b = L_{xy}/L_{xx} \quad \text{and} \quad a = \bar{y} - b\bar{x} = \left( \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i \right) / n$$

Sometimes, the line  $y = a + bx$  is called the *estimated regression line* or, more briefly, the *regression line*.

### Example 11.8

#### Solution

**Obstetrics** Derive the estimated regression line for the data in Table 11.1.

First

$$\sum_{i=1}^{31} x_i \quad \sum_{i=1}^{31} x_i^2 \quad \sum_{i=1}^{31} y_i \quad \sum_{i=1}^{31} x_i y_i$$

must be obtained so as to compute the corrected sum of squares ( $L_{xx}$ ) and cross products ( $L_{xy}$ ). These quantities are given as follows:

$$\sum_{i=1}^{31} x_i = 534 \quad \sum_{i=1}^{31} x_i^2 = 9876 \quad \sum_{i=1}^{31} y_i = 992 \quad \sum_{i=1}^{31} x_i y_i = 17,500$$

Then compute  $L_{xy}$  and  $L_{xx}$ :

$$L_{xy} = \sum_{i=1}^{31} x_i y_i - \left( \sum_{i=1}^{31} x_i \right) \left( \sum_{i=1}^{31} y_i \right) / 31 = 17,500 - (534)(992) / 31 = 412$$

$$L_{xx} = \sum_{i=1}^{31} x_i^2 - \left( \sum_{i=1}^{31} x_i \right)^2 / 31 = 9876 - 534^2 / 31 = 677.42$$

Finally, compute the slope of the regression line:

$$b = L_{xy}/L_{xx} = 412/677.42 = 0.608$$

The intercept of the regression line can also be computed. Note from Equation 11.3 that

$$a = \left( \sum_{i=1}^{31} y_i - 0.608 \sum_{i=1}^{31} x_i \right) / 31 = [992 - 0.608(534)] / 31 = 21.52$$

Thus the regression line is given by  $y = 21.52 + 0.608x$ . This regression line is shown in Figure 11.1.

How can the regression line be used? One use is to *predict* values of  $y$  for given values of  $x$ .

**Definition 11.6**

The predicted, or average, value of  $y$  for a given value of  $x$ , as estimated from the fitted regression line, is denoted by  $\hat{y} = a + bx$ . Thus the point  $(x, a + bx)$  is always on the regression line.

**Example 11.9**

**Obstetrics** What is the estimated average birthweight if a pregnant woman has an estriol level of 15 mg/24 hr?

**Solution**

If the estriol level were 15 mg/24 hr, then the best prediction of the average value of  $y$  would be

$$\hat{y} = 21.52 + 0.608(15) = 30.65$$

Because  $y$  is in the units of birthweight (g)/100, the estimated average birthweight =  $30.65 \times 100 = 3065$  g.

One possible use of estriol levels is to identify women who are carrying a low-birthweight fetus. If such women can be identified prior to delivery, then drugs might be used to prolong the pregnancy until the fetus grows larger because low-birthweight infants are at greater risk than normal infants for mortality in the first year of life and for poor growth and development in childhood.

**Example 11.10**

**Obstetrics** Low birthweight is defined here as  $\leq 2500$  g. For what estriol level would the predicted birthweight be 2500 g?

**Solution**

Note that the predicted value of  $y$  (birthweight/100) is

$$\hat{y} = 21.52 + 0.608x$$

If  $\hat{y} = 2500/100 = 25$ , then  $x$  can be obtained from the equation

$$25 = 21.52 + 0.608x \quad \text{or} \quad x = (25 - 21.52)/0.608 = 3.48/0.608 = 5.72$$

Thus if a woman has an estriol level of 5.72 mg/24 hr, then the predicted birthweight is 2500 g. Furthermore, the predicted infant birthweight for all women with estriol levels of  $\leq 5$  mg/24 hr is  $< 2500$  g (assuming estriol can only be measured in increments of 1 mg/24 hr). This level could serve as a critical value for identifying high-risk women and trying to prolong their pregnancies.

How can the slope of the regression line be interpreted? The slope of the regression line tells us the amount  $y$  increases per unit increase in  $x$ .

**Example 11.11**

**Obstetrics** Interpret the slope of the regression line for the birthweight–estriol data in Example 11.1.

**Solution**

The slope of 0.608 tells us that the predicted  $y$  increases by about 0.61 units per 1 mg/24 hr. Thus the predicted birthweight increases by 61 g for every 1 mg/24 hr increase in estriol.

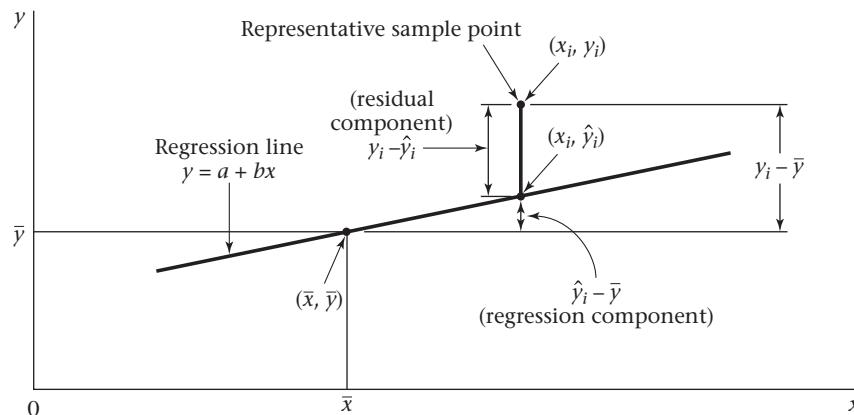
In this section, we learned how to fit regression lines using the method of least squares based on the linear-regression model in Equation 11.2. Note that the method of least squares is appropriate whenever the average residual for each given value of  $x$  is 0—that is, when  $E(e|X = x) = 0$  in Equation 11.2. Normality of the residuals is not strictly required. However, the normality assumption in Equation 11.2 is necessary to perform hypothesis tests concerning regression parameters, as discussed in the next section.

On the flowchart at the end of this chapter (Figure 11.32, p. 503), we answer yes to (1) interested in relationships between two variables? (2) both variables continuous? and (3) interested in predicting one variable from another? This leads us to the box labeled “Simple linear regression.”

## 11.4 Inferences About Parameters from Regression Lines

In Section 11.3, the fitting of regression lines using the method of least squares was discussed. Because this method can be used with any set of points, criteria for distinguishing regression lines that fit the data well from those that do not must be established. Consider the regression line in Figure 11.5.

**Figure 11.5** Goodness of fit of a regression line



A hypothetical regression line and a representative sample point have been drawn. First, notice that the point  $(\bar{x}, \bar{y})$  falls on the regression line. This feature is common to all estimated regression lines because a regression line can be represented as

$$y = a + bx = \bar{y} - b\bar{x} + bx = \bar{y} + b(x - \bar{x})$$

or, equivalently,

**Equation 11.4**

$$y - \bar{y} = b(x - \bar{x})$$

If  $\bar{x}$  is substituted for  $x$  and  $\bar{y}$  for  $y$  in Equation 11.4, then 0 is obtained on both sides of the equation, which shows that *the point  $(\bar{x}, \bar{y})$  must always fall on the estimated regression line*. If a typical sample point  $(x_i, y_i)$  is selected and a line is drawn through this point parallel to the  $y$ -axis, then the representation in Figure 11.5 is obtained.

---

**Definition 11.7** For any sample point  $(x_i, y_i)$ , the **residual**, or **residual component**, of that point about the regression line is defined by  $y_i - \hat{y}_i$ .

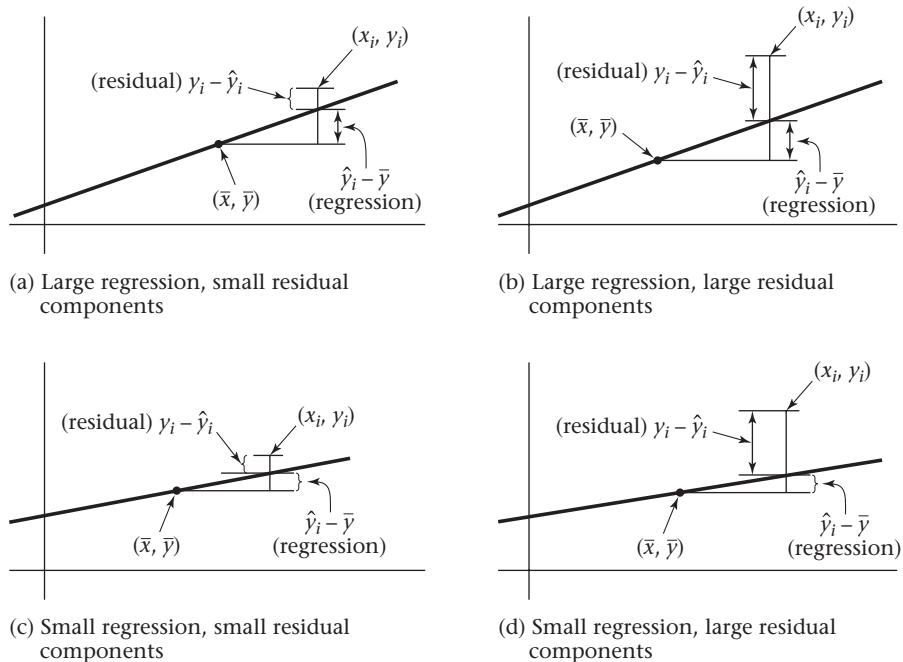
---



---

**Definition 11.8** For any sample point  $(x_i, y_i)$ , the **regression component** of that point about the regression line is defined by  $(\hat{y}_i - \bar{y})$ .

---

**Figure 11.6** Regression lines with varying goodness-of-fit relationships

In Figure 11.5 the deviation  $y_i - \bar{y}$  can be separated into residual ( $y_i - \hat{y}_i$ ) and regression ( $\hat{y}_i - \bar{y}$ ) components. Note that if the point  $(x_i, y_i)$  fell exactly on the regression line, then  $y_i = \hat{y}_i$  and the residual component  $y_i - \hat{y}_i$  would be 0 and  $y_i - \bar{y} = \hat{y}_i - \bar{y}$ . Generally speaking, a good-fitting regression line will have regression components large in absolute value relative to the residual components, whereas the opposite is true for poor-fitting regression lines. Some typical situations are shown in Figure 11.6.

The best-fitting regression line is depicted in Figure 11.6a, with large regression components and small residual components. The worst-fitting regression line is depicted in Figure 11.6d, which has small regression components and large residual components. Intermediate situations for goodness of fit are shown in Figures 11.6b and 11.6c.

How can the plots in Figure 11.6 be quantified? One strategy is to square the deviations about the mean  $y_i - \bar{y}$ , sum them up over all points, and decompose this sum of squares into regression and residual components.

**Definition 11.9**

The **total sum of squares**, or **Total SS**, is the sum of squares of the deviations of the individual sample points from the sample mean:

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

**Definition 11.10**

The **regression sum of squares**, or **Reg SS**, is the sum of squares of the regression components:

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

**Definition 11.11**

The **residual sum of squares**, or **Res SS**, is the sum of squares of the residual components:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

It can be shown that the following relationship is true.

**Equation 11.5****Decomposition of the Total Sum of Squares into Regression and Residual Components**

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

or    Total SS = Reg SS + Res SS

**F Test for Simple Linear Regression**

The criterion for goodness of fit used in this book is the ratio of the regression sum of squares to the residual sum of squares. A large ratio indicates a good fit, whereas a small ratio indicates a poor fit. In hypothesis-testing terms we want to test the hypothesis  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ , where  $\beta$  is the underlying slope of the regression line in Equation 11.2.

The following terms are introduced for ease of notation in describing the hypothesis test.

**Definition 11.12**

The **regression mean square**, or **Reg MS**, is the Reg SS divided by the number of predictor variables ( $k$ ) in the model (not including the constant). Thus Reg MS = Reg SS/ $k$ . For simple linear regression, which we have been discussing,  $k = 1$  and thus Reg MS = Reg SS. For multiple regression in Section 11.9,  $k$  is  $>1$ . We will refer to  $k$  as the degrees of freedom for the regression sum of squares, or Reg  $df$ .

**Definition 11.13**

The **residual mean square**, or **Res MS**, is the ratio of the Res SS divided by  $(n - k - 1)$ , or Res MS = Res SS/ $(n - k - 1)$ . For simple linear regression,  $k = 1$  and Res MS = Res SS/ $(n - 2)$ . We refer to  $n - k - 1$  as the degrees of freedom for the residual sum of squares, or Res  $df$ . Res MS is also sometimes denoted by  $s_{y-x}^2$  in the literature.

Under  $H_0$ ,  $F = \text{Reg MS}/\text{Res MS}$  follows an  $F$  distribution with 1 and  $n - 2$   $df$ , respectively.  $H_0$  should be rejected for large values of  $F$ . Thus, for a level  $\alpha$  test,  $H_0$  will be rejected if  $F > F_{1,n-2,1-\alpha}$  and accepted otherwise.

The expressions for the regression and residual sums of squares in Equation 11.5 simplify for computational purposes as follows.

**Equation 11.6****Short Computational Form for Regression and Residual SS**

$$\text{Regression SS} = bL_{xy} = b^2 L_{xx} = L_{xy}^2 / L_{xx}$$

$$\text{Residual SS} = \text{Total SS} - \text{Regression SS} = L_{yy} - L_{xy}^2 / L_{xx}$$

Thus the test procedure can be summarized as follows.

**Equation 11.7**
**F Test for Simple Linear Regression**

To test  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ , use the following procedure:

- (1) Compute the test statistic

$$F = \text{Reg MS}/\text{Res MS} = \left( L_{xy}^2/L_{xx} \right) / \left[ \left( L_{yy} - L_{xy}^2/L_{xx} \right) / (n-2) \right]$$

that follows an  $F_{1,n-2}$  distribution under  $H_0$ .

- (2) For a two-sided test with significance level  $\alpha$ , if

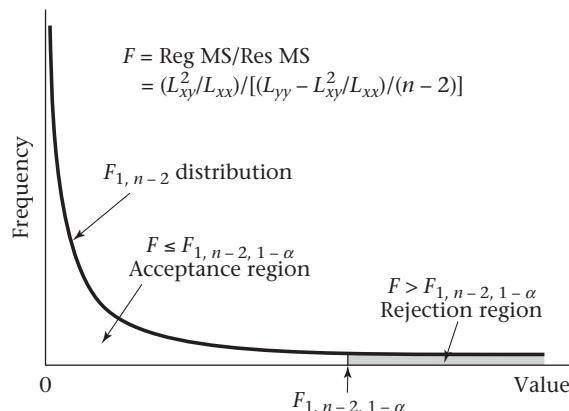
$F > F_{1,n-2,1-\alpha}$ , then reject  $H_0$ ; if

$F \leq F_{1,n-2,\alpha}$ , then accept  $H_0$ .

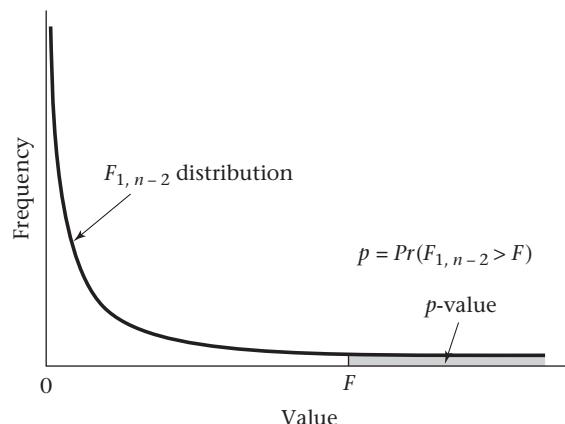
- (3) The exact  $p$ -value is given by  $\Pr(F_{1,n-2} > F)$ .

The acceptance and rejection regions for the regression  $F$  test are illustrated in Figure 11.7. The computation of the  $p$ -value for the regression  $F$  test is shown in Figure 11.8. These results are typically summarized in an analysis-of-variance (ANOVA) table, as in Table 11.2.

**Figure 11.7 Acceptance and rejection regions for the simple linear-regression  $F$  test**



**Figure 11.8 Computation of the  $p$ -value for the simple linear-regression  $F$  test**



**Table 11.2** ANOVA table for displaying regression results

	SS	df	MS	F statistic	p-value
Regression	(a) <sup>a</sup>	1	(a)/1	$F = [(a)/1]/[(b)/(n - 2)]$	$Pr(F_{1,n-2} > F)$
Residual	(b) <sup>b</sup>	$n - 2$	$(b)/(n - 2)$		
Total	$(a) + (b)$				

<sup>a</sup>(a) = Regression SS.<sup>b</sup>(b) = Residual SS.**Example 11.12**

**Obstetrics** Test for the significance of the regression line derived for the birthweight–estriol data in Example 11.8.

**Solution**

From Example 11.8,

$$L_{xy} = 412, L_{xx} = 677.42$$

Furthermore,

$$\sum_{i=1}^{31} y_i^2 = 32,418 \quad L_{yy} = \sum_{i=1}^{31} y_i^2 - \left( \sum_{i=1}^{31} y_i \right)^2 / 31 = 32,418 - 992^2 / 31 = 674$$

Therefore,

$$\text{Reg SS} = L_{xy}^2 / L_{xx} = \text{Reg MS} = 412^2 / 677.42 = 250.57$$

$$\text{Total SS} = L_{yy} = 674$$

$$\text{Res SS} = \text{Total SS} - \text{Reg SS} = 674 - 250.57 = 423.43$$

$$\text{Res MS} = \text{Res SS} / (31 - 2) = \text{Res SS} / 29 = 423.43 / 29 = 14.60$$

$$F = \text{Reg MS} / \text{Res MS} = 250.57 / 14.60 = 17.16 \sim F_{1,29} \text{ under } H_0$$

From Table 9 in the Appendix,

$$F_{1,29,999} < F_{1,20,999} = 14.82 < 17.16 = F$$

Therefore,  $p < .001$

and  $H_0$  is rejected and the alternative hypothesis, namely that the slope of the regression line is significantly different from 0, is accepted, implying a *significant linear relationship* between birthweight and estriol level. These results are summarized in the ANOVA table (Table 11.3) using the MINITAB REGRESSION program.

**Table 11.3** ANOVA results for the birthweight–estriol data in Example 11.12**Analysis of Variance**

Source	DF	SS	MS	F	P
Regression	1	250.57	250.57	17.16	0.000
Residual Error	29	423.43	14.60		
Total	30	674.00			

A summary measure of goodness of fit frequently referred to in the literature is  $R^2$ .

**Definition 11.14**  $R^2$  is defined as Reg SS/Total SS.

$R^2$  can be thought of as the proportion of the variance of  $y$  that is explained by  $x$ . If  $R^2 = 1$ , then all variation in  $y$  can be explained by variation in  $x$ , and all data points fall

on the regression line. In other words, once  $x$  is known  $y$  can be predicted exactly, with no error or variability in the prediction. If  $R^2 = 0$ , then  $x$  gives no information about  $y$ , and the variance of  $y$  is the same with or without knowing  $x$ . If  $R^2$  is between 0 and 1, then for a given value of  $x$ , the variance of  $y$  is lower than it would be if  $x$  were unknown but is still greater than 0. In particular, the best estimate of the variance of  $y$  given  $x$  (or  $\sigma^2$  in the regression model in Equation 11.2) is given by Res MS (or  $s_{y \cdot x}^2$ ). For large  $n$ ,  $s_{y \cdot x}^2 \approx s_y^2(1 - R^2)$ . Thus  $R^2$  represents the proportion of the variance of  $y$  that is explained by  $x$ .

**Example 11.13**

**Obstetrics** Compute and interpret  $R^2$  and  $s_{y \cdot x}^2$  for the birthweight–estriol data in Example 11.12.

**Solution**

From Table 11.3, the  $R^2$  for the birthweight–estriol regression line is given by  $250.57/674 = .372$ . Thus about 37% of the variance of birthweight can be explained by estriol level. Furthermore,  $s_{y \cdot x}^2 = 14.60$ , as compared with

$$s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1) = 674 / 30 = 22.47$$

Thus, for the subgroup of women with a specific estriol level, such as 10 mg/24 hr, the variance of birthweight is 14.60, whereas for *all* women with any estriol level, the variance of birthweight is 22.47. Note that

$$s_{y \cdot x}^2 / s_y^2 = 14.60 / 22.47 = .650 \approx 1 - R^2 = 1 - .372 = .628$$

In some computer packages (e.g., Stata) the expression  $1 - s_{y \cdot x}^2 / s_y^2$  is referred to as *adjusted R*<sup>2</sup>. Thus, for the estriol data,  $R^2 = 0.372$ , while adjusted  $R^2 = 0.350$ . For large  $n$ , adjusted  $R^2 \approx R^2$ . For small  $n$ , a better measure of % variation of  $y$  explained by  $x$  is given by the adjusted  $R^2$ .

**Example 11.14**

**Pulmonary Disease** Forced expiratory volume (FEV) is a standard measure of pulmonary function. To identify people with abnormal pulmonary function, standards of FEV for normal people must be established. One problem here is that FEV is related to both age and height. Let us focus on boys who are ages 10–15 and postulate a regression model of the form  $FEV = \alpha + \beta(\text{height}) + e$ . Data were collected on FEV and height for 655 boys in this age group residing in Tecumseh, Michigan [2]. Table 11.4 presents the mean FEV in liters for each of twelve 4-cm height groups. Find the best-fitting regression line, and test it for statistical significance. What proportion of the variance of FEV can be explained by height?

**Table 11.4**

**Mean FEV by height group for boys ages 10–15 in Tecumseh, Michigan**

Height (cm)	Mean FEV (L)	Height (cm)	Mean FEV (L)
134 <sup>a</sup>	1.7	158	2.7
138	1.9	162	3.0
142	2.0	166	3.1
146	2.1	170	3.4
150	2.2	174	3.8
154	2.5	178	3.9

<sup>a</sup>The middle value of each 4-cm height group is given here.

Source: Reprinted with permission of the *American Review of Respiratory Disease*, 108, 258–272, 1973.

**Solution**

A linear-regression line is fitted to the points in Table 11.4:

$$\sum_{i=1}^{12} x_i = 1872 \quad \sum_{i=1}^{12} x_i^2 = 294,320 \quad \sum_{i=1}^{12} y_i = 32.3$$

$$\sum_{i=1}^{12} y_i^2 = 93.11 \quad \sum_{i=1}^{12} x_i y_i = 5156.20$$

Therefore,

$$L_{xy} = 5156.20 - \frac{1872(32.3)}{12} = 117.4$$

$$L_{xx} = 294,320 - \frac{1872^2}{12} = 2288$$

$$b = L_{xy}/L_{xx} = 0.0513$$

$$a = \left( \sum_{i=1}^{12} y_i - b \sum_{i=1}^{12} x_i \right) / 12 = [32.3 - 0.0513(1872)] / 12 = -5.313$$

Thus the fitted regression line is

$$\text{FEV} = -5.313 + 0.0513 \times \text{height}$$

Statistical significance is assessed by computing the  $F$  statistic in Equation 11.7 as follows:

$$\text{Reg SS} = L_{xy}^2 / L_{xx} = 117.4^2 / 2288 = 6.024 = \text{Reg MS}$$

$$\text{Total SS} = L_{yy} = 93.11 - 32.3^2 / 12 = 6.169$$

$$\text{Res SS} = 6.169 - 6.024 = 0.145$$

$$\text{Res MS} = \text{Res SS} / (n - 2) = 0.145 / 10 = 0.0145$$

$$F = \text{Reg MS} / \text{Res MS} = 414.8 \sim F_{1,10} \text{ under } H_0$$

Clearly, the fitted line is statistically significant because from Table 9 in the Appendix,  $F_{1,10,999} = 21.04$ , so  $p < .001$ . These results can be displayed in an ANOVA table (Table 11.5).

**Table 11.5**

**ANOVA table for the FEV–height regression results in Example 11.14**

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	6.0239	6.0239	414.78	0.000
Residual Error	10	0.1452	0.0145		
Total	11	6.1692			

Finally, the proportion of the variance of FEV that is explained by height is estimated by adjusted  $R^2 = 1 - 0.0145/(6.1692/11) = .974$ . Thus differences in height explain most of the variability in FEV among boys in this age group.

### *t* Test for Simple Linear Regression

In this section an alternative method for testing the hypothesis  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$  is presented. This method is based on the *t* test and is equivalent to the *F* test presented in Equation 11.7. The procedure is widely used and also provides interval estimates for  $\beta$ .

The hypothesis test here is based on the sample regression coefficient  $b$  or, more specifically, on  $b/se(b)$ , and  $H_0$  will be rejected if  $|b|/se(b) > c$  for some constant  $c$  and will be accepted otherwise.

The sample regression coefficient  $b$  is an **unbiased estimator** of the population regression coefficient  $\beta$  and, in particular, under  $H_0$ ,  $E(b) = 0$ . Furthermore, the variance of  $b$  is given by

$$\sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2 = \sigma^2 / L_{xx}$$

In general,  $\sigma^2$  is unknown. However, the best estimate of  $\sigma^2$  is given by  $s_{y \cdot x}^2$ . Hence

$$se(b) \approx s_{y \cdot x} / (L_{xx})^{1/2}$$

Finally, under  $H_0$ ,  $t = b/se(b)$  follows a  $t$  distribution with  $n - 2$  df. Therefore, the following test procedure for a two-sided test with significance level  $\alpha$  is used.

### Equation 11.8

#### ***t* Test for Simple Linear Regression**

To test the hypothesis  $H_0: \beta = 0$  vs.

$H_1: \beta \neq 0$ , use the following procedure:

(1) Compute the test statistic

$$t = b / (s_{y \cdot x}^2 / L_{xx})^{1/2}$$

(2) For a two-sided test with significance level  $\alpha$ ,

If  $t > t_{n-2, 1-\alpha/2}$  or  $t < t_{n-2, \alpha/2} = -t_{n-2, 1-\alpha/2}$

then reject  $H_0$ ;

if  $-t_{n-2, 1-\alpha/2} \leq t \leq t_{n-2, 1-\alpha/2}$

then accept  $H_0$ .

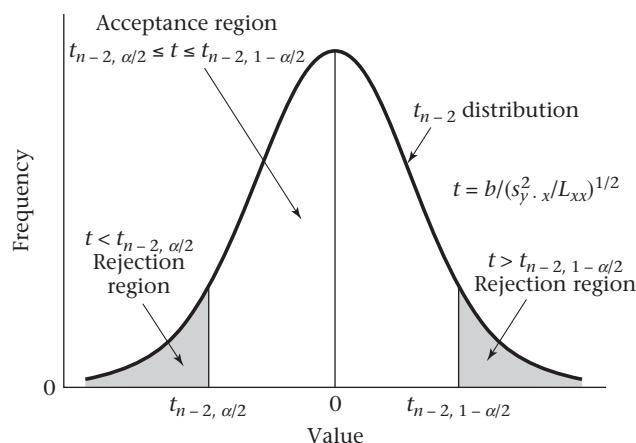
(3) The  $p$ -value is given by

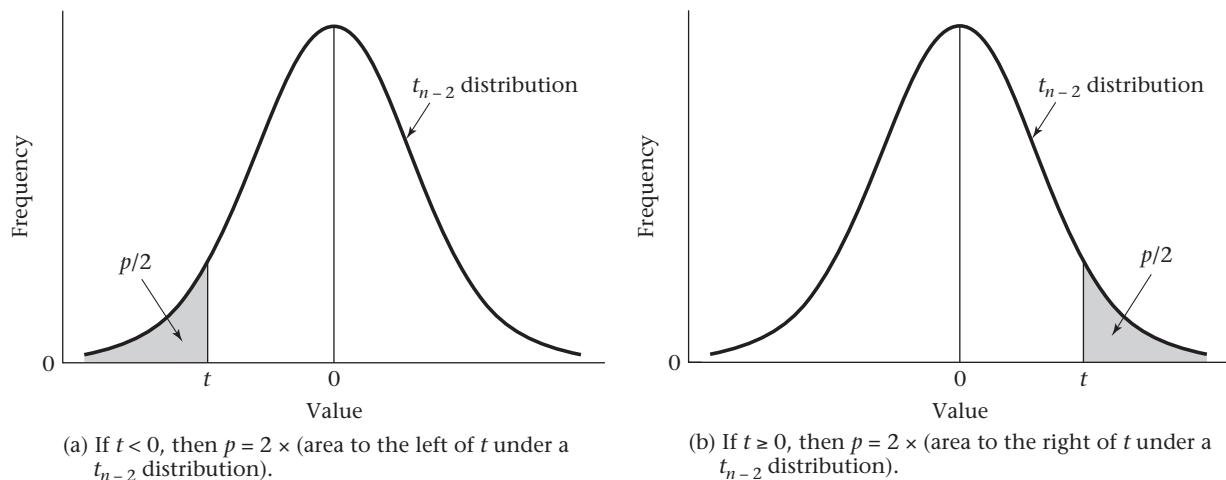
$p = 2 \times (\text{area to the left of } t \text{ under a } t_{n-2} \text{ distribution}) \text{ if } t < 0$

$p = 2 \times (\text{area to the right of } t \text{ under a } t_{n-2} \text{ distribution}) \text{ if } t \geq 0$

The acceptance and rejection regions for this test are shown in Figure 11.9. Computation of the  $p$ -value is illustrated in Figure 11.10.

**Figure 11.9 Acceptance and rejection regions for the  $t$  test for simple linear regression**



**Figure 11.10 Computation of the  $p$ -value for the  $t$  test for simple linear regression**

The  $t$  test in this section and the  $F$  test in Equation 11.7 are equivalent in that they always provide the same  $p$ -values. Which test is used is a matter of personal preference; both appear in the literature.

**Example 11.15**

**Obstetrics** Assess the statistical significance for the birthweight–estriol data using the  $t$  test in Equation 11.8.

**Solution**

From Example 11.8,  $b = L_{xy}/L_{xx} = 0.608$ . Furthermore, from Table 11.3 and Example 11.12,

$$se(b) = \left( s_{y-x}^2 / L_{xx} \right)^{1/2} = (14.60/677.42)^{1/2} = 0.147$$

Thus  $t = b/se(b) = 0.608/0.147 = 4.14 \sim t_{29}$  under  $H_0$

Because  $t_{29, .9995} = 3.659 < 4.14 = t$

we have  $p < 2 \times (1 - .9995) = .001$

This information is summarized in Table 11.6. Note that the  $p$ -values based on the  $F$  test in Table 11.3 and the  $t$  test in Table 11.6 are the same ( $p = .000$ ).

**Table 11.6****The  $t$  test approach for the birthweight–estriol example**

The regression equation is brthwgt = 21.5 + 0.608 estriol				
Predictor	Coef	SE Coef	T	P
Constant	21.523	2.620	8.21	0.000
estriol	0.6082	0.1468	4.14	0.000

**11.5 Interval Estimation for Linear Regression****Interval Estimates for Regression Parameters**

Standard errors and interval estimates for the parameters of a regression line are often computed to obtain some idea of the precision of the estimates. Furthermore, if we want to compare our regression coefficients with previously published regression coefficients  $\beta_0$  and  $\alpha_0$ , where these estimates are based on much larger samples

than ours, then, based on our data, we can check whether  $\beta_0$  and  $\alpha_0$  fall within the 95% confidence intervals for  $\beta$  and  $\alpha$ , respectively, to decide whether the two sets of results are comparable.

The standard errors of the estimated regression parameters are given as follows.

**Equation 11.9**
**Standard Errors of Estimated Parameters in Simple Linear Regression**

$$se(b) = \sqrt{\frac{s_{y-x}^2}{L_{xx}}}$$

$$se(a) = \sqrt{s_{y-x}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{L_{xx}} \right)}$$

Furthermore, the two-sided  $100\% \times (1 - \alpha)$  confidence intervals for  $\beta$  and  $\alpha$  are given by

**Equation 11.10**
**Two-Sided  $100\% \times (1 - \alpha)$  Confidence Intervals for the Parameters of a Regression Line**

If  $b$  and  $a$  are, respectively, the estimated slope and intercept of a regression line as given in Equation 11.3 and  $se(b)$ ,  $se(a)$  are the estimated standard errors as given in Equation 11.9, then the two-sided  $100\% \times (1 - \alpha)$  confidence intervals for  $\beta$  and  $\alpha$  are given by

$$b \pm t_{n-2, 1-\alpha/2} se(b) \quad \text{and} \quad a \pm t_{n-2, 1-\alpha/2} se(a), \quad \text{respectively.}$$

**Example 11.16**

**Obstetrics** Provide standard errors and 95% confidence intervals for the regression parameters of the birthweight–estriol data in Table 11.1.

**Solution**

From Example 11.15, the standard error of  $b$  is given by

$$\sqrt{14.60 / 677.42} = 0.147$$

Thus a 95% confidence interval for  $\beta$  is obtained from

$$0.608 \pm t_{29, .975}(0.147) = 0.608 \pm 2.045(0.147) = 0.608 \pm 0.300 = (0.308, 0.908)$$

Compute  $\bar{x}$  to obtain the standard error of  $a$ . From Example 11.8,

$$\bar{x} = \frac{\sum_{i=1}^{31} x_i}{31} = \frac{534}{31} = 17.23$$

Thus the standard error of  $a$  is given by

$$\sqrt{14.60 \left( \frac{1}{31} + \frac{17.23^2}{677.42} \right)} = 2.62$$

It follows that a 95% confidence interval for  $\alpha$  is provided by

$$21.52 \pm t_{29, .975}(2.62) = 21.52 \pm 2.045(2.62) = 21.52 \pm 5.36 = (16.16, 26.88)$$

These intervals are rather wide, which is not surprising due to the small sample size.

Suppose another data set based on 500 pregnancies, where the birthweight–estriol regression line is estimated as  $y = 25.04 + 0.52x$ , is found in the literature.

Because 0.52 is within the 95% confidence interval for the slope and 25.04 is within the 95% confidence interval for the intercept, our results are comparable to those of the earlier study. We assume in this analysis that the variability in the slope (0.52) and intercept (25.04) estimates from the sample of 500 pregnancies is negligible compared with the error from the data set with 31 pregnancies.

## Interval Estimation for Predictions Made from Regression Lines

One important use for regression lines is in making predictions. Frequently, the accuracy of these predictions must be assessed.

### Example 11.17

**Pulmonary Function** Suppose we want to use the FEV–height regression line computed in Example 11.14 to develop normal ranges for 10- to 15-year-old boys of specific heights. In particular, consider John H., who is 12 years old and 160 cm tall and whose FEV is 2.5 L. Can his FEV be considered abnormal for his age and height?

In general, if all boys of height  $x$  are considered, then the average FEV for such boys can be best estimated from the regression equation by  $\hat{y} = a + bx$ . How accurate is this estimate? The answer to this question depends on whether we are making predictions for *one specific boy* or for the *mean value of all boys of a given height*. The first estimate would be useful to a pediatrician interested in assessing the lung function of a particular patient, whereas the second estimate would be useful to a researcher interested in relationships between pulmonary function and height over large populations of boys. The standard error ( $se_1$ ) of the first type of estimate and the resulting interval estimate are given as follows.

### Equation 11.11

#### Predictions Made from Regression Lines for Individual Observations

Suppose we wish to make predictions from a regression line for an individual observation with independent variable  $x$  that was not used in constructing the regression line. The distribution of observed  $y$  values for the subset of individuals with independent variable  $x$  is normal with mean  $= \hat{y} = a + bx$  and standard deviation given by

$$se_1(\hat{y}) = \sqrt{s_{y|x}^2 \left[ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}} \right]}$$

Furthermore,  $100\% \times (1 - \alpha)$  of the observed values will fall within the interval

$$\hat{y} \pm t_{n-2, 1-\alpha/2} se_1(\hat{y})$$

This interval is sometimes called a  $100\% \times (1 - \alpha)$  prediction interval for  $y$ .

### Example 11.18

**Pulmonary Function** Can the FEV of John H. in Example 11.17 be considered abnormal for his age and height?

#### Solution

John's observed FEV is 2.5 L. The regression equation relating FEV and height was computed in Example 11.14 and is given by  $y = -5.313 + 0.0513 \times \text{height}$ . Thus the estimated average FEV for 12-year-old boys of height 160 cm is

$$\hat{y} = -5.313 + 160 \times 0.0513 = 2.90 \text{ L}$$

Before computing the  $se_1(\hat{y})$ , we need to obtain  $\bar{x}$ . From Example 11.14,

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{1872}{12} = 156.0$$

Thus  $se_1(\hat{y})$  is given by

$$se_1(\hat{y}) = \sqrt{0.0145 \left[ 1 + \frac{1}{12} + \frac{(160 - 156)^2}{2288} \right]} = \sqrt{0.0145(1.090)} = 0.126$$

Furthermore, 95% of boys of this age and height will have an FEV between

$$2.90 \pm t_{10, .975}(0.126) = 2.90 \pm 2.228(0.126) = 2.90 \pm 0.28 = (2.62, 3.18)$$

How can this prediction interval be used? Because the observed FEV for John H. (2.5 L) does not fall within this interval, we can say that John's lung function is abnormally low for a boy of his age and height; to find a reason for this abnormality, further exploration, if possible, is needed.

The magnitude of the standard error in Equation 11.11 depends on how far the observed value of  $x$  for the new sample point is from the mean value of  $x$  for the data points used in computing the regression line ( $\bar{x}$ ). The standard error is smaller when  $x$  is close to  $\bar{x}$  than when  $x$  is far from  $\bar{x}$ . In general, making predictions from a regression line for values of  $x$  that are very far from  $\bar{x}$  is risky because the predictions are likely to be more inaccurate.

### Example 11.19

**Pulmonary Function** Suppose Bill is 190 cm tall, with an FEV of 3.5 L. Compare the standard error of his predicted value with that for John, given in Example 11.18.

### Solution

From Equation 11.11,

$$se_1(\hat{y}) = \sqrt{0.0145 \left[ 1 + \frac{1}{12} + \frac{(190 - 156)^2}{2288} \right]} \\ = \sqrt{0.0145(1.589)} = 0.152 > 0.126 = se_1 \quad (\text{computed in Example 11.18})$$

This result is expected because 190 cm is further than 160 cm from  $\bar{x} = 156$  cm.

Suppose we want to assess the mean value of FEV for a large number of boys of a particular height rather than for one particular boy. This parameter might interest a researcher working with growth curves of pulmonary function in children. How can the estimated mean FEV and the standard error of the estimate be found? The procedure is as follows.

### Equation 11.12

#### Standard Error and Confidence Interval for Predictions Made from Regression Lines for the Average Value of $y$ for a Given $x$

The best estimate of the average value of  $y$  for a given  $x$  is  $\hat{y} = a + bx$ . Its standard error, denoted by  $se_2(\hat{y})$ , is given by

$$se_2(\hat{y}) = \sqrt{s_{y \cdot x}^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}} \right]}$$

Furthermore, a two-sided  $100\% \times (1 - \alpha)$  confidence interval for the average value of  $y$  is

$$\hat{y} \pm t_{n-2, 1-\alpha/2} se_2(\hat{y})$$

**Example 11.20**

**Pulmonary Function** Compute the standard error and 95% confidence interval for the average value of FEV over a large number of boys with height of 160 cm.

**Solution**

See the results of Example 11.18 for the necessary raw data to perform the computations. The best estimate of the mean value of FEV is the same as given in Example 11.18, which was 2.90 L. However, the standard error is computed differently. From Equation 11.12,

$$se_2(\hat{y}) = \sqrt{0.0145 \left[ \frac{1}{12} + \frac{(160 - 156)^2}{2288} \right]} = \sqrt{0.0145(0.090)} = 0.036$$

Therefore, a 95% confidence interval for the mean value of FEV over a large number of boys with height 160 cm is given by

$$2.90 \pm t_{10,975}(0.036) = 2.90 \pm 2.228(0.036) = 2.90 \pm 0.08 = (2.82, 2.98)$$

Notice that this interval is much narrower than the interval computed in Example 11.18 (2.62, 3.18), which is a range encompassing approximately 95% of individual boys' FEVs. This disparity reflects the intuitive idea that there is much more precision in estimating the mean value of  $y$  for a large number of boys with the same height  $x$  than in estimating  $y$  for one particular boy with height  $x$ .

Note again that the standard error for the average value of  $y$  for a given value of  $x$  is not the same for all values of  $x$  but gets larger the further  $x$  is from the mean value of  $x$  ( $\bar{x}$ ) used to estimate the regression line.

**Example 11.21**

**Pulmonary Function** Compare the standard error of the average FEV for boys of height 190 cm with that for boys of 160 cm.

**Solution**

From Equation 11.12,

$$\begin{aligned} se_2(\hat{y}) &= \sqrt{0.0145 \left[ \frac{1}{12} + \frac{(190 - 156)^2}{2288} \right]} = \sqrt{0.0145(0.589)} \\ &= 0.092 > 0.036 = se_2(\hat{y}) \text{ for } x = 160 \text{ cm} \end{aligned}$$

This result is expected because 190 cm is further than 160 cm from  $\bar{x} = 156$  cm.

**REVIEW QUESTIONS 11A**

- 1 What is a residual? Why are residuals important in regression analysis?
- 2 A 79-year-old man was admitted to the hospital with coronary-artery disease, abdominal pain, and worsening intermittent claudication (which roughly means loss of circulation in the legs, making walking difficult and/or painful) [3]. As part of the patient's workup, his lab values were followed over time while in the hospital. His hematocrit (%) values over the first 7 days in the hospital are shown in Table 11.7.

**Table 11.7** Hematocrit (%) values over the first 7 days in the hospital for a patient with intermittent claudication

Day 0	Day 3	Day 4	Day 5	Day 6	Day 7
28.9	28.7	26.4	30.4	30.3	33.2

- (a) Fit a linear-regression line to the hematocrit values over the 7-day period.
- (b) Is there a statistically significant change in his hematocrit values over time?
- (c) Suppose we want to predict his hematocrit on the eighth hospital day.
  - (i) What is the best estimate of this value?
  - (ii) What is the standard error of this estimate?
  - (iii) What is a 95% confidence interval associated with this estimate?

## 11.6 Assessing the Goodness of Fit of Regression Lines

A number of assumptions were made in using the methods of simple linear regression in the previous sections of this chapter. What are some of these assumptions, and what possible situations could be encountered that would make these assumptions not viable?

### Equation 11.13

#### Assumptions Made in Linear-Regression Models

- (1) For any given value of  $x$ , the corresponding value of  $y$  has an average value  $\alpha + \beta x$ , which is a linear function of  $x$ .
- (2) For any given value of  $x$ , the corresponding value of  $y$  is normally distributed about  $\alpha + \beta x$  with the same variance  $\sigma^2$  for any  $x$ .
- (3) For any two data points  $(x_1, y_1), (x_2, y_2)$ , the error terms  $e_1, e_2$  are independent of each other.

Let us now reassess the birthweight–estriol data for possible violation of linear regression assumptions. To assess whether these assumptions are reasonable, we can use several different kinds of plots. The simplest plot is the  $x - y$  scatter plot. Here we plot the dependent variable  $y$  vs. the independent variable  $x$  and superimpose the regression line  $y = a + bx$  on the same plot. We have constructed a scatter plot of this type for the birthweight–estriol data in Figure 11.1. The linearity assumption appears reasonable in that there is no obvious curvilinearity in the raw data. However, there is a hint that there is more variability about the regression line for higher estriol values than for lower estriol values. To focus more clearly on this issue, we can compute the residuals about the fitted regression line and then construct a scatter plot of the residuals vs. either the estriol values ( $x$ ) or the predicted birthweights ( $\hat{y} = a + bx$ ).

From Equation 11.2, we see that the errors ( $e$ ) about the true regression line ( $y = \alpha + \beta x$ ) have the same variance  $\sigma^2$ . However, it can be shown that the residuals about the fitted regression line ( $y = a + bx$ ) have different variances depending on how far an individual  $x$  value is from the mean  $x$  value used to generate the regression line. Specifically, residuals for points  $(x_i, y_i)$  where  $x_i$  is close to the mean  $x$  value for all points used in constructing the regression line (i.e.,  $|x_i - \bar{x}|$  is small) will tend to be larger than residuals where  $|x_i - \bar{x}|$  is large. Interestingly, if  $|x_i - \bar{x}|$  is very large, then the regression line is forced to go through the point  $(x_i, y_i)$  (or nearly through it) with a small residual for this point. The standard deviation of the residuals is given in Equation 11.14.

### Equation 11.14

#### Standard Deviation of Residuals About the Fitted Regression Line

Let  $(x_i, y_i)$  be a sample point used in estimating the regression line,  $y = \alpha + \beta x$ .

If  $y = a + bx$  is the estimated regression line, and

$\hat{e}_i$  = residual for the point  $(x_i, y_i)$  about the estimated regression line, then

$\hat{e}_i = y_i - (a + bx_i)$  and

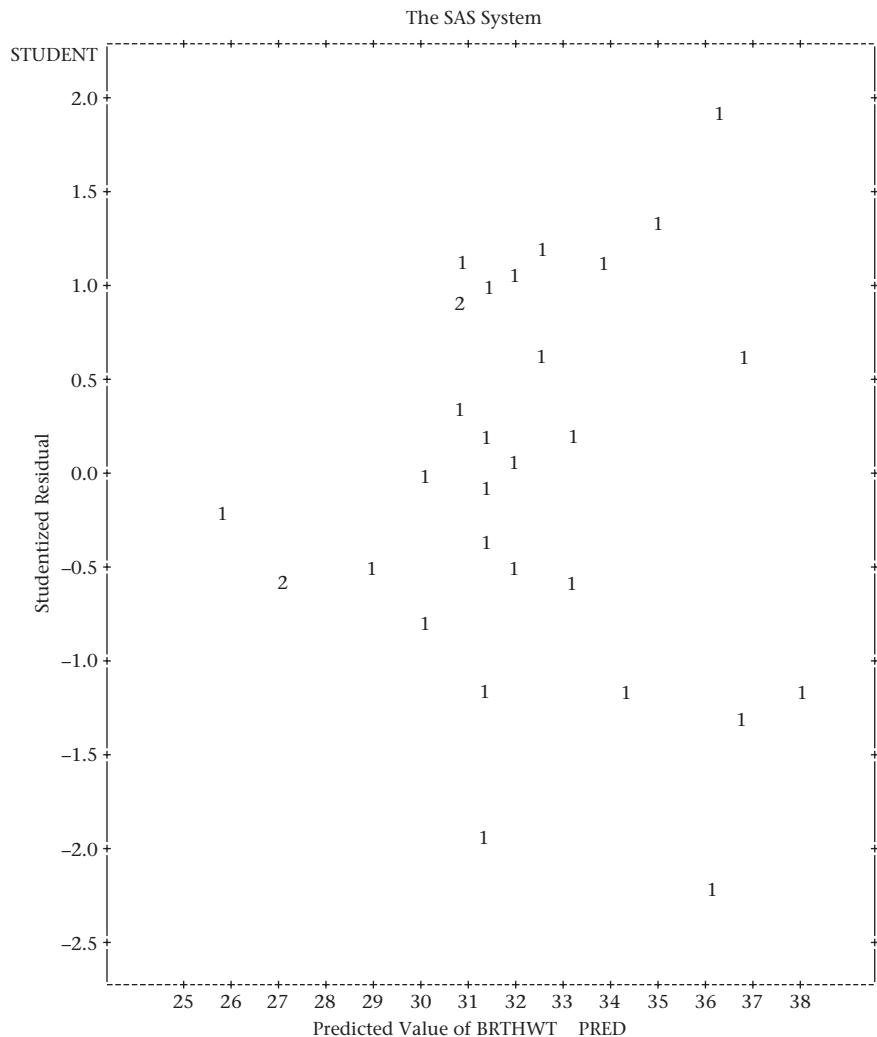
$$sd(\hat{e}_i) = \sqrt{\hat{\sigma}^2 \left[ 1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}} \right]}$$

The **Studentized residual** corresponding to the point  $(x_i, y_i)$  is  $\hat{e}_i/sd(\hat{e}_i)$ .

In Figure 11.11, we have plotted the Studentized residuals (the individual residuals divided by their standard deviations) vs. the predicted birthweights (g/100) ( $\hat{y} = 21.52 + 0.608 \times \text{estriol}$ ).

A point labeled 2 indicates that there are two identical data points—for example, the second and third points in Table 11.1 are both (9, 25). There is still a hint that

**Figure 11.11** Plot of Studentized residuals vs. the predicted value of birthweight for the birthweight–estriol data in Table 11.1



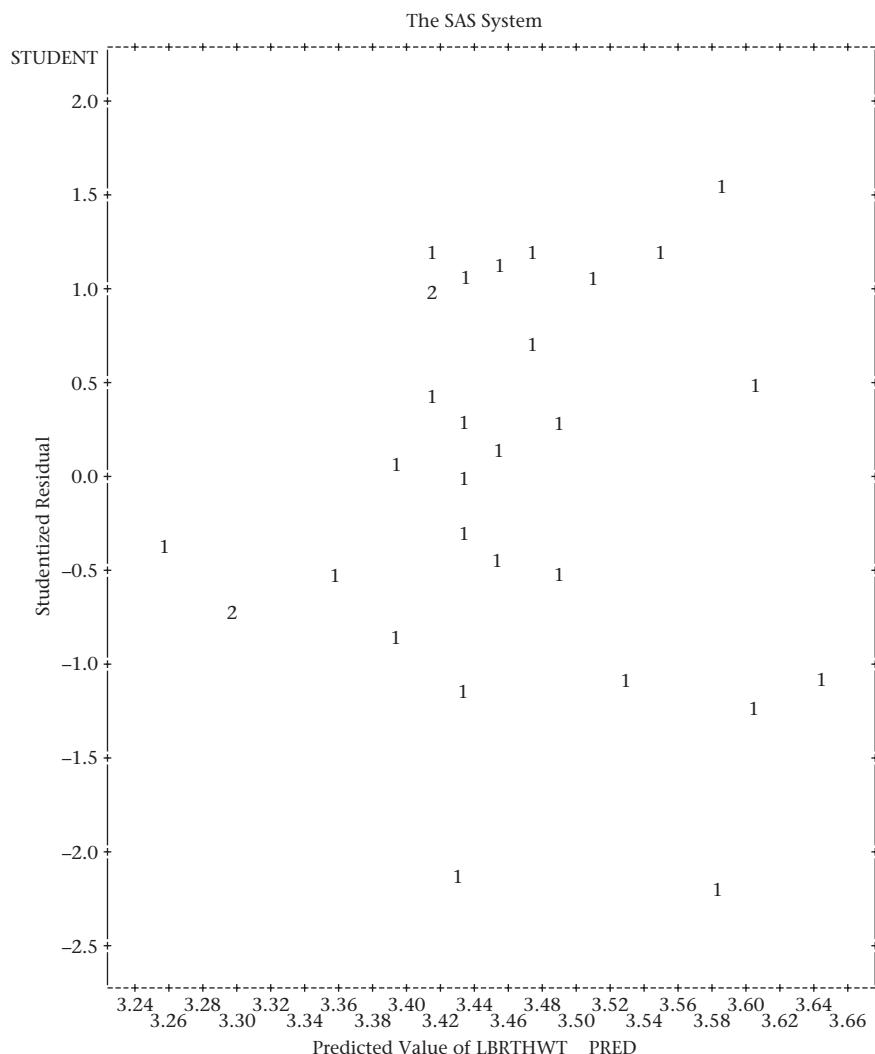
the spread increases slightly as the predicted birthweight increases. However, this impression is mainly due to the four data points with the lowest predicted values, all of which have residuals that are close to 0. One commonly used strategy that can be employed if unequal residual variances are present is to transform the dependent variable ( $y$ ) to a different scale. This type of transformation is called a **variance-stabilizing transformation**. The goal of using such a transformation is to make the residual variances approximately the same for each level of  $x$  (or, equivalently, each level of the predicted value). The most common transformations when the residual variance is an increasing function of  $x$  are either the ln or square-root transformations. The square-root transformation is useful when the residual variance is proportional to the average value of  $y$  (e.g., if the average value goes up by a factor of 2, then the residual variance goes up by a factor of 2 also). The log transformation is useful when the residual variance is proportional to the square of the average values (e.g., if the average value goes up by a factor of 2, then the residual variance goes up by a factor of 4). For purposes of illustration, we have computed the regression using the ln transformation for birth-weight (i.e.,  $y = \ln \text{birthweight}$ ). The residual plot is shown in Figure 11.12.

The plots in Figures 11.11 and 11.12 look similar. The plot using the square-root transformation for birthweight is also similar. Therefore, we would probably choose to keep the data in the original scale for the sake of simplicity. However, in other data sets the use of the appropriate transformation is crucial and each of the linearity, equal-variance, and normality assumptions can be made more plausible using a transformed scale. However, occasionally a transformation may make the equal-variance assumption more plausible but the linearity assumption less plausible. Another possibility is to keep the data in the original scale but employ a weighted regression in which the weight is approximately inversely proportional to the residual variance. This may be reasonable if the data points consist of averages over varying numbers of individuals (e.g., people living in different cities, where the weight is proportional to the size of the city). Weighted regression is beyond the scope of this text (see Draper & Smith [4] for a more complete discussion of this technique).

Other issues of concern in judging the goodness of fit of a regression line are **outliers** and **influential points**. In Section 8.9, we discussed methods for the detection of outliers in a sample, where only a single variable is of interest. However, it is more difficult to detect outliers in a regression setting than in univariate problems, particularly if multiple outliers are present in a data set. *Influential points* are defined heuristically as points that have an important influence on the coefficients of the fitted regression lines. Suppose we delete the  $i$ th sample point and refit the regression line from the remaining  $n - 1$  data points. If we denote the estimated slope and intercept for the reduced data set by  $b^{(i)}$  and  $a^{(i)}$ , respectively, then the sample point will be influential if either  $|b - b^{(i)}|$  or  $|a - a^{(i)}|$  is large. Outliers and influential points are not necessarily the same. An outlier  $(x_i, y_i)$  may or may not be influential, depending on its location relative to the remaining sample points. For example, if  $|x_i - \bar{x}|$  is small, then even a gross outlier will have a relatively small influence on the slope estimate but will have an important influence on the intercept estimate. Conversely, if  $|x_i - \bar{x}|$  is large, then even a data point that is not a gross outlier may be influential. See Draper & Smith [4] and Weisberg [5] for a more complete description of residual analysis, detection of outliers, and influential points in a regression setting.

We have discussed using residual analysis to assess the validity of the linearity assumption (assumption 1 in Equation 11.13), and the validity of the equal-variance

**Figure 11.12** Plot of Studentized residuals vs. the predicted value of  $\ln(\text{birthweight})$  for the birthweight-estriol data in Table 11.1



assumption (assumption 2 in Equation 11.13). The normality assumption is most important in small samples. In large samples, an analog to the central-limit theorem can be used to establish the unbiasedness of  $b$  as an estimator of  $\beta$  and the appropriateness of test of significance concerning  $\beta$  (such as the  $F$  test for simple linear regression in Equation 11.7 or the  $t$  test for simple linear regression in Equation 11.8), or formulas for confidence-interval width of  $\beta$  (Equation 11.10), even if the error terms are not normally distributed. The independence assumption (assumption 3, Equation 11.13) is important to establish the validity of  $p$ -values and confidence-interval width from simple linear regression. Specifically, if multiple data points from the same individual are used in fitting a regression line, then  $p$ -values will generally be too low, and confidence-interval width will generally be too narrow using standard methods of regression analysis (which assume independence). We discuss this type of *clustered data* in more detail in Chapter 13.

## 11.7 The Correlation Coefficient

The discussion of linear-regression analysis in Sections 11.2–11.6 primarily focused on methods of predicting one dependent variable ( $y$ ) from an independent variable ( $x$ ). Often we are interested not in predicting one variable from another but rather in investigating whether or not there is a relationship between two variables. The **correlation coefficient**, introduced in Definition 5.13, is a useful tool for quantifying the relationship between variables and is better suited for this purpose than the regression coefficient.

### Example 11.22

**Cardiovascular Disease** Serum cholesterol is an important risk factor in the etiology of cardiovascular disease. Much research has been devoted to understanding the environmental factors that cause elevated cholesterol levels. For this purpose, cholesterol levels were measured on 100 genetically unrelated spouse pairs. We are not interested in predicting the cholesterol level of a husband from that of his wife but instead would like some quantitative measure of the relationship between their levels. We will use the correlation coefficient for this purpose.

In Definition 5.13, we defined the population correlation coefficient  $\rho$ . In general,  $\rho$  is unknown and we have to estimate  $\rho$  by the sample correlation coefficient  $r$ .

### Definition 11.15

The sample (Pearson) correlation coefficient ( $r$ ) is defined by

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

The correlation is not affected by changes in location or scale in either variable and must lie between  $-1$  and  $+1$ . The sample correlation coefficient can be interpreted in a similar manner to the population correlation coefficient  $\rho$ .

### Equation 11.15

#### Interpretation of the Sample Correlation Coefficient

- (1) If the correlation is greater than 0, such as for birthweight and estriol, then the variables are said to be **positively correlated**. Two variables ( $x, y$ ) are positively correlated if as  $x$  increases,  $y$  tends to increase, whereas as  $x$  decreases,  $y$  tends to decrease.
- (2) If the correlation is less than 0, such as for pulse rate and age, then the variables are said to be **negatively correlated**. Two variables ( $x, y$ ) are negatively correlated if as  $x$  increases,  $y$  tends to decrease, whereas as  $x$  decreases,  $y$  tends to increase.
- (3) If the correlation is exactly 0, such as for birthweight and birthday, then the variables are said to be **uncorrelated**. Two variables ( $x, y$ ) are uncorrelated if there is no linear relationship between  $x$  and  $y$ .

Thus the correlation coefficient provides a *quantitative* measure of the dependence between two variables: the closer  $|r|$  is to 1, the more closely related the variables are; if  $|r| = 1$ , then one variable can be predicted exactly from the other.

As was the case for the population correlation coefficient ( $\rho$ ), interpreting the sample correlation coefficient ( $r$ ) in terms of degree of dependence is only correct if the variables  $x$  and  $y$  are normally distributed and in certain other

special cases. If the variables are not normally distributed, then the interpretation may not be correct (see Example 5.30 for an example of two random variables that have correlation coefficient = 0 but are completely dependent).

**Example 11.23** Suppose the two variables under study are temperature in °F ( $y$ ) and temperature in °C ( $x$ ). The correlation between these two variables must be 1 because one variable is a linear function of the other ( $y = \frac{9}{5}x + 32$ ).

**Example 11.24** **Obstetrics** Compute the sample correlation coefficient for the birthweight–estriol data presented in Table 11.1.

**Solution** From Examples 11.8 and 11.12,

$$L_{xy} = 412 \quad L_{xx} = 677.42 \quad L_{yy} = 674$$

$$\text{Therefore, } r = L_{xy} / \sqrt{L_{xx} L_{yy}} = 412 / \sqrt{677.42(674)} = 412/675.71 = .61$$

### Relationship Between the Sample Correlation Coefficient ( $r$ ) and the Population Correlation Coefficient ( $\rho$ )

We can relate the sample correlation coefficient  $r$  and the population correlation coefficient  $\rho$  more clearly by dividing the numerator and denominator of  $r$  by  $(n - 1)$  in Definition 11.15, whereby

**Equation 11.16**

$$r = \frac{L_{xy} / (n - 1)}{\sqrt{\left(\frac{L_{xx}}{n - 1}\right)\left(\frac{L_{yy}}{n - 1}\right)}}$$

We note that  $s_x^2 = L_{xx} / (n - 1)$  and  $s_y^2 = L_{yy} / (n - 1)$ . Furthermore, if we define the *sample covariance* by  $s_{xy} = L_{xy} / (n - 1)$ , then we can re-express Equation 11.16 in the following form.

**Equation 11.17**

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\text{sample covariance between } x \text{ and } y}{(\text{sample standard deviation of } x)(\text{sample standard deviation of } y)}$$

This is completely analogous to the definition of the population correlation coefficient  $\rho$  given in Definition 5.13 with the population quantities,  $Cov(X, Y)$ ,  $\sigma_x$ , and  $\sigma_y$  replaced by their sample estimates  $s_{xy}$ ,  $s_x$ , and  $s_y$ .

### Relationship Between the Sample Regression Coefficient ( $b$ ) and the Sample Correlation Coefficient ( $r$ )

What is the relationship between the sample regression coefficient ( $b$ ) and the sample correlation coefficient ( $r$ )? Note from Equation 11.3 that  $b = L_{xy} / L_{xx}$  and from Definition 11.15 that  $r = L_{xy} / \sqrt{L_{xx} L_{yy}}$ . Therefore, if  $r$  is multiplied by  $\sqrt{L_{yy} / L_{xx}}$ , we obtain

**Equation 11.18**

$$r \sqrt{\frac{L_{yy}}{L_{xx}}} = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} \times \frac{\sqrt{L_{yy}}}{\sqrt{L_{xx}}} = \frac{L_{xy}}{L_{xx}} = b$$

Furthermore, from Definition 11.4,

$$s_y^2 = \frac{L_{yy}}{n-1}$$

$$s_x^2 = \frac{L_{xx}}{n-1}$$

$$s_y^2 / s_x^2 = L_{yy} / L_{xx}$$

$$\text{or } s_y / s_x = \sqrt{L_{yy} / L_{xx}}$$

Substituting  $s_y / s_x$  for  $\sqrt{L_{yy} / L_{xx}}$  on the left-hand side of Equation 11.18 yields the following relationship.

**Equation 11.19**

$$b = \frac{rs_y}{s_x}$$

How can Equation 11.19 be interpreted? The regression coefficient ( $b$ ) can be interpreted as a rescaled version of the correlation coefficient ( $r$ ), where the scale factor is the ratio of the standard deviation of  $y$  to that of  $x$ . Note that  $r$  will be unchanged by a change in the units of  $x$  or  $y$  (or even by which variable is designated as  $x$  and which is designated as  $y$ ), whereas  $b$  is in the units of  $y/x$ .

### Example 11.25

**Pulmonary Function** Compute the correlation coefficient between FEV and height for the pulmonary-function data in Example 11.14.

#### Solution

From Example 11.14,

$$L_{xy} = 117.4 \quad L_{xx} = 2288 \quad L_{yy} = 6.169$$

$$\text{Therefore, } r = \frac{117.4}{\sqrt{2288(6.169)}} = \frac{117.4}{118.81} = .988$$

Thus a very strong positive correlation exists between FEV and height. The sample regression coefficient  $b$  was calculated as 0.0513 in Example 11.14. Furthermore, the sample standard deviation of  $x$  and  $y$  can be computed as follows:

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{L_{xx}}{n-1}} = \sqrt{\frac{2288}{11}} = \sqrt{208} = 14.42$$

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{L_{yy}}{n-1}} = \sqrt{\frac{6.169}{11}} = \sqrt{0.561} = 0.749$$

and their ratio is thus given by

$$s_y / s_x = 0.749 / 14.42 = .0519$$

Finally,  $b$  can be expressed as a rescaled version of  $r$  as

$$b = r(s_y / s_x) \quad \text{or} \quad .0513 = .988(.0519)$$

Notice that if height is re-expressed in inches rather than centimeters (1 in. = 2.54 cm), then  $s_x$  is divided by 2.54, and  $b$  is multiplied by 2.54; that is,

$$b_{\text{in.}} = b_{\text{cm}} \times 2.54 = .0513 \times 2.54 = .130$$

where  $b_{\text{in.}}$  is in the units of liters per inch and  $b_{\text{cm}}$  is in the units of liters per centimeter. However, the correlation coefficient remains the same at .988.

When should the regression coefficient be used, and when should the correlation coefficient be used? The regression coefficient is used when we specifically want to predict one variable from another. The correlation coefficient is used when we simply want to describe the linear relationship between two variables but do not want to make predictions. In cases in which it is not clear which of these two aims is primary, both a regression and a correlation coefficient can be reported.

### Example 11.26

**Obstetrics, Pulmonary Disease, Cardiovascular Disease** For the birthweight–estriol data in Example 11.1, the obstetrician is interested in using a regression equation to predict birthweight from estriol levels. Thus the regression coefficient is more appropriate. Similarly, for the FEV–height data in Example 11.14, the pediatrician is interested in using a growth curve relating a child’s pulmonary function to height, and again the regression coefficient is more appropriate. However, in collecting data on cholesterol levels in spouse pairs in Example 11.22, the geneticist is interested simply in describing the relationship between cholesterol levels of spouse pairs and is not interested in prediction. Thus the correlation coefficient is more appropriate here.

In this section, we have introduced the concept of a correlation coefficient. In the next section, we discuss various hypothesis tests concerning correlation coefficients. Correlation coefficients are used when we are interested in studying the association between two variables but are not interested in predicting one variable from another. On the flowchart at the end of this chapter (Figure 11.32, p. 503), we answer yes to (1) interested in relationships between two variables? and (2) both variables continuous? no to (3) interested in predicting one variable from another? yes to (4) interested in studying the correlation between two variables? and yes to (5) both variables normal? This leads us to the box “Pearson correlation methods.”

### REVIEW QUESTIONS 11B

- 1 What is the difference between a regression coefficient and a correlation coefficient?
- 2 Refer to the data in Table 2.11.
  - (a) Compute the correlation coefficient between white-blood count following admission and duration of hospital stay.
  - (b) Discuss what this correlation coefficient means.

## 11.8 Statistical Inference for Correlation Coefficients

In the previous section, we defined the sample correlation coefficient. Based on Equation 11.17, if every unit in the reference population could be sampled, then the sample correlation coefficient ( $r$ ) would be the same as the population correlation coefficient, denoted by  $\rho$ , which was introduced in Definition 5.13.

In this section, we will use  $r$ , which is computed from finite samples, to test various hypotheses concerning  $\rho$ .

### One-Sample $t$ Test for a Correlation Coefficient

### Example 11.27

**Cardiovascular Disease** Suppose serum-cholesterol levels in spouse pairs are measured to determine whether there is a correlation between cholesterol levels in spouses. Specifically, we wish to test the hypothesis  $H_0: \rho = 0$  vs.  $H_1: \rho \neq 0$ . Suppose that  $r = .25$  based on 100 spouse pairs. Is this evidence enough to warrant rejecting  $H_0$ ?

In this instance, the hypothesis test would naturally be based on the sample correlation coefficient  $r$  and  $H_0$  would be rejected if  $|r|$  is sufficiently far from 0. Assuming that each of the random variables  $x$  = serum-cholesterol level for the husband and  $y$  = serum-cholesterol level for the wife is normally distributed, then the best procedure for testing the hypothesis is given as follows:

**Equation 11.20****One-Sample  $t$  Test for a Correlation Coefficient**

To test the hypothesis  $H_0: \rho = 0$  vs.  $H_1: \rho \neq 0$ , use the following procedure:

- (1) Compute the sample correlation coefficient  $r$ .
- (2) Compute the test statistic

$$t = r(n - 2)^{1/2} / (1 - r^2)^{1/2}$$

which under  $H_0$  follows a  $t$  distribution with  $n - 2$  df.

- (3) For a two-sided level  $\alpha$  test,

$$\text{if } t > t_{n-2, 1-\alpha/2} \quad \text{or} \quad t < -t_{n-2, 1-\alpha/2}$$

then reject  $H_0$ .

$$\text{If } -t_{n-2, 1-\alpha/2} \leq t \leq t_{n-2, 1-\alpha/2}$$

then accept  $H_0$ .

- (4) The  $p$ -value is given by

$$p = 2 \times (\text{area to the left of } t \text{ under a } t_{n-2} \text{ distribution}) \quad \text{if } t < 0$$

$$p = 2 \times (\text{area to the right of } t \text{ under a } t_{n-2} \text{ distribution}) \quad \text{if } t \geq 0$$

- (5) We assume an underlying normal distribution for each of the random variables used to compute  $r$ .

The acceptance and rejection regions for this test are shown in Figure 11.13. Computation of the  $p$ -value is illustrated in Figure 11.14.

**Example 11.28**

Perform a test of significance for the data in Example 11.27.

**Solution**

We have  $n = 100$ ,  $r = .25$ . Thus in this case,

$$t = .25\sqrt{98}/\sqrt{1 - .25^2} = 2.475/.968 = 2.56$$

From Table 5 in the Appendix,

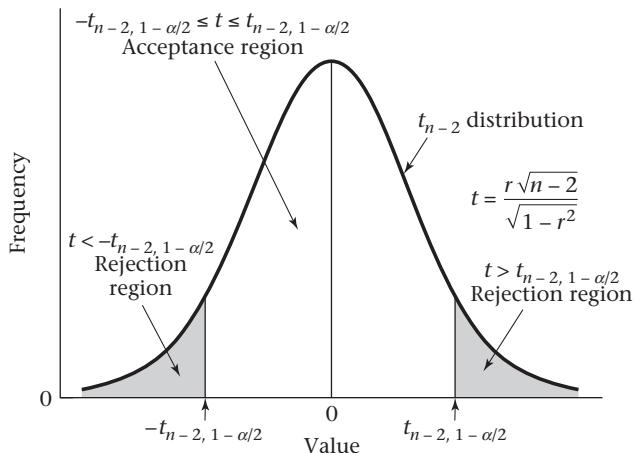
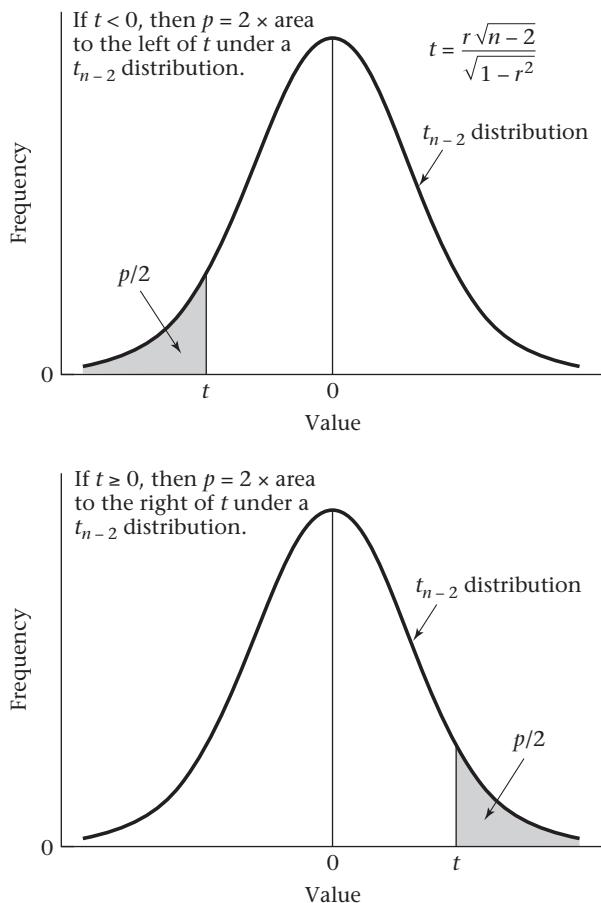
$$t_{60, .99} = 2.39 \quad t_{60, .995} = 2.66 \quad t_{120, .99} = 2.358 \quad t_{120, .995} = 2.617$$

Therefore, because  $60 < 98 < 120$ ,

$$.005 < p/2 < .01 \quad \text{or} \quad .01 < p < .02$$

and  $H_0$  is rejected. Alternatively, using Excel 2007, the exact  $p$ -value = TDIST(2.56, 98, 2) = .012. We conclude there is a significant aggregation of cholesterol levels between spouses. This result is possibly due to common environmental factors such as diet. But it could also be due to the tendency for people of similar body build to marry each other, and their cholesterol levels may have been correlated at the time of marriage.

Interestingly, the one-sample  $t$  test for correlation coefficients in Equation 11.20 is mathematically equivalent to the  $F$  test in Equation 11.7 and the  $t$  test in Equation 11.8 for simple linear regression, in that they always yield the same  $p$ -values. The question as to which test is more appropriate is best answered by whether a regression or a correlation coefficient is the parameter of primary interest.

**Figure 11.13 Acceptance and rejection regions for the one-sample  $t$  test for a correlation coefficient****Figure 11.14 Computation of the  $p$ -value for the one-sample  $t$  test for a correlation coefficient**

### One-Sample $z$ Test for a Correlation Coefficient

In the previous section, a test of the hypothesis  $H_0: \rho = 0$  vs.  $H_1: \rho \neq 0$  was considered. Sometimes the correlation between two random variables is expected to be some quantity  $\rho_0$  other than 0 and we want to test the hypothesis  $H_0: \rho = \rho_0$  vs.  $H_1: \rho \neq \rho_0$ .

**Example 11.29**

Suppose the body weights of 100 fathers ( $x$ ) and first-born sons ( $y$ ) are measured and a sample correlation coefficient  $r$  of .38 is found. We might ask whether or not this sample correlation is compatible with an underlying correlation of .5 that might be expected on genetic grounds. How can this hypothesis be tested?

In this case, we want to test the hypothesis  $H_0: \rho = .5$  vs.  $H_1: \rho \neq .5$ . The problem with using the  $t$  test formation in Equation 11.20 is that the sample correlation coefficient  $r$  has a skewed distribution for nonzero  $\rho$  that cannot be easily approximated by a normal distribution. Fisher considered this problem and proposed the following transformation to better approximate a normal distribution:

**Equation 11.21****Fisher's  $z$  Transformation of the Sample Correlation Coefficient  $r$** 

The  $z$  transformation of  $r$  given by

$$z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$$

is approximately normally distributed under  $H_0$  with mean

$$z_0 = \frac{1}{2} \ln[(1+\rho_0) / (1-\rho_0)]$$

and variance  $1/(n - 3)$ . The  $z$  transformation is very close to  $r$  for small values of  $r$  but tends to deviate substantially from  $r$  for larger values of  $r$ . A table of the  $z$  transformation is given in Table 13 in the Appendix.

**Example 11.30**

Compute the  $z$  transformation of  $r = .38$ .

**Solution**

The  $z$  transformation can be computed from Equation 11.21 as follows:

$$z = \frac{1}{2} \ln \left( \frac{1+0.38}{1-0.38} \right) = \frac{1}{2} \ln \left( \frac{1.38}{0.62} \right) = \frac{1}{2} \ln(2.226) = \frac{1}{2}(0.800) = 0.400$$

Alternatively, we could refer to Table 13 in the Appendix with  $r = .38$  to obtain  $z = 0.400$ .

Fisher's  $z$  transformation can be used to conduct the hypothesis test as follows: Under  $H_0$ ,  $Z$  is approximately normally distributed with mean  $z_0$  and variance  $1/(n - 3)$  or, equivalently,

$$\lambda = (Z - z_0)\sqrt{n-3} \sim N(0, 1)$$

$H_0$  will be rejected if  $z$  is far from  $z_0$ . Thus the following test procedure for a two-sided level  $\alpha$  test is used.

**Equation 11.22****One-Sample  $z$  Test for a Correlation Coefficient**

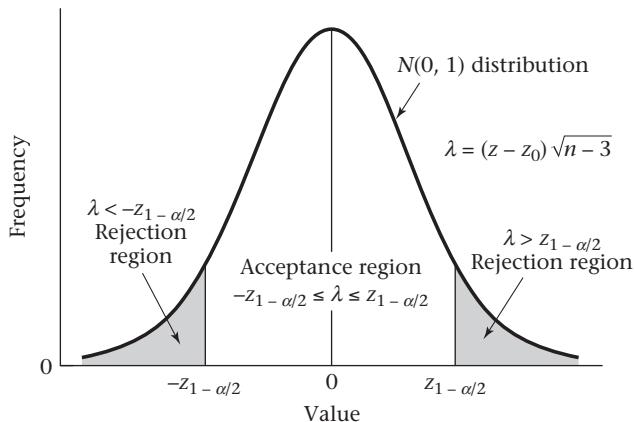
To test the hypothesis  $H_0: \rho = \rho_0$  vs.  $H_1: \rho \neq \rho_0$ , use the following procedure:

- (1) Compute the sample correlation coefficient  $r$  and the  $z$  transformation of  $r$ .
- (2) Compute the test statistic

$$\lambda = (z - z_0)\sqrt{n-3}$$

- (3) If  $\lambda > z_{1-\alpha/2}$  or  $\lambda < -z_{1-\alpha/2}$  reject  $H_0$ .
- If  $-z_{1-\alpha/2} \leq \lambda \leq z_{1-\alpha/2}$  accept  $H_0$ .

**Figure 11.15** Acceptance and rejection regions for the one-sample z test for a correlation coefficient



(4) The exact  $p$ -value is given by

$$\begin{aligned} p &= 2 \times \Phi(\lambda) && \text{if } \lambda \leq 0 \\ p &= 2 \times [1 - \Phi(\lambda)] && \text{if } \lambda > 0 \end{aligned}$$

(5) Assume an underlying normal distribution for each of the random variables used to compute  $r$  and  $z$ .

The acceptance and rejection regions for this test are shown in Figure 11.15. Computation of the  $p$ -value is illustrated in Figure 11.16.

### Example 11.31

Perform a test of significance for the data in Example 11.29.

#### Solution

In this case  $r = .38$ ,  $n = 100$ ,  $\rho_0 = .50$ . From Table 13 in the Appendix,

$$z_0 = \frac{1}{2} \ln\left(\frac{1+.5}{1-.5}\right) = .549 \quad z = \frac{1}{2} \ln\left(\frac{1+.38}{1-.38}\right) = .400$$

Hence

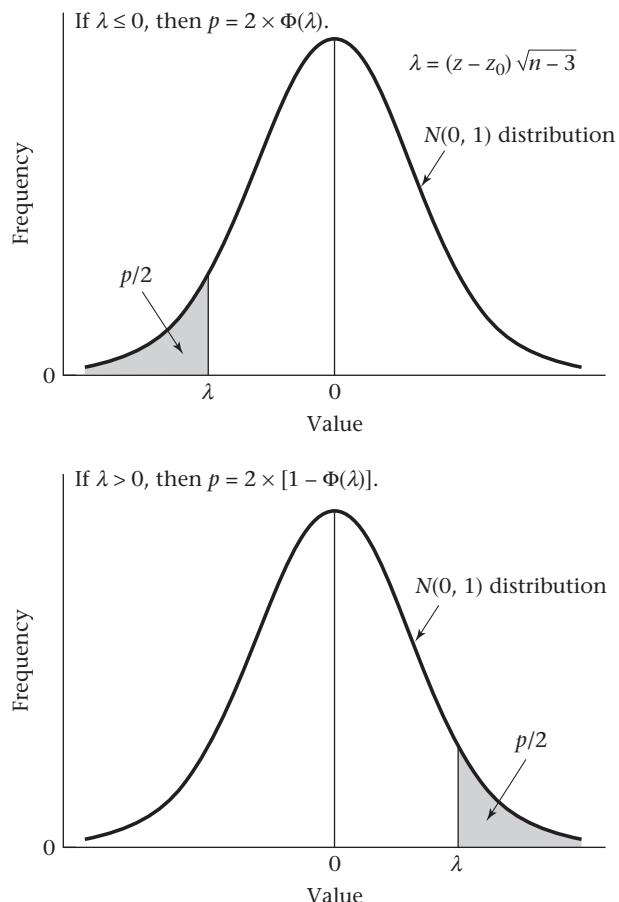
$$\lambda = (0.400 - 0.549)\sqrt{97} = (-0.149)(9.849) = -1.47 \sim N(0, 1)$$

Thus the  $p$ -value is given by

$$2 \times [1 - \Phi(-1.47)] = 2 \times (1 - .9292) = .142$$

Therefore, we accept  $H_0$  that the sample estimate of .38 is compatible with an underlying correlation of .50; this would be expected on purely genetic grounds.

To sum up, the  $z$  test in Equation 11.22 is used to test hypotheses about nonzero null correlations, whereas the  $t$  test in Equation 11.20 is used to test hypotheses about null correlations of zero. The  $z$  test can also be used to test correlations of zero under the null hypothesis, but the  $t$  test is slightly more powerful in this case and is preferred. However, if  $\rho_0 \neq 0$ , then the one-sample  $z$  test is very sensitive to

**Figure 11.16 Computation of the  $p$ -value for the one-sample  $z$  test for a correlation coefficient**

non-normality of either  $x$  or  $y$ . This is also true for the two-sample correlation test presented later in this section (see p. 464).

### Interval Estimation for Correlation Coefficients

In the previous two sections, we learned how to estimate a correlation coefficient  $\rho$  and how to perform appropriate hypothesis tests concerning  $\rho$ . It is also of interest to obtain confidence limits for  $\rho$ . An easy method for obtaining confidence limits for  $\rho$  can be derived based on the approximate normality of Fisher's  $z$  transformation of  $r$ . This method is given as follows.

#### Equation 11.23

##### Interval Estimation of a Correlation Coefficient ( $\rho$ )

Suppose we have a sample correlation coefficient  $r$  based on a sample of  $n$  pairs of observations. To obtain a two-sided  $100\% \times (1 - \alpha)$  confidence interval for the population correlation coefficient ( $\rho$ ):

- (1) Compute Fisher's  $z$  transformation of  $r = z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$ .
- (2) Let  $z_\rho$  = Fisher's  $z$  transformation of  $\rho = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right)$ .

A two-sided  $100\% \times (1 - \alpha)$  confidence interval is given for  $z_\rho = (z_1, z_2)$  where

$$z_1 = z - z_{1-\alpha/2} / \sqrt{n-3}$$

$$z_2 = z + z_{1-\alpha/2} / \sqrt{n-3}$$

and  $z_{1-\alpha/2} = 100\% \times (1 - \alpha/2)$  percentile of an  $N(0, 1)$  distribution

- (3) A two-sided  $100\% \times (1 - \alpha)$  confidence interval for  $\rho$  is then given by  $(\rho_1, \rho_2)$  where

$$\rho_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$$

$$\rho_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$$

The interval  $(z_1, z_2)$  in Equation 11.23 can be derived in a similar manner to the confidence interval for the mean of a normal distribution with known variance (see Equation 6.7), which is given by

**Equation 11.24** 
$$(z_1, z_2) = z \pm z_{1-\alpha/2} / \sqrt{n-3}$$

We then solve Equation 11.23 for  $r$  in terms of  $z$ , whereby

**Equation 11.25** 
$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

We now substitute the confidence limits for  $z_\rho$ —that is,  $(z_1, z_2)$  in Equation 11.24—into Equation 11.25 to obtain the corresponding confidence limits for  $\rho$  given by  $(\rho_1, \rho_2)$  in Equation 11.23.

### Example 11.32

In Example 11.29, a sample correlation coefficient of .38 was obtained between the body weights of 100 pairs of fathers ( $x$ ) and first-born sons ( $y$ ). Provide a 95% confidence interval for the underlying correlation coefficient  $\rho$ .

### Solution

From Example 11.31, the  $z$  transformation of  $r = 0.400$ . From step 2 of Equation 11.23, a 95% confidence interval for  $z_\rho$  is given by  $(z_1, z_2)$ , where

$$z_1 = 0.400 - 1.96 / \sqrt{97} = 0.400 - 0.199 = 0.201$$

$$z_2 = 0.400 + 1.96 / \sqrt{97} = 0.400 + 0.199 = 0.599$$

From step 3 of Equation 11.23, a 95% confidence interval for  $\rho$  is given by  $(\rho_1, \rho_2)$  where

$$\begin{aligned} \rho_1 &= \frac{e^{2(0.201)} - 1}{e^{2(0.201)} + 1} \\ &= \frac{e^{402} - 1}{e^{402} + 1} \\ &= \frac{1.4950 - 1}{1.4950 + 1} \\ &= \frac{0.4950}{2.4950} = .198 \end{aligned}$$

$$\begin{aligned}\rho_2 &= \frac{e^{2(0.599)} - 1}{e^{2(0.599)} + 1} \\ &= \frac{e^{1.198} - 1}{e^{1.198} + 1} \\ &= \frac{2.3139}{4.3139} = .536\end{aligned}$$

Thus a 95% confidence interval for  $\rho = (.198, .536)$ .

Notice that the confidence interval for  $z_\rho$ , given by  $(z_1, z_2) = (0.201, 0.599)$ , is symmetric about  $z = 0.400$ . However, when the confidence limits are transformed back to the original scale (the scale of  $\rho$ ) the corresponding confidence limits for  $\rho$  are given by  $(\rho_1, \rho_2) = (.198, .536)$ , which are not symmetric around  $r = .380$ . The reason for this is that Fisher's  $z$  transformation is a nonlinear function of  $r$ , which only becomes approximately linear when  $r$  is small (i.e.,  $|r| \leq .2$ ).

## Sample-Size Estimation for Correlation Coefficients

### Example 11.33

**Nutrition** Suppose a new dietary questionnaire is constructed to be administered over the Internet, based on dietary recall over the past 24 hours. To validate this questionnaire, participants are given 3 days' worth of food diaries, in which they fill out in real time exactly what they eat for 3 days, spaced about 1 month apart. The average intake over 3 days will be considered a gold standard. The correlation between the 24-hour recall and the gold standard will be an index of validity. How large a sample is needed to have 80% power to detect a significant correlation between these measures, if it is expected that the true correlation is .5 and a one-sided test is used with  $\alpha = .05$ ?

To address this question, we use the Fisher's  $z$ -transform approach. Specifically, we want to test the hypothesis  $H_0: \rho = 0$  vs.  $H_1: \rho = \rho_0 > 0$ . Under  $H_0$ ,

$$z \sim N[0, 1 / (n - 3)]$$

We will reject  $H_0$  at level  $\alpha$  if  $z\sqrt{n-3} > z_{1-\alpha}$ . Suppose  $z_0$  is the Fisher's  $z$  transform of  $\rho_0$ . If we subtract  $z_0\sqrt{n-3}$  from both sides of the equation, it follows that

$$\lambda = \sqrt{n-3}(z - z_0) > z_{1-\alpha} - z_0\sqrt{n-3}$$

Furthermore, under  $H_1$ ,  $\lambda \sim N(0, 1)$ . Therefore,

$$\begin{aligned}Pr(\lambda > z_{1-\alpha} - z_0\sqrt{n-3}) &= 1 - \Phi(z_{1-\alpha} - z_0\sqrt{n-3}) \\ &= \Phi(z_0\sqrt{n-3} - z_{1-\alpha})\end{aligned}$$

If we require a power of  $1 - \beta$ , then we set the right-hand side to  $1 - \beta$  or, equivalently,

$$z_0\sqrt{n-3} - z_{1-\alpha} = z_{1-\beta}$$

If follows that

$$\text{Power} = 1 - \beta = \Phi(z_0\sqrt{n-3} - z_{1-\alpha})$$

The corresponding sample-size estimate is obtained by solving for  $n$ , whereby

$$n = \left[ \left( z_{1-\alpha} + z_{1-\beta} \right)^2 / z_0^2 \right] + 3$$

The procedure is summarized as follows.

**Equation 11.26****Power and Sample-Size Estimation for Correlation Coefficients**

Suppose we wish to test the hypothesis  $H_0: \rho = 0$  vs.  $H_1: \rho = \rho_0 > 0$ . For the specific alternative  $\rho = \rho_0$ , to test the hypothesis with a one-sided significance level of  $\alpha$  and specified sample size  $n$ , the power is given by

$$\text{Power} = \Phi(z_0 \sqrt{n-3} - z_{1-\alpha})$$

For the specific alternative  $\rho = \rho_0$ , to test the hypothesis with a one-sided significance level of  $\alpha$  and specified power of  $1 - \beta$ , we require a sample size of

$$n = \left[ (z_{1-\alpha} + z_{1-\beta})^2 / z_0^2 \right] + 3$$

**Solution to Example 11.33**

In this case, we have  $\rho_0 = .5$ . Therefore, from Table 13 in the Appendix,  $z_0 = .549$ . Also,  $\alpha = .05$ ,  $1 - \beta = .80$ . Thus, from Equation 11.26, we have

$$\begin{aligned} n &= \left[ (z_{.95} + z_{.80})^2 / .549^2 \right] + 3 \\ &= \left[ (1.645 + 0.84)^2 / .549^2 \right] + 3 \\ &= 23.5 \end{aligned}$$

Therefore, to have 80% power, we need 24 participants in the validation study.

**Example 11.34**

**Nutrition** Suppose that 50 participants are actually enrolled in the validation study. What power will the study have if the true correlation is .5 and a one-sided test is used with  $\alpha = .05$ ?

**Solution**

We have  $\alpha = .05$ ,  $\rho_0 = .50$ ,  $z_0 = .549$ ,  $n = 50$ . Thus, from Equation 11.26,

$$\begin{aligned} \text{Power} &= \Phi(.549 \sqrt{47} - z_{.95}) \\ &= \Phi(3.764 - 1.645) \\ &= \Phi(2.12) = .983 \end{aligned}$$

Therefore, the study will have 98.3% power.

**Two-Sample Test for Correlations**

The use of Fisher's  $z$  transformation can be extended to two-sample problems.

**Example 11.35**

**Hypertension** Suppose there are two groups of children. Children in one group live with their natural parents, whereas children in the other group live with adoptive parents. One question that arises is whether or not the correlation between the blood pressure of a mother and a child is different in these two groups. A different correlation would suggest a genetic effect on blood pressure. Suppose there are 1000 mother-child pairs in the first group, with correlation .35, and 100 mother-child pairs in the second group, with correlation .06. How can this question be answered?

We want to test the hypothesis  $H_0: \rho_1 = \rho_2$  vs.  $H_1: \rho_1 \neq \rho_2$ . It is reasonable to base the test on the difference between the  $z$ 's in the two samples. If  $|z_1 - z_2|$  is large, then  $H_0$  will be rejected; otherwise,  $H_0$  will be accepted. This principle suggests the following test procedure for a two-sided level  $\alpha$  test.

**Equation 11.27****Fisher's z Test for Comparing Two Correlation Coefficients**

To test the hypothesis  $H_0: \rho_1 = \rho_2$  vs.  $H_1: \rho_1 \neq \rho_2$ , use the following procedure:

- (1) Compute the sample correlation coefficients ( $r_1, r_2$ ) and Fisher's  $z$  transformation ( $z_1, z_2$ ) for each of the two samples.
- (2) Compute the test statistic

$$\lambda = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0, 1) \text{ under } H_0$$

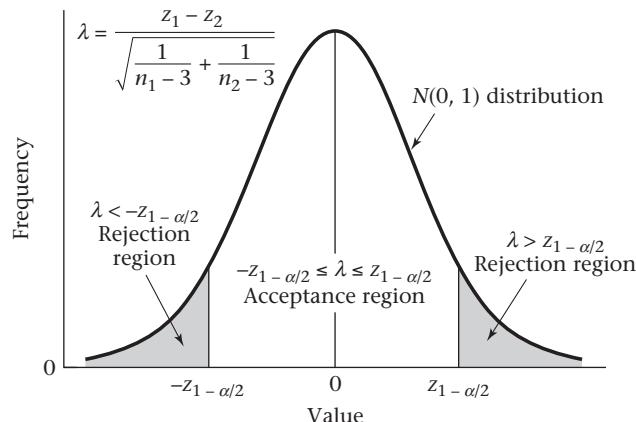
- (3) If  $\lambda > z_{1-\alpha/2}$  or  $\lambda < -z_{1-\alpha/2}$  reject  $H_0$ .  
If  $-z_{1-\alpha/2} \leq \lambda \leq z_{1-\alpha/2}$  accept  $H_0$ .
- (4) The exact  $p$ -value is given by

$$\begin{aligned} p &= 2\Phi(\lambda) && \text{if } \lambda \leq 0 \\ p &= 2 \times [1 - \Phi(\lambda)] && \text{if } \lambda > 0 \end{aligned}$$

- (5) Assume an underlying normal distribution for each of the random variables used to compute  $r_1, r_2$  and  $z_1, z_2$ .

The acceptance and rejection regions for this test are shown in Figure 11.17. Computation of the  $p$ -value is illustrated in Figure 11.18.

**Figure 11.17 Acceptance and rejection regions for Fisher's  $z$  test for comparing two correlation coefficients**

**Example 11.36**

Perform a significance test for the data in Example 11.35.

**Solution**

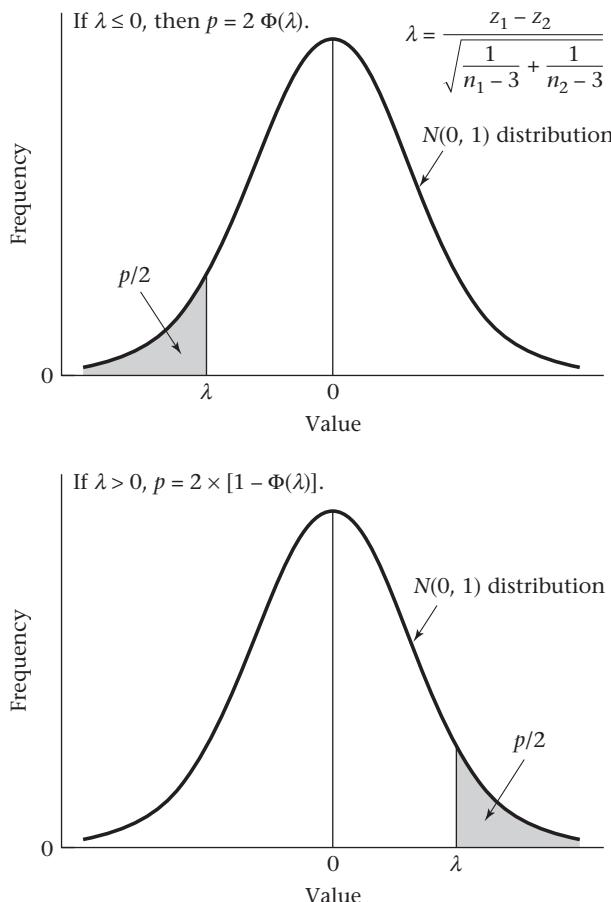
$$r_1 = .35 \quad n_1 = 1000 \quad r_2 = .06 \quad n_2 = 100$$

Thus, from Table 13 in the Appendix,

$$z_1 = 0.365 \quad z_2 = 0.060$$

$$\text{and} \quad \lambda = \frac{0.365 - 0.060}{\sqrt{\frac{1}{997} + \frac{1}{97}}} = 9.402(0.305) = 2.87 \sim N(0, 1) \text{ under } H_0$$

**Figure 11.18 Computation of the  $p$ -value for Fisher's  $z$  test for comparing two correlation coefficients**



Hence the  $p$ -value is given by

$$2 \times [1 - \Phi(2.87)] = .004$$

Therefore, there is a significant difference between the mother-child correlations in the two groups, implying a significant genetic effect on blood pressure.

### Comparison of Dependent Correlation Coefficients

The methods in Equation 11.27 pertain to the comparison of correlation coefficients obtained from two independent samples. In some cases, we are interested in comparing two correlation coefficients obtained from the same subjects.

#### Example 11.37

**Hypertension** In Data Set INFANTBP.DAT on the Companion Website, we have collected information on response to the taste of salt and sugar, respectively, as well as the blood pressure of the infants. The primary focus of the study was to assess whether the magnitude of the salt-taste response was related to blood pressure. The magnitude of the sugar-taste response was used as a control. To be more specific, we let

$X$  = mean sucks per burst (MSB) on exposure to high doses (15%) of salt (mean of trials 7 and 8 in the salt-taste component of the study) minus MSB on response to water (mean of trials 5 and 6 in the salt-taste component of the study)

$Y$  = MSB on exposure to 15% sucrose (trial 4 in the sugar-taste component of the study) minus MSB on exposure to water (trial 2 in the sugar-taste component of the study)

$Z$  = diastolic blood pressure (DBP)

The main goal of the study was to test the salt-taste hypothesis: that the correlation between  $X$  and  $Z$  was 0, or  $H_0: \rho_{XZ} = 0$  vs.  $H_1: \rho_{XZ} \neq 0$ . We could also test the sugar-taste hypothesis:  $H_0: \rho_{YZ} = 0$  vs.  $H_1: \rho_{YZ} \neq 0$ . Each of these hypotheses can be tested using the one-sample  $t$  test for correlation in Equation 11.20.

However, if we specifically want to assess whether the effect of exposure to salt on DBP differs from the effect of exposure to sugar on DBP, then we might want to test the hypothesis  $H_0: \rho_{XZ} = \rho_{YZ}$  vs.  $H_1: \rho_{XZ} \neq \rho_{YZ}$ . Unfortunately, we cannot use the two-sample correlation test in Equation 11.27 because this test assumes these two correlations are obtained from two independent samples, whereas the estimates of  $\rho_{XZ}$  and  $\rho_{YZ}$  are from the same subjects.

To address this issue, we use the method of Wolfe [6]. Specifically, we assume  $\sigma_X = \sigma_Y$  and use the following procedure.

### Equation 11.28

#### Wolfe's Test for Comparing Dependent Correlation Coefficients

Suppose we want to test the hypothesis  $H_0: \rho_{XZ} = \rho_{YZ}$  vs.  $H_1: \rho_{XZ} \neq \rho_{YZ}$ , where  $X$ ,  $Y$ , and  $Z$  are obtained from the same subjects. We assume  $\sigma_X = \sigma_Y$ . Under this assumption, these hypotheses are equivalent to the hypothesis:  $H_0: \rho_{X-Y,Z} = 0$  vs.  $H_1: \rho_{X-Y,Z} \neq 0$ . Hence,

- (1) We use the one-sample  $t$  test for correlation in Equation 11.20 based on the following test statistic:

$$t = r \sqrt{n-2} / \sqrt{1-r^2} \sim t_{n-2} \text{ under } H_0$$

where  $r = \text{Corr}(X_i - Y_i, Z_i)$ .

- (2) We reject  $H_0$  if  $t > t_{n-2, 1-\alpha/2}$  or if  $t < t_{n-2, \alpha/2}$ .

We accept  $H_0$  if  $t_{n-2, \alpha/2} \leq t \leq t_{n-2, 1-\alpha/2}$ .

- (3) The  $p$ -value is given by

$$\begin{aligned} 2 \times Pr(t_{n-2} > t) &\quad \text{if } t \geq 0, \\ 2 \times Pr(t_{n-2} < t) &\quad \text{if } t < 0. \end{aligned}$$

To see why this formulation works, we write

$$\rho_{XZ} = \text{Cov}(X, Z) / (\sigma_X \sigma_Z)$$

$$\rho_{YZ} = \text{Cov}(Y, Z) / (\sigma_Y \sigma_Z).$$

If  $\sigma_X = \sigma_Y$ , then

$$\begin{aligned} \rho_{XZ} - \rho_{YZ} &= \frac{\text{Cov}(X, Z) - \text{Cov}(Y, Z)}{\sigma_X \sigma_Z} = \frac{\text{Cov}(X - Y, Z)}{\sigma_X \sigma_Z} \\ &= \frac{\text{Cov}(X - Y, Z)}{\sigma_{X-Y} \sigma_Z} (\sigma_{X-Y} / \sigma_X) = \rho_{X-Y,Z} (\sigma_{X-Y} / \sigma_X) \end{aligned}$$

However, because  $\sigma_{X-Y}/\sigma_X \neq 0$ , it follows that  $\rho_{XZ} - \rho_{YZ} = 0$  if and only if  $\rho_{X-Y,Z} = 0$ . Thus, as with the paired *t*-test, we have reduced a two-sample problem with dependent correlation coefficients to an equivalent one-sample problem based on  $\text{Corr}(X - Y, Z)$ .

### Solution to Example 11.37

We compute  $X$  and  $Y$  from Data Set INFANTBP ( $n = 100$ ). There are 96 subjects who had nonmissing values for  $X$ ,  $Y$ , and  $Z$ . The mean and  $sd$  for each of these variables is given in Table 11.8.

We wish to test the hypothesis  $H_0: \rho_{XZ} = \rho_{YZ}$  vs.  $H_1: \rho_{XZ} \neq \rho_{YZ}$ . Because  $X$  and  $Y$  have different standard deviations we compute standardized scores for each of the variables, where

$$\text{Score}(X_i) = [X_i - \text{mean}(X_i)] / \text{sd}(X_i)$$

$$\text{Score}(Y_i) = [Y_i - \text{mean}(Y_i)] / \text{sd}(Y_i)$$

**Table 11.8** Salt and sugar taste data

	$X$ (salt taste)	$Y$ (sugar taste)	$Z$ (DBP)
Mean	-2.77	6.53	42.62
$sd$	6.85	26.44	7.31
$n$	96	96	96

By the definition of  $\text{Score}(X_i)$  and  $\text{Score}(Y_i)$ , each score variable has mean = 0 and  $sd = 1$ . Also, the linear transformation from  $X$  to  $\text{Score}(X)$  and  $Y$  to  $\text{Score}(Y)$  will not affect the correlation of each variable, respectively, with  $Z$ ; that is,  $\text{Corr}(X, Z) = \text{Corr}[\text{Score}(X), Z]$  and  $\text{Corr}(Y, Z) = \text{Corr}[\text{Score}(Y), Z]$ . Thus we can restate the hypotheses in the form

$$H_0: \text{Corr}[\text{Score}(X), Z] = \text{Corr}[\text{Score}(Y), Z]$$

$$\text{vs. } H_1: \text{Corr}[\text{Score}(X), Z] \neq \text{Corr}[\text{Score}(Y), Z]$$

We compute  $\text{Corr}[\text{Score}(X), Z] = .299$ ,  $\text{Corr}[\text{Score}(Y), Z] = .224$ . Thus, to test whether each correlation coefficient is significantly different from 0, we use the one-sample *t*-test for correlation, where

$$t_X = 0.299 \sqrt{94} / \sqrt{1 - .299^2} = 3.066 \sim t_{94}, p = .003$$

$$\text{and } t_Y = 0.244 \sqrt{94} / \sqrt{1 - .244^2} = 2.248 \sim t_{94}, p = .027$$

Hence, both the salt-taste and sugar-taste variables are significantly correlated with DBP. However, to determine if these correlation coefficients differ significantly from each other, we compute:  $\text{Corr}[\text{Score}(X) - \text{Score}(Y), Z] = .052$ . From Equation 11.28, the test statistic is

$$t = .052 \sqrt{94} / \sqrt{1 - .052^2} = 0.51 \sim t_{94}$$

The  $p$ -value =  $2 \times Pr(t_{94} > 0.51) = .61$ . Thus there is no significant difference between these two correlations. Therefore, it is possible that there is some other factor (such as obesity) related to both the salt-taste and sugar-taste scores causing the positive associations of each score with DBP.

### REVIEW QUESTIONS 11C

- What is the difference between the one-sample *t* test for correlation coefficients and the one-sample *z* test for correlation coefficients?
- Refer to the data in Table 2.11.

- (a) What test can be used to assess whether there is a significant association between the first white-blood count following admission and the duration of hospital stay?
- (b) Implement the test in Review Question 11C.2a, and report a two-tailed  $p$ -value.
- (c) Provide a 95% confidence interval for the correlation coefficient in Review Question 11C.2a.
- 3 Refer to the data in Table 2.11.
- (a) Suppose we want to compare the correlation coefficient between duration of hospital stay and white-blood count for males vs. females. What test can we use to accomplish this?
- (b) Perform the test in Review Question 11C.3a, and report a two-tailed  $p$ -value.
- 4 What is the difference between the two-sample correlation tests in Equations 11.27 and 11.28? When do we use each?

## 11.9 Multiple Regression

In Sections 11.2 through 11.6 problems in linear-regression analysis in which there is one independent variable ( $x$ ), one dependent variable ( $y$ ), and a linear relationship between  $x$  and  $y$  were discussed. In practice, there is often more than one independent variable and we would like to look at the relationship between each of the independent variables ( $x_1, \dots, x_k$ ) and the dependent variable ( $y$ ) after taking into account the remaining independent variables. This type of problem is the subject matter of **multiple-regression analysis**.

### Example 11.38

**Hypertension, Pediatrics** A topic of interest in hypertension research is how the relationship between the blood-pressure levels of newborns and infants relate to subsequent adult blood pressure. One problem that arises is that the blood pressure of a newborn is affected by several extraneous factors that make this relationship difficult to study. In particular, newborn blood pressures are affected by (1) birthweight and (2) the day of life on which blood pressure is measured. In this study, the infants were weighed at the time of the blood-pressure measurements. We refer to this weight as the “birthweight,” although it differs somewhat from their actual weight at birth. Because the infants grow in the first few days of life, we would expect that infants seen at 5 days of life would on average have a greater weight than those seen at 2 days of life. We would like to be able to adjust the observed blood pressure for these two factors before we look at other factors that may influence newborn blood pressure.

### Estimation of the Regression Equation

Suppose a relationship is postulated between systolic blood pressure (SBP) ( $y$ ), birth-weight ( $x_1$ ), and age in days ( $x_2$ ), of the form

#### Equation 11.29

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where  $\epsilon$  is an error term that is normally distributed with mean 0 and variance  $\sigma^2$ . We would like to estimate the parameters of this model and test various hypotheses concerning it. The same method of least squares that was introduced in Section 11.3

for simple linear regression will be used to fit the parameters of this multiple-regression model. In particular,  $\alpha$ ,  $\beta_1$ ,  $\beta_2$  will be estimated by  $a$ ,  $b_1$  and  $b_2$ , respectively, where we choose  $a$ ,  $b_1$  and  $b_2$  to minimize the sum of

$$[y - (a + b_1x_1 + b_2x_2)]^2$$

over all the data points.

In general, if we have  $k$  independent variables  $x_1, \dots, x_k$ , then a linear-regression model relating  $y$  to  $x_1, \dots, x_k$  is of the form

**Equation 11.30**

$$y = \alpha + \sum_{j=1}^k \beta_j x_j + e$$

where  $e$  is an error term that is normally distributed with mean 0 and variance  $\sigma^2$ .

We estimate  $\alpha, \beta_1, \dots, \beta_k$  by  $a, b_1, \dots, b_k$  using the method of least squares, where we minimize the sum of

$$\left[ y - \left( a + \sum_{j=1}^k b_j x_j \right) \right]^2$$

**Example 11.39**

**Hypertension, Pediatrics** Suppose SBP, birthweight (oz), and age (days) are measured for 16 infants and the data are as shown in Table 11.9. Estimate the parameters of the multiple-regression model in Equation 11.29.

**Table 11.9**

**Sample data for infant blood pressure, age, and birthweight for 16 infants**

$i$	Birthweight (oz) ( $x_1$ )	Age (days) ( $x_2$ )	SBP (mm Hg) ( $y$ )
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	92
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88

**Solution**

Use the SAS PROC REG program to obtain the least-squares estimates. The results are given in Table 11.10.

According to the parameter-estimate column, the regression equation is given by

$$y = 53.45 + 0.126x_1 + 5.89x_2$$

**Table 11.10 Least-squares estimates of the regression parameters for the newborn blood-pressure data in Table 11.9 using the SAS PROC REG program**

The REG Procedure					
<b>Model:</b> MODEL1					
<b>Dependent Variable:</b> sysbp					
<b>Number of Observations Read</b> 16					
<b>Number of Observations Used</b> 16					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	591.03564	295.51782	48.08	<.0001
Error	13	79.90186	6.14630		
Corrected Total	15	670.93750			
Root MSE		2.47917	R-Square	0.8809	
Dependent Mean		88.06250	Adj R-Sq	0.8626	
Coeff Var		2.81524			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t  Standardized Estimate
Intercept	1	53.45019	4.53189	11.79	<.0001 0
brthwgt	1	0.12558	0.03434	3.66	0.0029 0.35208
agedys	1	5.88772	0.68021	8.66	<.0001 0.83323
Squared Partial					
Corr Type II					

The regression equation tells us that for a newborn the average blood pressure increases by an estimated 0.126 mm Hg per ounce of birthweight and 5.89 mm Hg per day of age.

**Example 11.40** **Hypertension, Pediatrics** Calculate the predicted average SBP of a baby with birth-weight 8 lb (128 oz) measured at 3 days of life.

**Solution** The average SBP is estimated by

$$53.45 + 0.126(128) + 5.89(3) = 87.2 \text{ mm Hg}$$

The regression coefficients in Table 11.10 are called *partial-regression coefficients*.

**Definition 11.16** Suppose we consider the multiple-regression model

$$y = \alpha + \sum_{j=1}^k \beta_j x_j + e$$

where  $e$  follows a normal distribution with mean 0 and variance  $\sigma^2$ . The  $\beta_j$ ,  $j = 1, 2, \dots, k$  are referred to as **partial-regression coefficients**.  $\beta_j$  represents the average increase in  $y$  per unit increase in  $x_j$ , with all other variables held constant (or stated another way, after adjusting for all other variables in the model), and is estimated by the parameter  $b_j$ .

Partial-regression coefficients differ from simple linear-regression coefficients as given in Equation 11.2. The latter represent the average increase in  $y$  per unit increase in  $x$ , without considering any other independent variables. If there are strong relationships among the independent variables in a multiple-regression model, then the partial-regression coefficients may differ considerably from the simple linear-regression coefficients obtained from considering each independent variable separately.

**Example 11.41****Solution**

**Hypertension** Interpret the regression coefficients in Table 11.10.

The partial-regression coefficient for birthweight =  $b_1 = 0.126$  mm Hg/oz represents the estimated average increase in SBP per 1 oz increase in birthweight for infants of the same age. The regression coefficient for age =  $b_2 = 5.89$  mm Hg/day represents the estimated average increase in SBP per 1-day increase in age for infants of the same birthweight.

We are often interested in ranking the independent variables according to their predictive relationship with the dependent variable  $y$ . It is difficult to rank the variables based on the magnitude of the partial-regression coefficients because the independent variables are often in different units. Specifically, from the multiple-regression model in Equation 11.30, we see that  $b$  estimates the increase in  $y$  per unit increase in  $x$ , while holding the values of all other variables in the model constant. If  $x$  is increased by 1 standard deviation unit ( $s_x$ ) to  $x + s_x$ , then  $y$  would be expected to increase by  $b \times s_x$  raw units or  $(b \times s_x)/s_y$  standard deviation units of  $y$  ( $s_y$ ).

**Definition 11.17**

**The standardized regression coefficient ( $b_s$ )** is given by  $b \times (s_x/s_y)$ . It represents the estimated average increase in  $y$  (expressed in standard deviation units of  $y$ ) per standard deviation increase in  $x$ , after adjusting for all other variables in the model.

Thus the standardized regression coefficient is a useful measure for comparing the predictive value of several independent variables because it tells us the predicted increase in standard-deviation units of  $y$  per standard-deviation increase in  $x$ . By expressing change in standard-deviation units of  $x$ , we can control for differences in the units of measurement for different independent variables.

**Example 11.42**

Compute the standardized regression coefficients for birthweight and age in days using the data in Tables 11.9 and 11.10.

**Solution**

From Table 11.9,  $s_y = 6.69$ ,  $s_{x_1} = 18.75$ ,  $s_{x_2} = 0.946$ . Therefore, referring to the standardized-estimate column in Table 11.10,

$$b_s(\text{birthweight}) = \frac{0.1256 \times 18.75}{6.69} = 0.352$$

$$b_s(\text{age in days}) = \frac{5.888 \times 0.946}{6.69} = 0.833$$

Thus the average increase in SBP is 0.352 standard-deviation units of blood pressure per standard-deviation increase in birthweight, holding age constant, and 0.833 standard-deviation units of blood pressure per standard-deviation increase in age, holding birthweight constant. Thus age appears to be the more important variable after controlling for both variables simultaneously in the multiple-regression model.

## Hypothesis Testing

**Example 11.43**

**Hypertension, Pediatrics** We would like to test various hypotheses concerning the data in Table 11.9. First, we would like to test the overall hypothesis that birthweight and age when considered together are significant predictors of blood pressure. How can this be done?

Specifically, we will test the hypothesis  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  vs.  $H_1$ : at least one of  $\beta_1, \dots, \beta_k \neq 0$ . The test of significance is similar to the  $F$  test in Section 11.4. The test procedure for a level  $\alpha$  test is given as follows.

**Equation 11.31**

**$F$  Test for Testing the Hypothesis  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  vs.**

**$H_1$ : At Least One of the  $\beta_j \neq 0$  in Multiple Linear Regression**

- (1) Estimate the regression parameters using the method of least squares, and compute Reg SS and Res SS,

$$\text{where } \text{Res SS} = \sum_{i=1}^n (\gamma_i - \hat{\gamma}_i)^2$$

$$\text{Reg SS} = \text{Total SS} - \text{Res SS}$$

$$\text{Total SS} = \sum_{i=1}^n (\gamma_i - \bar{\gamma})^2$$

$$\hat{\gamma}_i = a + \sum_{j=1}^k b_j x_{ij}$$

$x_{ij}$  =  $j$ th independent variable for the  $i$ th subject,  $j = 1, \dots, k$ ;  $i = 1, \dots, n$

- (2) Compute Reg MS = Reg SS/ $k$ , Res MS = Res SS/( $n - k - 1$ ).

- (3) Compute the test statistic

$$F = \text{Reg MS}/\text{Res MS}$$

which follows an  $F_{k,n-k-1}$  distribution under  $H_0$ .

- (4) For a level  $\alpha$  test,

if  $F > F_{k,n-k-1,1-\alpha}$  then reject  $H_0$

if  $F \leq F_{k,n-k-1,1-\alpha}$  then accept  $H_0$

- (5) The exact  $p$ -value is given by the area to the right of  $F$  under an  $F_{k,n-k-1}$  distribution =  $Pr(F_{k,n-k-1} > F)$ .

The acceptance and rejection regions for this test procedure are shown in Figure 11.19. Computation of the exact  $p$ -value is illustrated in Figure 11.20.

**Example 11.44**

**Hypertension, Pediatrics** Test the hypothesis  $H_0: \beta_1 = \beta_2 = 0$  vs.  $H_1$ : either  $\beta_1 \neq 0$  or  $\beta_2 \neq 0$  using the data in Tables 11.9 and 11.10.

**Solution**

Refer to Table 11.10 and note that

$$\text{Reg SS} = 591.04 \text{ (called Model SS)}$$

$$\text{Reg MS} = 591.04/2 = 295.52 \text{ (called Model MS)}$$

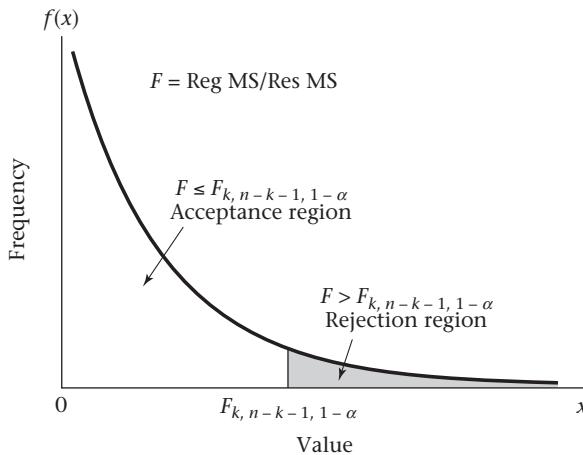
$$\text{Res SS} = 79.90 \text{ (called Error SS)}$$

$$\text{Res MS} = 79.90/13 = 6.146 \text{ (called Error MS)}$$

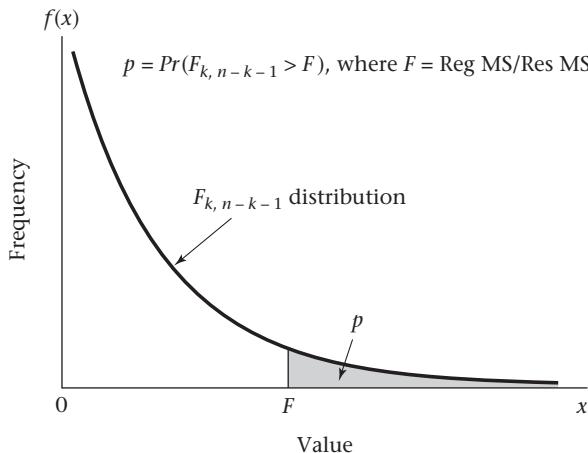
$$F = \text{Reg MS}/\text{Res MS} = 48.08 \sim F_{2,13} \text{ under } H_0$$

Because  $F_{2,13,999} < F_{2,12,999} = 12.97 < 48.08 = F$  it follows that

**Figure 11.19** Acceptance and rejection regions for testing the hypothesis  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  vs.  $H_1: \text{at least one of the } \beta_j \neq 0$  in multiple linear regression



**Figure 11.20** Computation of the  $p$ -value for testing the hypothesis  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  vs.  $H_1: \text{at least one of the } \beta_j \neq 0$  in multiple linear regression



$p < .001$ . Thus we can conclude that the two variables, when considered together, are significant predictors of blood pressure.

The significant  $p$ -value for this test could be attributed to either variable. We would like to perform significance tests to identify the independent contributions of each variable. How can this be done?

In particular, to assess the independent contribution of birthweight, we assume age is making a contribution under either hypothesis, and we test the hypothesis  $H_0: \beta_1 = 0, \beta_2 \neq 0$  vs.  $H_1: \beta_1 \neq 0, \beta_2 \neq 0$ . Similarly, to assess the independent contribution of age, we assume birthweight is making a contribution under either hypothesis and test the hypothesis  $H_0: \beta_2 = 0, \beta_1 \neq 0$  vs.  $H_1: \beta_2 \neq 0, \beta_1 \neq 0$ . In general, if we have  $k$  independent variables, then to assess the specific effect of the  $l$ th independent variable ( $x_l$ ), on  $y$  after controlling for the effects of all other variables, we wish to test the hypothesis  $H_0: \beta_l = 0, \text{ all other } \beta_j \neq 0$  vs.  $H_1: \text{all } \beta_j \neq 0$ . We focus on assessing the independent contribution of birthweight. Our approach is to compute the standard error of the partial-regression coefficient for birthweight and base our test on  $t = b_l / se(b_l)$ , which will follow a  $t$  distribution with  $n - k - 1$  df under  $H_0$ . Specifically, the following test procedure for a level  $\alpha$  test is used.

**Equation 11.32**

**t Test for Testing the Hypothesis  $H_0: \beta_\ell = 0$ , All Other  $\beta_j \neq 0$  vs.  $H_1: \beta_\ell \neq 0$ , All Other  $\beta_j \neq 0$  in Multiple Linear Regression**

(1) Compute

$$t = b_\ell / se(b_\ell)$$

which should follow a  $t$  distribution with  $n - k - 1$  df under  $H_0$ .

(2) If  $t < t_{n-k-1, \alpha/2}$  or  $t > t_{n-k-1, 1-\alpha/2}$  then reject  $H_0$

If  $t_{n-k-1, \alpha/2} \leq t \leq t_{n-k-1, 1-\alpha/2}$  then accept  $H_0$

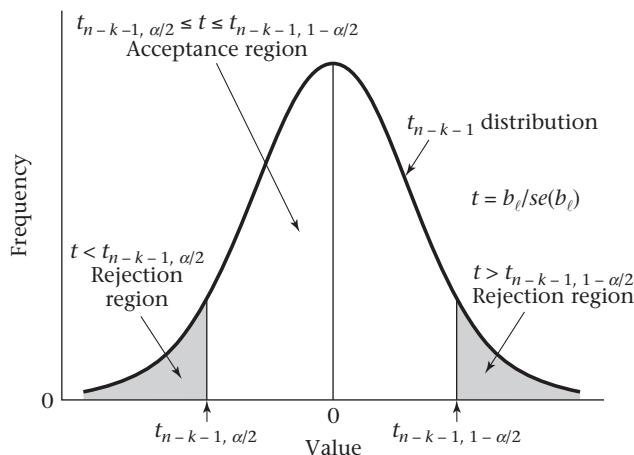
(3) The exact  $p$ -value is given by

$$2 \times Pr(t_{n-k-1} > t) \quad \text{if } t \geq 0$$

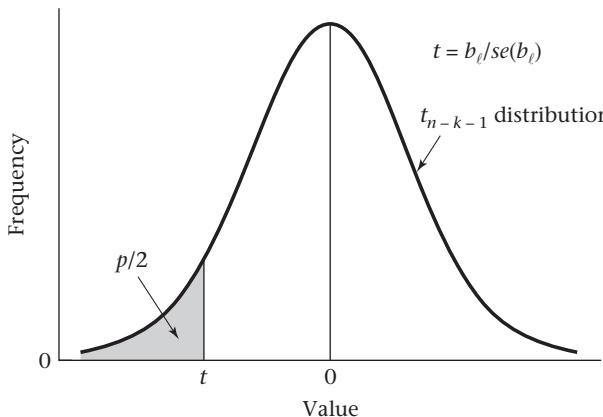
$$2 \times Pr(t_{n-k-1} \leq t) \quad \text{if } t < 0$$

The acceptance and rejection regions for this test are depicted in Figure 11.21. The computation of the exact  $p$ -value is illustrated in Figure 11.22.

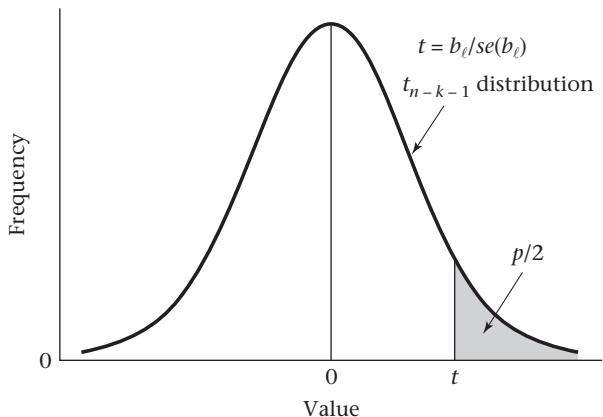
**Figure 11.21 Acceptance and rejection regions for the  $t$  test for multiple linear regression**



**Figure 11.22 Computation of the exact  $p$ -value for the  $t$  test for multiple linear regression**



(a) If  $t < 0$ , then  $p = 2 \times$  (area to the left of  $t$  under a  $t_{n-k-1}$  distribution).



(b) If  $t \geq 0$ , then  $p = 2 \times$  (area to the right of  $t$  under a  $t_{n-k-1}$  distribution).

**Example 11.45**

**Hypertension, Pediatrics** Test for the independent contributions of birthweight and age in predicting SBP in infants, using the output in Table 11.10.

**Solution**

From Table 11.10,

$$\begin{aligned} b_1 &= 0.1256 \\ se(b_1) &= 0.0343 \\ t(\text{birthweight}) &= b_1 / se(b_1) = 3.66 \\ p &= 2 \times Pr(t_{13} > 3.66) = .003 \\ b_2 &= 5.888 \\ se(b_2) &= 0.6802 \\ t(\text{age}) &= b_2 / se(b_2) = 8.66 \\ p &= 2 \times Pr(t_{13} > 8.66) < .001 \end{aligned}$$

Therefore, both birthweight and age have highly significant associations with SBP, even after controlling for the other variable.

It is possible that an independent variable ( $x_1$ ) will seem to have an important effect on a dependent variable ( $y$ ) when considered by itself but will not be significant after adjusting for another independent variable ( $x_2$ ). This usually occurs when  $x_1$  and  $x_2$  are strongly related to each other and when  $x_2$  is also related to  $y$ . We refer to  $x_2$  as a confounder of the relationship between  $y$  and  $x_1$ . We discuss confounding in more detail in Chapter 13. Indeed, one of the advantages of multiple-regression analysis is that it lets us identify which few variables among a large set of independent variables have a significant relationship to the dependent variable *after adjusting for other important variables*.

**Example 11.46**

**Hypertension, Pediatrics** Suppose we consider the two independent variables,  $x_1$  = birthweight,  $x_2$  = body length and try to use these variables to predict SBP in newborns ( $y$ ). Perhaps both  $x_1$  and  $x_2$ , *when considered separately* in a simple linear-regression model as given in Equation 11.2, have a significant relationship to blood pressure. However, because birthweight and body length are closely related to each other, after adjusting for birthweight, body length may not be significantly related to blood pressure based on the test procedure in Equation 11.32. One possible interpretation of this result is that the effect of body length on blood pressure can be explained by its strong relationship to birthweight.

In some instances, two strongly related variables are entered into the same multiple-regression model and, after controlling for the effect of the other variable, neither variable is significant. Such variables are referred to as collinear. It is best to avoid using highly collinear variables in the same multiple-regression model because their simultaneous presence can make it impossible to identify the specific effects of each variable.

**Example 11.47**

**Hypertension** A commonly used measure of obesity is body-mass index (BMI), which is defined as weight/(height)<sup>2</sup>. It is well known that both weight and BMI, considered separately, are strongly related to level of blood pressure. However, if they are entered simultaneously in the same multiple-regression model, then it is

possible that neither will be significant because they are strongly related to each other. Thus, if we control for weight, there may be no additional predictive power for BMI, and vice versa.

In Equation 11.32, we have considered the test of the hypothesis  $H_0$ : that a specific partial-regression coefficient  $\beta_\ell = 0$  vs. the alternative hypothesis  $H_1$ : that  $\beta_\ell \neq 0$ . Under both  $H_0$  and  $H_1$ , all other partial-regression coefficients are allowed to be different from 0. We used a  $t$  statistic to test these hypotheses. Another way to perform this test is in the form of a partial  $F$  test, which is given as follows.

### Equation 11.33

#### Partial $F$ Test for Partial-Regression Coefficients in Multiple Linear Regression

To test the hypothesis  $H_0: \beta_\ell = 0$ , all other  $\beta_j \neq 0$  vs.  $H_1: \beta_\ell \neq 0$ , all other  $\beta_j \neq 0$  in multiple linear regression, we

- (1) Compute

$$F = \frac{\text{Regr SS}_{\text{full model}} - \text{Regr SS}_{\text{all variables except } \beta_\ell \text{ in the model}}}{\text{Res MS}_{\text{full model}}}$$

which should follow an  $F_{1,n-k-1}$  distribution under  $H_0$ .

- (2) The exact  $p$ -value is given by  $Pr(F_{1,n-k-1} > F)$ .
- (3) It can be shown that the  $p$ -value from using the partial  $F$  test given in (2) is the same as the  $p$ -value obtained from using the  $t$  test in Equation 11.32.

Many statistical packages use variable selection strategies such as forward and backward selection based on a succession of partial  $F$  tests. A complete discussion of variable selection strategies is provided in [4] and [5].

## Criteria for Goodness of Fit

In Section 11.6, we discussed criteria for goodness of fit in simple linear-regression models, based on residual analysis. Similar criteria can be used in a multiple-regression setting.

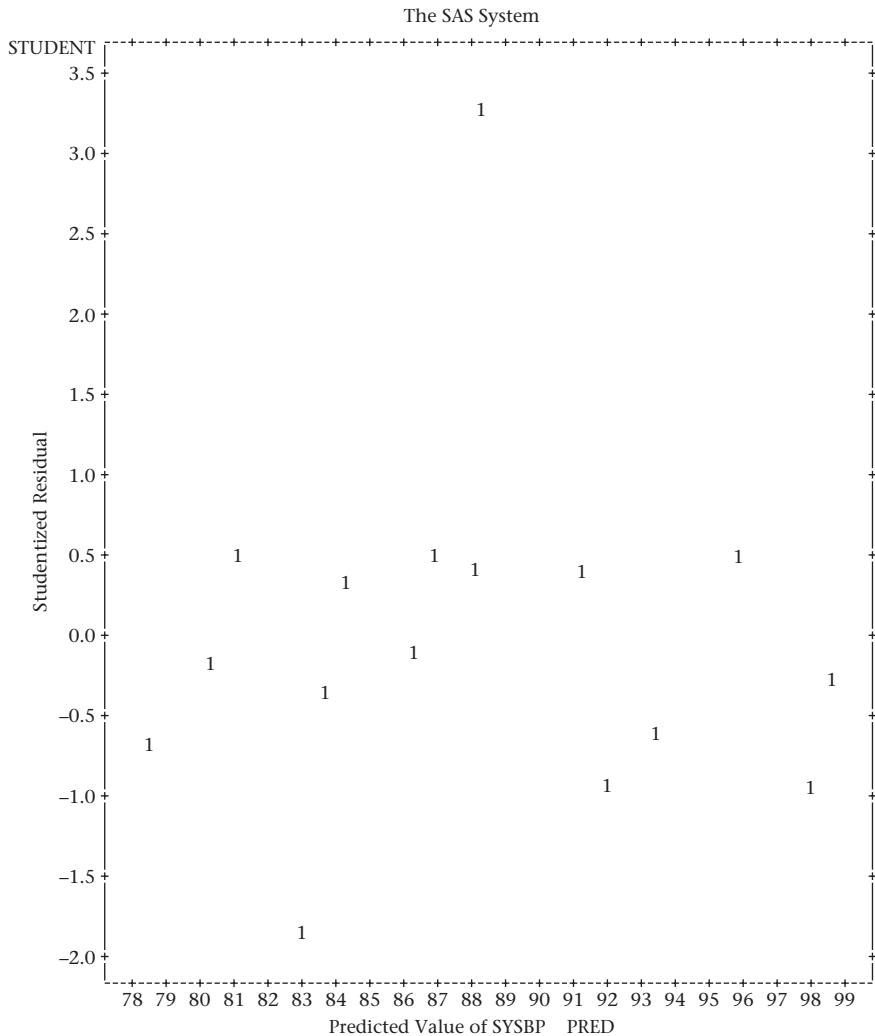
### Example 11.48

**Hypertension, Pediatrics** Assess the goodness of fit of the multiple-regression model in Table 11.10 fitted to the infant blood-pressure data in Table 11.9.

### Solution

We compute the residual for each of the 16 sample points in Table 11.9. The standard error of each of the fitted residuals is different and depends on the distance of the corresponding sample point from the average of the sample points used in fitting the regression line. Thus we will usually be interested in the Studentized residuals = STUDENT( $i$ ) =  $\hat{e}_i / sd(\hat{e}_i)$ . (See [4] or [5] for the formulas used to compute  $sd(\hat{e}_i)$  in a multiple-regression setting.) We have plotted the Studentized residuals against the predicted blood pressure (Figure 11.23a) and each of the independent variables (Figures 11.23b and 11.23c). This lets us identify any outlying values as well as violations of the linearity and equal-variance assumptions in the multiple-regression model.

**Figure 11.23a Plot of Studentized residuals vs. predicted values of SBP for the multiple-regression model in Table 11.10**



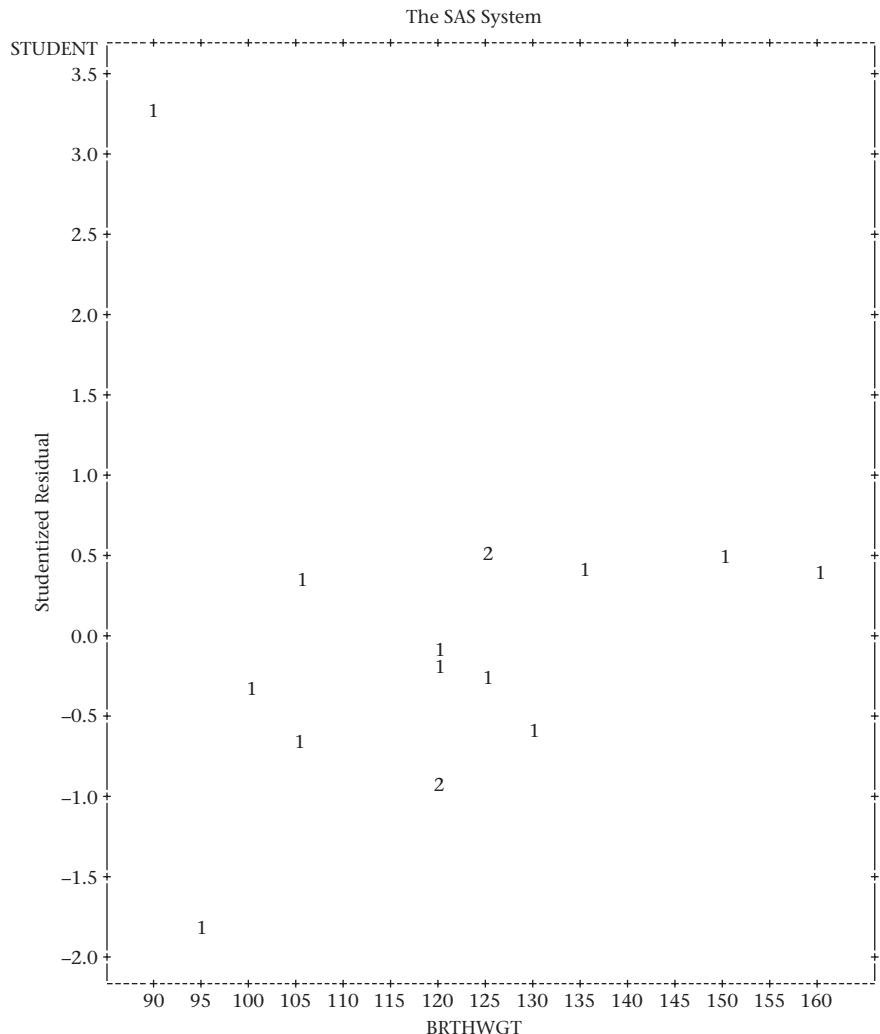
There seems to be a possible outlying value with a Studentized residual  $\approx 3.0$  corresponding to a birthweight of 90 oz and an age in days = 4.0 (observation 10). To focus more clearly on outlying values, some computer packages let the user delete an observation, recompute the regression model from the remaining data points, and compute the residual of the deleted observation based on the recomputed regression. The rationale for this procedure is that the outlying value may have affected the estimates of the regression parameters. Let

$$y = a^{(i)} + b_1^{(i)}x_1 + \cdots + b_k^{(i)}x_k$$

denote the estimated regression model with the  $i$ th sample point deleted. The residual of the deleted point from this regression line is

$$\hat{e}^{(i)} = y_i - [a^{(i)} + b_1^{(i)}x_{i1} + \cdots + b_k^{(i)}x_{ik}]$$

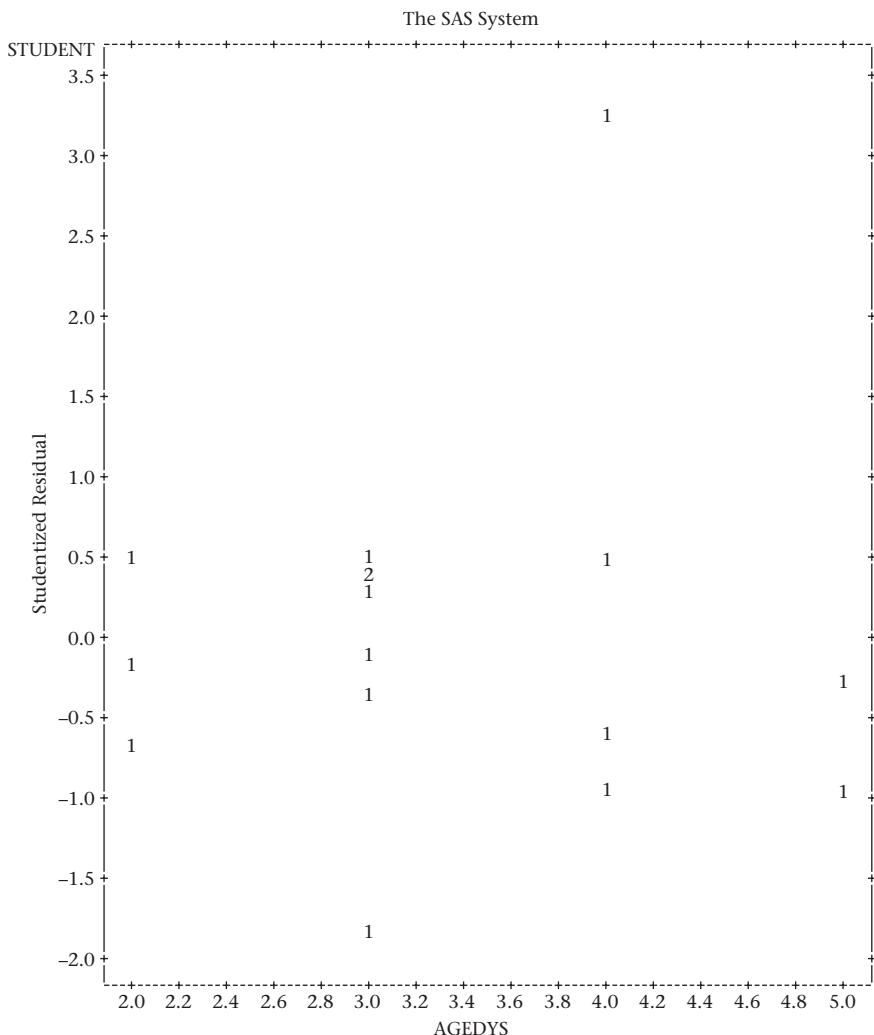
**Figure 11.23b Plot of Studentized residuals vs. birthweight for the multiple-regression model in Table 11.10**



with standard error  $\text{sd}(\hat{\epsilon}^{(i)})$ . The corresponding Studentized residual is  $\hat{\epsilon}^{(i)}/\text{sd}(\hat{\epsilon}^{(i)})$  and is denoted by  $\text{RSTUDENT}(i)$ . It is sometimes called an **externally Studentized residual** because the  $i$ th data point was not used in estimating the regression parameters, as opposed to  $\text{STUDENT}(i)$ , which is sometimes called an **internally Studentized residual** because the  $i$ th data point was used in estimating the regression parameters. We have plotted the externally Studentized residuals [ $\text{RSTUDENT}(i)$ ] against the predicted blood pressure (Figure 11.24a) and each of the independent variables (Figures 11.24b and 11.24c). These plots really highlight the outlying value. Data point 10 has a value of  $\text{RSTUDENT}$  that is approximately 7 standard deviations above zero, which indicates a gross outlier.

The plots in Figures 11.24a–11.24c do not really reflect the multivariate nature of the data. Specifically, under the multiple-regression model in Equation 11.30, the relationship between  $y$  and a specific independent variable  $x_i$  is characterized as follows.

**Figure 11.23c Plot of Studentized residuals vs. age in days for the multiple-regression model in Table 11.10**



**Equation 11.34**

$y$  is normally distributed with expected value  $= \alpha_\ell + \beta_\ell x_\ell$  and variance  $\sigma^2$  where

$$\alpha_\ell = \alpha + \beta_1 x_1 + \cdots + \beta_{\ell-1} x_{\ell-1} + \beta_{\ell+1} x_{\ell+1} + \cdots + \beta_k x_k$$

Thus, given the values of all other independent variables ( $x_1, \dots, x_{\ell-1}, x_{\ell+1}, \dots, x_k$ ),

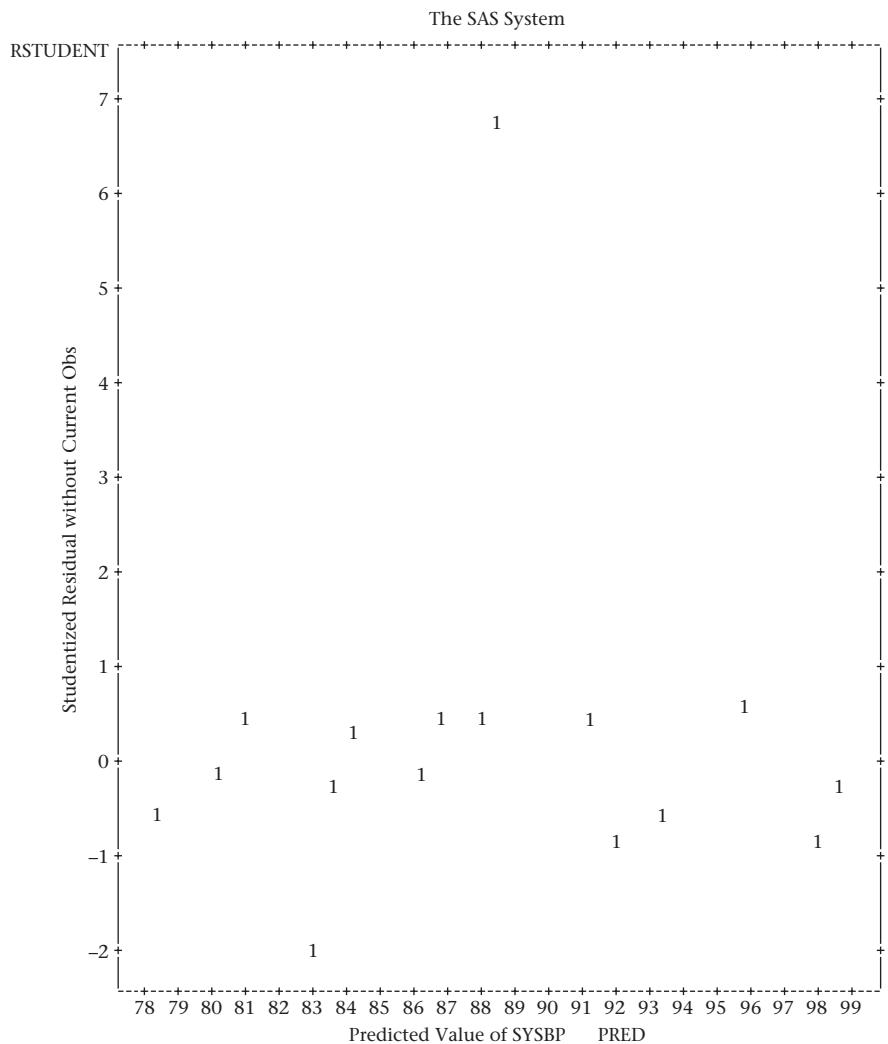
- (1) The average value of  $y$  is linearly related to  $x_\ell$ .
- (2) The variance of  $y$  is constant (i.e.,  $\sigma^2$ ).
- (3)  $y$  is normally distributed.

A partial-residual plot is a good way to check the validity of the assumptions in Equation 11.34.

**Definition 11.18**

A **partial-residual plot** characterizing the relationship between the dependent variable  $y$  and a specific independent variable  $x_i$  in a multiple-regression setting is constructed as follows:

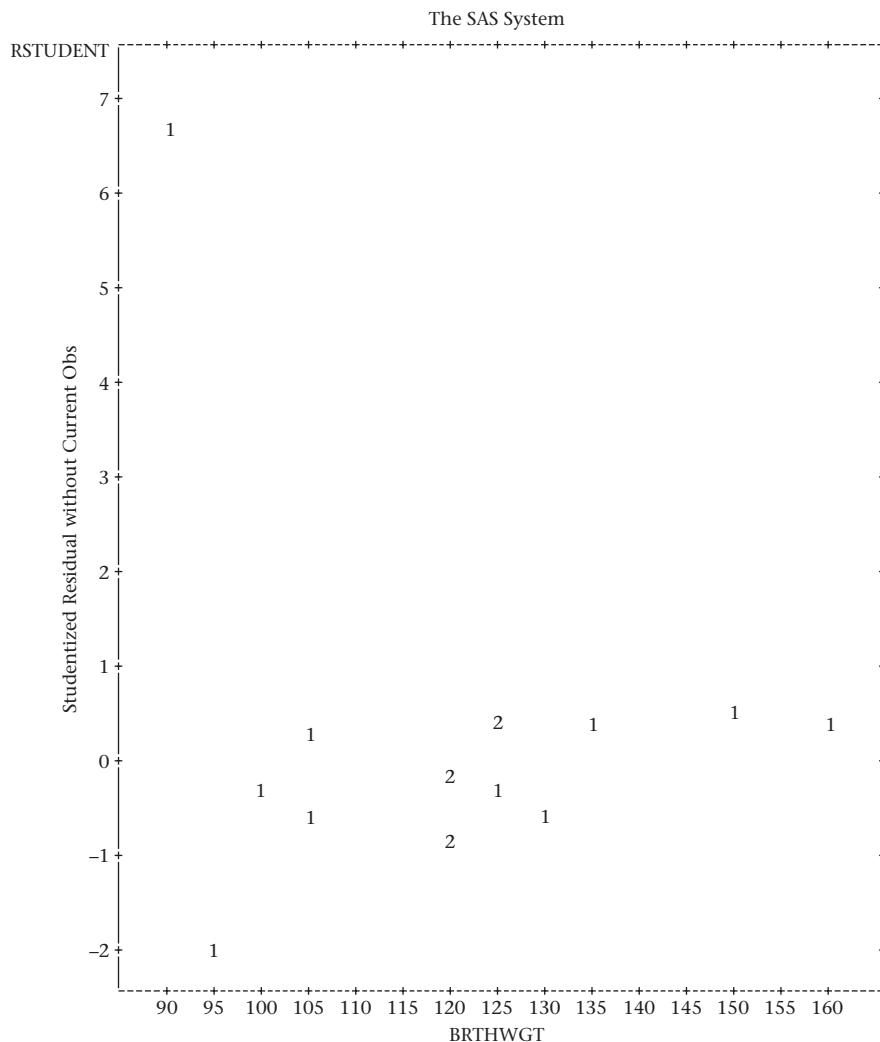
**Figure 11.24a Plot of RSTUDENT vs. the predicted SBP for the multiple-regression model in Table 11.10**



- (1) A multiple regression is performed of  $y$  on all predictors other than  $x_t$  (i.e.,  $x_1, \dots, x_{t-1}, x_{t+1}, \dots, x_k$ ), and the residuals are saved.
- (2) A multiple regression is performed of  $x_t$  on all other predictors (i.e.,  $x_1, \dots, x_{t-1}, x_{t+1}, \dots, x_k$ ), and the residuals are saved.
- (3) The partial-residual plot is a scatter plot of the residuals in step 1 on the  $y$ -axis against the residuals in step 2 on the  $x$ -axis.

Many computer packages compute partial-residual plots as an option in their multiple-regression routines, so the user need not perform the individual steps 1 to 3. The partial-residual plot reflects the relationship between  $y$  and  $x_t$  after each variable is adjusted for all other predictors in the multiple-regression model, which is a primary goal of performing a multiple-regression analysis. It can be shown that if the multiple-regression model in Equation 11.30 holds, then the residuals in step 1 should be linearly related to the residuals in step 2 with slope =  $\beta_t$  (i.e., the partial-regression coefficient pertaining to  $x_t$  in the multiple-regression model in Equation 11.30) and

**Figure 11.24b Plot of RSTUDENT vs. birthweight for the multiple-regression model in Table 11.10**



constant residual variance  $\sigma^2$ . A separate partial-residual plot can be constructed relating  $y$  to each predictor  $x_1, \dots, x_k$ .

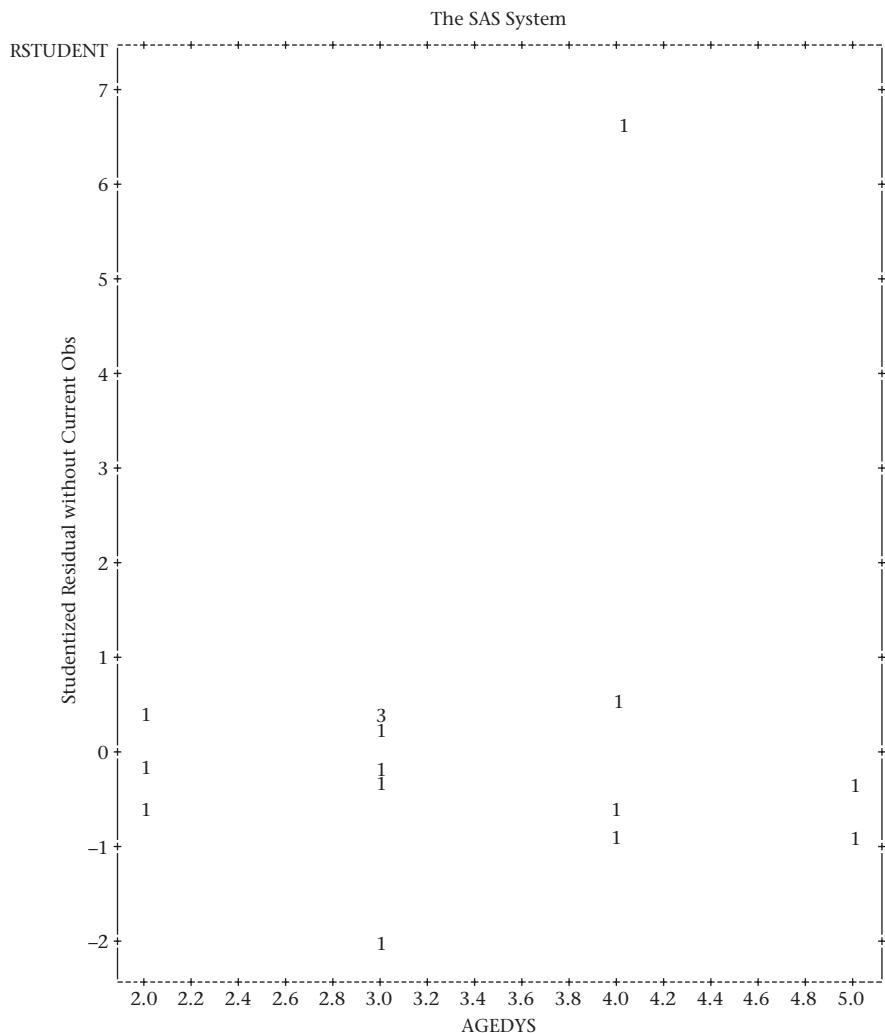
#### Example 11.49

**Hypertension, Pediatrics** Construct a separate partial-residual plot relating SBP to birthweight and age for the data in Table 11.9.

#### Solution

We refer to the SAS output in Figures 11.25a and 11.25b. The  $y$ -axis in Figure 11.25a corresponds to residuals of SBP after adjusting for age in days. The  $x$ -axis corresponds to residuals of birthweight after adjusting for age in days. Figure 11.25b is defined similarly. Hence, the  $x$ - and  $y$ -axes correspond to residuals and are not in the familiar units of blood pressure and birthweight in Figure 11.25a, for example. In Figure 11.25a, we notice that the relationship between SBP and birthweight is approximately linear (perhaps slightly curvilinear) with the exception of observation 10, which we previously identified as an outlier. In Figure 11.25b, the relationship between SBP and age appears to be linear with the exception of observation 10.

**Figure 11.24c** Plot of RSTUDENT vs. age in days for the multiple-regression model in Table 11.10

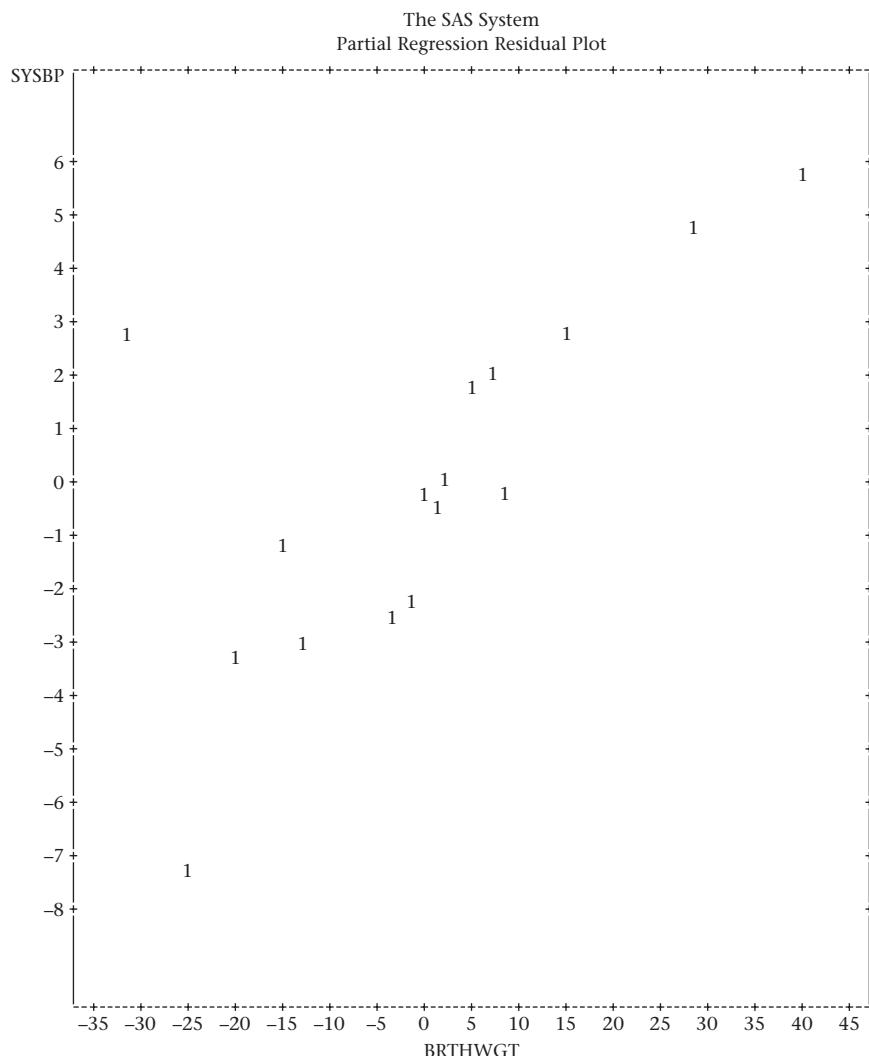


Notice that the three  $x$  values clustered in the lower left of the plot all correspond to age = 2 days. However, they have different abscissas in this plot because they reflect the residual of age after correcting for birthweight and the three birthweights are all different (see observations 4, 7, and 11 with birthweights = 105, 125, and 120 oz, respectively). In these data, the fitted regression line of age on birthweight is given by  $\text{age} = 2.66 + 0.0054 \times \text{birthweight}$ .

Because we identified observation 10 as an outlier, we deleted this observation and reran the regression analysis based on the reduced sample of size 15. The regression model is given in Table 11.11 and the partial-residual plots in Figures 11.26a and 11.26b. The estimated multiple-regression model is

$$\gamma = 47.94 + 0.183 \times \text{birthweight} + 5.28 \times \text{age}$$

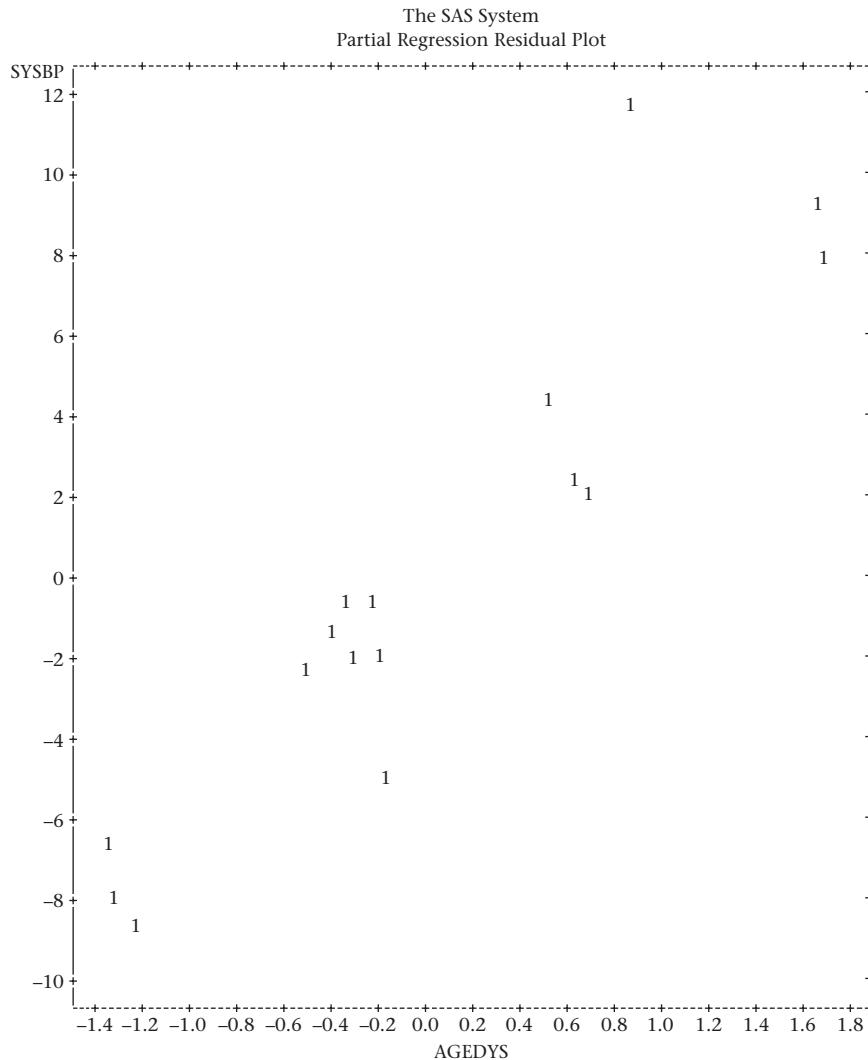
which differs considerably from the multiple-regression model in Table 11.10 ( $y = 53.45 + 0.126 \times \text{birthweight} + 5.89 \times \text{age}$ ), particularly for the estimated regression coefficient for birthweight, which increased by about 50%. No outliers are evident in either of the partial-residual plots in Figure 11.26. There is only a slight hint of curvilinearity in Figure 11.26a, which relates SBP to birthweight (after controlling for age).

**Figure 11.25a Partial-residual plot of SBP vs. birthweight for the model in Table 11.10**

In Section 11.9, we were introduced to multiple linear regression. This technique is used when we wish to relate a normally distributed outcome variable  $y$  (called the dependent variable) to several (more than one) independent variables  $x_1, \dots, x_k$ . The independent variables need not be normally distributed. Indeed, the independent variables can even be categorical, as discussed further in the Case Study in Sections 11.10 and 12.5. On the flowchart at the end of this chapter (Figure 11.32, p. 503), we answer no to (1) interested in relationships between two variables? and we answer continuous to (2) outcome variable continuous or binary? This leads us to the box labeled “multiple-regression methods.”

### REVIEW QUESTIONS 11D

- 1 What is the difference between a univariate-regression model and a multiple-regression model?
- 2 What is the difference between a simple-regression coefficient and a partial-regression coefficient?

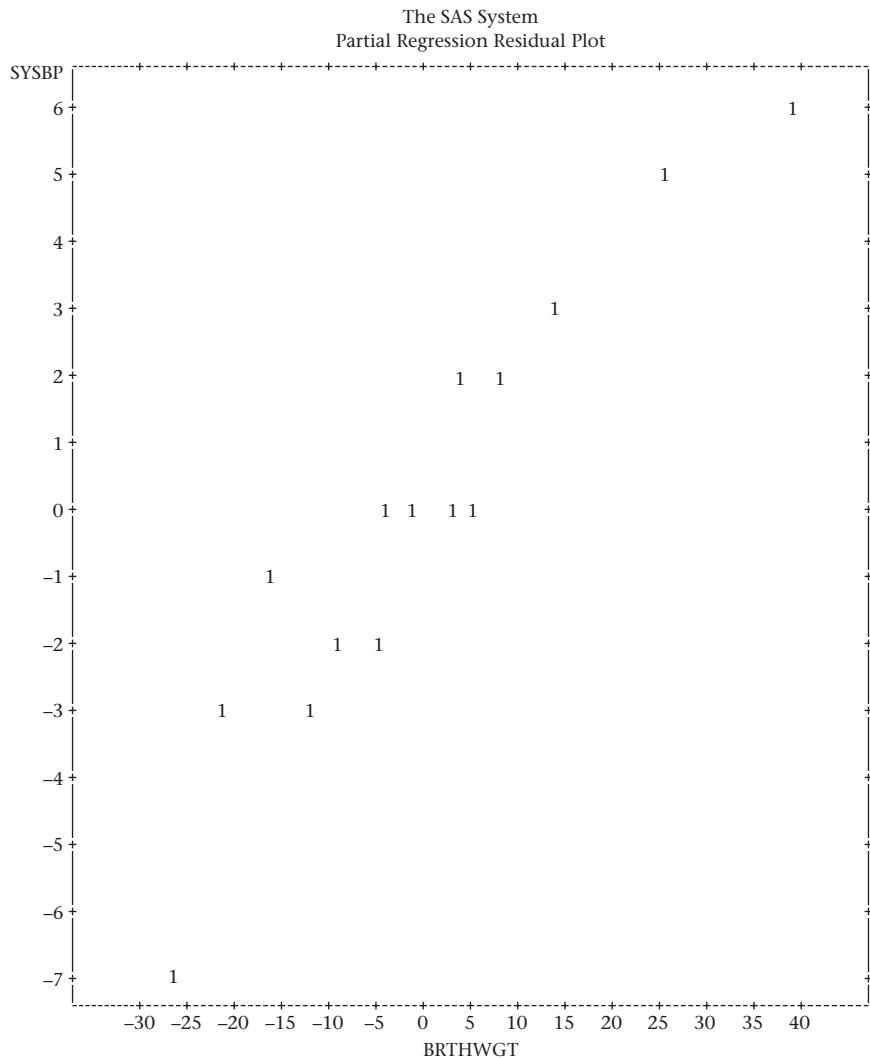
**Figure 11.25b Partial-residual plot of SBP vs. age in days for the model in Table 11.10**

- 3 Refer to the data in Table 2.11.
  - (a) Run a multiple-regression model of  $\ln(\text{duration of hospital stay})$  on age, sex, and white-blood count on admission to the hospital and service (1 = medical/2 = surgical).
  - (b) Interpret the results in a meaningful way.

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

In Table 8.13, we compared the mean finger-wrist tapping score (MAXFWT) between exposed and control children using a two-sample  $t$  test. However, another approach would be to use regression methods to compare the two groups using dummy-variable coding to denote group membership.

**Figure 11.26a Partial-residual plot of SBP vs. birthweight based on the data in Table 11.9 after deleting one outlier (observation 10) ( $n = 15$ )**



### Definition 11.19

A **dummy variable** is a binary variable used to represent a categorical variable with two categories (say A and B). The dummy variable is set to the value  $c_1$  if a subject is in category A and to  $c_2$  if a subject is in category B. The most common choices for the values  $c_1$  and  $c_2$  are 1 and 0, respectively.

### Example 11.50

**Environmental Health, Pediatrics** Use regression methods to compare the mean MAXFWT between exposed and control children.

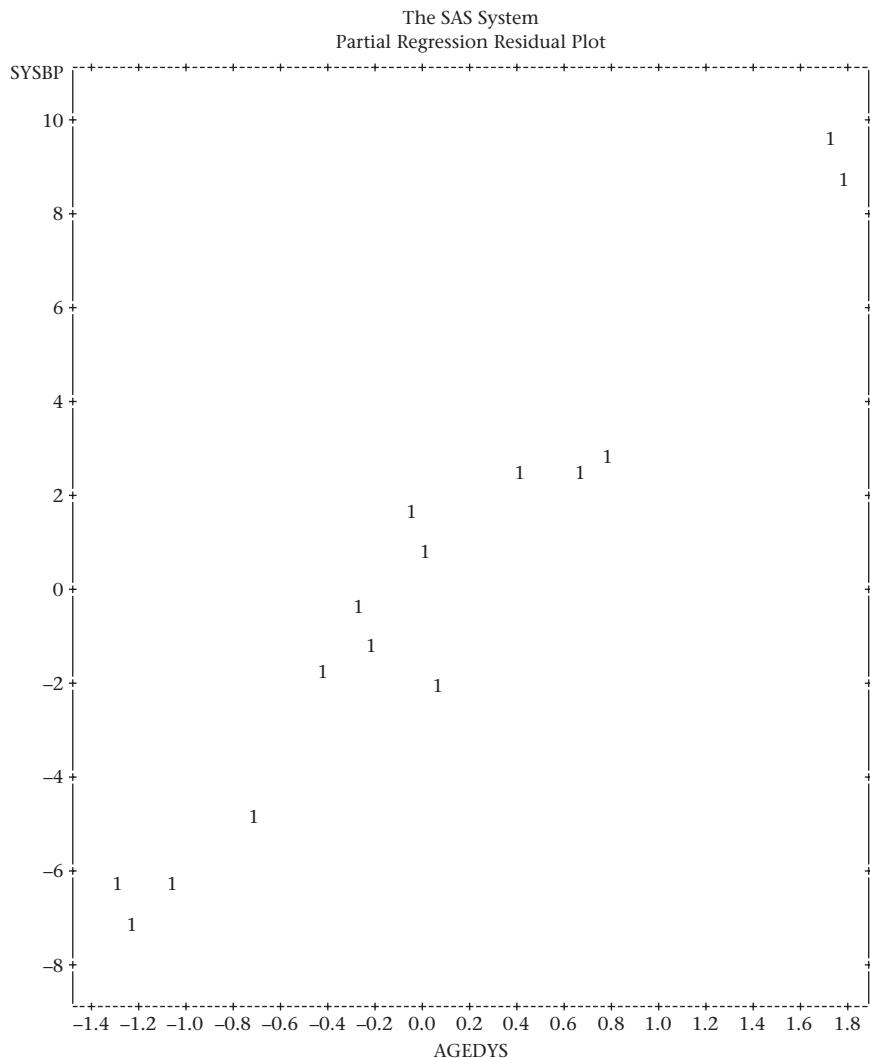
#### Solution

We will represent group membership by the dummy variable CSCN2 defined by

$$\text{CSCN2} = \begin{cases} 1 & \text{if child is exposed} \\ 0 & \text{if child is control} \end{cases}$$

We then can run a simple linear-regression model of the following form.

**Figure 11.26b Partial-residual plot of SBP vs. age in days based on the data in Table 11.9 after deleting one outlier (observation 10) ( $n = 15$ )**



**Equation 11.35**

$$\text{MAXFWT} = \alpha + \beta \times \text{CSCN2} + e$$

What do the parameters of this model mean? If a child is in the exposed group, then the average value of MAXFWT for that child is  $\alpha + \beta$ ; if a child is in the control group, then the average value of MAXFWT for that child is  $\alpha$ . Thus  $\beta$  represents the difference between the average value of MAXFWT for children in the exposed group vs. the control group. Our best estimate of  $\alpha + \beta$  is given by the sample mean of MAXFWT for children in the exposed group; our best estimate of  $\alpha$  is given by the sample mean of MAXFWT for children in the control group. Thus our best estimate of  $\beta$  is given by the mean difference in MAXFWT between the exposed and control groups. Another way to interpret  $\beta$  is as the average increase in MAXFWT per 1-unit increase in CSCN2. However, a 1-unit increase in CSCN2 corresponds to the difference between the exposed and control groups for CSCN2. We have run the regression model in Equation 11.35 using the SAS PROC

**Table 11.11** Multiple-regression model of SBP on birthweight and age based on data in Table 11.9 after deleting one outlier (observation 10) ( $n = 15$ )

The REG Procedure					
<b>Model:</b> MODEL1					
<b>Dependent Variable:</b> sysbp					
		Number of Observations Read	15		
		Number of Observations Used	15		
Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	2	602.96782	301.48391	217.52	<.0001
Error	12	16.63218	1.38601		
Corrected Total	14	619.60000			
Root MSE		1.17729	R-Square	0.9732	
Dependent Mean		87.60000	Adj R-Sq	0.9687	
Coeff Var		1.34394			
Parameter Estimates					
		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr >  t
Intercept	1	47.93769	2.30154	20.83	<.0001
brthwgt	1	0.18316	0.01840	9.96	<.0001
agedys	1	5.28248	0.33520	15.76	<.0

REG procedure. The results are given in Table 11.12. We see there is a significant difference between the mean MAXFWT for the two groups ( $p = .003$ ). The estimated mean difference between the MAXFWT scores for the two groups is  $-6.66$  taps/10 seconds with standard error =  $2.22$  taps per 10 seconds. The regression coefficient  $b$  corresponds exactly to the mean difference reported in Table 8.13 (mean for exposed group =  $48.44$  taps per 10 seconds; mean for control group =  $55.10$  taps per 10 seconds, mean difference =  $-6.66$  taps per 10 seconds). The standard error for  $b$  corresponds exactly to the standard error of the mean difference given by a two-sample  $t$  test with *equal* variances. The absolute value of the  $t$  statistic and the  $p$ -value for the two procedures are also the same.

This leads to the following general principle.

### Equation 11.36

### Relationship Between Simple Linear Regression and $t$ Test Approaches

Suppose we wish to compare the underlying mean between two groups where the observations in group 1 are assumed to be normally distributed with mean  $\mu_1$  and variance  $\sigma^2$  and the observations in group 2 are assumed to be normally distributed with mean  $\mu_2$  and variance  $\sigma^2$ . To test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ , we can use one of two equivalent procedures:

- (1) We can perform a two-sample  $t$  test with equal variances.
- (2) We can set up a linear-regression model of the form

$$y = \alpha + \beta x + e$$

**Table 11.12 Simple linear-regression model comparing exposed and control children for MAXFWT ( $n = 95$ )**

The REG Procedure					
<b>Model: MODEL1</b>					
<b>Dependent Variable: maxfwt</b>					
		Number of Observations Read		120	
		Number of Observations Used		95	
		Number of Observations with Missing Values		25	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	940.63327	940.63327	9.02	0.0034
Error	93	9697.30357	104.27208		
Corrected Total	94	10638			
Root MSE		10.21137	R-Square	0.0884	
Dependent Mean		52.85263	Adj R-Sq	0.0786	
Coeff Var		19.32046			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	55.09524	1.28651	42.83	<.0001
cscn2	1	-6.65774	2.21667	-3.00	0.0034

where  $y$  is the outcome variable and  $x = 1$  if a subject is in group 1 and 0 if a subject is in group 2 and  $e \sim N(0, \sigma^2)$ .

The  $t$  statistic in procedure 1 =  $(\bar{y}_1 - \bar{y}_2)/\sqrt{s^2(1/n_1 + 1/n_2)}$  is the same as the  $t$  statistic in procedure 2 =  $b/se(b)$ . The estimated mean difference between the groups in procedure 1 =  $\bar{y}_1 - \bar{y}_2$  is the same as the estimated slope ( $b$ ) in procedure 2. The standard error of the mean difference in procedure 1 =  $\sqrt{s^2(1/n_1 + 1/n_2)}$  is the same as the standard error of  $b$  in procedure 2. The  $t$  statistics and  $p$ -values are the same in procedure 1 and procedure 2.

One issue with neurologic-function data in children is that they are often strongly related to age and in some cases to gender as well. Even slight age differences between the exposed and control groups could explain differences between the groups in neurologic function. The first issue is to determine whether MAXFWT is related to age and/or to gender. For this purpose, we might construct a multiple-regression model of the form

**Equation 11.37**

$$\text{MAXFWT} = \alpha + \beta_1 \text{age} + \beta_2 \text{sex} + e$$

where age is in years and sex is coded as 1 if male and 2 if female. Note: A slightly more accurate measure of age could be obtained if we used age in years + (age in months)/12 rather than simply age in years. The results from fitting the model in Equation 11.37 are given in Table 11.13. We see that MAXFWT is strongly related to age ( $p < .001$ ) and is slightly, but not significantly, related to gender. Older children

**Table 11.13** Multiple-regression model of MAXFWT on age and sex (*n* = 95)

The REG Procedure					
<b>Model:</b> MODEL1					
<b>Dependent Variable:</b> maxfwt					
		Number of Observations Read		120	
		Number of Observations Used		95	
		Number of Observations with Missing Values		25	
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5438.14592	2719.07296	48.11	<.0001
Error	92	5199.79092	56.51947		
Corrected Total	94	10638			
Root MSE		7.51794	R-Square	0.5112	
Dependent Mean		52.85263	Adj R-Sq	0.5006	
Coeff Var		14.22435			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	31.59139	3.16011	10.00	<.0001
ageyr	1	2.52068	0.25706	9.81	<.0001
sex	1	-2.36574	1.58722	-1.49	0.1395

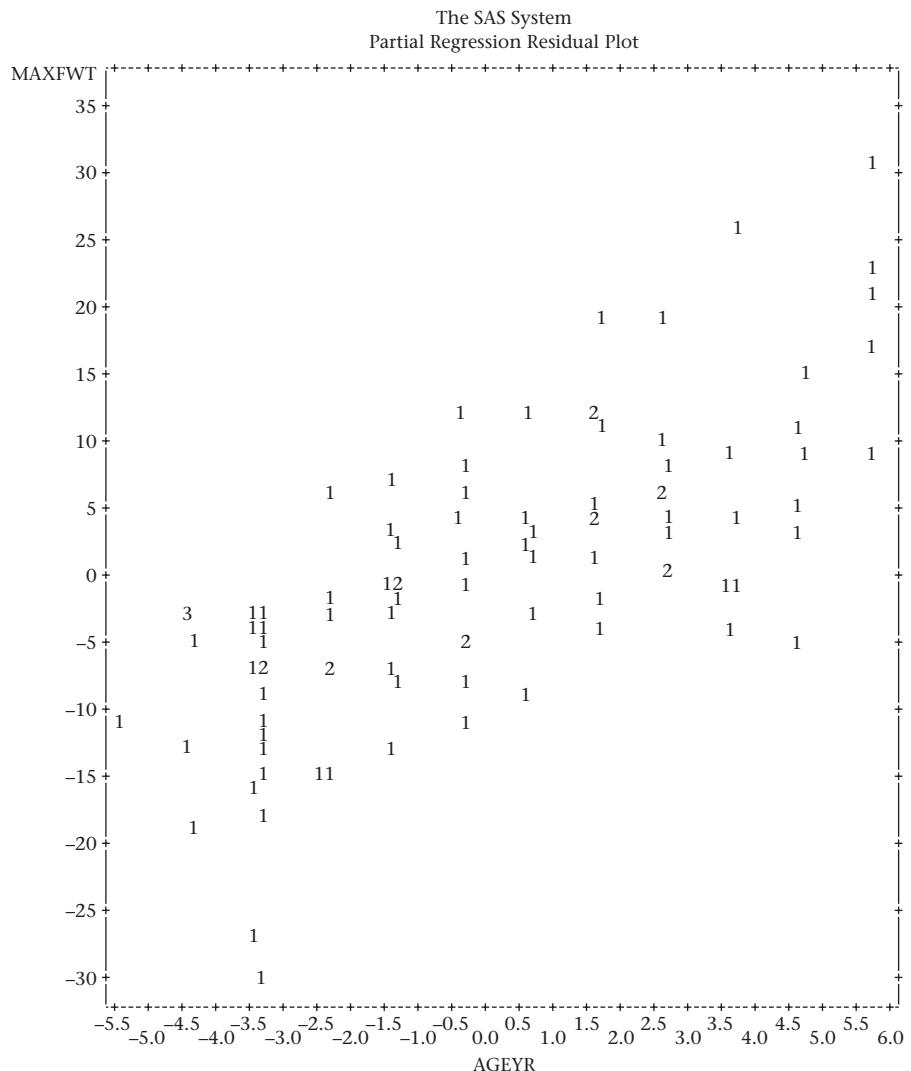
and males have higher values of MAXFWT (mean MAXFWT increases by 2.5 taps per 10 seconds for every 1 year increase in age for children of the same sex and is 2.4 taps per 10 seconds higher in boys than in girls of the same age). The partial-residual plots of MAXFWT vs. age and sex are given in Figures 11.27a and 11.27b, respectively. From Figure 11.27a, we see that MAXFWT is strongly and approximately linearly related to age. No obvious outlying values are apparent. From Figure 11.27b, we see that males (corresponding to the left cloud of points) tend to have slightly higher values of MAXFWT (after correcting for age) than females. The Studentized residual plot vs. the predicted MAXFWT is given in Figure 11.28. No outliers are apparent from this plot, either. Also, the variance of MAXFWT looks similar for different values of age, sex, and the predicted MAXFWT.

From Table 11.12, mean MAXFWT differs between the exposed and control groups by 6.66 taps per 10 seconds. Thus even a 1-year difference in mean age between the two groups would account for approximately  $(2.52/6.66) \times 100\%$  or 38% of the observed mean difference in MAXFWT. Thus it is essential to redo the crude analyses in Table 11.12, adjusting for possible age and sex differences between groups. We will use the multiple-regression model.

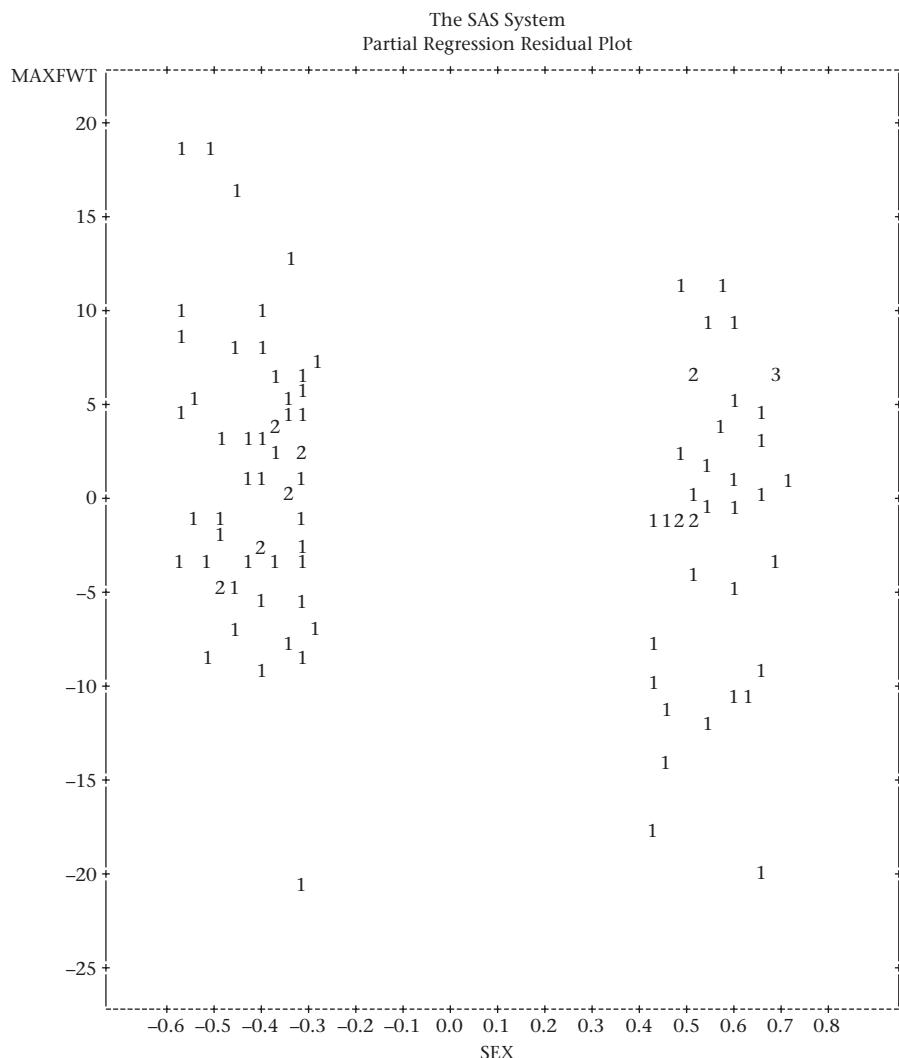
**Equation 11.38**

$$\text{MAXFWT} = \alpha + \beta_1 \times \text{CSCN2} + \beta_2 \times \text{age} + \beta_3 \times \text{sex} + e$$

where CSCN2 = 1 if in the exposed group and 0 if in the control group and sex = 1 if male and 2 if female. The fitted model is given in Table 11.14. We see that the estimated mean difference between the groups is  $-5.15 \pm 1.56$  taps/10 sec ( $p = .001$ ),

**Figure 11.27a** Partial-residual plot of MAXFWT vs. age (in years) ( $n = 95$ )

after controlling for age and sex. The effects of age and sex are similar to those seen in Table 11.13. The partial-residual plot of MAXFWT on group (CSCN2) is given in Figure 11.29. The left cloud of points corresponds to the control group and the right cloud to the exposed group. The control group does appear to have generally higher values than the exposed group, although there is considerable overlap between the two groups. Also, the range of residual values for the control group appears to be greater than for the exposed group. However, the control group ( $n = 63$ ) is larger than the exposed group ( $n = 32$ ), which would account, at least in part, for the difference in the range. In Table 8.13, when we performed a two-sample  $t$  test comparing the two groups we found that the within-group standard deviation was 10.9 for the control group and 8.6 for the exposed group ( $p = .14$  using the  $F$  test). Another interesting finding is that the age- and sex-adjusted mean difference between the groups ( $-5.15$  taps/10 sec) is smaller than the crude difference ( $-6.66$  taps/10 seconds). This difference is explained in part by differences in the age-sex distribution between the two groups. The model in Equation 11.38 is called an **analysis-of-covariance**

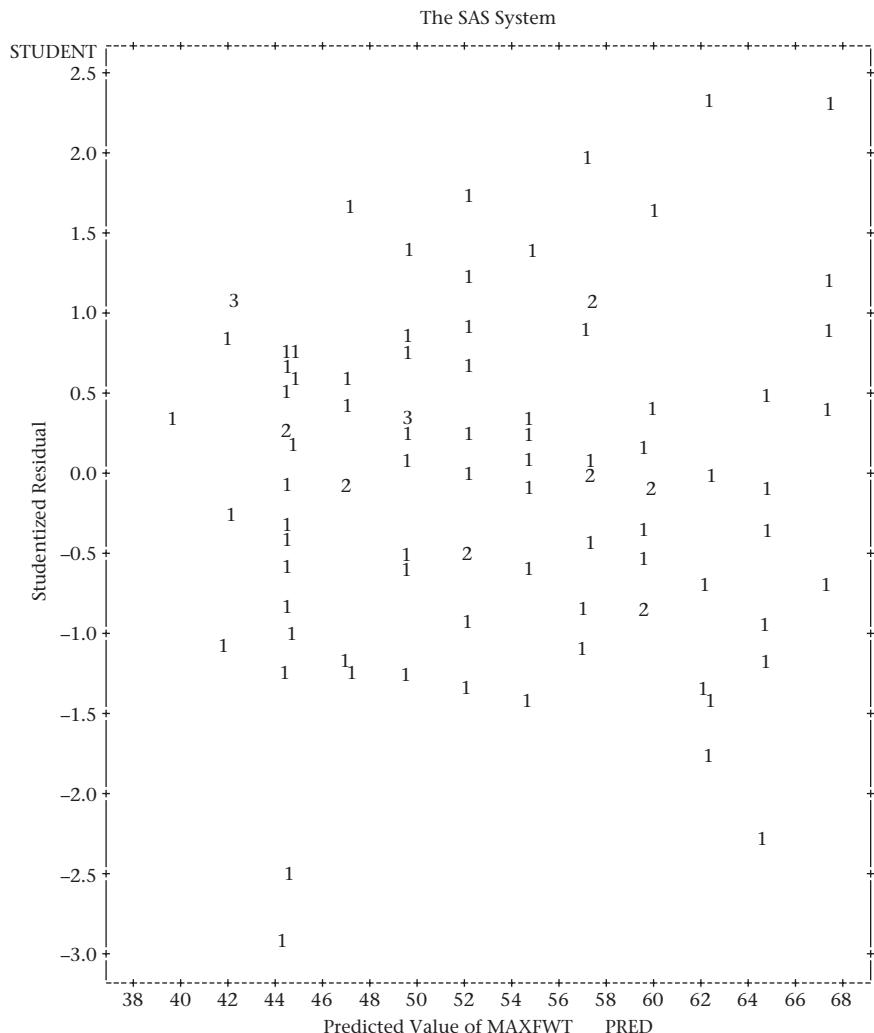
**Figure 11.27b** Partial-residual plot of MAXFWT vs. sex ( $n = 95$ )

model. This model is a general procedure for assessing the mean difference in a normally distributed outcome variable between groups after controlling for one or more confounding variables (sometimes called *covariates*). Groups are defined by a categorical variable, which may have two or more categories. The covariates may be any combination of either continuous (e.g., age) or categorical (e.g., sex) variables. We discuss analysis-of-covariance models in more detail in Chapter 12.

## 11.11 Partial and Multiple Correlation

### Partial Correlation

The correlation coefficient is a measure of linear association between two variables  $x$  and  $y$ . In some instances, it is important to assess the degree of association between two variables after controlling for other covariates. The partial correlation accomplishes this goal.

**Figure 11.28** Studentized residual plotted against predicted MAXFWT ( $n = 95$ )**Definition 11.20**

Suppose we are interested in the association between two variables  $x$  and  $y$  but want to control for other covariates  $z_1, \dots, z_k$ . The **partial correlation** is defined to be the Pearson correlation between two derived variables  $e_x$  and  $e_y$ , where

$e_x$  = the residual from the linear regression of  $x$  on  $z_1, \dots, z_k$

$e_y$  = the residual from the linear regression of  $y$  on  $z_1, \dots, z_k$

**Example 11.51**

**Hypertension, Pediatrics** Consider the pediatric blood-pressure data given in Table 11.9, where we related SBP to birthweight and age for 16 infants. Estimate the partial correlation between SBP and each risk factor after controlling for the other risk factor.

**Solution**

Look at Table 11.10, under the last column, labeled Squared Partial Corr Type II. The partial correlation between SBP and birthweight after correcting for age is  $\sqrt{.5071} = .71$ . The partial correlation between SBP and age after controlling for birthweight is  $\sqrt{.8521} = .92$ . These relationships are displayed in the partial-residual plots of SBP on birthweight (Figure 11.25a) and age (Figure 11.25b), respectively.

**Table 11.14** Multiple-regression model comparing mean MAXFWT between exposed and control children after controlling for age and sex ( $n = 95$ )

The REG Procedure					
<b>Model:</b> MODEL1					
<b>Dependent Variable:</b> maxfwt					
		Number of Observations Read		120	
		Number of Observations Used		95	
		Number of Observations with Missing Values		25	
Analysis of Variance					
Sum of Mean					
<b>Source</b>	<b>DF</b>	<b>Squares</b>	<b>Square</b>	<b>F Value</b>	<b>Pr &gt; F</b>
Model	3	5994.81260	1998.27087	39.16	<.0001
Error	91	4643.12424	51.02334		
Corrected Total	94	10638			
Root MSE		7.14306	R-Square	0.5635	
Dependent Mean		52.85263	Adj R-Sq	0.5491	
Coeff Var		13.51506			
Parameter Estimates					
Parameter Standard					
<b>Variable</b>	<b>DF</b>	<b>Estimate</b>	<b>Error</b>	<b>t Value</b>	<b>Pr &gt;  t </b>
Intercept	1	34.12129	3.09869	11.01	<.0001
cscn2	1	-5.14717	1.55831	-3.30	0.0014
ageyr	1	2.44202	0.24540	9.95	<.0001
sex	1	-2.38521	1.50808	-1.58	0.1172

## Multiple Correlation

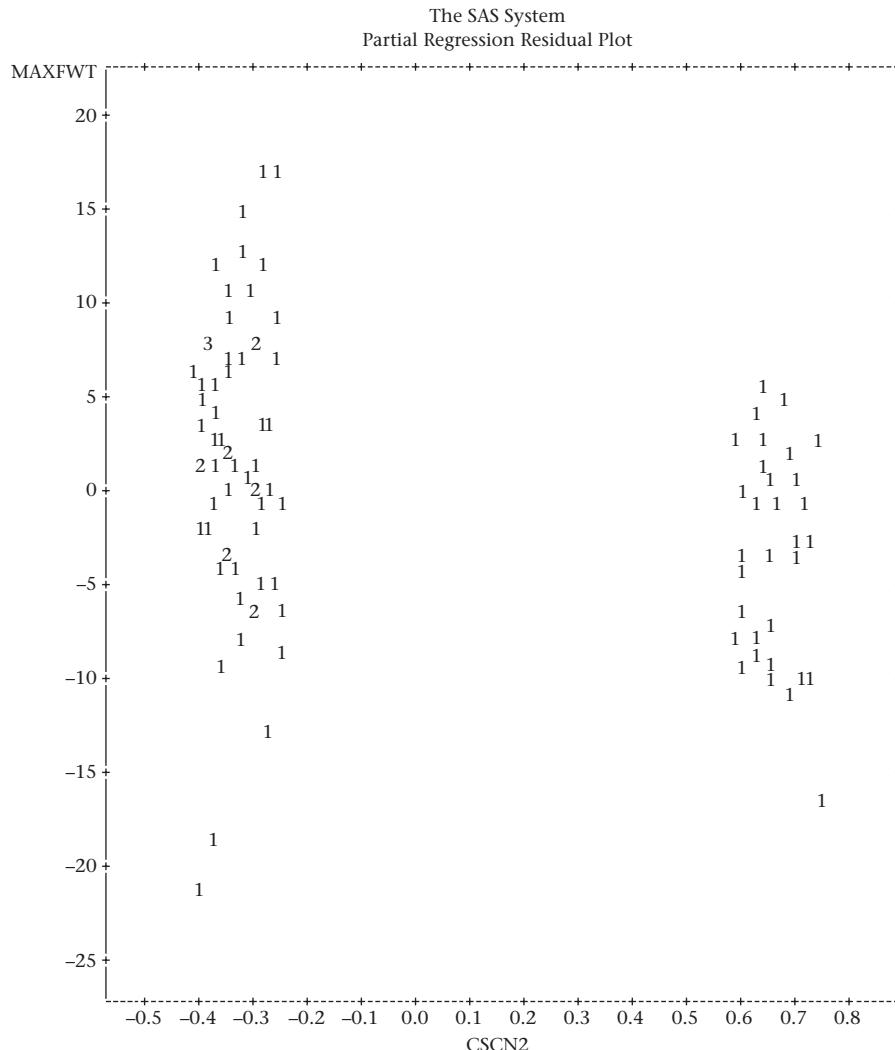
The partial-correlation coefficient provides a measure of association between two variables while controlling for the effects of one or more other covariates. In a multiple-regression setting, where we have an outcome variable ( $y$ ) and there are two or more predictor variables ( $x_1, \dots, x_k$ ), then the partial correlation between  $y$  and a single predictor  $x_j$  is a measure of the specific association between  $y$  and  $x_j$ , while controlling for all other predictors  $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k$ . However, we are also often interested in the association between  $y$  and *all* the predictors when considered as a group. This measure of association is given by the multiple correlation.

### Definition 11.21

Suppose we have an outcome variable  $y$  and a set of predictors  $x_1, \dots, x_k$ . The maximum possible correlation between  $y$  and a linear combination of the predictors  $c_1x_1 + \dots + c_kx_k$  is given by the correlation between  $y$  and the regression function  $\beta_1x_1 + \dots + \beta_kx_k$  and is called the **multiple correlation** between  $y$  and  $\{x_1, \dots, x_k\}$ . It is estimated by the Pearson correlation between  $y$  and  $b_1x_1 + \dots + b_kx_k$ , where  $b_1, \dots, b_k$  are the least-squares estimates of  $\beta_1, \dots, \beta_k$ . The multiple correlation can also be shown to be equivalent to  $\sqrt{\text{Reg SS}/\text{Total SS}} = \sqrt{R^2}$  from Equation 11.31.

### Example 11.52

**Hypertension, Pediatrics** Compute the multiple correlation between SBP and the predictors (birthweight, age) based on the data in Table 11.9.

**Figure 11.29** Partial-residual plot of MAXFWT on CSCN2 after correcting for age and sex ( $n = 95$ )

**Solution** Refer to Table 11.10. The  $R^2$  for the regression model is  $591.04/670.94 = .8809$ . The multiple correlation is  $\sqrt{.8809} = .94$ . This indicates a strong association between  $y$  and the set of predictors {birthweight, age}.

## 11.12 Rank Correlation

Sometimes we may want to look at the relationship between two variables, but one or both of the variables are either ordinal or have a distribution that is far from normal. Significance tests based on the Pearson correlation coefficient will then no longer be valid, and nonparametric analogs to these tests are needed.

### Example 11.53

**Obstetrics** The Apgar score was developed in 1952 as a measure of the physical condition of an infant at 1 and 5 minutes after birth [7]. The score is obtained by summing five components, each of which is rated as 0, 1, or 2 and represents

different aspects of the condition of an infant at birth [8]. The method of scoring is displayed in Table 11.15. The score is routinely calculated for most newborn infants in U.S. hospitals. Suppose we are given the data in Table 11.16. We wish to relate the Apgar scores at 1 and 5 minutes and to assess the significance of this relationship. How should this be done?

Text not available due to copyright restrictions

**Table 11.16 Apgar scores at 1 and 5 minutes for 24 newborns**

Infant	Apgar score, 1 min	Apgar score, 5 min	Infant	Apgar score, 1 min	Apgar score, 5 min
1	10	10	13	6	9
2	3	6	14	8	10
3	8	9	15	9	10
4	9	10	16	9	10
5	8	9	17	9	10
6	9	10	18	9	9
7	8	9	19	8	10
8	8	9	20	9	9
9	8	9	21	3	3
10	8	9	22	9	9
11	7	9	23	7	10
12	8	9	24	10	10

The ordinary correlation coefficient developed in Section 11.7 should not be used because the significance of this measure can be assessed only if the distribution of each Apgar score is assumed to be normally distributed. Instead, a nonparametric analog to the correlation coefficient based on ranks is used.

#### Definition 11.22

The **Spearman rank-correlation coefficient ( $r_s$ )** is an ordinary correlation coefficient based on ranks.

$$\text{Thus } r_s = \frac{L_{xy}}{\sqrt{L_{xx} \times L_{yy}}}$$

where the  $L$ 's are computed from the ranks rather than from the actual scores.

The rationale for this estimator is that if there were a perfect positive correlation between the two variables, then the ranks for each person on each variable would be the same and  $r_s = 1$ . The less perfect the correlation, the closer to zero  $r_s$  would be.

**Example 11.54**

**Obstetrics** Compute the Spearman rank-correlation coefficient for the Apgar-score data in Table 11.16.

**Solution**

We use MINITAB to rank the Apgar 1-minute and 5-minute scores as shown in Table 11.17 under APGAR\_1R and APGAR\_5R, respectively. Average ranks are used for tied values. We then compute the correlation coefficient between APGAR\_1R and APGAR\_5R and obtain  $r_s = .593$ .

We would now like to test the rank correlation for statistical significance. A similar test to that given in Equation 11.20 for the Pearson correlation coefficient can be performed, as follows.

**Equation 11.39****t Test for Spearman Rank Correlation**

- (1) Compute the test statistic

$$t_s = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$$

which under the null hypothesis of no correlation follows a  $t$  distribution with  $n - 2$  degrees of freedom.

- (2) For a two-sided level  $\alpha$  test,

if  $t_s > t_{n-2, 1-\alpha/2}$  or  $t_s < t_{n-2, \alpha/2} = -t_{n-2, 1-\alpha/2}$   
then reject  $H_0$ ; otherwise, accept  $H_0$ .

- (3) The exact  $p$ -value is given by

$$\begin{aligned} p &= 2 \times (\text{area to the left of } t_s \text{ under a } t_{n-2} \text{ distribution}) && \text{if } t_s < 0 \\ p &= 2 \times (\text{area to the right of } t_s \text{ under a } t_{n-2} \text{ distribution}) && \text{if } t_s \geq 0 \end{aligned}$$

- (4) This test is valid only if  $n \geq 10$ .

The acceptance and rejection regions for this test are given in Figure 11.30. The computation of the exact  $p$ -value is illustrated in Figure 11.31.

**Example 11.55**

**Obstetrics** Perform a significance test for the Spearman rank-correlation coefficient based on the Apgar-score data in Table 11.16.

**Solution**

Note that  $r_s = .593$  from Example 11.54. The test statistic is given by

$$t_s = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{.593 \sqrt{22}}{\sqrt{1-.593^2}} = \frac{2.780}{.805} = 3.45$$

which follows a  $t_{22}$  distribution under  $H_0$ . Note that

$$t_{22,.995} = 2.819 \quad t_{22,.9995} = 3.792$$

Thus, the two-tailed  $p$ -value is given by

$$2 \times (1 - .9995) < p < 2 \times (1 - .995) \text{ or } .001 < p < .01$$

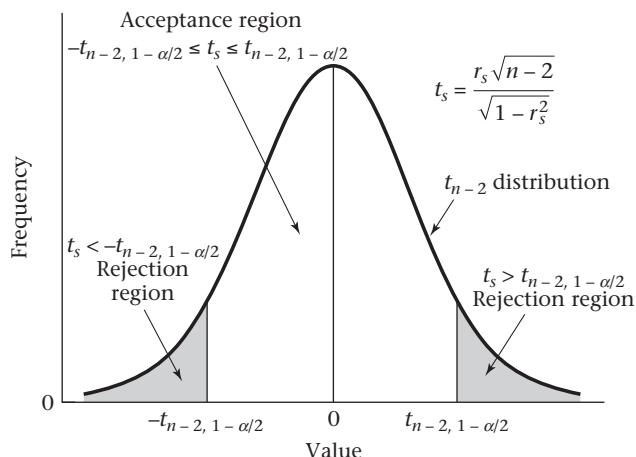
Thus there is a significant rank correlation between the two scores.

Note that the test procedure given in Equation 11.39 is valid only for  $n \geq 10$ . If  $n < 10$ , then the  $t$  distribution is not a good approximation to the distribution of  $t_s$ , and a table giving exact significance levels must be used. For this purpose, exact

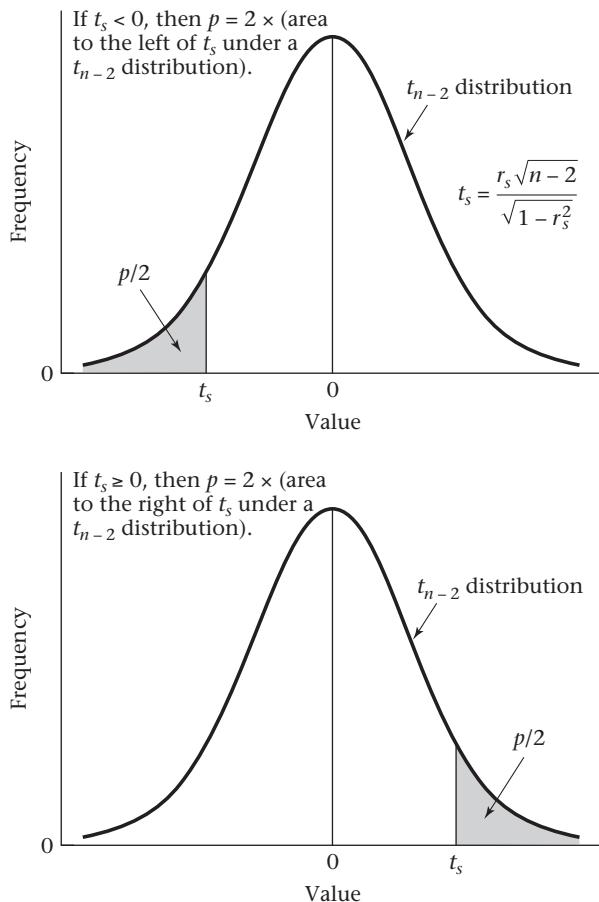
**Table 11.17** Ranks of Apgar-score data from Table 11.16

Row	APGAR_1M	APGAR_5M	APGAR_1R	APGAR_5R
1	10	10	23.5	19.5
2	3	6	1.5	2.0
3	8	9	10.0	8.5
4	9	10	18.5	19.5
5	8	9	10.0	8.5
6	9	10	18.5	19.5
7	8	9	10.0	8.5
8	8	9	10.0	8.5
9	8	9	10.0	8.5
10	8	9	10.0	8.5
11	7	9	4.5	8.5
12	8	9	10.0	8.5
13	6	9	3.0	8.5
14	8	10	10.0	19.5
15	9	10	18.5	19.5
16	9	10	18.5	19.5
17	9	10	18.5	19.5
18	9	9	18.5	8.5
19	8	10	10.0	19.5
20	9	9	18.5	8.5
21	3	3	1.5	1.0
22	9	9	18.5	8.5
23	7	10	4.5	19.5
24	10	10	23.5	19.5

**Correlations:** APGAR\_1R, APGAR\_5R  
**Pearson correlation of APGAR\_1R and APGAR\_5R = 0.593**  
**P-Value = 0.002**

**Figure 11.30** Acceptance and rejection regions for the  $t$  test for a Spearman rank-correlation coefficient

**Figure 11.31 Computation of the exact  $p$ -value for the  $t$  test for a Spearman rank-correlation coefficient**



two-sided critical values for  $r_s$  when  $n \leq 9$  are presented in Table 14 in the Appendix. This table can be used in the following way:

- (1) Suppose the critical value in the table for significance level  $\alpha$  is  $c$ .
- (2) Reject  $H_0$  using a two-sided test with significance level  $\alpha$  if  $r_s \geq c$  or  $r_s \leq -c$ , and accept  $H_0$  otherwise.

**Example 11.56** Suppose  $r_s = .750$  based on a sample of size 9. Assess the statistical significance of the results.

**Solution**

From Table 14 in the Appendix, the critical value for  $\alpha = .05$ ,  $n = 9$  is  $.683$  and for  $\alpha = .02$ ,  $n = 9$  is  $.783$ . Because  $.683 < .750 < .783$ , it follows that the two-tailed  $p$ -value is given by  $.02 \leq p < .05$ .

In this section, we have discussed rank-correlation methods. These methods are used when we are interested in studying the association between two continuous variables, where at least one of the variables is not normally distributed. On the flowchart at the end of this chapter (Figure 11.32, p. 503), we answer yes to (1) interested in relationships between two variables? and (2) both variables continuous? no

to (3) interested in predicting one variable from another? which leads to (4) interested in studying the correlation between two variables, and no to (5) both variables normal? This leads us to the box labeled “rank-correlation methods.”

Rank-correlation methods are also useful for studying the association between two ordinal variables. Again referring to the flowchart at the end of this chapter (Figure 11.32, p. 503), we answer yes to (1) interested in relationships between two variables? no to (2) both variables continuous? no to (3) one variable continuous and one categorical? and yes to (4) ordinal data? This also leads us to the box labeled “rank-correlation methods.”

### REVIEW QUESTIONS 11E

- 1 What is the difference between an ordinary (Pearson) correlation coefficient and a partial-correlation coefficient?
- 2 Suppose that the Pearson correlation coefficient between  $y$  and  $x_1$  is .30 but that the partial-correlation coefficient between  $y$  and  $x_1$ , controlling for  $x_2$ , is 0. How do you interpret the results?
- 3 What is the difference between an ordinary (Pearson) correlation coefficient and a multiple-correlation coefficient?
- 4 Under what circumstances do we prefer to use rank correlation as a measure of association instead of ordinary (Pearson) correlation?

## 11.13 Interval Estimation for Rank-Correlation Coefficients

It is worthwhile considering what underlying parameter we wish to estimate with a rank-correlation coefficient. Suppose  $X_i, Y_i$  are the scores for the  $i$ th subject,  $i = 1, \dots, n$  with corresponding sample ranks denoted by  $\text{rank}(X_i)$  and  $\text{rank}(Y_i)$ , respectively. Let us define the sample percentile of  $X_i$  and  $Y_i$  by  $\hat{P}_i = \text{rank}(X_i) / (n+1)$ ,  $\hat{P}_i^* = \text{rank}(Y_i) / (n+1)$ . These are sample estimates of the cumulative distribution function (cdf) for  $X$  and  $Y$  for the  $i^{\text{th}}$  subject. Since dividing by a constant ( $n + 1$ ) does not affect the rank of an individual observation, the Spearman rank-correlation coefficient  $r_s$  can be defined by

$$r_s = \text{corr}[\text{rank}(X_i), \text{rank}(Y_i)] = \text{corr}(\hat{P}_i, \hat{P}_i^*)$$

If  $n \rightarrow \infty$ ,  $\hat{P}_i$  and  $\hat{P}_i^*$  will approach  $P_i$  and  $P_i^*$  where

$$F_X(X_i) = P_i, F_Y(Y_i) = P_i^*$$

and  $F_X, F_Y$  are the true cdfs (or percentiles) of  $X$  and  $Y$  in the reference population.  $P_i, P_i^*$  are sometimes referred to as the *grades* of  $X_i, Y_i$  in the reference population. The underlying rank correlation we are trying to estimate is the correlation between the grades of  $X$  and  $Y$  for the  $i$ th subject given by

$$\rho_s = \text{corr}(P_i, P_i^*)$$

where  $r_s$  is the sample (point) estimate of  $\rho_s$ . It is also of interest to obtain interval estimates for  $\rho_s$ .

We cannot use the method used for interval estimates for Pearson correlation coefficients given in Equation 11.23 because the sampling distribution of the Spearman rank-correlation coefficient ( $r_s$ ) is different from that of the Pearson correlation coefficient ( $r$ ). However, a valid method based on sample probits can be used. [9]

**Equation 11.40****Interval Estimation for Spearman Rank-Correlation Coefficients**

Suppose we have an estimated Spearman rank-correlation  $r_s$  based on a sample of size  $n$ . To obtain an approximate two-sided  $100\% \times (1 - \alpha)$  confidence interval for  $\rho_s$  (the underlying rank correlation) we proceed as follows:

- (1) Compute the sample probit  $\hat{H}_i$  and  $\hat{H}_i^*$  corresponding to  $X_i, Y_i$ , where  $\hat{H}_i = \Phi^{-1}(\hat{P}_i)$ ,  $\hat{H}_i^* = \Phi^{-1}(\hat{P}_i^*)$  and  $\hat{P}_i = \text{rank}(X_i)/(n+1)$  and  $\hat{P}_i^* = \text{rank}(X_i^*)/(n+1)$ . The probit has previously been referred to as the inverse normal distribution in Chapter 5. Thus, probit(0.5) =  $z_{.5} = 0$ , probit(0.975) =  $z_{.975} = 1.96$ , etc.
- (2) Compute the Pearson correlation  $r$  between sample probits given by

$$r_h = \text{corr}(\hat{H}_i, \hat{H}_i^*)$$

which is a sample estimate of the probit correlation  $\rho_h = \text{corr}(H_i, H_i^*)$  where  $H_i = \Phi^{-1}(P_i)$ ,  $H_i^* = \Phi^{-1}(P_i^*)$ .

- (3) Because  $r_h$  is a slightly negatively biased estimate of  $\rho_h$  (Olkin and Pratt [10]), we compute the bias-corrected estimator of  $\rho_h$  given by

$$r_{\text{cor},h} = r_h \left\{ 1 + (1 - r_h^2) / [2(n - 4)] \right\}.$$

- (4) Let  $z_h$  = Fisher's z-transform of  $\rho_h \equiv 0.5 \ln[(1 + \rho_h) / (1 - \rho_h)]$ .
- (5) Compute a  $100\% \times (1 - \alpha)$  confidence interval for  $z_h$  given by

$$(z_{1h}, z_{2h}) = \hat{z}_h \pm z_{1-\alpha/2} / \sqrt{n-3}$$

where  $\hat{z}_h$  = Fisher's z-transform of  $r_{\text{cor},h} = 0.5 \ln[(1 - r_{\text{cor},h}) / (1 - r_{\text{cor},h})]$ .

- (6) The corresponding  $100\% \times (1 - \alpha)$  confidence interval for  $\rho_h$  is  $(r_{1h}, r_{2h})$ , where

$$r_{1h} = [\exp(2z_{1h}) - 1] / [\exp(2z_{1h}) + 1], \quad r_{2h} = [\exp(2z_{2h}) - 1] / [\exp(2z_{2h}) + 1].$$

- (7) Furthermore, a  $100\% \times (1 - \alpha)$  confidence interval for  $\rho_s$  is given by  $(r_{s1}, r_{s2})$ , where

$$(r_{s1}, r_{s2}) = [(6 / \pi) \sin^{-1}(r_{1h} / 2), (6 / \pi) \sin^{-1}(r_{2h} / 2)].$$

- (8) This procedure is valid for  $n \geq 10$ . The rationale for this procedure is that for normally distributed scales such as  $H$  and  $H^*$ , there is a relationship between the underlying rank correlation and Pearson correlation given by Moran [11]

$$\rho_{s,h} = (6 / \pi) \sin^{-1}(\rho_h / 2)$$

where  $\rho_h = \text{corr}(H_i, H_i^*)$  and  $\rho_{s,h} = \text{corr}(P_i, P_i^*)$ . However, because the probit transformation is rank-preserving,  $P_i$  and  $P_i^*$  are the same in the probit scale and the original scale. Thus,  $\rho_{s,h} = \rho_s = \text{corr}(P_i, P_i^*)$ .

**Example 11.57**

**Obstetrics** Obtain a 95% confidence interval for the underlying Spearman rank-correlation  $\rho_s$  for the Apgar score data in Tables 11.16 and 11.17.

**Solution**

We begin with the table of ranks in Table 11.17 and compute the sample cdfs  $(\hat{P}_i, \hat{P}_i^*)$  and the corresponding sample probits  $\hat{H}_i, \hat{H}_i^*$ . We use Excel where the probit

**Table 11.18 Computation of 95% confidence limits for the Apgar score data in Table 11.16**

observation	APGAR_1R	APGAR_5R	P_i	P_i*	H_i	H_i*	APGAR_1M	APGAR_5M
1	23.5	19.5	0.94	0.78	1.555	0.772	10	10
2	1.5	2	0.06	0.08	-1.555	-1.405	3	6
3	10	8.5	0.4	0.34	-0.253	-0.412	8	9
4	18.5	19.5	0.74	0.78	0.643	0.772	9	10
5	10	8.5	0.4	0.34	-0.253	-0.412	8	9
6	18.5	19.5	0.74	0.78	0.643	0.772	9	10
7	10	8.5	0.4	0.34	-0.253	-0.412	8	9
8	10	8.5	0.4	0.34	-0.253	-0.412	8	9
9	10	8.5	0.4	0.34	-0.253	-0.412	8	9
10	10	8.5	0.4	0.34	-0.253	-0.412	8	9
11	4.5	8.5	0.18	0.34	-0.915	-0.412	7	9
12	10	8.5	0.4	0.34	-0.253	-0.412	8	9
13	3	8.5	0.12	0.34	-1.175	-0.412	6	9
14	10	19.5	0.4	0.78	-0.253	0.772	8	10
15	18.5	19.5	0.74	0.78	0.643	0.772	9	10
16	18.5	19.5	0.74	0.78	0.643	0.772	9	10
17	18.5	19.5	0.74	0.78	0.643	0.772	9	10
18	18.5	8.5	0.74	0.34	0.643	-0.412	9	9
19	10	19.5	0.4	0.78	-0.253	0.772	8	10
20	18.5	8.5	0.74	0.34	0.643	-0.412	9	9
21	1.5	1	0.06	0.04	-1.555	-1.751	3	3
22	18.5	8.5	0.74	0.34	0.643	-0.412	9	9
23	4.5	19.5	0.18	0.78	-0.915	0.772	7	10
24	23.5	19.5	0.94	0.78	1.555	0.772	10	10

$$H_i = \Phi^{-1}(P_i) = \text{NORMSINV}(P_i)$$

$$H_i^* = \Phi^{-1}(P_i^*) = \text{NORMSINV}(P_i^*)$$

We then computed the sample probit correlation  $r_h = \text{corr}(H_i, H_i^*) = 0.645$ , the corrected sample probit correlation  $= r_{\text{cor}}_h = r_h \{1 + (1 - r_h^2)/[2(20)]\} = 0.654$  and its z transform

$$z_h = 0.5 \times \ln [(1 + r_{\text{cor}}_h)/(1 - r_{\text{cor}}_h)] = 0.783$$

We then computed the 95% confidence limits for  $z_h$  (Step 5) given by

$$z1_h = z_h - 1.96 / \sqrt{21} = 0.355$$

$$z2_h = z_h + 1.96 / \sqrt{21} = 1.210$$

We then obtain the corresponding 95% confidence limits for  $p_h$  (Step 6) given by  $(r1_h, r2_h)$ , where

$$r1_h = \{\exp[2(0.355)] - 1\}/\{\exp[2(0.355)] + 1\} = 0.341$$

$$r2_h = \{\exp[2(1.210)] - 1\}/\{\exp[2(1.210)] + 1\} = 0.837$$

Finally, we obtain the 95% confidence limits for the rank correlation  $\rho_s$  (Step 7) given by  $(rs1, rs2)$ , where

$$rs1 = (6/\pi) \sin^{-1}(0.341/2) = (6/\pi) \text{ARSIN}(0.341/2) = 0.327$$

$$rs2 = (6/\pi) \sin^{-1}(0.837/2) = (6/\pi) \text{ARSIN}(0.837/2) = 0.824$$

The corresponding point estimate of  $\rho_s$  is the Spearman rank correlation  $r_s = \text{corr}(\text{APGAR\_1R}, \text{APGAR\_5R}) = 0.593$ .

It is instructive to compare the Pearson and Spearman correlations in this example. Thus, the original Apgar scores at 1 and 5 minutes are given in the columns labeled APGAR\_1M, APGAR\_5M. The Pearson correlation between these scores is  $r_{\text{raw}} = \text{corr}(\text{APGAR\_1M}, \text{APGAR\_5M}) = 0.845$ . Using the Fisher's z transform method of obtaining confidence intervals for Pearson correlations (Equation 11.23), we obtain the 95% confidence interval for  $\rho$  given by

$$(r1_{\text{raw}}, r2_{\text{raw}}) = (0.670, 0.931), \text{ where}$$

$$r1_{\text{raw}} = \{\exp[2(z1_{\text{raw}})] - 1\}/\{\exp[2(z1_{\text{raw}})] + 1\}$$

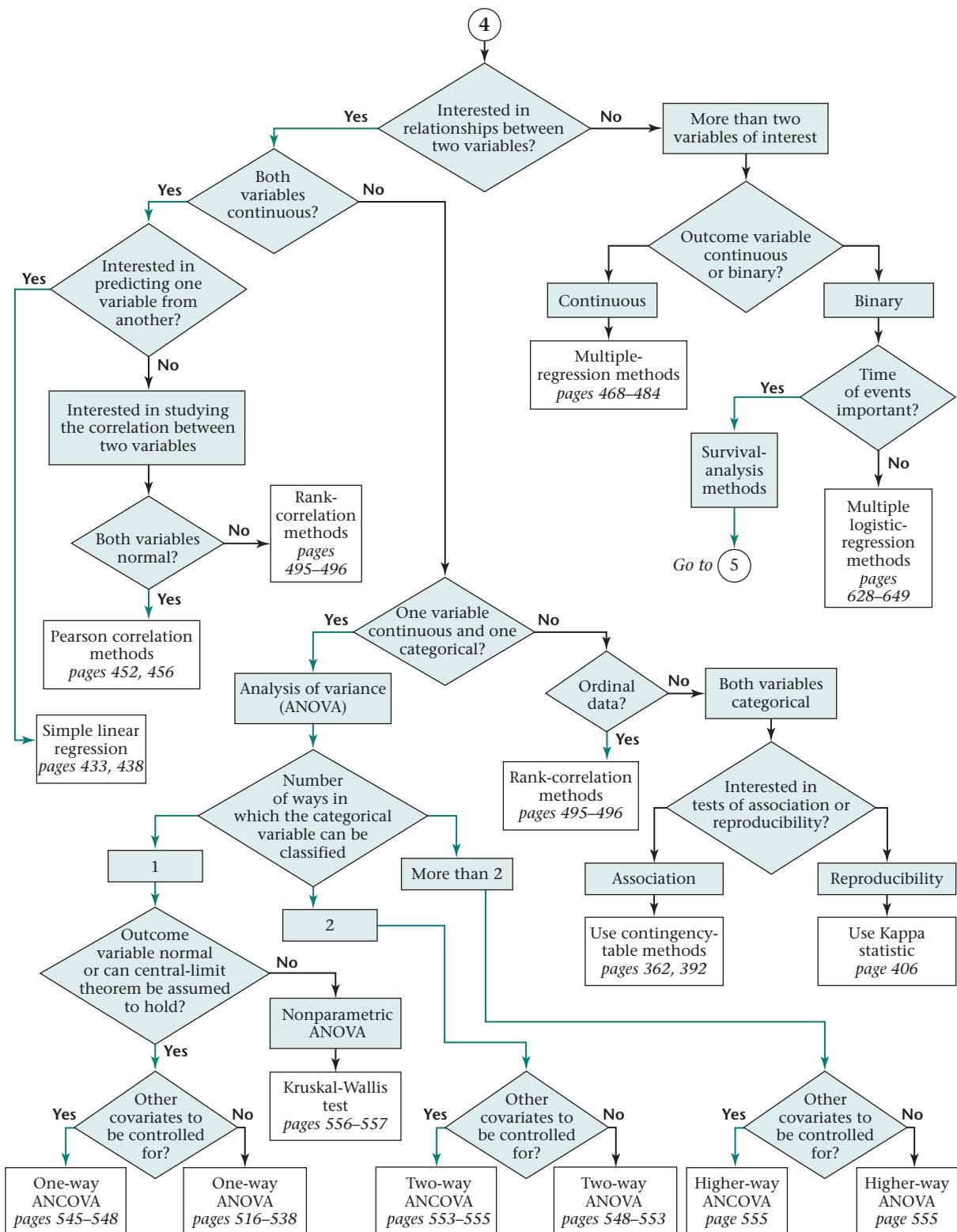
$$r2_{\text{raw}} = \{\exp[2(z2_{\text{raw}})] - 1\}/\{\exp[2(z2_{\text{raw}})] + 1\}$$

$$\text{and } (z1_{\text{raw}}, z2_{\text{raw}}) = z_{\text{raw}} \pm 1.96/\sqrt{21} = (0.810, 1.666)$$

$$z_{\text{raw}} = 0.5 \ln [(1 + r_{\text{raw}})/(1 - r_{\text{raw}})] = 1.238$$

There is a large difference between the point estimates for the Pearson (0.845) and Spearman (0.593) correlations. The corresponding interval estimates (Pearson, 0.670–0.931; Spearman, 0.327–0.824) are also very different. Based on the ordinal nature of the rankings, the Spearman rank correlation is preferable in this case.

A SAS macro implementing the procedures for obtaining interval estimates for Spearman rank-correlation coefficients is available at the following website: <https://sites.google.com/a/channing.harvard.edu/bernardrosner/home/>.

**Figure 11.32** Flowchart for appropriate methods of statistical inference

## 11.14 Summary

In this chapter, we studied methods of statistical inference that are appropriate for investigating the relationship between two or more variables. If only two variables, both of which are continuous, are being studied, and we wish to predict one variable (the dependent variable) as a function of the other variable (the independent variable), then simple linear-regression analysis is used. If we simply want to look at the association between two normally distributed variables without distinguishing between dependent and independent variables, then Pearson correlation methods are more appropriate. If both variables are continuous but are not normally distributed, or are ordinal variables, then rank correlation can be used instead. If both variables of interest are categorical and we are interested in the association between the two variables, then the contingency-table methods of Chapter 10 can be used. If, in contrast, we are almost certain there is some association between the two variables and we want to quantify the degree of association, then the Kappa statistic can be used.

In many instances we are interested in more than two variables and we want to predict the value of one variable (the dependent variable) as a function of several independent variables. If the dependent variable is normally distributed, then multiple-regression methods can be used. Multiple-regression methods can be very powerful because the independent variables can be either continuous or categorical, or a combination of both.

In many situations we have a continuous outcome variable that we want to relate to one or more categorical variables. In general, this situation can be handled with ANOVA methods. However, in many instances the formulation is easier if multiple-regression methods are used. We will discuss these alternative approaches in Chapter 12. The preceding methods are summarized in the flowchart in Figure 11.32 (p. 503) and again in the back of the book.

## PROBLEMS

### Hematology

The data in Table 11.19 are given for 9 patients with aplastic anemia [12].

\***11.1** Fit a regression line relating the percentage of reticulocytes ( $x$ ) to the number of lymphocytes ( $y$ ).

\***11.2** Test for the statistical significance of this regression line using the  $F$  test.

\***11.3** What is  $R^2$  for the regression line in Problem 11.1?

\***11.4** What does  $R^2$  mean in Problem 11.3?

\***11.5** What is  $s_{y,x}^2$ ?

\***11.6** Test for the statistical significance of the regression line using the  $t$  test.

\***11.7** What are the standard errors of the slope and intercept for the regression line in Problem 11.1?

**11.8** What is the  $z$  transformation of .34?

**Table 11.19** Hematologic data for patients with aplastic anemia

Patient number	% Reticulocytes	Lymphocytes (per mm <sup>3</sup> )
1	3.6	1700
2	2.0	3078
3	0.3	1820
4	0.3	2706
5	0.2	2086
6	3.0	2299
7	0.0	676
8	1.0	2088
9	2.2	2013

Source: Reprinted with permission of *The New England Journal of Medicine*, 312(16), 1015–1022, 1985.

### Pulmonary Function

Suppose the correlation coefficient between FEV for 100 sets of identical twins is .7, whereas the comparable correlation for 120 sets of fraternal twins is .38.

\***11.9** What test procedure can be used to compare the two correlation coefficients?

\***11.10** Perform the procedure in Problem 11.9 using the critical-value method.

\***11.11** What is the *p*-value of the test?

Suppose the correlation coefficient between weight is .78 for the 100 sets of identical twins and .50 for the 120 sets of fraternal twins.

\***11.12** Test for whether the true correlation coefficients differ between these groups. Report a *p*-value.

### Obstetrics

The data in Table 11.20 give the infant-mortality rates per 1000 livebirths in the United States for the period 1960–2005 [13].

**Table 11.20 U.S. infant-mortality rates per 1000 livebirths, 1960–2005**

<i>x</i> *	<i>y</i> *	<i>x</i>	<i>y</i>
1960	26.0	1985	10.6
1965	24.7	1990	9.2
1970	20.0	1995	7.6
1975	16.1	2000	6.9
1980	12.6	2005	6.9

\**x* = year, *y* = infant-mortality rate per 1000 live births.

\***11.13** Fit a linear-regression line relating infant-mortality rate to chronological year using these data. Use a data transformation if necessary.

\***11.14** Test for the significance of the linear relationship developed in Problem 11.13.

\***11.15** If the present trends continue for the next 5 years, then what would be the predicted infant-mortality rate in 2010?

\***11.16** Provide a standard error for the estimate in Problem 11.15.

\***11.17** Can the linear relationship developed in Problem 11.13 be expected to continue indefinitely? Why or why not?

### Hypertension

The Update to the Task Force Report on Blood Pressure Control in Children [14] reported the observed 90th percentile of SBP in single years of age from age 1 to 17 based

on prior studies. The data for boys of average height are given in Table 11.21.

Suppose we seek a more efficient way to display the data and choose linear regression to accomplish this task.

**11.18** Fit a regression line relating age to SBP, using the data in Table 11.21.

**Table 11.21 90th percentile of SBP in boys ages 1–17 of average height**

Age ( <i>x</i> )	SBP <sup>a</sup> ( <i>y</i> )	Age ( <i>x</i> )	SBP <sup>a</sup> ( <i>y</i> )
1	99	10	115
2	102	11	117
3	105	12	120
4	107	13	122
5	108	14	125
6	110	15	127
7	111	16	130
8	112	17	132
9	114		

<sup>a</sup>90th percentile for each 1-year age group.

**11.19** Provide a 95% confidence interval for the parameters of the regression line.

**11.20** What is the predicted blood pressure for an average 13-year-old boy as estimated from the regression line?

**11.21** What is the standard error of the estimate in Problem 11.20?

**11.22** Answer Problems 11.20 and 11.21 for a 17-year-old boy.

**11.23** Do you think the linear regression provides a good fit to the data? Why or why not? Use residual analysis to justify your answer.

### Cancer

The following statistics are taken from an article by Burch relating cigarette smoking to lung cancer [15]. The article presents data relating mortality from lung cancer to average cigarette consumption (lb/person) for females in England and Wales over a 40-year period. The data are given in Table 11.22.

$$\left( \begin{array}{l} \sum_{i=1}^8 x_i = 2.38, \quad \sum_{i=1}^8 x_i^2 = 1.310, \quad \sum_{i=1}^8 y_i = -15.55, \\ \sum_{i=1}^8 y_i^2 = 30.708, \quad \sum_{i=1}^8 x_i y_i = -4.125 \end{array} \right)$$

**11.24** Compute the correlation between 5-year lung-cancer mortality and annual cigarette consumption when each is expressed in the  $\log_{10}$  scale.

**11.25** Test this correlation for statistical significance, and report a *p*-value.

**Table 11.22 Cigarette consumption and lung-cancer mortality in England and Wales, 1930–1969**

Period	$\log_{10}$ mortality (over 5 years), <i>y</i>	$\log_{10}$ annual cigarette consumption (lb/person), <i>x</i>
1930–1934	−2.35	−0.26
1935–1939	−2.20	−0.03
1940–1944	−2.12	0.30
1945–1949	−1.95	0.37
1950–1954	−1.85	0.40
1955–1959	−1.80	0.50
1960–1964	−1.70	0.55
1965–1969	−1.58	0.55

Source: Reprinted with permission of the *Journal of the Royal Statistical Society, A*, 141, 437–477, 1978.

**11.26** Fit a regression line relating 5-year lung-cancer mortality to annual cigarette consumption.

**11.27** To test the significance of this regression line, is it necessary to perform any additional tests other than those in Problem 11.25? If so, perform them.

**11.28** What is the expected lung-cancer mortality rate with an annual cigarette consumption of 1 lb/person?

**11.29** Why are the variables mortality rate and annual cigarette consumption expressed in the log scale?

### Hypertension

Refer to the data in Table 3.10. Another method for relating measures of reactivity for the automated and manual blood pressures is the correlation coefficient. Suppose the correlation coefficient relating these two measures of reactivity is .19, based on 79 people having reactivity measured by each type of blood-pressure monitor.

**11.30** What is the appropriate procedure to test if there is a relationship between reactivity as measured by the automated and manual monitors?

**11.31** Conduct the test procedure in Problem 11.30, and report a *p*-value. What do the results mean, in words?

**11.32** Provide a 95% confidence interval for the correlation coefficient between these two measures of reactivity.

### Nutrition

Refer to Data Set VALID.DAT on the Companion Website.

**11.33** Assess the agreement between the food-frequency questionnaire and the dietary record with regard to total fat intake, saturated fat intake, alcohol intake, and total caloric

intake. Quantify the level of agreement by representing the dietary intake both in the original continuous scale and on a quintile scale.

### Pulmonary Disease

Refer to Data Set FEV.DAT on the Companion Website.

**11.34** Use regression methods to look at the relationship between level of pulmonary function and other factors (e.g., age, height, and personal smoking) when considered separately and simultaneously. Assess the goodness of fit of the models you develop. Perform analyses for males and females combined (where gender is controlled for in the analysis), as well as gender-specific analyses.

### Hepatic Disease

Refer to Data Set HORMONE.DAT on the Companion Website.

**11.35** Use methods of linear-regression analysis to assess whether these are dose–response relationships with regard to biliary and pancreatic secretion levels. Perform separate analyses for each of the four active hormones tested.

**11.36** Use methods of linear-regression analysis to assess whether there are dose–response relationships with regard to biliary and pancreatic pH levels. Perform separate analyses for each of the four active hormones tested.

### Pediatrics, Cardiovascular Disease

An important area of investigation is the development of strategies for altering cardiovascular risk factors in children. Low-density lipoprotein (LDL) cholesterol has consistently been shown to be related to cardiovascular disease in adults. A study was conducted in Bogalusa, Louisiana, and Brooks County, Texas, to identify modifiable variables that are related to LDL cholesterol in children [16]. It was found that the correlation coefficient between LDL cholesterol and ponderal index [weight (kg)/height<sup>3</sup> (cm<sup>3</sup>)], which is a measure of obesity, was .28 among 903 white boys and .14 among 474 black boys.

\***11.37** What test can be used to assess whether there is a significant association between LDL cholesterol and ponderal index?

\***11.38** Implement the test in Problem 11.37 for white boys, and report a two-sided *p*-value.

\***11.39** What test can be used to compare correlations between white and black boys?

\***11.40** Implement the test in Problem 11.39, and report a *p*-value.

\***11.41** Provide a 95% confidence interval for the true correlation among white boys ( $\rho_1$ ) and black boys ( $\rho_2$ ), respectively.

### Hypertension, Pediatrics

In Problems 6.56–6.58, we described Data Set INFANTBP.DAT (see also on the Companion Website). The data set concerns the possible relationship between infant blood pressure and responsiveness to salt and sugar, respectively. Indices were constructed summarizing responsiveness to salt and sugar.

**11.42** Use linear-regression methods to relate each salt and sugar index proposed to systolic blood pressure (SBP). Assess the goodness of fit of the proposed models; use appropriate data transformations, if necessary.

**11.43** Answer Problem 11.42 for diastolic blood pressure.

### Environmental Health, Pediatrics

Refer to Data Set LEAD.DAT on the Companion Website. In Section 11.10, we used regression methods to relate MAXFWT to lead exposure group.

**11.44** Use regression methods to assess the relationship between full-scale IQ (IQF) and lead-exposure group, where lead exposure is quantified as a categorical variable with two categories (exposed and control). Control for the possible confounding effects of age and gender in your analysis. Assess the goodness of fit of the model(s) you propose.

**11.45** LD72 and LD73 are variable names for the actual blood-lead levels in 1972 and 1973, respectively. Use regression methods to assess the relationship between MAXFWT and the actual blood-lead level(s), while controlling for age and sex. Assess the goodness of fit of the model(s) you propose.

**11.46** Answer the same questions posed in Problem 11.45 for IQF.

**11.47** Let  $\rho_2$  = correlation coefficient between MAXFWT and LD72, and  $\rho_3$  = correlation coefficient between MAXFWT and LD73. Perform a hypothesis test to statistically compare  $\rho_2$  and  $\rho_3$ . Report a two-tailed  $p$ -value.

### Pediatrics, Endocrinology

Transient hypothyroxinemia, a common finding in premature infants, is not thought to have long-term consequences or to require treatment. A study was performed to investigate whether hypothyroxinemia in premature infants is a cause of subsequent motor and cognitive abnormalities [17]. Blood thyroxine values were obtained on routine screening in the first week of life from 536 infants who weighed 2000 g or less at birth and were born at 33 weeks gestation or earlier. The data in Table 11.23 were presented concerning the relationship between mean thyroxine level and gestational age.

**11.48** What is the best-fitting regression line relating mean thyroxine level to gestational age?

**Table 11.23 Relationship between mean thyroxine level and gestational age among 536 premature infants**

(x) Gestational age (weeks)	(y) Mean thyroxine level ( $\mu\text{g}/\text{dL}$ )
$\leq 24^a$	6.5
25	7.1
26	7.0
27	7.1
28	7.2
29	7.1
30	8.1
31	8.7
32	9.5
33	10.1

<sup>a</sup>Treated as 24 in subsequent analyses.

*Hint:*

$$\sum_{i=1}^{10} x_i = 285, \quad \sum_{i=1}^{10} x_i^2 = 8205, \quad \sum_{i=1}^{10} y_i = 78.4$$

$$\sum_{i=1}^{10} y_i^2 = 627.88, \quad \sum_{i=1}^{10} x_i y_i = 2264.7$$

**11.49** Is there a significant association between mean thyroxine level and gestational age? Report a  $p$ -value.

**11.50** Assess the goodness of fit of the regression line fitted in Problem 11.48.

To test the primary hypothesis of the study, infants were categorized by gestational age and were designated as having severe hypothyroxinemia if their thyroxine concentration was 2.6  $sd$  below the mean score for the assay that their specimen came from. Thyroxine assays typically were done in batches of 240 specimens; a mean and  $sd$  were calculated for each batch, based on a sample size of 240. Children in the study were given the Bayley Mental Development Index at <30 months of age. The Bayley test is a commonly used test of mental development in young children. The results for the subgroup of children with gestational age 30–31 weeks are shown in Table 11.24.

**Table 11.24 Bayley Mental Development Index by presence or absence of severe hypothyroxinemia**

Severe hypothyroxinemia	Mean score $\pm sd$	<i>n</i>
No	$106 \pm 21$	138
Yes	$88 \pm 25$	17

**11.51** Perform a test to compare the mean Bayley score between children with and without severe hypothyroxinemia (report a *p*-value).

**11.52** Suppose we wanted to use data on children of all gestational ages in the study. Suggest a type of analysis that could be used to relate the Bayley score to severe hypothyroxinemia, while controlling for gestational age. (Do not actually carry out the analysis.)

### Hypertension

Endothelin is a powerful vasoconstrictor peptide derived from the endothelium. The contribution of endothelin to blood-pressure regulation in patients with hypertension was assessed by studying the effect of an endothelin-receptor antagonist, Bosentan. In the study, 293 patients with mild to moderate hypertension were randomized to receive one of four oral doses of Bosentan (100, 500, 1000, or 2000 mg daily), a placebo, or the ACE (angiotensin-converting enzyme) inhibitor enalapril (an established antihypertensive drug) [18]. The reported mean changes in systolic blood pressure (SBP) over a 24-hour period are shown in Table 11.25.

**Table 11.25 Mean change in SBP over 24 hours**

Group	Mean change	Bosentan dose (mg)	ln(dose)
Placebo	-0.9	1	0
100 mg bosentan	-2.5	100	4.61
500 mg bosentan	-8.4	500	6.21
1000 mg bosentan	-7.4	1000	6.91
2000 mg bosentan	-10.3	2000	7.60

**11.53** Fit a regression line relating the mean change in SBP to the ln(dose) of bosentan. (Note: For the placebo group, assume the dose of Bosentan = 1 mg; hence the ln(dose) = 0.)

**11.54** What test can be used to assess whether the mean change in SBP is significantly related to ln(dose) of Bosentan?

**11.55** Implement the method in Problem 11.54, and report a two-tailed *p*-value.

**11.56** What is the estimated mean change in SBP for an average patient taking 2000 mg of bosentan? Provide a 95% confidence interval corresponding to this estimate.

### Endocrinology

Refer to Data Set BONEDEN.DAT on the Companion Website.

**11.57** Use regression analysis to relate the number of pack-years smoked to the bone density of the lumbar spine. Assess the goodness of fit of the regression line. (*Hint:* For a twinship, relate the difference in bone density between the heavier- and lighter-smoking twin to the difference in the number of pack-years of smoking.)

**11.58** Answer the question in Problem 11.57 for bone density at the femoral neck.

**11.59** Answer the question in Problem 11.57 for bone density at the femoral shaft.

One of the issues in relating bone density to smoking is that smokers and nonsmokers differ in many other characteristics that may be related to bone density; these differences are referred to as confounders.

**11.60** Compare the weight of the heavier- vs. the lighter-smoking twin using hypothesis-testing methods.

**11.61** Repeat the analyses in Problem 11.57 controlling for weight differences between the heavier- and the lighter-smoking twins.

**11.62** Answer the question in Problem 11.61 for bone density at the femoral neck.

**11.63** Answer the question in Problem 11.61 for bone density at the femoral shaft.

**11.64** Consider other possible confounding variables in comparing the heavier- vs. lighter-smoking twin. Repeat the analyses in Problems 11.60–11.63, adjusting for these confounding variables. What is your overall conclusion regarding the possible association between bone density and smoking?

### Health Promotion

Refer to Data Set SMOKE.DAT on the Companion Website.

**11.65** Use rank-correlation methods to test whether the number of days abstinent from smoking is related to age. Compare your results with those obtained in Problem 9.29.

**11.66** Use rank-correlation methods to test whether the amount previously smoked is related to the number of days abstinent from smoking. Compare your results with those obtained in Problem 9.30.

**11.67** Use rank-correlation methods to test whether the adjusted CO level is related to the number of days abstinent from smoking. Provide a 95% confidence interval for the underlying rank correlation. Compare your results with those obtained in Problem 9.31.

### Nutrition

Refer to Data Set VALID.DAT on the Companion Website.

**11.68** Use rank-correlation methods to relate alcohol intake as reported on the diet record with alcohol intake as reported on the food-frequency questionnaire.

**11.69** Answer Problem 11.68 for total fat intake.

**11.70** Answer Problem 11.68 for saturated fat intake.

**11.71** Answer Problem 11.68 for total caloric intake.

**11.72** Do you think parametric or nonparametric methods are better suited to analyze the data in VALID.DAT? Explain.

**11.73** Suppose we have an estimated rank correlation of .45 based on a sample of size 24. Assess the significance of the results.

**11.74** Suppose we have an estimated rank correlation of .75 based on a sample of size 8. Assess the significance of the results.

### Endocrinology

A 65-year-old woman with low bone density in 1992 was treated with alendronate through the year 1999. Bone density was measured irregularly over this period. The results for change in bone density of the lumbar spine are shown in Table 11.26.

**11.75** What is the estimated rate of increase in bone density of the lumbar spine *per year*? What is the standard error of the estimated rate of increase *per year*?

**11.76** Provide a significance test to assess whether the mean bone density has changed significantly over time. Provide a two-tailed *p*-value.

**11.77** The normal change in bone density over time from age 40 to age 80 is a decrease of 0.15 g/cm<sup>2</sup>. Does the rate of change in this woman differ significantly from the expected age-related change?

Another parameter measured was bone density at the femoral neck (hip). The results are shown in Table 11.27.

**Table 11.26 Change in bone density, lumbar spine, over time**

Visit ( <i>i</i> )	Time from baseline (months) ( <i>t</i> )	Bone density, lumbar spine (g/cm <sup>2</sup> ) ( <i>x</i> )
1	0	0.797
2	8	0.806
3	18	0.817
4	48	0.825
5	64	0.837
6	66	0.841
7	79	0.886
8	92	0.881

$$\text{Note: } \sum_{i=1}^8 t_i = 375, \quad \sum_{i=8}^8 x_i = 6.69, \quad \sum_{i=1}^8 t_i^2 = 25,849,$$

$$\sum_{i=1}^8 x_i^2 = 5.60197, \quad \sum_{i=1}^8 x_i t_i = 320.874.$$

**11.78** Provide a measure of association between bone density of the lumbar spine and bone density of the femoral neck.

**11.79** Assess whether there is a significant relationship between bone density of the lumbar spine and bone density of the femoral neck. Please provide a two-tailed *p*-value.

**Table 11.27 Change in bone density, femoral neck, over time**

Visit ( <i>i</i> )	Time from baseline (months) ( <i>t</i> )	Bone density, femoral neck (g/cm <sup>2</sup> ) ( <i>y</i> )
1	0	0.643
2	8	0.638
3	18	0.648
4	48	0.674
5	64	0.640
6	66	0.676
7	79	0.651
8	92	0.680

**11.80** Assess whether the correlation coefficient between bone density and time is significantly different for the lumbar spine vs. the femoral neck. Report a two-tailed *p*-value.

### Ophthalmology

Lutein, an important carotenoid in the maintenance of ocular health, has been found postmortem in the macula of eyes. Hence, a study is planned to supplement patients with high doses of lutein in capsule form to possibly prevent age-related macular degeneration, an important eye disease that can cause partial or total blindness in large numbers of elderly people.

To assess compliance in study participants, a blood sample will be drawn. It is estimated that a serum lutein > 5 µg/dL would indicate that a participant is taking study medication.

The study began in 1999. A test sample of 9 participants had their lutein level measured in 1999 and again in 2003. The researchers found a calibration error in the 1999 assays, but the 2003 assays were correct. The data are shown in Table 11.28.

**Table 11.28 Serum-lutein data analyzed in 1999 and 2003**

Sample	1999 Serum-lutein value (µg/dL)	2003 Serum-lutein value (µg/dL)
1	3.5	6.4
2	2.9	7.5
3	4.1	8.4
4	5.1	9.6
5	6.4	12.0
6	1.9	4.2
7	1.3	3.1
8	4.1	6.3
9	2.3	4.4
Mean	3.511	6.878
sd	1.616	2.839

**11.81** Using regression methods, derive a calibration formula predicting the 2003 value as a function of the 1999 value.

**11.82** Suppose a participant had a 1999 serum-lutein value of  $5.0 \mu\text{g}/\text{dL}$ . What is that person's predicted serum lutein based on the 2003 assay? What is a 95% confidence interval around this estimate?

In 1999, 100 participants were randomized to lutein and 100 participants were randomized to placebo. Each participant had his or her blood analyzed using the old assay method in 2000 after taking the study capsules for 1 year.

The mean serum-lutein level in the active group was  $7.0 \pm 4.0$  (mean  $\pm$  *sd*) based on the old assay and is assumed to be normally distributed. The mean serum-lutein level in the placebo group was  $2.0 \pm 1.5$  (mean  $\pm$  *sd*) based on the old assay and is assumed to be normally distributed.

**11.83** If a serum-lutein value  $>5.0 \mu\text{g}/\text{dL}$  indicates that a participant is taking lutein capsules, then what percentage of the active group is complying? (Assume the old assay is correct and that the serum values can be measured exactly; that is, no continuity correction is needed.)

**11.84** If a serum-lutein value  $<5.0 \mu\text{g}/\text{dL}$  indicates that a participant is not taking lutein capsules, then what percentage of the placebo group is complying (not taking lutein capsules outside the study)? Assume that the old assay is correct and that the serum values can be measured exactly.

## Diabetes

The Mayo Clinic is based in Rochester, Minnesota. Residents of Rochester and surrounding areas get almost all their medical care at the Mayo Clinic. In addition, migration in and out of Rochester is relatively low. This makes it feasible to track disease history over time and to assess whether the incidence of disease has changed over time. The data in Table 11.29 were presented concerning the incidence of diabetes over time among Rochester, Minnesota residents ages 30 and over [19].

**Table 11.29 Age-adjusted incidence of diabetes mellitus among Rochester, Minnesota residents, age  $\geq 30$  years from 1970–1994**

Time period	Time-period score	Annual incidence <sup>a</sup>
1970–1974	1	240.4
1975–1979	2	243.1
1980–1984	3	256.7
1985–1989	4	315.9
1990–1994	5	371.8
Mean	3.0	285.6
<i>sd</i>	1.58	57.1
<i>N</i>	5	5

<sup>a</sup>Incidence rates are per 100,000 participants and are age- and sex-adjusted to the 1980 U.S. Caucasian population. For simplicity, express annual incidence per  $10^5$  participants (e.g., as 240.4 rather than .002404 for the time period 1970–1974).

**11.85** Fit a linear-regression model relating annual incidence of diabetes to time period. (For this purpose, score the time period as 1 if 1970–1974, 2 if 1975–1979, ..., 5 if 1990–1994.)

**11.86** Test for whether there has been a significant change in diabetes incidence over time.

**11.87** Based on your answers to Problems 11.85 and 11.86, make a projection of the incidence rate of diabetes mellitus during the period 1995–1999, and provide a 95% confidence interval about this estimate.

## Diabetes

A group of 10-year-old boys were first ascertained in a camp for diabetic boys. They had their weight measured at baseline and again when they returned to camp 1 year later. Each time, a serum sample was obtained from which a determination of hemoglobin A1c (HgbA1c) was made. HgbA1c (also called *glycosylated hemoglobin*) is routinely used to monitor compliance with taking insulin injections. Usually, the poorer the compliance, the higher the HgbA1c level will be. The hypothesis is that the level HgbA1c is related to weight. The data in Table 11.30 were obtained.

**11.88** What test can be performed to assess the relationship between weight and HgbA1c at the initial visit?

**11.89** Please perform the test in Problem 11.88, and report a two-tailed *p*-value.

**11.90** Do the results in Problem 11.89 imply a relationship between change in HgbA1c and change in weight for an individual boy? Why or why not?

**11.91** Compute a rank correlation between change in weight and change in HgbA1c, each over 1 year. Use this measure to directly test the hypothesis that change in weight over 1 year is related to change in HgbA1c. Report a two-tailed *p*-value, and provide a 95% confidence interval for the underlying rank correlation.

## Cancer, Endocrinology

Obesity is very common in American society and is a risk factor for breast cancer in postmenopausal women. One mechanism explaining why obesity is a risk factor is that it may raise estrogen levels in women. In particular, one type of estrogen, serum estradiol, is a strong risk factor for breast cancer. To better assess these relationships, researchers studied a group of 151 African-American and 60 Caucasian premenopausal women. Adiposity was measured in two different ways:  $\text{BMI} = \text{weight} (\text{kg})/\text{height}^2 (\text{m}^2)$  and waist-hip ratio (WHR) = waist circumference/hip circumference. BMI is a measure of overall adiposity, whereas WHR is a measure of abdominal adiposity. In addition, a complete hormonal profile was obtained, including serum estradiol (ES\_1). Finally, other breast-cancer risk factors were also assessed among these women, including (1) ethnicity (ETHNIC = 1 if African-American, = 0 if Caucasian), (2) age (ENTAGE), (3) parity (NUMCHILD = number of children), (4) age at first

**Table 11.30 Relationship between weight change and change in HgbA1c among diabetic boys**

ID no.	1st visit		2nd visit		Weight 2nd visit – weight 1st visit (kg) (y)	HgbA1c 2nd visit – HgbA1c 1st visit (%) (x)
	Weight (kg) (y)	HgbA1c % (x)	Weight (kg) (y)	HgbA1c % (x)		
1	40.5	9.9	45.5	8.4	+5.0	-1.5
2	43.2	11.3	47.0	9.2	+3.8	-2.1
3	36.3	10.4	42.0	9.6	+5.7	-0.8
4	30.8	7.4	35.3	8.1	+4.5	+0.7
5	29.9	9.8	33.2	7.9	+3.3	-1.9
6	39.7	8.5	46.1	7.7	+6.4	-0.8
7	36.9	7.3	37.8	7.7	+0.9	+0.4
8	36.4	7.8	37.0	8.4	+0.6	+0.6
9	34.5	7.8	34.3	9.6	-0.2	+1.8
10	35.2	5.7	38.4	6.5	+3.2	+0.8
11	43.0	7.4	48.6	7.4	+5.6	0
12	38.4	6.5	42.7	7.0	+4.3	+0.5
13	42.0	7.1	48.0	7.4	+6.0	+0.3
14	34.1	8.9	41.3	8.1	+7.2	-0.8
15	43.5	9.3	51.4	7.4	+7.9	-1.9
Mean	37.63	8.34	41.91	8.03	4.28	-0.31
sd	4.35	1.56	5.73	0.90	2.40	1.18
n	15	15	15	15	15	15
$\sum x_i y_i$	4722.09		5030.35		-42.18	

birth (AGEFBO), (5) any children (ANYKIDS = 1 if yes, = 0 if no), (6) age at menarche (AGEMNRCH = age when menstrual periods began). The data are provided in Data Set ESTRADL.DAT on the Companion Website.

**11.92** Is there a crude association between either measure of adiposity (BMI, WHR), considered separately, and serum estradiol?

**11.93** Are these relationships similar for Caucasian and African-American women?

**11.94** Do the relationships between the adiposity measures and serum estradiol persist after controlling for the other breast-cancer risk factors in list items 1 to 6?

**11.95** One debate in the breast-cancer literature is whether overall adiposity (BMI) or central (abdominal) adiposity (WHR) is a better indicator of breast-cancer risk. Perform analyses to inform the debate as to which measure of adiposity is more closely related to serum estradiol either crudely or after adjusting for other breast-cancer risk factors.

**11.96** It is well known that African-American women have higher levels of obesity than Caucasian women. Are there differences between estradiol levels for African-American women and Caucasian women after controlling for obesity?

## Diabetes

Refer to Data Set DIABETES.DAT on the Companion Website. Another approach to addressing Problem 8.152 is to calculate  $b$  = the rate of growth of weight over time for each individual boy, using linear-regression analysis, and relate  $b$  to  $h$  = the mean HgbA1c over all follow-up visits.

**11.97** Perform the analysis just suggested for weight. Is there a significant association between the rate of growth ( $b$ ) and mean HgbA1c ( $h$ )?

**11.98** Repeat the analysis in Problem 11.97 for height.

**11.99** Repeat the analysis in Problem 11.97 for BMI.

## Ophthalmology

Retinitis pigmentosa (RP) is a hereditary ocular disease in which patches of pigment appear on the retina, potentially resulting in substantial vision loss and in some cases complete blindness. An important issue is how fast the subjects decline. Visual field is an important measure of area of vision which is measured in degree<sup>2</sup>. A visual field area for a normal person is around 11,000 degree<sup>2</sup>. The longitudinal data in Table 11.31 were provided by an individual patient.

Suppose the rate of change of  $In$  (visual field) is a linear function of follow-up time.

**Table 11.31 Longitudinal visual field data for one RP patient**

Visit	Time (yr)	Visual field area (degree <sup>2</sup> )	<i>In</i> (visual field area)
1	0	3059	8.03
2	1	3053	8.02
3	2	1418	7.26
4	3	1692	7.43
5	4	1978	7.59
6	5	1567	7.36
7	6	1919	7.56
8	7	1998	7.60
9	11	1648	7.41
10	13	1721	7.45
11	15	1264	7.14
mean	6.09	1938	7.532
sd	4.97	597	0.280

**11.100** Write down a linear regression model that summarizes this relationship.

**11.101** Fit the regression line using the method of least squares, and assess whether there is a significant change in visual field over time for this subject. Report a two-sided *p*-value.

**11.102** What does the intercept mean in this context? What is the estimated % decline in visual field per year?

**11.103** What is the estimated visual field for this subject after 20 years of follow-up? Provide a 95% confidence interval associated with this estimate.

### Cardiology

Coronary flow reserve (CFR) was estimated in 31 patients with hypertension. Investigators measured the myocardial velocity ratio (MVR) following high-dose dobutamine, and obtained the statistical data shown in Table 11.32

**Table 11.32 Descriptive statistics for CFR and MVR for 31 hypertensive patients**

	Mean	sd
CFR	2.0	0.4
MVR	0.82	0.2

**11.104** If the corrected sum of cross products ( $L_{xy}$ ) is 0.744, estimate the slope in the regression of CFR(*y*) on MVR(*x*).

**11.105** Estimate the intercept of this regression line.

**11.106** Test whether there is a significant relation between these two variables at the 5% level.

**11.107** What is the residual variance around the regression line?

**11.108** What is the value of adjusted  $R^2$  for this line? How do you interpret it?

### Nutrition

Another part of the European Prospective Investigation of Cancer (EPIC) study described in Problem 8.182 was to assess the relationship between dietary and plasma vitamin C levels [20]). The data in Table 11.33 were obtained.

**Table 11.33 Relationship between diet record (DR) vitamin C\* and plasma vitamin C in the EPIC-Norfolk Study**

	Mean	sd	N
DR vitamin C (mg/day)	90.6	50.1	323
Plasma vitamin C (μmol/L)	57.5	21.2	323
correlation			
DR vit C vs. plasma vit C	0.40		

\*Based on food intake only, exclusive of vitamin supplements.

**11.109** Suppose we wish to predict plasma vitamin C as a function of DR vitamin C. If we assume a linear relationship between plasma vitamin C and DR vitamin C, what is the slope and intercept of the least squares line? (*Hint:* Recall the relationship between a correlation coefficient and a regression coefficient.)

**11.110** Test the hypothesis that there is a significant relationship between plasma vitamin C and DR vitamin C. Report a *p*-value (two-tail).

**11.111** What percentage of the variance of plasma vitamin C is explained by DR vitamin C? What is another name for this quantity?

**11.112** One problem with studying the relationship between plasma vitamin C and DR vitamin C is that there may be factors other than DR vitamin C that influence plasma vitamin C. One such factor is vitamin C supplement use. Write down a regression model relating plasma vitamin C to both DR vitamin C and vitamin C supplement use (yes/no), and interpret the coefficients of that model.

### Pulmonary Disease

The Data Set FEV.DAT on the Companion Website contains pulmonary function measures on 654 children ages 3–19 seen in East Boston, MA as part of Childhood Respiratory Disease (CRD) Study. The data set contains data on age, sex, height (inches), FEV = volume of air expelled in 1 second (liters), and smoking status. The code sheet for these data is given in Table 11.34.

**Table 11.34 FEV.DOC**

Column	Variable	Format or Code
1-5	ID number	
7-8	Age (yrs)	
10-15	FEV (liters)	X.XXX
17-20	Height (inches)	XX.X
22	Sex	0=female/1=male
24	Smoking Status	0=non-current smoker/ 1=current smoker

Some descriptive statistics of the variables in the data set using Minitab are given in Table 11.35.

**Table 11.35 Descriptive statistics for FEV.DAT**

Results for: FEV					
Descriptive Statistics: Age, FEV, Hgt					
Variable	N	N*	Mean	SE Mean	StDev
Age	654	0	9.931	0.116	2.954
FEV	654	0	2.6368	0.0339	0.8671
Hgt	654	0	61.144	0.223	5.704
Variable	Minimum		Q1	Median	Q3
Age	3.000		8.000	10.000	12.000
FEV	0.7910		1.9770	2.5475	3.1205
Hgt	46.000		57.000	61.500	65.500
Variable	Maximum				
Age	19.000				
FEV	5.7930				
Hgt	74.000				
Tally for Discrete Variables: Sex, Smoke					
Sex	Count	Percent	Smoke	Count	Percent
0	318	48.62	0	589	90.06
1	336	51.38	1	65	9.94
N =	654		N =	654	

We first ran a regression of FEV on smoking as shown in Table 11.36.

**Table 11.36 Regression of FEV on smoking in FEV.DAT**

#### Regression Analysis: FEV vs. Smoke

The regression equation is

$$\text{FEV} = 2.57 + 0.711 \text{ Smoke}$$

Predictor	Coef	SE Coef	T	P
Constant	2.56614	0.03466	74.04	0.000
Smoke	0.7107	0.1099	6.46	0.000
<b>S = 0.841185 R-Sq = 6.0% R-Sq(adj) = 5.9%</b>				
<b>Analysis of Variance</b>				
Source	DF	SS	MS	F
Regression	1	29.570	29.570	41.79 0.000
Residual	652	461.350	0.708	
<b>Error</b>				
Total	653	490.920		

**11.113** The regression coefficient for smoking is  $0.7 \pm 0.1$ ,  $p < .001$ . Does this mean that smokers have higher pulmonary function than nonsmokers? Why or why not?

It is well known that age, height, and sex are important predictors of pulmonary function. Hence, we ran the regression model in Table 11.37.

**Table 11.37 Regression of FEV on age, height, sex and smoking in FEV.DAT**

#### Regression Analysis: FEV vs. Age, Hgt, Sex, Smoke

The regression equation is

$$\text{FEV} = -4.46 + 0.0655 \text{ Age} + 0.104 \text{ Hgt} + 0.157 \text{ Sex} - 0.0872 \text{ Smoke}$$

Predictor	Coef	SE Coef	T	P
Constant	-4.4570	0.2228	-20.00	0.000
Age	0.065509	0.009489	6.90	0.000
Hgt	0.104199	0.004758	21.90	0.000
Sex	0.15710	0.03321	4.73	0.000
Smoke	-0.08725	0.05925	-1.47	0.141
<b>S = 0.412216 R-Sq = 77.5% R-Sq(adj) = 77.4%</b>				

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	380.640	95.160	560.02 0.000	
Residual	649	110.280	0.170		
<b>Error</b>					
Total	653	490.920			

**11.114** What does the regression coefficient for smoking mean in Table 11.37? How does it differ from the regression coefficient in Table 11.36?

There was some previous literature that the effect of age and height on FEV might not be linear. Hence, we ran the regression model in Table 11.38.

**Table 11.38** Quadratic regression of FEV on age, height, sex, and smoking

## Regression Analysis: FEV vs. Age, Hgt, ...

The regression equation is

$$\text{FEV} = -4.84 + 0.0633 \text{ Age} + 0.110 \text{ Hgt} + 0.0952 \text{ Sex} + 0.00177 (\text{Age}-10)^2 + 0.00284 (\text{Hgt}-61)^2 - 0.140 \text{ Smoke}$$

Predictor	Coef	SE Coef	T	P
Constant	-4.8360	0.2392	-20.22	0.000
Age	0.06334	0.01095	5.78	0.000
Hgt	0.109594	0.005239	20.92	0.000
Sex	0.09515	0.03287	2.90	0.004
(Age-10) <sup>2</sup>	0.001767	0.001754	1.01	0.314
(Hgt-61) <sup>2</sup>	0.0028438	0.0004949	5.75	0.000
Smoke	-0.13959	0.05746	-2.43	0.015

$$S = 0.395080 \quad R-\text{Sq} = 79.4\% \quad R-\text{Sq}(\text{adj}) = 79.2\%$$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	389.931	64.988	416.36	0.000
Residual	647	100.989	0.156		
Error					
Total	653	490.920			

**11.115** Is there evidence of nonlinearity for age or for height?

**11.116** What is the estimated mean FEV for a nonsmoking 15-year-old boy with height = 5'6" (66 inches)?

Finally, we were concerned that the effect of smoking might not be the same for boys and girls. Hence, we ran the model in Table 11.39.

**Table 11.39** Quadratic regression analysis of FEV on age, height, sex, smoking, and sex x smoking

## Results for: FEV1.mtw

## Regression Analysis: FEV vs. Age, Hgt, ...

The regression equation is

$$\text{FEV} = -4.80 + 0.0639 \text{ Age} + 0.109 \text{ Hgt} + 0.0832 \text{ Sex} + 0.00167 (\text{Age}-10)^2 + 0.00280 (\text{Hgt}-61)^2 - 0.196 \text{ Smoke} + 0.141 \text{ Sex*Smoke}$$

Predictor	Coef	SE Coef	T	P
Constant	-4.8040	0.2403	-19.99	0.000
Age	0.06387	0.01095	5.83	0.000
Hgt	0.109121	0.005247	20.80	0.000
Sex	0.08325	0.03404	2.45	0.015

Predictor	Coef	SE Coef	T	P
(Age-10) <sup>2</sup>	0.001673	0.001754	0.95	0.341
(Hgt-61) <sup>2</sup>	0.0027970	0.0004959	5.64	0.000
Smoke	-0.19623	0.07142	-2.75	0.006
Sex*Smoke	0.1414	0.1060	1.33	0.183

$$S = 0.394842 \quad R-\text{Sq} = 79.5\% \quad R-\text{Sq}(\text{adj}) = 79.3\%$$

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	7	390.208	55.744	357.56	0.000
Residual Error	646	100.712	0.156		
Total	653	490.920			

**11.117** What is the estimated mean difference in FEV between a male smoker and male nonsmoker of the same age and height? What is the estimated mean difference in FEV between a female smoker and a female nonsmoker of the same age and height?

## Nutrition

The assessment of diet is an important exposure for many disease outcomes. However, there is often a lot of imprecision in dietary recall. In one study, 70- to 79-year-old women were asked about the preschool diet of their children (ages 2–4) using a food frequency questionnaire (FFQ). A unique aspect of the study is that simultaneous diet record data exist on the same children recorded in real time by their mothers when they were ages 2–4 and their mothers were 20 to 40 years old. The data in Table 11.40 were available on average servings of margarine per week.

**Table 11.40** Margarine intake assessed by two different recording methods (servings per week, [n = 12])

ID	FFQ	DR
340	7	0
399	7	0.5
466	0	0
502	0	0
541	0	0
554	7	2.5
558	7	3
605	7	0.5
611	21	3.7
618	0	2.5
653	21	4.1
707	7	8.5

The Pearson correlation between intake from the two recording methods was 0.448. Assume that FFQ and DR margarine intake are normally distributed.

**11.118** Test the hypothesis that the true Pearson correlation ( $\rho$ ) is significantly different from zero [provide a  $p$ -value (two-tail)].

**11.119** Provide a 95% confidence interval for  $\rho$ .

The distribution of dietary intake for individual food items is often not very normally distributed.

An alternative measure of correlation between the FFQ and DR that does not depend on the assumption of normality is the Spearman rank correlation ( $r_s$ ). For the margarine data,  $r_s = .679$ .

**11.120** Test the hypothesis that the Spearman rank correlation is significantly different from 0 [provide a  $p$ -value (two-tail)].

**11.121** Provide a 95% confidence interval for the true Spearman rank correlation ( $\rho_s$ ).

## REFERENCES

- [1] Greene, J., & Touchstone, J. (1963). Urinary tract estriol: An index of placental function. *American Journal of Obstetrics and Gynecology*, 85(1), 1–9.
- [2] Higgins, M., & Keller, J. (1973). Seven measures of ventilatory lung function. *American Review of Respiratory Disease*, 108, 258–272.
- [3] Kelsey, P. B., Chen, S., & Lauwers, G. Y. (2003). Case 37-2003—a 79-year-old man with coronary artery disease, peripheral vascular disease, end-stage renal disease, and abdominal pain and distention. *New England Journal of Medicine*, 349, 2147–2155.
- [4] Draper, N., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.
- [5] Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: Wiley.
- [6] Wolfe, D. A. (1976). On testing equality of related correlation coefficients. *Biometrika*, 63, 214–215.
- [7] Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current Research in Anesthesia and Analgesia*, 32, 260–267.
- [8] Apgar, V., et al. (1958). Evaluation of the newborn infant—second report. *JAMA*, 168(15), 1985–1988.
- [9] Rosner, B., & Glynn, R. J. (2007). Interval estimation for rank correlation coefficients based on the probit transformation with extension to measurement error correction of correlated ranked data. *Statistics in Medicine*, 26(3), 633–646.
- [10] Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, 29, 202–211.
- [11] Moran, P. A. P. (1948). Rank correlation and product-moment correlation. *Biometrika*, 42, 203–206.
- [12] Torok-Storb, B., Doney, K., Sale, G., Thomas, E. D., & Storb, R. (1985). Subsets of patients with aplastic anemia identified by flow microfluorometry. *New England Journal of Medicine*, 312(16), 1015–1022.
- [13] National Center for Health Statistics. (2009). *Monthly vital statistics report, annual summary*, 57(14).
- [14] The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. (2004). *Pediatrics*, 114(2), 555–576.
- [15] Burch, P. R. B. (1978). Smoking and lung cancer: The problem of inferring cause. *Journal of the Royal Statistical Society*, 141, 437–477.
- [16] Webber, L. S., Harsha, D. W., Phillips, G. T., Srinivasan, S. R., Simpson, J. W., & Berenson, G. S. (1991, April). Cardiovascular risk factors in Hispanic, white and black children: The Brooks County and Bogalusa Heart Studies. *American Journal of Epidemiology*, 133(7), 704–714.
- [17] Reuss, M. L., Paneth, N., Pinto-Martin, J. A., Lorenz, J. M., & Susser, M. (1996). Relation of transient hypoxinemia in preterm infants to neurologic development at 2 years of age. *New England Journal of Medicine*, 334(13), 821–827.
- [18] Krum, H., Viskoper, R. J., Lacourciere, V., Budde, M., & Charlon, V. for the Bosentan Hypertension Investigators. (1998). The effect of an endothelin-receptor antagonist, Bosentan, on blood pressure in patients with essential hypertension. *New England Journal of Medicine*, 338(12), 784–790.
- [19] Burke, J. P., O'Brien, P., Ransom, J., Palumbo, P. J., Lydick, E., Yawn, B. P., Melton, L. J., III, & Leibson, C. L. (2002). Impact of case ascertainment on recent trends in diabetes incidence in Rochester, Minnesota. *American Journal of Epidemiology*, 155, 859–865.
- [20] Rosner, B., Michels, K. B., Chen, Y.-H., & Day, N. E. (2008). Measurement error correction for nutritional exposures with correlated measurement error: use of the method of triads in a longitudinal setting. *Statistics in Medicine*, 27(18), 3466–3489.

# 12

## Multisample Inference

### 12.1 Introduction to the One-Way Analysis of Variance

In Chapter 8 we were concerned with comparing the means of two normal distributions using the two-sample  $t$  test for independent samples. Frequently, the means of more than two distributions need to be compared.

#### Example 12.1

**Pulmonary Disease** A topic of ongoing public-health interest is whether *passive smoking* (exposure among nonsmokers to cigarette smoke in the atmosphere) has a measurable effect on pulmonary health. White and Froeb studied this question by measuring pulmonary function in several ways in the following six groups [1]:

- (1) *Nonsmokers (NS)*: People who themselves did not smoke and were not exposed to cigarette smoke either at home or on the job.
- (2) *Passive smokers (PS)*: People who themselves did not smoke and were not exposed to cigarette smoke in the home but were employed for 20 or more years in an enclosed working area that routinely contained tobacco smoke.
- (3) *Noninhaling smokers (NI)*: People who smoked pipes, cigars, or cigarettes but who did not inhale.
- (4) *Light smokers (LS)*: People who smoked and inhaled 1–10 cigarettes per day for 20 or more years. (*Note*: There are 20 cigarettes in a pack.)
- (5) *Moderate smokers (MS)*: People who smoked and inhaled 11–39 cigarettes per day for 20 or more years.
- (6) *Heavy smokers (HS)*: People who smoked and inhaled 40 or more cigarettes per day for 20 or more years.

A principal measure used by White and Froeb to assess pulmonary function was forced mid-expiratory flow (FEF). They were interested in comparing mean FEF among the six groups.

The  $t$  test methodology generalizes nicely in this case to a procedure called the *one-way analysis of variance* (ANOVA).

### 12.2 One-Way ANOVA—Fixed-Effects Model

#### Example 12.2

**Pulmonary Disease** Refer to Example 12.1. The authors identified 200 males and 200 females in each of the six groups except for the NI group, which was limited

to 50 males and 50 females because of the small number of such people available. The mean and standard deviation of FEF for each of the six groups for males are presented in Table 12.1. How can the means of these six groups be compared?

**Table 12.1** FEF data for smoking and nonsmoking males

Group number, $i$	Group name	Mean FEF (L/s)	sd FEF (L/s)	$n_i$
1	NS	3.78	0.79	200
2	PS	3.30	0.77	200
3	NI	3.32	0.86	50
4	LS	3.23	0.78	200
5	MS	2.73	0.81	200
6	HS	2.59	0.82	200

Source: Reprinted by permission of *The New England Journal of Medicine*, 302(13), 720–723, 1980.

Suppose there are  $k$  groups with  $n_i$  observations in the  $i$ th group. The  $j$ th observation in the  $i$ th group will be denoted by  $y_{ij}$ . Let's assume the following model.

**Equation 12.1**

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

where  $\mu$  is a constant,  $\alpha_i$  is a constant specific to the  $i$ th group, and  $e_{ij}$  is an error term, which is normally distributed with mean 0 and variance  $\sigma^2$ . Thus, a typical observation from the  $i$ th group is normally distributed with mean  $\mu + \alpha_i$  and variance  $\sigma^2$ .

It is not possible to estimate both the overall constant  $\mu$  as well as the  $k$  constants  $\alpha_i$ , which are specific to each group. The reason is that we only have  $k$  observed mean values for the  $k$  groups, which are used to estimate  $k + 1$  parameters. As a result, we need to constrain the parameters so that only  $k$  parameters will be estimated. Some typical constraints are (1) the sum of the  $\alpha_i$ 's is set to 0, or (2) the  $\alpha_i$  for the last group ( $\alpha_k$ ) is set to 0. We use the former approach in this text. However, SAS uses the latter approach.

**Definition 12.1**

The model in Equation 12.1 is a **one-way analysis of variance**, or a **one-way ANOVA model**. With this model, the means of an arbitrary number of groups, each of which follows a normal distribution with the same variance, can be compared. Whether the variability in the data comes mostly from variability within groups or can truly be attributed to variability between groups can also be determined.

The parameters in Equation 12.1 can be interpreted as follows.

**Equation 12.2**

**Interpretation of the Parameters of a One-Way ANOVA Fixed-Effects Model**

- (1)  $\mu$  represents the underlying mean of all groups taken together.
- (2)  $\alpha_i$  represents the difference between the mean of the  $i$ th group and the overall mean.
- (3)  $e_{ij}$  represents random error about the mean  $\mu + \alpha_i$  for an individual observation from the  $i$ th group.

Intuitively, in Table 12.1 FEF is predicted by an overall mean FEF plus the effect of each smoking group plus random variability within each smoking group. Group means are compared within the context of this model.

## 12.3 Hypothesis Testing in One-Way ANOVA—Fixed-Effects Model

The null hypothesis ( $H_0$ ) in this case is that the underlying mean FEF of each of the six groups is the same. This hypothesis is equivalent to stating that each  $\alpha_i = 0$  because the  $\alpha_i$  sum up to 0. The alternative hypothesis ( $H_1$ ) is that at least two of the group means are not the same. This hypothesis is equivalent to stating that at least one  $\alpha_i \neq 0$ . Thus we wish to test the hypothesis  $H_0$ : all  $\alpha_i = 0$  vs.  $H_1$ : at least one  $\alpha_i \neq 0$ .

### F Test for Overall Comparison of Group Means

The mean FEF for the  $i$ th group will be denoted by  $\bar{y}_i$ , and the mean FEF over all groups by  $\bar{\bar{y}}$ . The deviation of an individual observation from the overall mean can be represented as

**Equation 12.3**

$$y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})$$

The first term on the right-hand side ( $y_{ij} - \bar{y}_i$ ) represents the deviation of an individual observation from the group mean for that observation and is an indication of *within-group variability*. The second term on the right-hand side ( $\bar{y}_i - \bar{\bar{y}}$ ) represents the deviation of a group mean from the overall mean and is an indication of *between-group variability*. These terms are depicted in Figure 12.1.

Generally speaking, if the between-group variability is large and the within-group variability is small, as in Figure 12.1a, then  $H_0$  is rejected and the underlying group means are declared significantly different. Conversely, if the between-group variability is small and the within-group variability is large, as in Figure 12.1b, then  $H_0$ , the hypothesis that the underlying group means are the same, is accepted.

If both sides of Equation 12.3 are squared and the squared deviations are summed over all observations over all groups, then the following relationship is obtained:

**Equation 12.4**

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2$$

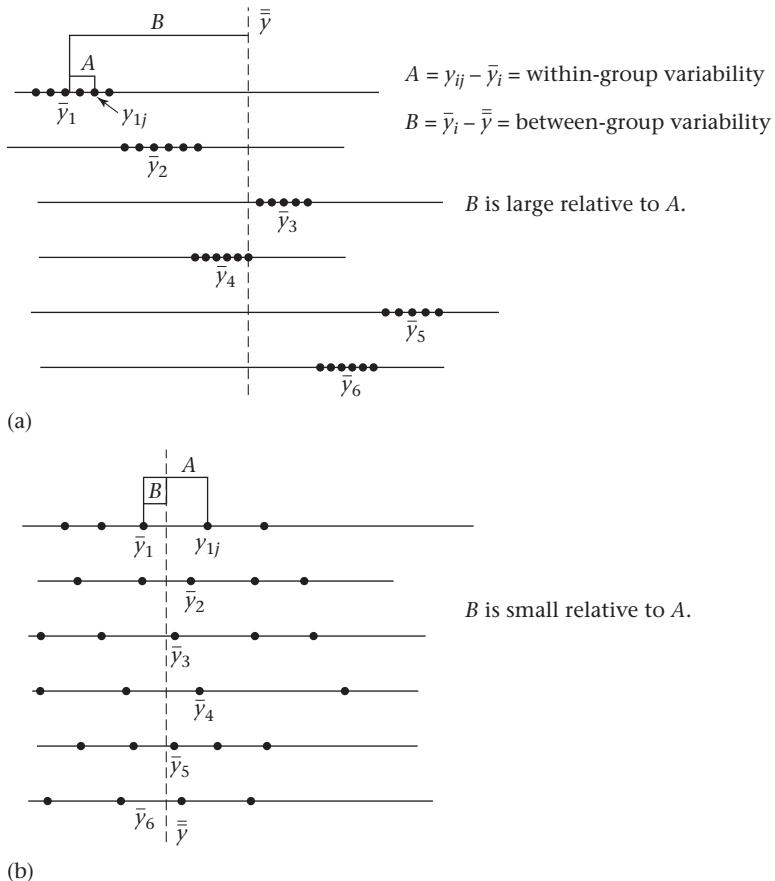
because the cross-product term can be shown to be zero.

**Definition 12.2**

The term

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$$

is called the **Total Sum of Squares (Total SS)**.

**Figure 12.1 Comparison of between-group and within-group variability**

**Definition 12.3** The term

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\gamma_{ij} - \bar{y}_i)^2$$

is called the **Within Sum of Squares (Within SS)**.

**Definition 12.4** The term

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2$$

is called the **Between Sum of Squares (Between SS)**.

Thus the relationship in Equation 12.4 can be written as Total SS = Within SS + Between SS.

To perform the hypothesis test, it is easier to use the short computational form for the Within SS and Between SS in Equation 12.5.

**Equation 12.5****Short Computational Form for the Between SS and Within SS**

$$\text{Between SS} = \sum_{i=1}^k n_i \bar{y}_i^2 - \frac{\left( \sum_{i=1}^k n_i \bar{y}_i \right)^2}{n} = \sum_{i=1}^k n_i \bar{y}_i^2 - \frac{y_{..}^2}{n}$$

$$\text{Within SS} = \sum_{i=1}^k (n_i - 1) s_i^2$$

where  $y_{..}$  = sum of the observations across all groups—i.e., the grand total of all observations over all groups—and  $n$  = total number of observations over all groups.

**Example 12.3**

**Pulmonary Disease** Compute the Within SS and Between SS for the FEF data in Table 12.1.

**Solution**

We use Equation 12.5 as follows:

$$\begin{aligned}\text{Between SS} &= [200(3.78)^2 + 200(3.30)^2 + \dots + 200(2.59)^2] \\ &\quad - \frac{[200(3.78) + 200(3.30) + \dots + 200(2.59)]^2}{1050} \\ &= 10,505.58 - 3292^2/1050 = 10,505.58 - 10,321.20 = 184.38 \\ \text{Within SS} &= 199(0.79)^2 + 199(0.77)^2 + 49(0.86)^2 + 199(0.78)^2 \\ &\quad + 199(0.81)^2 + 199(0.82)^2 \\ &= 124.20 + 117.99 + 36.24 + 121.07 + 130.56 + 133.81 = 663.87\end{aligned}$$

Finally, the following definitions are important.

**Definition 12.5**

**Between Mean Square = Between MS = Between SS/ $(k - 1)$**

**Definition 12.6**

**Within Mean Square = Within MS = Within SS/ $(n - k)$**

The significance test will be based on the ratio of the **Between MS** to the **Within MS**. If this ratio is large, then we reject  $H_0$ ; if it is small, we accept (or fail to reject)  $H_0$ . Furthermore, under  $H_0$ , the ratio of Between MS to Within MS follows an  $F$  distribution with  $k - 1$  and  $n - k$  degrees of freedom. Thus the following test procedure for a level  $\alpha$  test is used.

**Equation 12.6****Overall  $F$  Test for One-Way ANOVA**

To test the hypothesis  $H_0: \alpha_i = 0$  for all  $i$  vs.  $H_1: \text{at least one } \alpha_i \neq 0$ , use the following procedure:

- (1) Compute the Between SS, Between MS, Within SS, and Within MS using Equation 12.5 and Definitions 12.5 and 12.6.
- (2) Compute the test statistic  $F = \text{Between MS}/\text{Within MS}$ , which follows an  $F$  distribution with  $k - 1$  and  $n - k$  df under  $H_0$ .
- (3) If  $F > F_{k-1,n-k,1-\alpha}$  then reject  $H_0$   
If  $F \leq F_{k-1,n-k,1-\alpha}$  then accept  $H_0$

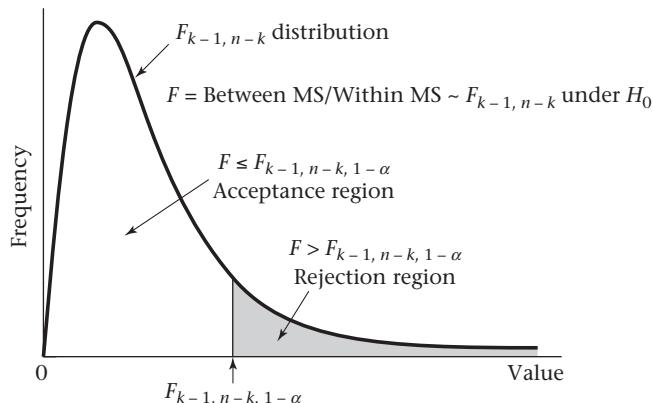
- (4) The exact  $p$ -value is given by the area to the right of  $F$  under an  $F_{k-1, n-k}$  distribution =  $Pr(F_{k-1, n-k} > F)$ .

The acceptance and rejection regions for this test are shown in Figure 12.2. Computation of the exact  $p$ -value is illustrated in Figure 12.3. The results from the ANOVA are typically displayed in an ANOVA table, as in Table 12.2.

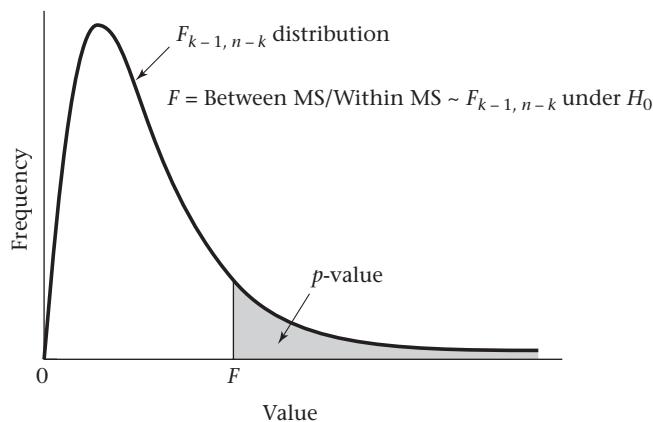
**Table 12.2** Display of one-way ANOVA results

Source of variation	SS	df	MS	F statistic	p-value
Between	$\sum_{i=1}^k n_i \bar{y}_i^2 - \frac{\bar{y}_{..}^2}{n} = A$	$k - 1$	$\frac{A}{k - 1}$	$\frac{A/(k-1)}{B/(n-k)} = F$	$Pr(F_{k-1, n-k} > F)$
Within	$\sum_{i=1}^k (n_i - 1)s_i^2 = B$	$n - k$	$\frac{B}{n-k}$		
Total	Between SS + Within SS				

**Figure 12.2** Acceptance and rejection regions for the overall  $F$  test for one-way ANOVA



**Figure 12.3** Computation of the exact  $p$ -value for the overall  $F$  test for one-way ANOVA



**Example 12.4**

**Pulmonary Disease** Test whether the mean FEF scores differ significantly among the six groups in Table 12.1.

**Solution**

From Example 12.3, Between SS = 184.38 and Within SS = 663.87. Therefore, because there are 1050 observations combined over all 6 groups, it follows that

$$\text{Between MS} = 184.38/5 = 36.875$$

$$\text{Within MS} = 663.87/(1050 - 6) = 663.87/1044 = 0.636$$

$$F = \text{Between MS}/\text{Within MS} = 36.875/0.636 = 58.0 \sim F_{5,1044} \text{ under } H_0$$

Refer to Table 9 in the Appendix and find that

$$F_{5,120,999} = 4.42$$

$$\text{Because } F_{5,1044,999} < F_{5,120,999} = 4.42 < 58.0 = F$$

it follows that  $p < .001$ . Therefore, we can reject  $H_0$ , that all the means are equal, and can conclude that at least two of the means are significantly different. These results are displayed in an ANOVA table (Table 12.3).

**Table 12.3** ANOVA table for FEF data in Table 12.1

	SS	df	MS	F statistic	p-value
Between	184.38	5	36.875	58.0	$p < .001$
Within	663.87	1044	0.636		
Total	848.25				

## 12.4 Comparisons of Specific Groups in One-Way ANOVA

In the previous section a test of the hypothesis  $H_0$ : all group means are equal, vs.  $H_1$ : at least two group means are different, was presented. This test lets us detect when at least two groups have different underlying means, but it does not let us determine which of the groups have means that differ from each other. The usual practice is to perform the overall  $F$  test just discussed. If  $H_0$  is rejected, then specific groups are compared, as discussed in this section.

### t Test for Comparison of Pairs of Groups

Suppose at this point we want to test whether groups 1 and 2 have means that are significantly different from each other. From the underlying model in Equation 12.1, under either hypothesis,

**Equation 12.7**

$\bar{Y}_1$  is normally distributed with mean  $\mu + \alpha_1$  and variance  $\sigma^2/n_1$   
 $\bar{Y}_2$  is normally distributed with mean  $\mu + \alpha_2$  and variance  $\sigma^2/n_2$

The difference of the sample means ( $\bar{y}_1 - \bar{y}_2$ ) will be used as a test criterion. Thus, from Equation 12.7, because the samples are independent it follows that

**Equation 12.8**

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left[\alpha_1 - \alpha_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$$

However, under  $H_0$ ,  $\alpha_1 = \alpha_2$  and Equation 12.8 reduces to

**Equation 12.9**

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left[0, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]$$

If  $\sigma^2$  were known, then we could divide by the standard error

$$\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and obtain the test statistic:

**Equation 12.10**

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The test statistic  $Z$  would follow an  $N(0, 1)$  distribution under  $H_0$ . Because  $\sigma^2$  is generally unknown, the best estimate of it, denoted by  $s^2$ , is substituted, and the test statistic is revised accordingly.

How should  $\sigma^2$  be estimated? Recall from Equation 12.1 that  $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$ . Thus the underlying variance of each group is the same. Therefore, a pooled estimate of group-specific variances is reasonable. Recall that when a pooled estimate of the variance from two independent samples was obtained in Chapter 8, we used a weighted average of the sample variances from the individual samples, where the weights were the number of degrees of freedom in each sample. In particular, from Equation 8.10,

$$s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] / (n_1 + n_2 - 2)$$

For the one-way ANOVA, there are  $k$  sample variances and a similar approach is used to estimate  $\sigma^2$  by computing a weighted average of  $k$  individual sample variances, where the weights are the number of degrees of freedom in each of the  $k$  samples. This formula is given as follows.

**Equation 12.11****Pooled Estimate of the Variance for One-Way ANOVA**

$$s^2 = \sum_{i=1}^k (n_i - 1)s_i^2 / \sum_{i=1}^k (n_i - 1) = \left[ \sum_{i=1}^k (n_i - 1)s_i^2 \right] / (n - k) = \text{Within MS}$$

However, note from Equations 12.5, 12.11, and Definition 12.6 that this weighted average is the same as the Within MS. Thus the Within MS is used to estimate  $\sigma^2$ . Note that  $s^2$  had  $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$  df in the two-sample case. Similarly, for the one-way ANOVA,  $s^2$  has

$$(n_1 - 1) + (n_2 - 1) + \dots + (n_k - 1) \text{df} = (n_1 + n_2 + \dots + n_k) - k = n - k \text{ df}$$

**Example 12.5**

**Pulmonary Disease** What is the best estimate of  $\sigma^2$  for the FEF data in Table 12.1? How many  $df$  does it have?

**Solution**

From Table 12.3, the best estimate of the variance is the Within MS = 0.636. It has  $n - k$   $df = 1044$   $df$ .

Hence the test statistic  $Z$  in Equation 12.10 will be revised, substituting  $s^2$  for  $\sigma^2$ , with the new test statistic  $t$  distributed as  $t_{n-k}$  rather than  $N(0, 1)$ . The test procedure is given as follows.

**Equation 12.12****t Test for the Comparison of Pairs of Groups in One-Way ANOVA (LSD Procedure)**

Suppose we wish to compare two specific groups, arbitrarily labeled as group 1 and group 2, among  $k$  groups. To test the hypothesis  $H_0: \alpha_1 = \alpha_2$  vs.  $H_1: \alpha_1 \neq \alpha_2$ , use the following procedure:

- (1) Compute the pooled estimate of the variance  $s^2 = \text{Within MS}$  from the one way ANOVA.
- (2) Compute the test statistic

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

which follows a  $t_{n-k}$  distribution under  $H_0$ .

- (3) For a two-sided level  $\alpha$  test,

if  $t > t_{n-k, 1-\alpha/2}$  or  $t < t_{n-k, \alpha/2}$

then reject  $H_0$

if  $t_{n-k, \alpha/2} \leq t \leq t_{n-k, 1-\alpha/2}$

then accept  $H_0$

- (4) The exact  $p$ -value is given by

$$\begin{aligned} p &= 2 \times \text{the area to the left of } t \text{ under a } t_{n-k} \text{ distribution if } t < 0 \\ &= 2 \times Pr(t_{n-k} < t) \end{aligned}$$

$$\begin{aligned} p &= 2 \times \text{the area to the right of } t \text{ under a } t_{n-k} \text{ distribution if } t \geq 0 \\ &= 2 \times Pr(t_{n-k} > t) \end{aligned}$$

The acceptance and rejection regions for this test are given in Figure 12.4. The computation of the exact  $p$ -value is illustrated in Figure 12.5. This test is often referred to as the *least significant difference (LSD) method*.

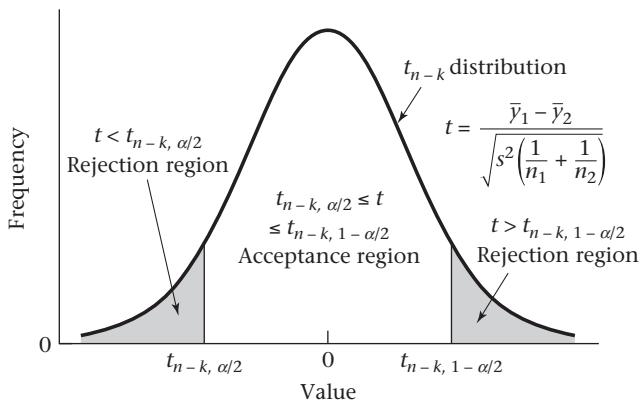
**Example 12.6**

**Pulmonary Disease** Compare each pair of groups for the FEF data in Table 12.1, and report any significant differences.

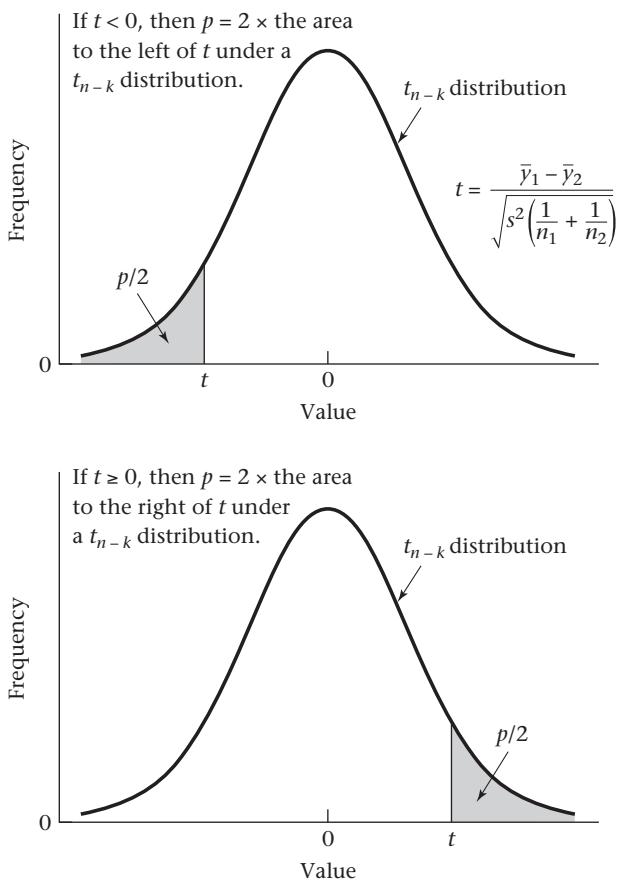
**Solution**

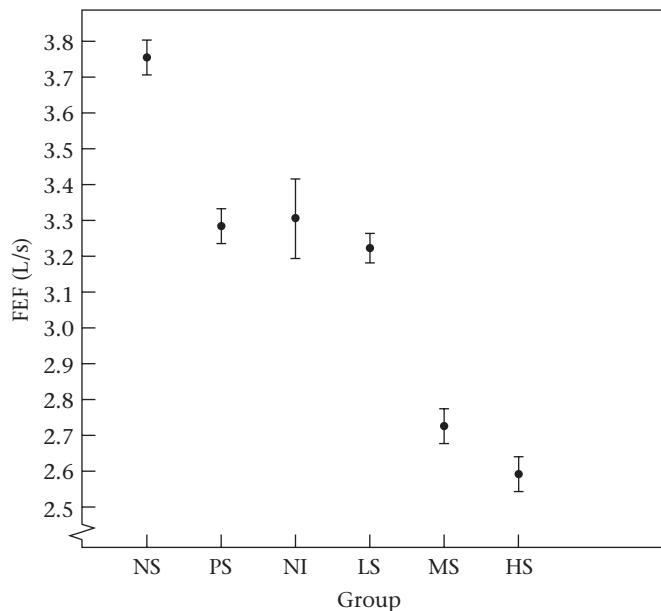
First plot the mean  $\pm se$  of the FEF values for each of the six groups in Figure 12.6 to obtain some idea of the magnitude of the differences between groups. The standard error for an individual group mean is estimated by  $s/\sqrt{n_i}$ , where  $s^2 = \text{Within MS}$ . Notice that the nonsmokers have the best pulmonary function; the passive smokers, noninhaling smokers, and light smokers have about the same pulmonary function and are worse off than the nonsmokers; and the moderate and heavy smokers have the poorest pulmonary function. Note also that the standard-error bars are wider for

**Figure 12.4** Acceptance and rejection regions for the  $t$  test for the comparison of pairs of groups in one-way ANOVA (LSD approach)



**Figure 12.5** Computation of the exact  $p$ -value for the  $t$  test for the comparison of pairs of groups in one-way ANOVA (LSD approach)



**Figure 12.6** Mean  $\pm$  se for FEF for each of six smoking groups

Source: Reprinted by permission of *The New England Journal of Medicine*, 302(13), 720–723, 1980.

the noninhaling smokers than for the other groups because this group has only 50 people compared with 200 for all other groups. Are the observed differences in the figure statistically significant as assessed by the LSD procedure in Equation 12.12? The results are presented in Table 12.4.

There are very highly significant differences (1) between the nonsmokers and all other groups, (2) between the passive smokers and the moderate and heavy smokers, (3) between the noninhaling smokers and the moderate and heavy smokers, and (4) between the light smokers and the moderate and heavy smokers. There are no significant differences between the passive smokers, noninhaling smokers, and light smokers and no significant differences between the moderate and heavy smokers, although there is a trend toward significance with the latter comparison. Thus these results tend to confirm what Figure 12.6 shows. They are very interesting because they show that the pulmonary function of passive smokers is significantly worse than that of nonsmokers and is essentially the same as that of noninhaling and light smokers ( $\leq 1/2$  pack cigarettes per day).

A frequent error in performing the  $t$  test in Equation 12.12 when comparing groups 1 and 2 is to use only the sample variances from *these two groups* rather than from *all k groups* to estimate  $\sigma^2$ . If the sample variances from only two groups are used, then different estimates of  $\sigma^2$  are obtained for each pair of groups considered, which is not reasonable because *all* the groups are assumed to have the same underlying variance  $\sigma^2$ . Furthermore, the estimate of  $\sigma^2$  obtained by using all  $k$  groups will be more accurate than that obtained from using any two groups because the estimate of the variance will be based on more information. This is the principal advantage of

**Table 12.4 Comparisons of specific pairs of groups for the FEF data in Table 12.1 using the LSD *t* test approach**

Groups compared	Test statistic	<i>p</i> -value
NS, PS	$t = \frac{3.78 - 3.30}{\sqrt{0.636 \left( \frac{1}{200} + \frac{1}{200} \right)}} = \frac{0.48}{0.08} = 6.02^a$	< .001
NS, NI	$t = \frac{3.78 - 3.32}{\sqrt{0.636 \left( \frac{1}{200} + \frac{1}{50} \right)}} = \frac{0.46}{0.126} = 3.65$	< .001
NS, LS	$t = \frac{3.78 - 3.23}{\sqrt{0.636 \left( \frac{1}{200} + \frac{1}{200} \right)}} = \frac{0.55}{0.08} = 6.90$	< .001
NS, MS	$t = \frac{3.78 - 2.73}{0.080} = \frac{1.05}{0.08} = 13.17$	< .001
NS, HS	$t = \frac{3.78 - 2.59}{0.080} = \frac{1.19}{0.08} = 14.92$	< .001
PS, NI	$t = \frac{3.30 - 3.32}{0.126} = \frac{-0.02}{0.126} = -0.16$	NS
PS, LS	$t = \frac{3.30 - 3.23}{0.080} = \frac{0.07}{0.08} = 0.88$	NS
PS, MS	$t = \frac{3.30 - 2.73}{0.080} = \frac{0.57}{0.08} = 7.15$	< .001
PS, HS	$t = \frac{3.30 - 2.59}{0.080} = \frac{0.71}{0.08} = 8.90$	< .001
NI, LS	$t = \frac{3.32 - 3.23}{0.126} = \frac{0.09}{0.126} = 0.71$	NS
NI, MS	$t = \frac{3.32 - 2.73}{0.126} = \frac{0.59}{0.126} = 4.68$	< .001
NI, HS	$t = \frac{3.32 - 2.59}{0.126} = \frac{0.73}{0.126} = 5.79$	< .001
LS, MS	$t = \frac{3.23 - 2.73}{0.08} = \frac{0.50}{0.08} = 6.27$	< .001
LS, HS	$t = \frac{3.23 - 2.59}{0.08} = \frac{0.64}{0.08} = 8.03$	< .001
MS, HS	$t = \frac{2.73 - 2.59}{0.08} = \frac{0.14}{0.08} = 1.76$	NS

<sup>a</sup>All test statistics follow a  $t_{1044}$  distribution under  $H_0$ .

performing the  $t$  tests in the framework of a one-way ANOVA rather than by considering each pair of groups separately and performing  $t$  tests for two independent samples as given in Equation 8.11 for each pair of samples. However, if there is reason to believe that not all groups have the same underlying variance ( $\sigma^2$ ), then the one-way ANOVA should not be performed, and  $t$  tests based on pairs of groups should be used instead.

## Linear Contrasts

In Equation 12.12, methods for comparing specific groups within the context of the ANOVA were developed. More general comparisons, such as the comparison of a collection of  $\ell_1$  groups with another collection of  $\ell_2$  groups, are frequently desired.

### Example 12.7

**Pulmonary Disease** Suppose we want to compare the pulmonary function of the group of smokers who inhale cigarettes with that of the group of nonsmokers. The three groups of inhaling smokers in Table 12.1 could just be combined to form one group of 600 inhaling smokers. However, these three groups were selected so as to be of the same size, whereas in the general population the proportions of light, moderate, and heavy smokers are not likely to be the same. Suppose large population surveys report that 70% of inhaling smokers are moderate smokers, 20% are heavy smokers, and 10% are light smokers. How can inhaling smokers as a group be compared with nonsmokers?

The estimation and testing of hypotheses for linear contrasts is used for this type of question.

### Definition 12.7

A **linear contrast** ( $L$ ) is any linear combination of the individual group means such that the linear coefficients add up to 0. Specifically,

$$L = \sum_{i=1}^k c_i \bar{y}_i$$

$$\text{where } \sum_{i=1}^k c_i = 0$$

Notice that the comparison of two means that was considered earlier in this section is a special case of a linear contrast.

### Example 12.8

**Pulmonary Disease** Suppose we want to compare the pulmonary function of the nonsmokers and the passive smokers. Represent this comparison as a linear contrast.

### Solution

Because the nonsmokers are the first group and the passive smokers are the second group, this comparison can be represented by the linear contrast

$$L = \bar{y}_1 - \bar{y}_2 \quad \text{that is,} \quad c_1 = +1 \quad c_2 = -1$$

### Example 12.9

**Pulmonary Disease** Suppose we want to compare the pulmonary function of nonsmokers with that of inhaling smokers, assuming that 10% of inhaling smokers are light smokers, 70% are moderate smokers, and 20% are heavy smokers. Represent this comparison as a linear contrast.

**Solution** This comparison can be represented by the linear contrast

$$\bar{y}_1 - 0.1\bar{y}_4 - 0.7\bar{y}_5 - 0.2\bar{y}_6$$

because the nonsmokers are group 1, the light smokers group 4, the moderate smokers group 5, and the heavy smokers group 6.

How can we test whether the underlying mean of a linear contrast is different from 0? In general, for any linear contrast,

$$L = c_1\bar{y}_1 + c_2\bar{y}_2 + \cdots + c_k\bar{y}_k$$

we wish to test the hypothesis  $H_0:\mu_L = 0$  vs.  $H_1:\mu_L \neq 0$ , where  $\mu_L$  is the mean of the linear contrast  $L$ :

$$c_1\alpha_1 + c_2\alpha_2 + \cdots + c_k\alpha_k$$

Because  $Var(\bar{y}_i) = s^2/n_i$ , we can derive  $Var(L)$  using Equation 5.9, where

$$Var(L) = s^2 \sum_{i=1}^k c_i^2 / n_i$$

Thus the following test procedure, which is analogous to the LSD  $t$  test for pairs of groups in Equation 12.12, can be used.

### Equation 12.13

#### ***t* Test for Linear Contrasts in One-Way ANOVA**

Suppose we want to test the hypothesis  $H_0:\mu_L = 0$  vs.  $H_1:\mu_L \neq 0$ , using a two-sided test with significance level  $= \alpha$ ,

where  $y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$ ,  $\mu_L = \sum_{i=1}^k c_i\alpha_i$  and  $\sum_{i=1}^k c_i = 0$ .

- (1) Compute the pooled estimate of the variance  $= s^2 = \text{Within MS}$  from the one-way ANOVA.
- (2) Compute the linear contrast

$$L = \sum_{i=1}^k c_i\bar{y}_i$$

- (3) Compute the test statistic

$$t = \frac{L}{\sqrt{s^2 \sum_{i=1}^k \frac{c_i^2}{n_i}}}$$

- (4) If  $t > t_{n-k,1-\alpha/2}$  or  $t < t_{n-k,\alpha/2}$  then reject  $H_0$ .  
If  $t_{n-k,\alpha/2} \leq t \leq t_{n-k,1-\alpha/2}$  then accept  $H_0$ .

- (5) The exact  $p$ -value is given by

$$p = 2 \times \text{the area to the left of } t \text{ under a } t_{n-k} \text{ distribution}$$

$$= 2 \times Pr(t_{n-k} < t), \text{ if } t < 0$$

$$p = 2 \times \text{the area to the right of } t \text{ under a } t_{n-k} \text{ distribution}$$

$$= 2 \times Pr(t_{n-k} > t), \text{ if } t \geq 0$$

**Example 12.10**

**Pulmonary Disease** Test the hypothesis that the underlying mean of the linear contrast defined in Example 12.9 is significantly different from 0.

**Solution**

From Table 12.3,  $s^2 = 0.636$ . Furthermore, the linear contrast  $L$  is given by

$$L = \bar{y}_1 - 0.1\bar{y}_4 - 0.7\bar{y}_5 - 0.2\bar{y}_6 = 3.78 - 0.1(3.23) - 0.7(2.73) - 0.2(2.59) = 1.03$$

The standard error of this linear contrast is given by

$$se(L) = \sqrt{s^2 \sum_{i=1}^k \frac{c_i^2}{n_i}} = \sqrt{0.636 \left[ \frac{(1)^2}{200} + \frac{(-0.1)^2}{200} + \frac{(-0.7)^2}{200} + \frac{(-0.2)^2}{200} \right]} = 0.070$$

Thus

$$t = L/se(L) = 1.03/0.070 = 14.69 \sim t_{1044} \text{ under } H_0$$

Clearly, this linear contrast is very highly significant ( $p < .001$ ), and the inhaling smokers as a group have strikingly worse pulmonary function than the nonsmokers.

Another useful application of linear contrasts is when the different groups correspond to different dose levels of a particular quantity, and the coefficients of the contrast are chosen to reflect a particular dose-response relationship. This application is particularly useful if the sample sizes of the individual groups are small and a comparison of any pair of groups does not show a significant difference, but the overall trend is consistent in one direction.

**Example 12.11**

**Pulmonary Disease** Suppose we want to study whether or not the number of cigarettes smoked is related to the level of FEF among those smokers who inhale cigarette smoke. Perform a test of significance for this trend.

**Solution**

Focus on the light smokers, moderate smokers, and heavy smokers in this analysis. We know that the light smokers smoke 1 to 10 cigarettes per day, and we will assume they smoke an average of  $(1 + 10)/2 = 5.5$  cigarettes per day. The moderate smokers smoke 11 to 39 cigarettes per day, and we will assume they smoke an average of  $(11 + 39)/2 = 25$  cigarettes per day. The heavy smokers smoke at least 40 cigarettes per day. We will assume they smoke exactly 40 cigarettes per day, which will underestimate the trend but is the best we can do with the information presented. We want to test the contrast

$$L = 5.5\bar{y}_4 + 25\bar{y}_5 + 40\bar{y}_6$$

for statistical significance. The problem is that the coefficients of this contrast do not add up to 0; indeed, they add up to  $5.5 + 25 + 40 = 70.5$ . However, if  $70.5/3 = 23.5$  is subtracted from each coefficient, then they will add up to 0. Thus we wish to test the contrast

$$L = (5.5 - 23.5)\bar{y}_4 + (25 - 23.5)\bar{y}_5 + (40 - 23.5)\bar{y}_6 = -18\bar{y}_4 + 1.5\bar{y}_5 + 16.5\bar{y}_6$$

for statistical significance. This contrast represents the increasing number of cigarettes smoked per day in the three groups. From Equation 12.13,

$$L = -18(3.23) + 1.5(2.73) + 16.5(2.59) = -58.14 + 4.10 + 42.74 = -11.31$$

$$se(L) = \sqrt{0.636 \left[ \frac{(-18)^2}{200} + \frac{1.5^2}{200} + \frac{16.5^2}{200} \right]} = \sqrt{0.636(2.99)} = \sqrt{1.903} = 1.38$$

Thus  $t = L/se(L) = -11.31/1.38 = -8.20 \sim t_{1044}$  under  $H_0$

Clearly, this trend is very highly significant ( $p < .001$ ), and we can say that among smokers who inhale, the greater the number of cigarettes smoked per day, the worse the pulmonary function will be.

## Multiple Comparisons—Bonferroni Approach

In many studies, comparisons of interest are specified before looking at the actual data, in which case the  $t$  test procedure in Equation 12.12 and the linear-contrast procedure in Equation 12.13 are appropriate. In other instances, comparisons of interest are only specified after looking at the data. In this case a large number of potential comparisons are often possible. Specifically, if there are a large number of groups and every pair of groups is compared using the  $t$  test procedure in Equation 12.12, then some significant differences are likely to be found just by chance.

### Example 12.12

Suppose there are 10 groups. Thus there are  $\binom{10}{2} = 45$  possible pairs of groups to be compared. Using a 5% level of significance would imply that  $.05(45)$ , or about two comparisons, are likely to be significant by chance alone. How can we protect ourselves against the detection of falsely significant differences resulting from making too many comparisons?

Several procedures, referred to as **multiple-comparisons procedures**, ensure that too many falsely significant differences are not declared. The basic idea of these procedures is to ensure that the *overall probability of declaring any significant differences between all possible pairs of groups* is maintained at some fixed significance level (say  $\alpha$ ). One of the simplest and most widely used such procedure is the method of *Bonferroni adjustment*. This method is summarized as follows.

### Equation 12.14

#### Comparison of Pairs of Groups in One-Way ANOVA—Bonferroni Multiple-Comparisons Procedure

Suppose we wish to compare two specific groups, arbitrarily labeled as group 1 and group 2, among  $k$  groups. To test the hypothesis  $H_0: \alpha_1 = \alpha_2$  vs.  $H_1: \alpha_1 \neq \alpha_2$ , use the following procedure:

- (1) Compute the pooled estimate of the variance  $s^2 = \text{Within MS}$  from the one-way ANOVA.
- (2) Compute the test statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

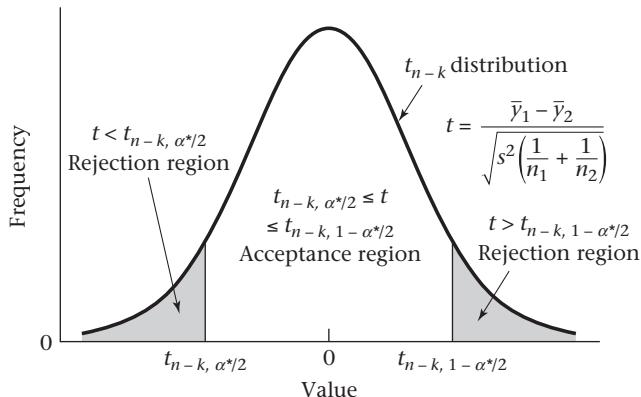
- (3) For a two-sided level  $\alpha$  test, let  $\alpha^* = \alpha / \binom{k}{2}$

If  $t > t_{n-k, 1-\alpha^*/2}$  or  $t < t_{n-k, \alpha^*/2}$  then reject  $H_0$

If  $t_{n-k, \alpha^*/2} \leq t \leq t_{n-k, 1-\alpha^*/2}$  then accept  $H_0$

The acceptance and rejection regions for this test are given in Figure 12.7. This test is called the *Bonferroni multiple-comparisons procedure*.

**Figure 12.7** Acceptance and rejection regions for the comparison of pairs of groups in one-way ANOVA (Bonferroni approach)



The rationale behind this procedure is that in a study with  $k$  groups, there are  $\binom{k}{2}$  possible two-group comparisons. Suppose each two-group comparison is conducted at the  $\alpha^*$  level of significance. Let  $E$  be the event that at least one of the two-group comparisons is statistically significant.  $Pr(E)$  is sometimes referred to as the “experiment-wise type I error.” We wish to determine the value  $\alpha^*$  such that  $Pr(E) = \alpha$ . To find  $\alpha^*$ , we note that

$Pr(\bar{E}) = Pr(\text{none of the two-group comparisons is statistically significant}) = 1 - \alpha$ . If each of the two-group comparisons were independent, then from the multiplication law of probability,  $Pr(\bar{E}) = (1 - \alpha^*)^c$ , where  $c = \binom{k}{2}$ . Therefore,

**Equation 12.15**

$$1 - \alpha = (1 - \alpha^*)^c$$

If  $\alpha^*$  is small, then it can be shown that the right-hand side of Equation 12.15 can be approximated by  $1 - c\alpha^*$ . Thus

$$1 - \alpha \approx 1 - c\alpha^*$$

or

$$\alpha^* \approx \alpha/c = \alpha/\binom{k}{2} \text{ as given in Equation 12.14.}$$

Usually all the two-group comparisons are *not* statistically independent, whereby the appropriate value  $\alpha^*$  is greater than  $\alpha/\binom{k}{2}$ . Thus the Bonferroni procedure is conservative in the sense that  $Pr(E) < \alpha$ .

**Example 12.13**

Apply the Bonferroni multiple-comparisons procedure to the FEF data in Table 12.1.

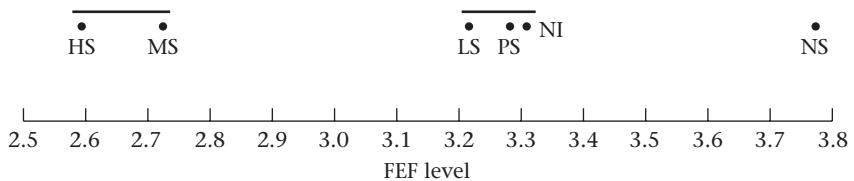
**Solution**

We wish to conduct a test with experiment-wise type I error = .05. We have a total of  $n = 1050$  subjects and  $k = 6$  groups. Thus  $n - k = 1044$  and  $c = \binom{6}{2} = 15$ . Thus  $\alpha^* = .05/15 = .0033$ . Therefore, we conduct  $t$  tests between each pair of groups using the

.0033 level of significance. From Equation 12.14, the critical value for each of these  $t$  tests is  $t_{1044,1-.0033/2} = t_{1044,.99833}$ . We will approximate a  $t$  distribution with 1044  $df$  by an  $N(0, 1)$  distribution or,  $t_{1044,.99833} \approx z_{.99833}$ . From Table 3 in the Appendix,  $z_{.99833} = 2.935$ . We now refer to Table 12.4, which provides the  $t$  statistics for each two-group comparison. We notice that the absolute value of all  $t$  statistics for two-group comparisons that were statistically significant using the LSD approach are  $\geq 3.65$ . Because  $3.65 > 2.935$ , it follows that they will remain statistically significant under the Bonferroni procedure. Furthermore, the comparisons that were not statistically significant with the LSD procedure are also not significant under the Bonferroni procedure. This must be the case because the Bonferroni procedure is more conservative than the LSD procedure. In this example, the critical region using the LSD procedure with a two-sided test ( $\alpha = .05$ ) is  $t < -1.96$  or  $t > 1.96$ , whereas the comparable critical region using the Bonferroni procedure is  $t < -2.935$  or  $t > 2.935$ .

The results of the multiple-comparisons procedure are typically displayed as in Figure 12.8. A line is drawn between the names or numbers of each pair of means that is *not* significantly different. This plot allows us to visually summarize the results of many comparisons of pairs of means in one concise display.

**Figure 12.8** Display of results of Bonferroni multiple-comparisons procedure on FEF data in Table 12.1



Note that the results of the LSD procedure in Table 12.4 and the Bonferroni procedure in Example 12.13 are the same: There are three distinct groups, namely heavy and moderate smokers; light smokers, passive smokers, and noninhaling smokers; and nonsmokers. In general, multiple-comparisons procedures are more strict than ordinary  $t$  tests (LSD procedure) if more than two means are being compared. That is, there are comparisons between pairs of groups for which the  $t$  test would declare a significant difference but the multiple-comparisons procedure would not. This is the price paid for trying to fix the  $\alpha$  level of finding *any* significant difference among pairs of groups in using the multiple-comparisons procedure rather than for *particular* pairs of groups in using the  $t$  test. If only two means are being compared, then the  $p$ -values obtained from using the two procedures are identical.

Also note from Equation 12.14 that as the number of groups being compared ( $k$ ) increases, the critical value for declaring statistical significance becomes larger. This is because as  $k$  increases,  $c = \binom{k}{2}$  increases and therefore  $\alpha^* = \alpha/c$  decreases. The critical value,  $t_{n-k,1-\alpha^*/2}$ , therefore increases because as  $k$  increases, the degrees of freedom ( $n - k$ ) decreases and the percentile  $1 - \alpha^*/2$  increases, both of which result in a larger critical value. This is not true for the LSD procedure, where the critical value  $t_{n-k,1-\alpha/2}$  remains roughly the same as  $k$  increases.

When should the more conservative multiple-comparisons procedure in Equation 12.14 rather than the LSD procedure in Equation 12.12 be used to identify specific differences between groups? This area is controversial. Some researchers routinely use

multiple-comparisons procedures for all one-way ANOVA problems; others never use them. My opinion is that multiple-comparisons procedures should be used if there are many groups and not all comparisons between individual groups have been planned in advance. However, if there are relatively few groups and only specific comparisons of interest are intended, which have been planned in advance, preferably stated in a written set of procedures for a study (commonly called a *protocol*), then I prefer to use ordinary *t* tests (i.e., the LSD procedure) rather than multiple-comparisons procedures. In a sense, by first performing the overall *F* test for one-way ANOVA (Equation 12.6), and only comparing pairs of groups with the LSD procedure (Equation 12.12) if the overall *F* test is statistically significant, we have protected ourselves to some extent from the multiple-comparisons problem, even with the LSD procedure.

The multiple-comparisons issue also occurs when one is considering multiple endpoints within the same study as opposed to comparing multiple groups for the same endpoint, as we have discussed in Equation 12.14.

### Example 12.14

**Cardiovascular Disease** In the Physicians' Health Study, the primary goal was to compare the rate of cardiovascular death between physicians randomized to aspirin treatment vs. placebo. However, it became clear early in the study that there wasn't enough power to assess this endpoint. Instead, several broader endpoints (such as a combined endpoint of either nonfatal myocardial infarction or fatal coronary heart disease) were also considered. A strict interpretation of the Bonferroni principle would require dividing the  $\alpha$  error by  $c$  = number of endpoints considered. However, each endpoint was specified in advance and they all are also highly correlated with each other. In our opinion, there was no need to adjust for multiple comparisons in this setting. These issues are discussed in more detail in Michels and Rosner [2].

## Multiple-Comparisons Procedures for Linear Contrasts

The multiple-comparisons procedure in Equation 12.14 is applicable for comparing pairs of means. In some situations, linear contrasts involving more complex comparisons than simple contrasts based on pairs of means are of interest. In this context, if linear contrasts, which have not been planned in advance, are suggested by looking at the data, then a multiple-comparisons procedure might be used to ensure that under  $H_0$ , the probability of detecting any significant linear contrast is no larger than  $\alpha$ . Scheffé's multiple-comparisons procedure is applicable in this situation and is summarized as follows.

### Equation 12.16

#### Scheffé's Multiple-Comparisons Procedure

Suppose we want to test the hypothesis  $H_0: \mu_L = 0$  vs.  $H_1: \mu_L \neq 0$ , at significance level  $\alpha$ , where

$$L = \sum_{i=1}^k c_i \bar{y}_i, \quad \mu_L = \sum_{i=1}^k c_i \mu_i, \quad \text{and} \quad \sum_{i=1}^k c_i = 0$$

we have  $k$  groups, with  $n_i$  subjects in the  $i$ th group and a total of  $n = \sum_{i=1}^k n_i$  subjects overall. To use Scheffé's multiple-comparisons procedure in this situation, perform the following steps:

- (1) Compute the test statistic

$$t = \frac{L}{\sqrt{s^2 \sum_{i=1}^k \frac{c_i^2}{n_i}}}$$

as given in Equation 12.13.

(2) If

$$t > c_2 = \sqrt{(k-1)F_{k-1,n-k,1-\alpha}} \quad \text{or} \quad t < c_1 = -\sqrt{(k-1)F_{k-1,n-k,1-\alpha}}$$

then reject  $H_0$

If  $c_1 \leq t \leq c_2$ , then accept  $H_0$ .

### Example 12.15

**Pulmonary Disease** Test the hypothesis that the linear contrast defined in Example 12.11, representing the relationship between level of FEF and the number of cigarettes smoked among smokers who inhale cigarettes, is significantly different from 0 using Scheffé's multiple-comparisons procedure.

#### Solution

From Example 12.11,  $t = L/se(L) = -8.20$ . There are six groups and 1050 subjects. Thus, because  $t$  is negative, the critical value is given by  $c_1 = -\sqrt{(k-1)F_{k-1,n-k,1-\alpha}} = -\sqrt{5F_{5,1044,.95}} \cdot F_{5,1044,.95}$  is approximated by  $F_{5,\infty,.95} = 2.21$ . We have  $c_1 = -\sqrt{5(2.21)} = -3.32$ . Because  $t = -8.20 < c_1 = -3.32$ ,  $H_0$  is rejected at the 5% level and a significant trend among inhaling smokers, with pulmonary function decreasing as the number of cigarettes smoked per day increases, is declared.

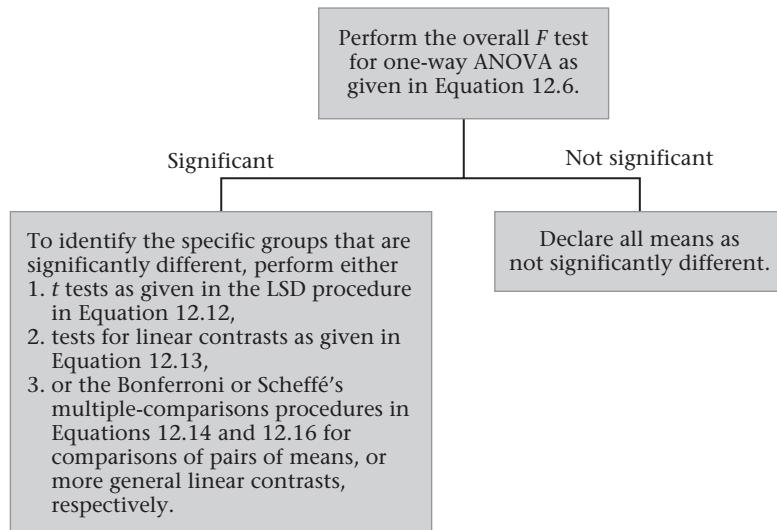
Scheffé's multiple-comparisons procedure could also have been used when pairs of means were being compared because a difference between means is a special case of a linear contrast. However, the Bonferroni procedure introduced in Equation 12.14 is preferable in this instance because if only pairs of means are being compared, then significant differences can appropriately be declared more often than with Scheffé's procedure (which is designed for a broader set of alternative hypotheses) when true differences exist in this situation. Indeed, from Example 12.13, the critical region using the Bonferroni procedure was  $t < -2.935$  or  $t > 2.935$ , whereas the corresponding critical region using Scheffé procedure is  $t < -3.32$  or  $t > 3.32$ .

Once again, if a few linear contrasts, which have been specified in advance, are to be tested, then it may not be necessary to use a multiple-comparisons procedure because if such procedures are used there is less power to detect differences for linear contrasts whose means are truly different from zero than the  $t$  tests introduced in Equation 12.13. Conversely, if many contrasts are to be tested, which have not been specified before looking at the data, then the multiple-comparisons procedure in this section may be useful in protecting against declaring too many significant differences.

Based on our work in Sections 12.1–12.4, Figure 12.9 summarizes the general procedure used to compare the means of  $k$  independent, normally distributed samples.

In this section, we have learned about one-way ANOVA methods. This technique is used to compare the means among several ( $>2$ ) normally distributed samples. To place these methods in a broader perspective, see the master flowchart at the end of the book (pp. 841–846). Beginning at the Start box, we answer no to (1) only one variable of interest? and proceed to (4). We then answer yes to (2) interested in relationships between two variables? no to (3) both variables continuous? and yes to (4) one variable continuous and one categorical? This leads us to the box labeled “analysis of variance.” We then answer 1 to (5) number of ways in which the categorical variable

**Figure 12.9 General procedure for comparing the means of  $k$  independent, normally distributed samples**



can be classified, yes to (6) outcome variable normal or can central-limit theorem be assumed to hold? and no to (7) other covariates to be controlled for? This leads us to the box labeled “one-way ANOVA.”

### The False-Discovery Rate

In some settings, particularly in genetic studies with many hypotheses, control of the experiment-wise type I error sometimes does not seem a reasonable approach to controlling for multiple comparisons because it results in very conservative inferential procedures.

#### Example 12.16

**Cardiovascular Disease, Genetics** A subsample of 520 cases of cardiovascular disease (CVD) and 1100 controls was obtained among men in a prospective cohort study. This type of study is called a *nested case-control study*. Baseline blood samples were obtained from men in the subsample and analyzed for 50 candidate single-nucleotide polymorphisms (SNPs). Each SNP was coded as 0 if homozygous wild type (the most common), 1 if heterozygote, and 2 if homozygous mutant. The association of each SNP with CVD was assessed using contingency-table methods. A chi-square test for trend was run for each SNP. This yielded 50 separate  $p$ -values. If the Bonferroni approach in Equation 12.14 were used, then  $\alpha^* = .05/50 = .001$ . Thus, with such a low value for  $\alpha^*$  it is likely that very few of the hypotheses would be rejected, resulting in a great loss in power. Instead, an alternative approach based on the **false-discovery rate (FDR)** was used to control for the problem of multiple testing.

The FDR approach was developed by Benjamini and Hochberg [3]. The primary goal is *not* to control the overall experiment-wise type I error rate. It is expected if many genes are being tested that there will be several (many) reported positive (statistically significant) results. The FDR attempts to control the proportion of false-positive results among reported statistically significant results.

**Equation 12.17****False-Discovery-Rate (FDR) Testing Procedure**

- (1) Suppose we have conducted  $k$  separate tests with  $p$ -values =  $p_1, \dots, p_k$ .
- (2) For convenience we will renumber the tests so that  $p_1 \leq p_2 \leq \dots \leq p_k$ .
- (3) Define  $q_i = kp_i/i, i = 1, \dots, k$ , where  $i$  = rank of the  $p$ -values among the  $k$  tests.
- (4) Let  $\text{FDR}_i$  = false-discovery rate for the  $i$ th test be defined by  $\min(q_i, \dots, q_k)$ .
- (5) Find the largest  $i$  such that  $\text{FDR}_i < \text{FDR}_0$  = critical level for the FDR (usually .05).
- (6) Reject  $H_0$  for the hypotheses  $1, \dots, i$ , and accept  $H_0$  for the remaining hypotheses.

An advantage of the FDR approach is that it is less conservative than the Bonferroni procedure and as a result yields more power to detect genuine positive effects.

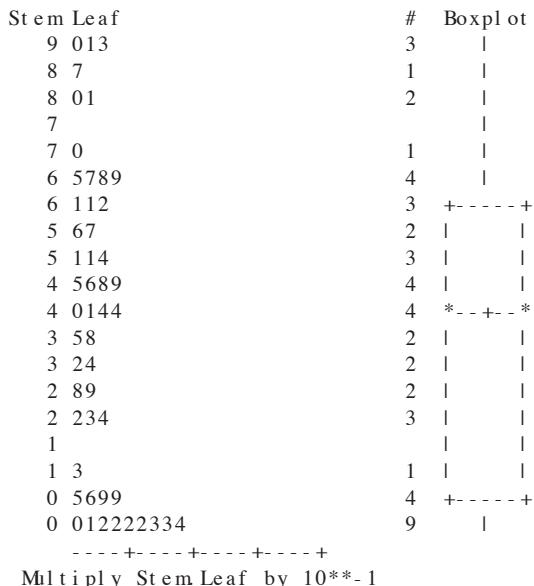
**Example 12.17**

**Cardiovascular Disease, Genetics** Apply the FDR approach to the genetics data described in Example 12.16.

**Solution**

In Figure 12.10, we present a stem-and-leaf plot and box plot of the  $p$ -values from the tests of each of the 50 SNPs. The  $p$ -values for the nominally significant genes are provided in Table 12.5. Note that 10 of the genes are statistically significant, with nominal  $p$ -values ranging from <.0001 to .048. We then applied the FDR approach, with results given in Table 12.6. The naive (nominal)  $p$ -values from Table 12.5 are given in the second column. The “Bonferroni  $p$ -value” =  $\min\{50 \times \text{nominal } p\text{-value}, 1.0\}$  is given in the third column. It represents the level of significance at which the results for a specific SNP would be just statistically significant if a Bonferroni correction were made. The  $q_i$  in step 3 of Equation 12.17 are given in the fourth column. Notice that the  $q_i$  are not necessarily in the same order as the original nominal  $p$ -values. The FDR for each gene is given in the last column.

**Figure 12.10** *p*-Values from tests of 50 SNPs



**Table 12.5** Ordered *p*-values for 10 most significant SNPs

	SNP	<i>p</i> -Value
1	gene30	<.00001
2	gene20	.011
3	gene48	.017
4	gene50	.017
5	gene4	.018
6	gene40	.019
7	gene7	.026
8	gene14	.034
9	gene26	.042
10	gene47	.048

**Table 12.6** Use of the FDR approach to analyzing the CVD data

	SNP	Naïve <i>p</i> -value	Bonferroni <i>p</i> -value	$q_i$	FDR $_i$
1	gene30	<.0001	.0035	.0035	.0035
2	gene20	.011	.54	.28	.16
3	gene48	.017	.86	.28	.16
4	gene50	.017	.87	.22	.16
5	gene4	.018	.92	.18	.16
6	gene40	.019	.94	.16	.16
7	gene7	.026	1.00	.18	.18
8	gene14	.034	1.00	.21	.21
9	gene26	.042	1.00	.23	.23
10	gene47	.048	1.00	.24	.24

Because only gene 30 has  $\text{FDR}_i < .05$ , we reject  $H_0$  only for this gene. This procedure guarantees that no more than 5% of the reported positive results will be false positives. Note that both  $q_i$  and  $\text{FDR}_i$  are noticeably less conservative than the Bonferroni *p*-values.

### REVIEW QUESTIONS 12A

- 1 What is the one-way ANOVA? How does it differ from performing separate *t*-tests for each pair of groups in a study?
- 2 What is the LSD procedure?
- 3 What is the Bonferroni procedure? How does it differ from the LSD procedure? Is it easier or harder to reject the null hypothesis with the Bonferroni procedure than with the LSD procedure?

## 12.5 Case Study: Effects of Lead Exposure on Neurologic and Psychological Function in Children

### Application of One-Way ANOVA

In Section 8.8 (Table 8.10), we analyzed the difference in mean finger-wrist tapping score (MAXFWT) by lead-exposure group. The children were subdivided

into an exposed group who had elevated blood-lead levels ( $\geq 40 \text{ } \mu\text{g}/100 \text{ mL}$ ) in either 1972 or 1973 and a control group who had normal blood-lead levels ( $< 40 \text{ } \mu\text{g}/100 \text{ mL}$ ) in both 1972 and 1973. We first removed outliers from each group using the Extreme Studentized Deviate (ESD) procedure and then used a two-sample  $t$  test to compare mean scores between the two groups (see Table 8.13). However, because the neurologic and psychological tests were performed in 1973, one could argue that it would be better to define an exposed group based on blood-lead levels in 1973 only. For this purpose, the variable LEAD\_GRP in the data set lets us subdivide the exposed group into two subgroups. Specifically, we will consider three lead-exposure groups according to the variable LEAD\_GRP:

If LEAD\_GRP = 1, then the child had normal blood-lead levels ( $< 40 \text{ } \mu\text{g}/100 \text{ mL}$ ) in both 1972 and 1973 (control group).

If LEAD\_GRP = 2, then the child had elevated blood-lead levels ( $\geq 40 \text{ } \mu\text{g}/100 \text{ mL}$ ) in 1973 (the currently exposed group).

If LEAD\_GRP = 3, then the child had elevated blood-lead levels in 1972 and normal blood-lead levels in 1973 (the previously exposed group).

The mean and standard deviation of MAXFWT for each group are given in Table 12.7 and the corresponding box plots in Figure 12.11.

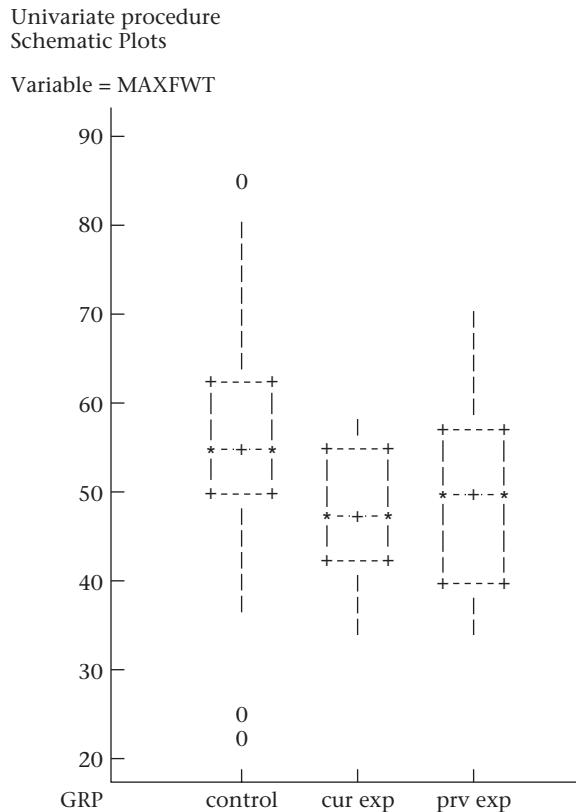
It appears the mean MAXFWT scores are similar in both the currently exposed and previously exposed groups (groups 2 and 3) and are lower than the corresponding mean score in the control group (group 1). To compare the mean scores in the three groups, we will use the one-way ANOVA. We begin by using the overall  $F$  test for one-way ANOVA given in Equation 12.6 to test the hypothesis  $H_0: \alpha_1 = \alpha_2 = \alpha_3$  vs.  $H_1$ : at least two of the  $\alpha_i$  are different. The results are given in Table 12.8.

We see there is an overall significant difference among the mean MAXFWT scores in the three groups. The  $F$  statistic is given under  $F$  value = 4.60. The  $p$ -value =  $Pr(F_{2,92} > 4.60)$  is listed under  $Pr > F$  and is .0125. Therefore, we will proceed to look at differences between each pair of groups. We will use the LSD procedure in Equation 12.12 because these comparisons are planned in advance. The results are given in Table 12.9.

**Table 12.7 Descriptive statistics of MAXFWT by group**

The MEANS Procedure						
Analysis Variable: MAXFWT						
group	N	Obs*	N*	Mean	Std Dev	Minimum
1	77	63	55.0952381	10.9348721	23.0000000	84.0000000
2	22	17	47.5882353	7.0804204	34.0000000	58.0000000
3	21	15	49.4000000	10.1966381	35.0000000	70.0000000

\*N Obs is the total number of subjects in each group. N is the number of subjects used in the analysis in each group (i.e., subjects who have a MAXFWT value that is not an outlier). Note: The test was only given to children ages  $\geq 5$ .

**Figure 12.11** Box plots of MAXFWT by group**Table 12.8** Overall *F* test for one-way ANOVA for MAXFWT

GLM Procedure					
Dependent Variable: MAXFWT					
Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Model	2	966.79062	483.39531	4.60	0.0125
Error	92	9671.14622	105.12115		
Corrected Total	94	10637.93684			

We see that there is a significant difference between the mean MAXFWT score for the currently exposed group and the control group ( $p = .0087$ , listed under  $Pr > |t|$  in the first row corresponding to the control group and the second column corresponding to the currently exposed group). There is a strong trend toward a significant difference between the previously exposed group and the control group ( $p = .0563$ ). There is clearly no significant difference between the mean MAXFWT scores for the currently and previously exposed groups ( $p$ -value = .6191). Other data given in Table 12.9 are the mean and standard error of the mean by group, listed under the MAXFWT LSMEAN and Standard Error LSMEAN columns, respectively. In this case, the MAXFWT LSMEAN column contains the ordinary arithmetic mean (same as Table 12.7). The standard error is (Error Mean

**Table 12.9 Comparison of group means for MAXFWT for pairs of specific groups (LSD procedure)**

The GLM Procedure				
Least Squares Means				
group	MAXFWT		Standard Error	
	LSMEAN	LSMEAN	Pr >  t	Number
Control	55.0952381	1.2917390	<.0001	1
cur exp	47.5882353	2.4866840	<.0001	2
prv exp	49.4000000	2.6472773	<.0001	3

Least Squares Means for effect group				
Pr >  t  for H0: LSMean(i) = LSMean(j)				
Dependent Variable: MAXFWT				
i/j	1	2	3	
1		0.0087	0.0563	
2	0.0087		0.6191	
3	0.0563	0.6191		

$\text{Square}/n_i)^{1/2}$  because our best estimate of the common within-group variance is the Error Mean Square (which we have previously referred to as the Within Mean Square). The third column provides a test of the hypothesis that the underlying mean in each specific group = 0 (specified as  $\text{Pr} > |t|$ ). This is not relevant in this example but would be if we were studying change scores over time. For uses of the general linear-model procedure other than for one-way ANOVA, the LSMEAN is different from the ordinary arithmetic mean. We discuss this issue in detail later in this section (see p. 546).

Another approach for analyzing these data is to look at the 95% confidence interval for the difference in underlying means for specific pairs of groups. This is given by

$$\bar{y}_{i_1} - \bar{y}_{i_2} \pm t_{n-k, .975} \sqrt{\text{Within MS} \left( \frac{1}{n_{i_1}} + \frac{1}{n_{i_2}} \right)}$$

with results in Table 12.10. We note again that there are significant differences between the control group (group 1) and the currently exposed group (group 2) for MAXFWT 95% CI = (1.9, 13.1), but not between groups 1 and 3, 95% CI = (-0.2, 11.5), or between groups 2 and 3, 95% CI = (-9.0, 5.4).

## Relationship Between One-Way ANOVA and Multiple Regression

In Table 12.7 we divided the exposed group into subgroups of currently exposed and previously exposed children. We then used a one-way ANOVA model to compare the mean MAXFWT among the currently exposed, previously exposed, and control groups. Another approach to this problem is to use a multiple-regression model with *dummy variables*. In Definition 11.19 we defined a single dummy variable to represent a categorical variable with two categories. This approach can be extended to represent a categorical variable with any number of categories.

**Table 12.10** 95% confidence intervals for mean difference in MAXFWT between pairs of groups

MAXFWT					
group		LSMEAN	95% Confidence Limits		
Control		55.095238	52.529733	57.660743	
cur exp		47.588235	42.649466	52.527004	
prv exp		49.400000	44.142279	54.657721	
Least Squares Means for Effect group					
Difference					
i		j	Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2		7.507003	1.941641	13.072365
1	3		5.695238	-0.155014	11.545490
2	3		-1.811765	-9.025299	5.401769

**Equation 12.18****Use of Dummy Variables to Represent a Categorical Variable with  $k$  Categories**

Suppose we have a categorical variable  $C$  with  $k$  categories. To represent that variable in a multiple-regression model, we construct  $k - 1$  dummy variables of the form

$$x_1 = \begin{cases} 1 & \text{if subject is in category 2} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if subject is in category 3} \\ 0 & \text{otherwise} \end{cases}$$

⋮

$$x_{k-1} = \begin{cases} 1 & \text{if subject is in category } k \\ 0 & \text{otherwise} \end{cases}$$

The category omitted (category 1) is referred to as the **reference group**.

It is arbitrary which group is assigned to be the reference group; the choice of a reference group is usually dictated by subject-matter considerations. In Table 12.11, we give the values of the dummy variables for subjects in different categories of  $C$ .

**Table 12.11** Representation of a categorical variable  $C$  by dummy variables

Category of $C$	Dummy variables			
	$x_1$	$x_2$	...	$x_{k-1}$
1	0	0	...	0
2	1	0	...	0
3	0	1	...	0
⋮	⋮	⋮	⋮	⋮
$k$	0	0	...	1

Notice that subjects in each category of  $C$  have a unique profile in terms of  $x_1, \dots, x_{k-1}$ . To relate the categorical variable  $C$  to an outcome variable  $y$ , we use the multiple-regression model

**Equation 12.19**

$$y = \alpha + \beta_1 x_2 + \beta_2 x_3 + \dots + \beta_{k-1} x_{k-1} + e$$

How can we use the multiple-regression model in Equation 12.19 to compare specific categories? From Equation 12.19, the average value of  $y$  for subjects in category 1 (the reference category) =  $\alpha$ , the average value of  $y$  for subjects in category 2 =  $\alpha + \beta_1$ . Thus  $\beta_1$  represents the difference between the average value of  $y$  for subjects in category 2 vs. the average value of  $y$  for subjects in the reference category. Similarly,  $\beta_j$  represents the difference between the average value of  $y$  for subjects in category  $(j+1)$  vs. the reference category,  $j = 1, \dots, k-1$ . In the fixed-effects one-way ANOVA model in Equation 12.1, we were interested in testing the hypothesis  $H_0$ : all underlying group means are the same vs.  $H_1$ : at least two underlying group means are different. An equivalent way to specify these hypotheses in a multiple-regression setting is  $H_0$ : all  $\beta_j = 0$  vs.  $H_1$ : at least one of the  $\beta_j \neq 0$ . The latter specification of the hypotheses is the same as those given in Equation 11.31 where we used the overall  $F$  test for multiple linear regression. Thus a fixed-effects one-way ANOVA model can be represented by a multiple linear-regression model based on a dummy-variable specification for the grouping variable. These results are summarized as follows.

**Equation 12.20****Relationship Between Multiple Linear Regression and One-Way ANOVA Approaches**

Suppose we wish to compare the underlying mean among  $k$  groups where the observations in group  $j$  are assumed to be normally distributed with mean =  $\mu_j = \mu + \alpha_j$  and variance =  $\sigma^2$ . To test the hypothesis  $H_0$ :  $\mu_j = 0$  for all  $j = 1, \dots, k$  vs.  $H_1$ : at least two  $\mu_j$  are different, we can use one of two equivalent procedures:

- (1) We can perform the overall  $F$  test for one-way ANOVA.
- (2) Or we can set up a multiple-regression model of the form

$$y = \alpha + \sum_{j=1}^{k-1} \beta_j x_j + e$$

where  $y$  is the outcome variable and  $x_j = 1$  if a subject is in group  $(j+1)$  and = 0 otherwise,  $j = 1, \dots, k-1$ .

The Between SS and Within SS for the one-way ANOVA model in procedure 1 are the same as the Regression SS and Residual SS for the multiple linear-regression model in procedure 2. The  $F$  statistics and  $p$ -values are the same as well.

To compare the underlying mean of the  $(j+1)$ th group vs. the reference group, we can use one of two equivalent procedures:

- (3) We can use the LSD procedure based on one-way ANOVA, where we compute the  $t$  statistic

$$t = \frac{\bar{y}_{j+1} - \bar{y}_1}{s\sqrt{1/n_{j+1} + 1/n_1}} \sim t_{n-k}$$

and  $s^2 = \text{Within MS}$ ,  $n = \sum_{j=1}^k n_j$

(4) Or we can compute the  $t$  statistic

$$t = \frac{b_j}{se(b_j)} \sim t_{n-k}$$

The test statistics and the  $p$ -values are the same under procedures 3 and 4.

To compare the underlying mean of the  $(j+1)$ th and  $(l+1)$ th groups, we can use one of two equivalent procedures:

(5) We can use the LSD procedure based on one-way ANOVA, whereby we compute the  $t$  statistic

$$t = \frac{\bar{y}_{j+1} - \bar{y}_{l+1}}{s\sqrt{1/n_{j+1} + 1/n_{l+1}}} \sim t_{n-k}$$

and  $s^2 = \text{Within MS}$

$$(6) \text{ Or } t = \frac{b_j - b_l}{se(b_j - b_l)} \sim t_{n-k}$$

The standard error of  $b_j - b_l$  and test statistic  $t$  can usually be obtained as a linear-contrast option for multiple-regression programs available in most statistical packages.

Another way to compute  $se(b_j - b_l)$  is to print out the **variance-covariance matrix** of the regression coefficients, which is an option in most statistical packages. If there are  $k$  regression coefficients, then the  $(j, j)$ th element of this matrix is the variance of  $b_j = \text{Var}(b_j)$ . The  $(j, l)$ th element of this matrix is the covariance between  $b_j$  and  $b_l = \text{Cov}(b_j, b_l)$ . Then using Equation 5.11, we have

$$\text{Var}(b_j - b_l) = \text{Var}(b_j) + \text{Var}(b_l) - 2\text{Cov}(b_j, b_l)$$

and

$$se(b_j - b_l) = \sqrt{\text{Var}(b_j - b_l)}$$

### Example 12.18

**Environmental Health, Pediatrics** Compare the mean MAXFWT among control children, currently exposed children, and previously exposed children, respectively, using a multiple-regression approach.

#### Solution

We set up dummy variables using the control group as the reference group and

$$\text{grp2} = \begin{cases} 1 & \text{if currently exposed} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{grp3} = \begin{cases} 1 & \text{if previously exposed} \\ 0 & \text{otherwise} \end{cases}$$

The multiple-regression model is

$$y = \alpha + \beta_1 \times \text{grp2} + \beta_2 \times \text{grp3} + e$$

This model is fitted in Table 12.12 using SAS PROC REG.

**Table 12.12** Multiple-regression model relating MAXFWT to group (control, currently exposed, previously exposed) ( $n = 95$ )

The REG Procedure					
Model: MODEL1					
Dependent Variable: MAXFWT					
Number of Observations Read					120
Number of Observations Used					95
Number of Observations with Missing Values					25
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	966.79062	483.39531	4.60	0.0125
Error	92	9671.14622	105.12115		
Corrected Total	94	10638			
Root MSE		10.25286	R-Square	0.0909	
Dependent Mean		52.85263	Adj R-Sq	0.0711	
Coeff Var		19.39896			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	55.09524	1.29174	42.65	<.0001
grp2	1	-7.50700	2.80218	-2.68	0.0087
grp3	1	-5.69524	2.94562	-1.93	0.0563

The analysis-of-variance table reveals that there are significant differences among the three groups ( $p = .0125$ ) and exactly matches the  $p$ -value in Table 12.8 based on the  $F$  test for one-way ANOVA. The parameter estimates reveal that currently exposed children ( $grp2 = 1$ ) have mean MAXFWT that is 7.51 taps/10 seconds slower than control children ( $p = .009$ ), whereas previously exposed children ( $grp3 = 1$ ) have mean MAXFWT that is 5.70 taps/10 seconds slower than control children, which is not quite statistically significant ( $p = .056$ ).

If we wish to compare currently and previously exposed children, then we can use either a linear-contrast option for SAS PROC REG or a least squares means (LSMEANS) option for SAS PROC GLM. We have used the latter, as shown in Table 12.9. We see that the currently and previously exposed children are not significantly different ( $p = .62$ ). In this case, the LSMEAN is the same as the ordinary arithmetic mean. The standard error of a group mean =  $s / \sqrt{n_j}$ ,  $j = 1, \dots, k$  where  $s = \sqrt{\text{Residual MS}} = \sqrt{\text{Error MS}}$ . For example, for the control group, the  $se = \sqrt{105.12115 / 63} = 1.292$ .

## One-Way Analysis of Covariance

In Equation 11.38, we compared the mean MAXFWT between exposed and control children after controlling for age and sex. We can also compare mean MAXFWT among the control group, the currently exposed group, and the previously exposed group, after controlling for age and gender, by using the model

**Equation 12.21**

$$y = \alpha + \beta_1 \times \text{grp2} + \beta_2 \times \text{grp3} + \beta_3 \times \text{age} + \beta_4 \times \text{sex} + e$$

The models in Equations 11.38 and 12.21 are referred to as **one-way analysis-of-covariance** (or **one-way ANCOVA**) **models**. In ANCOVA, we wish to compare the mean of a continuous outcome variable among two or more groups defined by a single categorical variable, after controlling for other potential confounding variables (also called *covariates*). We have fitted this model using PROC GLM of SAS (see Table 12.13).

Several hypotheses can be tested using Equation 12.21. First, we can test the hypothesis  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$  vs.  $H_1:$  at least one  $\beta_j \neq 0$ . In words, this is a test of whether any of the variables in Equation 12.21 have any relationship to MAXFWT. The results are given at the top of Table 12.13 ( $F$ -value = 29.06,  $p$ -value = .0001). Thus some of the variables are having a significant effect on MAXFWT. Second, to test for the effect of group after controlling for age and sex, we can test the hypothesis  $H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 \neq 0, \beta_4 \neq 0$  vs.  $H_1:$  all  $\beta_j \neq 0, j = 1, \dots, 4$ . This is given in the middle of Table 12.13, under the heading Type III SS. The type III SS provides an estimate of the effect of a specific risk factor, after controlling for the effects of all other variables in the model. In this case, the effect of group is significant after we control for age and sex ( $F$ -value = 5.40,  $p$  = .0061). Third, we are interested in comparing specific categories of the group variable as shown in the bottom of Table 12.13 under Least-Squares Means. We see that both the currently exposed ( $p$  = .009) and the previously exposed ( $p$  = .018) groups have significantly lower mean MAXFWT scores than the control group, after adjusting for age and sex, whereas there is no significant difference between the currently exposed and previously exposed groups ( $p$  = .909). To estimate the mean difference between groups, we refer to the MAXFWT LSMEAN column at the bottom of Table 12.13. In this case, the LSMEAN column is different from the ordinary arithmetic mean. Instead, it represents a mean value for each group that is, in a sense, adjusted for age and sex. Specifically, for each category of a categorical variable, the LSMEAN represents the average value of MAXFWT for a hypothetical sample of individuals who

- (1) For each continuous variable in the model have a mean value equal to the overall sample mean (over all categories) for that variable
- (2) And for any other categorical variable in the model (with  $k$  categories) have a proportion of  $1/k$  of individuals in each category

**Example 12.19**

**Environmental Health, Pediatrics** Compute the LSMEAN for the control, currently exposed, and previously exposed groups in the multiple-regression model in Table 12.13.

**Solution**

Besides group, the model has one continuous variable (age) and one other categorical variable (sex). The overall mean age for the entire sample ( $n = 95$ ) is 9.768 years. Also, in the PROC GLM analyses in Table 12.13, the previously exposed group is the reference group (the SAS convention is to use the last group as the reference group) and females are coded as 1 and males as 0. Thus the LSMEAN by group is

$$\text{Control : } 26.765 + 4.992 + 2.440(9.768) + \frac{1}{2}(-2.395)(1) = 54.4 \text{ taps / 10 seconds}$$

$$\text{Currently exposed : } 26.765 - 0.295 + 2.440(9.768) + \frac{1}{2}(-2.395)(1) = 49.1 \text{ taps / 10 seconds}$$

$$\text{Previously exposed : } 26.765 + 2.440(9.768) + \frac{1}{2}(-2.395)(1) = 49.4 \text{ taps / 10 seconds}$$

**Table 12.13 SAS PROC GLM output relating MAXFWT to group, age, and sex (n=95)**

The GLM Procedure					
Dependent Variable: MAXFWT					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5995.49850	1498.87462	29.06	<.0001
Error	90	4642.43834	51.58265		
Corrected Total	94	10637.93684			
		R-Square	Coeff Var	Root MSE	MAXFWT Mean
		0.563596	13.58893	7.182106	52.85263
Source	DF	Type I SS	Mean Square	F Value	Pr > F
group	2	966.790624	483.395312	9.37	0.0002
ageyr	1	4900.426329	4900.426329	95.00	<.0001
sex	1	128.281546	128.281546	2.49	0.1183
Source	DF	Type III SS	Mean Square	F Value	Pr > F
group	2	557.352577	278.676289	5.40	0.0061
ageyr	1	5027.978708	5027.978708	97.47	<.0001
sex	1	128.281546	128.281546	2.49	0.1183
Standard					
Parameter		Estimate	Error	t Value	Pr >  t
Intercept		26.76514260 B	3.02389880	8.85	<.0001
group	Control	4.99198543 B	2.06543844	2.42	0.0177
group	cur exp	-0.29455702 B	2.55441917	-0.12	0.9085
group	prv exp	0.00000000 B	.	.	.
ageyr		2.44032385	0.24717388	9.87	<.0001
sex	Female	-2.39491720 B	1.51865886	-1.58	0.1183
sex	Male	0.00000000 B	.	.	.
The GLM Procedure					
Least Squares Means					
Standard					
MAXFWT		Error			
group	LSMEAN	LSMEAN			
Control	54.3977803	0.9139840	<.0001	1	
cur exp	49.1112378	1.7622906	<.0001	2	
prv exp	49.4057948	1.8550823	<.0001	3	
Least Squares Means for effect group					
Pr >  t  for H0: LSMean(i) = LSMean(j)					
Dependent Variable: MAXFWT					
i/j	1	2	3		
1		0.0090	0.0177		
2	0.0090		0.9085		
3	0.0177	0.9085			

Thus both currently and previously exposed children have mean MAXFWT about 5 taps per 10 seconds slower than control children after adjusting for age and sex. Finally, we see that mean MAXFWT increases by 2.44 taps per 10 seconds for each year of age ( $p < .001$ ) and that females have a lower mean MAXFWT score than

males of the same age and group by 2.39 taps per 10 seconds, although the sex effect is not statistically significant ( $p = .12$ ).

In this section, we have discussed the one-way ANCOVA. The one-way ANCOVA is used to assess mean differences among several groups after controlling for other covariates (which can be either continuous or categorical). On the master flowchart (pp. 841–846), we answer (1) to (4) in the same way as for the ANOVA (see p. 535). This leads us to the box labeled “analysis of variance.” We could then answer 1 to (5) number of ways in which the categorical variable can be classified? yes to (6) outcome variable normal or can central-limit theorem be expected to hold? and yes to (7) other covariates to be controlled for? This leads us to the box labeled “one-way ANCOVA.”

### REVIEW QUESTIONS 12B

- 1** What is a dummy variable?
- 2** What is the ANCOVA? How does it differ from the ANOVA?
- 3** Suppose we want to study whether HgbA1c (a serum marker of compliance with diabetes medication) is related to ethnic group (assume the only ethnic groups are Caucasian/African-American/Hispanic/Asian) among diabetic patients, while controlling for age and sex.
  - (a)** Write down an ANCOVA model to perform this analysis.
  - (b)** Interpret the coefficients for the model in Review Question 12B.3a.

## 12.6 Two-Way ANOVA

In Sections 12.1–12.4, the relationship between pulmonary function and cigarette smoking was used to illustrate the fixed-effects one-way ANOVA. In this example, groups were defined by only one variable, cigarette smoking. In some instances, the groups being considered can be classified by two different variables and thus can be arranged in the form of an  $R \times C$  contingency table. We would like to be able to look at the effects of each variable after controlling for the effects of the other variable. The latter type of data is usually analyzed using a technique called the **two-way ANOVA**.

### Example 12.20

**Hypertension, Nutrition** A study was performed to look at the level of blood pressure in two different vegetarian groups, both compared with each other and with normals. A group of 226 strict vegetarians (SV), who eat no animal products of any kind, 63 lactovegetarians (LV), who eat dairy products but no other animal foods, and 460 normals (NOR), who eat a standard American diet, provided data for the study. Mean systolic blood pressure (SBP) by dietary group and sex is given in Table 12.14.

We are interested in the effects of sex and dietary group on SBP. The effects of sex and dietary group may be independent, or they may be related or “interact” with each other. One approach to this problem is to construct a two-way ANOVA model predicting mean SBP level as a function of sex and dietary group.

**Table 12.14** Mean SBP by dietary group and sex

Dietary group		Sex	
		Male	Female
SV	Mean	109.9	102.6
	n	138	88
LV	Mean	115.5	105.2
	n	26	37
NOR	Mean	128.3	119.6
	n	240	220

**Definition 12.8**

An **interaction effect** between two variables is defined as one in which the effect of one variable depends on the level of the other variable.

**Example 12.21**

**Hypertension, Nutrition** Suppose we hypothesize that SV males have mean SBP levels that are 10 mm Hg lower than those of normal males, whereas SV females have mean SBP levels identical to those of normal females. This relationship would be an example of an interaction effect between sex and dietary group because the effect of diet on blood pressure would be different for males and females.

In general, if an interaction effect is present, then it becomes difficult to interpret the separate (or main) effects of each variable because the effect of one factor (e.g., dietary group) depends on the level of the other factor (e.g., sex).

The general model for the two-way ANOVA is given as follows.

**Equation 12.22****Two-Way ANOVA—General Model**

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where

$y_{ijk}$  is the SBP of the  $k$ th person in the  $i$ th dietary group and the  $j$ th sex group

$\mu$  is a constant

$\alpha_i$  is a constant representing the effect of dietary group

$\beta_j$  is a constant representing the effect of sex

$\gamma_{ij}$  is a constant representing the interaction effect between dietary group and sex

$e_{ijk}$  is an error term, which is assumed to be normally distributed with mean 0 and variance  $\sigma^2$

By convention,

$$\sum_{i=1}^r \alpha_i = \sum_{j=1}^c \beta_j = 0, \quad \sum_{j=1}^c \gamma_{ij} = 0 \quad \text{for all } i$$

$$\sum_{i=1}^r \gamma_{ij} = 0 \quad \text{for all } j$$

Thus, from Equation 12.22,  $y_{ijk}$  is normally distributed with mean  $\mu + \alpha_i + \beta_j + \gamma_{ij}$  and variance  $\sigma^2$ .

## Hypothesis Testing in Two-Way ANOVA

Let us denote the mean SBP for the  $i$ th row and  $j$ th column by  $\bar{y}_{ij}$ , the mean SBP for the  $i$ th row by  $\bar{y}_{i\cdot}$ , the mean SBP for the  $j$ th column by  $\bar{y}_{\cdot j}$ , and the overall mean by  $\bar{y}_{..}$ . The deviation of an individual observation from the overall mean can be represented as follows.

**Equation 12.23**

$$y_{ijk} - \bar{y}_{..} = (y_{ijk} - \bar{y}_{ij}) + (\bar{y}_{i\cdot} - \bar{y}_{..}) + (\bar{y}_{\cdot j} - \bar{y}_{..}) + (\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{..})$$

**Definition 12.9**

The first term on the right-hand side  $(y_{ijk} - \bar{y}_{ij})$  represents the deviation of an individual observation from the group mean for that observation. The expression is an indication of *within-group variability* and is called the **error term**.

**Definition 12.10**

The second term on the right-hand side  $(\bar{y}_{i\cdot} - \bar{y}_{..})$  represents the deviation of the mean of the  $i$ th row from the overall mean and is called the **row effect**.

**Definition 12.11**

The third term on the right-hand side  $(\bar{y}_{\cdot j} - \bar{y}_{..})$  represents the deviation of the mean of the  $j$ th column from the overall mean and is called the **column effect**.

**Definition 12.12**

The fourth term on the right-hand side

$$(\bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{..}) = (\bar{y}_{ij} - \bar{y}_{i\cdot}) - (\bar{y}_{\cdot j} - \bar{y}_{..})$$

represents the deviation of the column effect in the  $i$ th row  $(\bar{y}_{ij} - \bar{y}_{i\cdot})$  from the overall column effect  $(\bar{y}_{\cdot j} - \bar{y}_{..})$  and is called the **interaction effect**.

We would like to test the following hypotheses concerning these data:

- (1) Test for the presence of row effects:  $H_0$ : all  $\alpha_i = 0$  vs.  $H_1$ : at least one  $\alpha_i \neq 0$ . This is a test for the effect of dietary group on SBP level after controlling for the effect of sex.
- (2) Test for the presence of column effects:  $H_0$ : all  $\beta_j = 0$  vs.  $H_1$ : at least one  $\beta_j \neq 0$ . This is a test for the effect of sex on SBP level after controlling for the effect of dietary group.
- (3) Test for the presence of interaction effects:  $H_0$ : all  $\gamma_{ij} = 0$  vs.  $H_1$ : at least one  $\gamma_{ij} \neq 0$ . This is a test of whether or not there is a differential effect of dietary group between males and females. For example, dietary group may have an effect on SBP only among men.

For simplicity, we have ignored the interaction term in subsequent analyses. The SAS General Linear Model procedure (PROC GLM) has been used to analyze the data. In particular, two “indicator” or “dummy” variables were set up to represent study group ( $x_1, x_2$ ), where

$$\begin{aligned}x_1 &= 1 \text{ if a person is in the first (SV) group} \\&= 0 \text{ otherwise}\end{aligned}$$

$$\begin{aligned}x_2 &= 1 \text{ if a person is in the second (LV) group} \\&= 0 \text{ otherwise}\end{aligned}$$

and the normal group is the reference group. A variable  $x_3$  is also included to represent sex, where

$$\begin{aligned}x_3 &= 1 \text{ if male} \\&= 0 \text{ if female}\end{aligned}$$

The multiple-regression model can then be written as

**Equation 12.24**

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

The results from using the SAS GLM procedure are shown in Table 12.15. The program first provides a test of the overall hypothesis  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$  vs.  $H_1$ : at least one of the  $\beta_j \neq 0$ , as given in Equation 11.31. The  $F$  statistic corresponding to this test is  $105.85 \sim F_{3,745}$  under  $H_0$ , with  $p$ -value  $< .001$ . Thus at least one of the effects (study group or sex) is significant. In the second part of the display, the program lists the type III SS and the corresponding  $F$  statistic ( $F$ -value) and  $p$ -value ( $Pr > F$ ). The type III SS provides an estimate of the effects of specific risk factors after controlling for the effects of all other variables in the model. Thus, to test the effect of study group after controlling for sex, we wish to test the hypothesis  $H_0: \beta_1 = \beta_2 = 0$ ,  $\beta_3 \neq 0$ , vs.  $H_1$ : at least one of  $\beta_1, \beta_2 \neq 0, \beta_3 \neq 0$ . The  $F$  statistic for this comparison is obtained by dividing the study MS =  $(51,806.42/2) = 25,903.21$  by the error MS = 195.89, yielding  $132.24 \sim F_{2,745}$  under  $H_0$ , and a  $p$ -value ( $Pr > F$ ) of = .0001. Thus there are highly significant effects of dietary group on SBP even after controlling for the effect of sex. Similarly, to test for the effect of sex, we test the hypothesis  $H_0: \beta_3 = 0$ , at least one  $\beta_1, \beta_2 \neq 0$ , vs.  $H_1$ :  $\beta_3 \neq 0$ , at least one  $\beta_1, \beta_2 \neq 0$ . The  $F$  statistic for the sex effect is given by  $(13,056/1)/195.89 = 66.65 \sim F_{1,745}$  under  $H_0$ ,  $p = .0001$ . Thus there are highly significant effects of sex after controlling for the effect of dietary group, with males having higher blood pressure than females. SAS also displays a type I SS as well as an associated  $F$  statistic and  $p$ -value. The purpose of the type I SS is to enter and test the variables in the order specified by the user. In this case, study group was specified first and sex was specified second. Thus the effect of study group is assessed first (without controlling for sex), yielding an  $F$  statistic of  $125.45 \sim F_{2,745}$  under  $H_0$ ,  $p = .0001$ . Second, the effect of sex is assessed after controlling for study group. This is the same hypothesis as was tested above using the type III SS. In general, except for the last user-specified risk factor, results from the type I SS (where all variables above the current variable on the user-specified variable list are controlled for) and the type III SS (where *all* other variables in the model are controlled for) will not necessarily be the same. Usually, unless we are interested in entering the variables in a prespecified order, hypothesis testing using the type III SS will be of greater interest.

Although there was a significant effect of study group after controlling for sex, this does not identify which specific dietary groups differ from one another on SBP. For this purpose,  $t$  tests are provided comparing specific dietary groups (1 = SV, 2 = LV, 3 = NOR) after controlling for sex. Refer to the  $3 \times 3$  table listed for STUDY under PROB  $> |T|$ . The (two-tailed)  $p$ -value comparing dietary group  $i$  with dietary group  $j$  is

**Table 12.15** SAS GLM procedure output illustrating the effects of study group and sex on SBP using the data set in Table 12.14

SAS  
GENERAL LINEAR MODELS PROCEDURE

**DEPENDENT VARIABLE: MSYS**

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR>F	R-SQUARE	C.V.	
MODEL	3	62202.79213079	20734.26404360	105.85	0.0001	0.298854	11.8858	
ERROR	745	145934.76850283	195.88559531		ROOT MSE		MSYS MEAN	
<b>CORRECTED</b>								
TOTAL	748	208137.56063362		13.99591352		117.75303516		
SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE III SS	F VALUE	PR > F
STUDY	2	49146.49426085	125.45	0.0001	2	51806.42069945	132.24	0.0001
SEX	1	13056.29786994	66.65	0.0001	1	13056.29786994	66.65	0.0001
STUDY				PROB >  T				
		SV	LV	NOR				
		.	0.0425	0.0001				
		0.0425	.	0.0001				
		0.0001	0.0001	.				
SEX				PROB >  T				
		MALE	FEMALE					
		.	0.0001					
		0.0001	.					
T FOR HO:								
PARAMETER	ESTIMATE		PARAMETERS	PR >  T		STD ERROR OF ESTIMATE		
INTERCEPT	119.75747587		141.53	0.0001		0.84614985		
STUDY	SV	-17.86546724	-15.66	0.0001		1.14061756		
	LV	-13.79147908	-7.32	0.0001		1.88356205		
	NOR	0.00000000	.	.		.		
SEX	MALE	8.42854624	8.16	0.0001		1.03239026		
	FEMALE	0.00000000	.	.		.		

given in the  $(i, j)$  cell of the table [as well as the  $(j, i)$  cell]. Thus, referring to the  $(1, 2)$  cell, we see that the mean SBP of people in group 1 (SV) differs significantly from that of people in group 2 (LV) after controlling for sex ( $p = .0425$ ). Similarly, referring to the  $(1, 3)$  or  $(3, 1)$  cells, we see that the mean SBP of people in group 1 (SV) differs significantly from that of people in group 3 (NOR) ( $p = .0001$ ). Similar results are obtained from a comparison of people in groups 2 (LV) and 3 (NOR). Furthermore, a  $2 \times 2$  table is listed for the sex effect, yielding a  $p$ -value for a comparison of the two sexes [refer to the  $(1, 2)$  or  $(2, 1)$  cell of the table] after controlling for the effect of study group ( $p = .0001$ ). This test is actually superfluous in this instance because there were only two groups under sex and thus the  $F$  test under type III SS for the sex effect is equivalent to the sex-effect  $t$  test.

Finally, note that at the bottom of the display are estimates of the regression parameters as well as their standard errors and associated  $t$  statistics. These have a similar interpretation to that of the multiple-regression parameters in Definition 11.16. In particular, the regression coefficient  $b_1 = -17.9$  mm Hg is an estimate of the difference in mean SBP between the SV and NOR groups after controlling for the effect of sex. Similarly, the regression coefficient  $b_2 = -13.8$  mm Hg is an estimate

of the difference in mean SBP between the LV and NOR groups after controlling for the effect of sex. Also, the estimated difference in mean SBP between the SV and LV groups is given by  $[-17.9 - (-13.8)] = -4.1$  mm Hg; thus the SVs on average have SBP 4.1 mm Hg lower than the LVs after controlling for the effect of sex. Because no explicit parameter was entered for the third study group (the normal group), the program lists the default value of 0. Finally, the regression coefficient  $b_3 = 8.4$  mm Hg tells us males have mean SBP 8.4 mm Hg higher than females, even after controlling for effect of study group.

It is possible to assess interaction effects for two-way ANOVA models (e.g., using the SAS PROC GLM program), but for the sake of simplicity these were not included in this example. Two-way and higher-way ANOVA are discussed in more detail in Kleinbaum et al. [4].

## Two-Way ANCOVA

We often want to look at the relationship between one or more categorical variables and a continuous outcome variable. If there is one categorical variable, then one-way ANOVA can be used; if there are two (or more) categorical variables, then two-way (higher-way) ANOVA can be used. However, other differences among the groups may make it difficult to interpret these analyses.

### Example 12.22

**Hypertension, Nutrition** In Example 12.20, differences in mean SBP by dietary group and sex were presented using a two-way ANOVA model. Highly significant differences were found among dietary groups after controlling for sex, with mean SBP of SV < mean SBP of LV < mean SBP of NOR. However, other important differences between these groups, such as differences in weight, and possibly age, may explain all or part of the apparent blood-pressure differences. How can we examine whether these blood-pressure differences persist, after accounting for the confounding variables?

The multiple-regression model in Equation 12.24 can be extended to allow for the effects of other covariates using the **two-way ANCOVA**. If weight is denoted by  $x_4$  and age by  $x_5$ , then we have the multiple-regression model

### Equation 12.25

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e$$

where  $e \sim N(0, \sigma^2)$ . We have fitted this model using the SAS PROC GLM program as shown in Table 12.16.

Note from the top of Table 12.16 that the overall model is highly significant ( $F$ -value = 103.16,  $p = .0001$ ), indicating that some of the variables are having a significant effect on SBP. To identify the effects of specific variables, refer to the type III SS. Note that each of the risk factors has a significant effect on SBP after controlling for the effects of all other variables in the model ( $p = .0001$ ). Finally, of principal importance is whether there are differences in mean blood pressure by dietary group after controlling for the effects of age, sex, and weight. In this regard, different conclusions are reached from those reached in Table 12.15. Referring to the  $t$  statistics for STUDY, we see that there is not a significant difference in mean SBP between the SVs (group 1) and the LVs (group 2) after controlling for the other variables ( $p = .7012$ ). There are still highly significant differences between each of the vegetarian groups

**Table 12.16 SAS GLM procedure output illustrating the effects of study group, age, sex and weight on SBP using the data set in Table 12.14**

SAS GENERAL LINEAR MODELS PROCEDURE								
DEPENDENT VARIABLE: MSYS								
SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.	
MODEL	5	85358.44910498	17071.68982100	103.16	0.0001	0.410402	10.9264	
ERROR	741	122628.85342226	165.49103026		ROOT MSE		MSYS MEAN	
CORRECTED TOTAL	746	207987.30252724		12.86433171			117.73630968	
SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE III SS	F VALUE	PR > F
STUDY GROUP	2	49068.28440076	148.25	0.0001	2	8257.21427825	24.95	0.0001
SEX	1	13092.51273176	79.11	0.0001	1	4250.57708379	25.68	0.0001
AGE	1	12978.84918739	78.43	0.0001	1	10524.41438768	63.60	0.0001
WGT	1	10218.80278507	61.75	0.0001	1	10218.80278507	61.75	0.0001
STUDY PROB >  T								
		SV	LV	NOR				
		SV	.	0.7012	0.0001			
		LV	0.7012	.	0.0001			
		NOR	0.0001	0.0001	.			
SEX PROB >  T								
		MALE	FEMALE					
		MALE	.	0.0001				
		FEMALE	0.0001	.				
T FOR H0:								
PARAMETER		ESTIMATE	PARAMETER=0		PR >  T		STD ERROR OF	
INTERCEPT		82.74987242		25.69	0.0001		ESTIMATE	
STUDY GROUP	SV	-8.22799340		-6.20	0.0001		3.22121552	
	LV	-8.95389632		-5.03	0.0001		1.32786689	
	NOR	0.00000000		.	.		1.78082376	
SEX	MALE	5.50352855		5.07	0.0001		1.08593669	
	FEMALE	0.00000000		.	.		.	
AGE		0.47488301		7.97	0.0001		0.05954906	
WGT		0.13011703		7.86	0.0001		0.01655851	

and normals ( $p = .0001$ ). Thus there must have been differences in either age and/or weight between the SV and LV groups that accounted for the significant blood-pressure difference between these groups in Table 12.15. Finally, the estimates of specific regression parameters are given at the bottom of Table 12.16. Note that after controlling for age, sex, and weight, the estimated differences in mean SBP between the SV and NOR groups =  $\beta_1 = -8.2$  mm Hg, between the LV and NOR groups =  $\beta_2 = -9.0$  mm Hg, and between the SV and LV groups =  $\beta_1 - \beta_2 = -8.23 - (-8.95) = 0.7$  mm Hg. These differences are all much smaller than the estimated differences in Table 12.15, of -17.9 mm Hg, -13.8 mm Hg, and -4.1 mm Hg, respectively, where age and weight were not controlled for. The difference in mean SBP between males and females is also much smaller in Table 12.16 after controlling for age and weight (5.5 mm Hg)

than in Table 12.15 (8.4 mm Hg), where these factors were not controlled for. Also, we see from Table 12.16 that the estimated effects of age and weight on mean SBP are 0.47 mm Hg per year and 0.13 mm Hg per lb, respectively. Thus it is important to control for the effects of possible explanatory variables in performing regression analyses.

In this section, we have learned about the two-way ANOVA and the two-way ANCOVA. The two-way ANOVA is used when we wish to simultaneously relate a normally distributed outcome variable to two categorical variables of primary interest. The two-way ANCOVA is used when we wish to simultaneously relate a normally distributed outcome variable to two categorical variables of primary interest and, in addition, wish to control for one or more other covariates, which may be continuous or categorical. We saw that both the two-way ANOVA and the two-way ANCOVA models can be represented as special cases of multiple-regression models.

On the master flowchart at the end of the book (pp. 841–846), we answer (1) to (4) in the same way as for the ANOVA (see p. 535). This leads us to the box labeled “analysis of variance.” We then answer 2 to (5) number of ways in which the categorical variable can be classified. If we have no other covariates to control for, then we answer no to (6) other covariates to be controlled for? and are led to the box labeled “two-way ANOVA.” If we have other covariates to be controlled for, then we answer yes to (6) and are led to the box labeled “two-way ANCOVA.”

If we want to study the primary effect of more than two categorical variables as predictors of a continuous outcome variable, then two-way ANOVA and two-way ANCOVA generalize to multiway ANOVA and multiway ANCOVA, respectively. This is beyond the scope of this book; see [4].

### REVIEW QUESTIONS 12C

- 1 What is the difference between a two-way ANOVA and a one-way ANOVA?
- 2 What is the difference between a two-way ANOVA and a two-way ANCOVA?
- 3 What do we mean by an interaction effect?
- 4 Refer to the data in HOSPITAL.DAT, on the Companion Website.
  - (a) Fit a model relating  $\ln(\text{duration of hospitalization})$  to service and antibiotic use. What type of model is this?
  - (b) Fit a model relating  $\ln(\text{duration of hospitalization})$  to service and antibiotic use, while controlling for differences in age and sex. What type of model is this?
  - (c) Interpret the results in Review Question 12C.4a and b.

## 12.7 The Kruskal-Wallis Test

In some instances we want to compare means among more than two samples, but either the underlying distribution is far from being normal or we have ordinal data. In these situations, a nonparametric alternative to the one-way ANOVA described in Sections 12.1–12.4 of this chapter must be used.

**Example 12.23**

**Ophthalmology** Arachidonic acid is well known to have an effect on ocular metabolism. In particular, topical application of arachidonic acid has caused lid closure, itching, and ocular discharge, among other effects. A study was conducted to compare the anti-inflammatory effects of four different drugs in albino rabbits after administration of arachidonic acid [5]. Six rabbits were studied in each group. Different rabbits were used in each of the four groups. For each animal in a group, one of the four drugs was administered to one eye and a saline solution was administered to the other eye. Ten minutes later arachidonic acid (sodium arachidonate) was administered to both eyes. Both eyes were evaluated every 15 minutes thereafter for lid closure. At each assessment the lids of both eyes were examined and a lid-closure score from 0 to 3 was determined, where 0 = eye completely open, 3 = eye completely closed, and 1, 2 = intermediate states. The measure of effectiveness ( $x$ ) is the change in lid-closure score (from baseline to follow-up) in the treated eye minus the change in lid-closure score in the saline eye. A high value for  $x$  is indicative of an effective drug. The results, after 15 minutes of follow-up, are presented in Table 12.17. Because the scale of measurement was ordinal (0, 1, 2, 3), the use of a nonparametric technique to compare the four treatment groups is appropriate.

**Table 12.17** Ocular anti-inflammatory effects of four drugs on lid closure after administration of arachidonic acid

Rabbit Number	Indomethacin		Aspirin		Piroxicam		BW755C	
	Score <sup>a</sup>	Rank	Score	Rank	Score	Rank	Score	Rank
1	+ 2	13.5	+ 1	9.0	+ 3	20.0	+ 1	9.0
2	+ 3	20.0	+ 3	20.0	+ 1	9.0	0	4.0
3	+ 3	20.0	+ 1	9.0	+ 2	13.5	0	4.0
4	+ 3	20.0	+ 2	13.5	+ 1	9.0	0	4.0
5	+ 3	20.0	+ 2	13.5	+ 3	20.0	0	4.0
6	0	4.0	+ 3	20.0	+ 3	20.0	- 1	1.0

<sup>a</sup>(Lid-closure score at baseline – lid-closure score at 15 minutes)<sub>drug eye</sub> – (lid-closure score at baseline – lid-closure score at 15 minutes)<sub>saline eye</sub>

We would like to generalize the Wilcoxon rank-sum test to enable us to compare more than two samples. To do so, the observations in all treatment groups are pooled and ranks are assigned to each observation in the combined sample. The average ranks ( $\bar{R}_i$ ) in the individual treatment groups are then compared. If the average ranks are close to each other, then  $H_0$ , that the treatments are equally effective, is accepted. If the average ranks are far apart, then  $H_0$  is rejected and we conclude at least some of the treatments are different. The test procedure for accomplishing this goal is known as the Kruskal-Wallis test.

**Equation 12.26****The Kruskal-Wallis Test**

To compare the means of  $k$  samples ( $k > 2$ ) using nonparametric methods, use the following procedure:

- (1) Pool the observations over all samples, thus constructing a combined sample of size  $N = \sum n_i$
- (2) Assign ranks to the individual observations, using the average rank in the case of tied observations.
- (3) Compute the rank sum  $R_i$  for each of the  $k$  samples.

(4) If there are no ties, compute the test statistic

$$H = H^* = \frac{12}{N(N+1)} \times \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

If there are ties, compute the test statistic

$$H = \frac{\frac{H^*}{\sum_{j=1}^g (t_j^3 - t_j)}}{1 - \frac{\sum_{j=1}^g (t_j^3 - t_j)}{N^3 - N}}$$

where  $t_j$  refers to the number of observations (i.e., the frequency) with the same value in the  $j$ th cluster of tied observations and  $g$  is the number of tied groups.

(5) For a level  $\alpha$  test,

if  $H > \chi_{k-1, 1-\alpha}^2$  then reject  $H_0$

if  $H \leq \chi_{k-1, 1-\alpha}^2$  then accept  $H_0$

(6) To assess statistical significance, the  $p$ -value is given by

$$p = Pr(\chi_{k-1}^2 > H)$$

(7) This test procedure should be used only if minimum  $n_i \geq 5$  (i.e., if the smallest sample size for an individual group is at least 5).

The acceptance and rejection regions for this test are shown in Figure 12.12. Computation of the exact  $p$ -value is given in Figure 12.13.

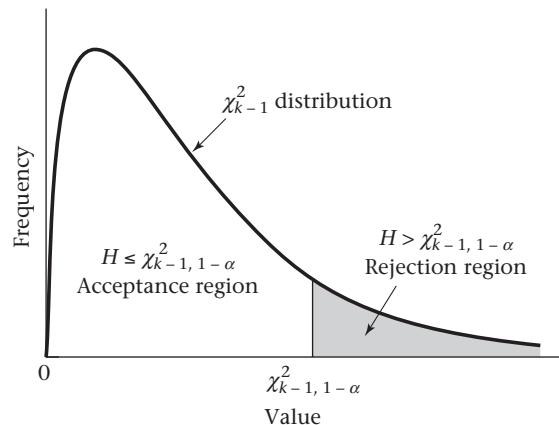
### Example 12.24

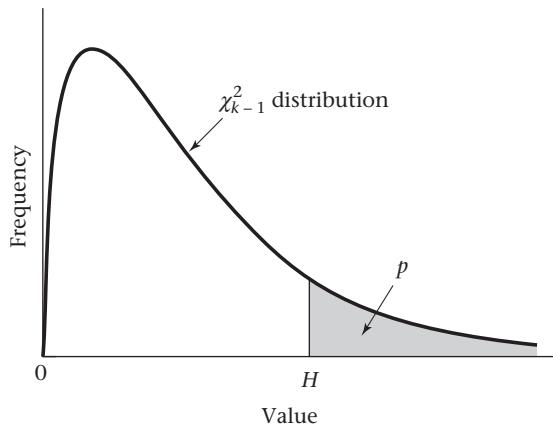
**Ophthalmology** Apply the Kruskal-Wallis test procedure to the ocular data in Table 12.17, and assess the statistical significance of the results.

### Solution

First pool the samples together and assign ranks to the individual observations. This procedure is performed in Table 12.18 with ranks given in Table 12.17.

**Figure 12.12 Acceptance and rejection regions for the Kruskal-Wallis test**



**Figure 12.13** Computation of the exact  $p$ -value for the Kruskal-Wallis test**Table 12.18** Assignment of ranks to the individual observations in Table 12.17

Lid-closure score	Frequency	Range of ranks	Average rank
-1	1	1	1.0
0	5	2–6	4.0
+1	5	7–11	9.0
+2	4	12–15	13.5
+3	9	16–24	20.0

Then compute the rank sum in the four treatment groups:

$$R_1 = 13.5 + 20.0 + \dots + 4.0 = 97.5$$

$$R_2 = 9.0 + 20.0 + \dots + 20.0 = 85.0$$

$$R_3 = 20.0 + 9.0 + \dots + 20.0 = 91.5$$

$$R_4 = 9.0 + 4.0 + \dots + 1.0 = 26.0$$

Because there are ties, compute the Kruskal-Wallis test statistic  $H$  as follows:

$$H = \frac{\frac{12}{24 \times 25} \times \left( \frac{97.5^2}{6} + \frac{85.0^2}{6} + \frac{91.5^2}{6} + \frac{26.0^2}{6} \right) - 3(25)}{1 - \frac{(5^3 - 5) + (5^3 - 5) + (4^3 - 4) + (9^3 - 9)}{24^3 - 24}}$$

$$= \frac{0.020 \times 4296.583 - 75}{1 - \frac{1020}{13,800}} = \frac{10.932}{0.926} = 11.804$$

To assess statistical significance, compare  $H$  with a chi-square distribution with  $k - 1 = 4 - 1 = 3$  df. Note from Table 6 in the Appendix that  $\chi^2_{3,99} = 11.34$ ,  $\chi^2_{3,995} = 12.84$ . Because  $11.34 < H < 12.84$ , it follows that  $.005 < p < .01$ . Thus there is a significant difference in the anti-inflammatory potency of the four drugs.

Note that although the sample sizes in the individual treatment groups were the same in Table 12.17, the Kruskal-Wallis test procedure can, in fact, be used for

samples of unequal size. Also, if there are no ties the Kruskal-Wallis test statistic  $H$  in Equation 12.26 can be written in the form

**Equation 12.27**

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{\bar{R}})^2$$

where  $\bar{R}_i$  = average rank in the  $i$ th sample and  $\bar{\bar{R}}$  = average rank over all samples combined. Thus, if the average rank is about the same in all samples, then  $|\bar{R}_i - \bar{\bar{R}}|$  will tend to be small and  $H_0$  will be accepted. On the contrary, if the average rank is very different across samples, then  $|\bar{R}_i - \bar{\bar{R}}|$  will tend to be large and  $H_0$  will be rejected.

The test procedure in Equation 12.26 is only applicable if minimum  $n_i \geq 5$ . If one of the sample sizes is smaller than 5, then either the sample should be combined with another sample or special small-sample tables should be used. Table 15 in the Appendix provides critical values for selected sample sizes for the case of three samples (i.e.,  $k = 3$ ). The procedure for using this table is as follows:

- (1) Reorder the samples so that  $n_1 \leq n_2 \leq n_3$ , that is, so that the first sample has the smallest sample size and the third sample has the largest sample size.
- (2) For a level  $\alpha$  test, see the  $\alpha$  column and the row corresponding to the sample sizes  $n_1, n_2, n_3$  to find the critical value  $c$ .
- (3) If  $H \geq c$ , then reject  $H_0$  at level  $\alpha$  (i.e.,  $p < \alpha$ ); if  $H < c$ , then accept  $H_0$  at level  $\alpha$  (i.e.,  $p \geq \alpha$ ).

**Example 12.25**

**Suppose there are three samples of size 2, 4, and 5 and  $H = 6.141$ . Assess the statistical significance of the results.**

**Solution**

Refer to the  $n_1 = 2, n_2 = 4, n_3 = 5$  row. The critical values for  $\alpha = .05$  and  $\alpha = .02$  are 5.273 and 6.541, respectively. Because  $H \geq 5.273$ , it follows that the results are statistically significant ( $p < .05$ ). Because  $H < 6.541$ , it follows that  $p \geq .02$ . Thus  $.02 \leq p < .05$ .

### Comparison of Specific Groups Under the Kruskal-Wallis Test

In Example 12.24 we determined that the treatments in Table 12.17 were not all equally effective. To determine which pairs of treatment groups are different, use the following procedure.

**Equation 12.28**

#### Comparison of Specific Groups Under the Kruskal-Wallis Test (Dunn Procedure)

To compare the  $i$ th and  $j$ th treatment groups under the Kruskal-Wallis test, use the following procedure:

- (1) Compute

$$z = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{\frac{N(N+1)}{12} \times \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

- (2) For a two-sided level  $\alpha$  test,

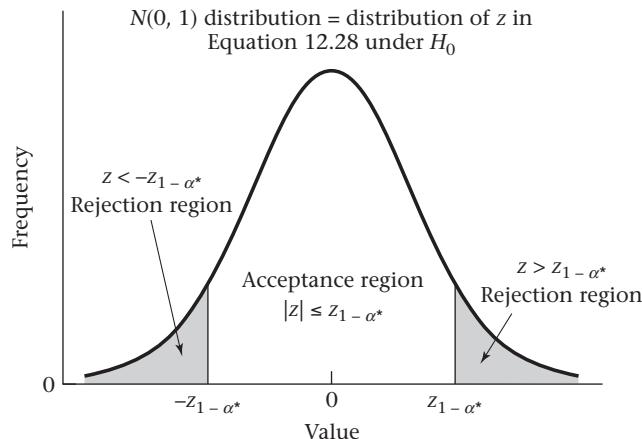
if  $|z| > z_{1-\alpha/2}$  then reject  $H_0$

if  $|z| \leq z_{1-\alpha/2}$  then accept  $H_0$

$$\text{where } \alpha^* = \frac{\alpha}{k(k-1)}$$

The acceptance and rejection regions for this test are shown in Figure 12.14.

**Figure 12.14** Acceptance and rejection regions for the Dunn procedure



**Example 12.26**

**Ophthalmology** Determine which specific groups are different, using the ocular data in Table 12.17.

**Solution**

From Example 12.24,

$$\bar{R}_1 = \frac{97.5}{6} = 16.25$$

$$\bar{R}_2 = \frac{85.0}{6} = 14.17$$

$$\bar{R}_3 = \frac{91.5}{6} = 15.25$$

$$\bar{R}_4 = \frac{26.0}{6} = 4.33$$

Therefore, the following test statistics are used to compare each pair of groups:

$$\text{Groups 1 and 2: } z_{12} = \frac{16.25 - 14.17}{\sqrt{\frac{24 \times 25}{12} \times \left(\frac{1}{6} + \frac{1}{6}\right)}} = \frac{2.08}{4.082} = 0.51$$

$$\text{Groups 1 and 3: } z_{13} = \frac{16.25 - 15.25}{4.082} = \frac{1.0}{4.082} = 0.24$$

$$\text{Groups 1 and 4: } z_{14} = \frac{16.25 - 4.33}{4.082} = \frac{11.92}{4.082} = 2.92$$

$$\text{Groups 2 and 3: } z_{23} = \frac{14.17 - 15.25}{4.082} = \frac{-1.08}{4.082} = -0.27$$

$$\text{Groups 2 and 4: } z_{24} = \frac{14.17 - 4.33}{4.082} = \frac{9.83}{4.082} = 2.41$$

$$\text{Groups 3 and 4: } z_{34} = \frac{15.25 - 4.33}{4.082} = \frac{10.92}{4.082} = 2.67$$

The critical value for  $\alpha = .05$  is given by  $z_{1-\alpha^*}$ , where

$$\alpha^* = \frac{.05}{4 \times 3} = .0042$$

From Table 3 in the Appendix,  $\Phi(2.635) = .9958 = 1 - .0042$ . Thus  $z_{1-.0042} = z_{.9958} = 2.635$  is the critical value. Because  $z_{14}$  and  $z_{34}$  are greater than the critical value, it follows that indomethacin (group 1) and piroxicam (group 3) have significantly better anti-inflammatory properties than BW755C (group 4), whereas the other treatment comparisons are not statistically significant.

In this section, we have discussed the Kruskal-Wallis test, which is a nonparametric test for the comparison of the distributions of several groups. It is used as an alternative to one-way ANOVA when the assumption of normality is questionable. On the master flowchart at the end of the book (pp. 841–846), we would answer (1) to (4) in the same way as for the ANOVA (see p. 535). This leads to the box labeled “analysis of variance.” We then answer 1 to (5) number of ways in which the categorical variable can be classified, and no to (6) outcome variable normal or can central-limit theorem be assumed to hold? This leads us to the successive boxes labeled “nonparametric ANOVA” and “Kruskal-Wallis test.”

### REVIEW QUESTIONS 12D

- 1** What is the difference between the Kruskal-Wallis test and the one-way ANOVA?
- 2** What is the Dunn procedure? What is it used for?
- 3** Suppose we have data on vitamin E intake (IU/day) (from both diet and vitamin supplements) at baseline in four treatment groups in a clinical trial of nutritional supplements. It is important to establish that the vitamin E intakes of the four treatment groups are comparable at baseline. The data for the first 10 participants in each group are given in Table 12.19.

**Table 12.19** Baseline vitamin E intake by treatment group in a clinical trial of nutritional supplements

Subject	Group 1	Group 2	Group 3	Group 4
1	5.92	5.22	4.33	5.37
2	8.24	3.29	16.31	6.39
3	7.27	3.67	6.19	4.90
4	6.24	4.29	7.95	4.75
5	5.21	109.17	4.02	3.07
6	8.25	5.82	6.12	10.64
7	8.33	7.17	5.60	6.50
8	4.12	4.42	12.20	159.90
9	6.27	5.29	3.33	6.00
10	5.38	55.99	7.33	7.31

- (a) Why might a nonparametric analysis of these data be useful?
- (b) Perform the Kruskal-Wallis test to assess whether there are any significant overall group differences in vitamin E intake. Please report a  $p$ -value.
- (c) If the results in Review Question 12D.3b are statistically significant, then identify which specific groups are significantly different.

## 12.8 One-Way ANOVA—The Random-Effects Model

In Example 12.1, we studied the effect of both active and passive smoking on the level of pulmonary function. We were specifically interested in the difference in pulmonary function between the PS and the NS groups. This is an example of the **fixed-effects analysis-of-variance model** because the subgroups being compared have been fixed by the design of the study. In other instances, we are interested in whether there are overall differences between groups and what percentage of the total variation is attributable to between-group vs. within-group differences but are not interested in comparing specific groups.

### Example 12.27

**Endocrinology** The Nurses' Health Study is a large prospective study of approximately 100,000 American nurses to whom a health-related questionnaire was mailed every 2 years starting in 1976. In one substudy, blood samples were obtained from a subset of nurses and serum levels of various hormones were related to the development of disease. As a first step in this process, blood samples were obtained from 5 postmenopausal women. Each blood sample was split into two equal aliquots, which were sent in a blinded fashion to one laboratory for analysis. The same procedure was followed for each of four different laboratories. The goal of the study was to assess how much variation in the analyses was attributable to between-person vs. within-person variation. Comparisons were made both between different hormones and between different laboratories [6].

Table 12.20 shows the reproducibility data for plasma estradiol from one laboratory. Can we estimate the degree of between-person and within-person variation from the data?

**Table 12.20 Reproducibility data for plasma estradiol (pg/mL), Nurses' Health Study**

Subject	Replicate		Absolute value of difference between replicates	Mean value
	1	2		
1	25.5	30.4	4.9	27.95
2	11.1	15.0	3.9	13.05
3	8.0	8.1	0.1	8.05
4	20.7	16.9	3.8	18.80
5	5.8	8.4	2.6	7.10

It appears from Table 12.20 that the variation between replicates depends to some extent on the mean level. Subject 1, who has the largest absolute difference between replicates (4.9), also has the highest mean value (27.95). This is common with many laboratory measures. For this reason, we will analyze the reproducibility data on the ln scale. The justification for using this transformation is that it can be shown

that if the within-person standard deviation is proportional to the mean level in the original scale, then the within-person standard deviation will be approximately independent of the mean level on the ln scale [7]. Using the ln scale will also enable us to estimate the coefficient of variation ( $CV$ ) (a frequently used index in reproducibility studies).

To assess between- and within-person variability, consider the following model:

**Equation 12.29**

$$y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \dots, k, \text{ and } j = 1, \dots, n_i$$

where

$y_{ij}$  =  $j$ th replicate for ln (plasma estradiol) for the  $i$ th subject

$\alpha_i$  is a random variable representing between-subject variability, which is assumed to follow an  $N(0, \sigma_A^2)$  distribution

$e_{ij}$  is a random variable representing within-subject variability, which follows an  $N(0, \sigma^2)$  distribution and is independent of  $\alpha_i$  and any of the other  $e_{ij}$

The model in Equation 12.29 is referred to as a **random-effects one-way analysis-of-variance model**. The underlying mean for the  $i$ th subject is given by  $\mu + \alpha_i$ , where  $\alpha_i$  is drawn from a normal distribution with mean 0 and variance  $\sigma_A^2$ . Thus two different individuals  $i_1, i_2$  will have different underlying means  $\mu + \alpha_{i_1}$  and  $\mu + \alpha_{i_2}$ , respectively. The extent of the between-subject variation is determined by  $\sigma_A^2$ . As  $\sigma_A^2$  increases, the between-subject variation increases as well. The within-subject variation is determined by  $\sigma^2$ . Thus, if we have two replicates  $y_{i1}, y_{i2}$  from the same individual ( $i$ ), they will be normally distributed with mean  $\mu + \alpha_i$  and variance  $\sigma^2$ . An important goal in the random-effects ANOVA is to test the hypothesis  $H_0: \sigma_A^2 = 0$  vs.  $H_1: \sigma_A^2 > 0$ . Under  $H_0$ , there is no underlying between-subject variation; all variation seen between individual subjects is attributable to within-person variation (or “noise”). Under  $H_1$  there is a true underlying difference among means for individual subjects. How can we test these hypotheses? It can be shown under either hypothesis that

**Equation 12.30**

$$E(\text{Within MS}) = \sigma^2$$

In the balanced case, where there are an equal number of replicates per subject, it can be shown that under either hypothesis

**Equation 12.31**

$$E(\text{Between MS}) = \sigma^2 + n\sigma_A^2$$

where  $n_1 = n_2 = \dots = n_k = n$  = number of replicates per subject

In the unbalanced case, where there may be an unequal number of replicates per subject, it can be shown that under either hypothesis

**Equation 12.32**

$$E(\text{Between MS}) = \sigma^2 + n_0\sigma_A^2$$

where

$$n_0 = \left( \sum_{i=1}^k n_i - \sum_{i=1}^k n_i^2 / \sum_{i=1}^k n_i \right) / (k-1)$$

If the number of replicates is the same for each subject, then  $n_1 = n_2 = \dots = n_k = n$  and

$$\begin{aligned} n_0 &= (kn - kn^2/kn)/(k-1) \\ &= (kn - n)/(k-1) \\ &= n(k-1)/(k-1) = n \end{aligned}$$

Thus the two formulas in Equations 12.31 and 12.32 agree in the balanced case. In general, in the unbalanced case,  $n_0$  is always less than the average number of replicates per subject ( $\bar{n} = \sum_{i=1}^k n_i/k$ ), but the difference between  $n_0$  and  $\bar{n}$  is usually small.

We use the same test statistic ( $F = \text{Between MS}/\text{Within MS}$ ) as was used for the fixed-effects model one-way ANOVA. Under the random-effects model, if  $H_1$  is true ( $\sigma_A^2 > 0$ ), then  $F$  will be large, whereas if  $H_0$  is true ( $\sigma_A^2 = 0$ ), then  $F$  will be small. It can be shown that under  $H_0$ ,  $F$  will follow an  $F$  distribution with  $k-1$  and  $N-k$  df, where  $N = \sum_{i=1}^k n_i$ . To estimate the variance components  $\sigma_A^2$  and  $\sigma^2$ , we use Equations 12.30–12.32.

From Equation 12.30, an unbiased estimate of  $\sigma^2$  is given by Within MS. From Equation 12.31, if we estimate  $\sigma^2$  by Within MS, then in the balanced case

$$\begin{aligned} E\left(\frac{\text{Between MS} - \text{Within MS}}{n}\right) \\ = \frac{E(\text{Between MS} - \text{Within MS})}{n} \\ = \frac{\sigma^2 + n\sigma_A^2 - \sigma^2}{n} = \sigma_A^2 \end{aligned}$$

Thus an unbiased estimate of  $\sigma_A^2$  is given by  $\hat{\sigma}_A^2 = (\text{Between MS} - \text{Within MS})/n$ . From Equation 12.32, an analogous result holds in the unbalanced case, where we replace  $n$  by  $n_0$  in the estimation of  $\sigma_A^2$ . These results are summarized as follows.

### Equation 12.33

#### One-Way ANOVA—Random-Effects Model

Suppose we have the model  $y_{ij} = \mu + \alpha_i + e_{ij}$ ,  $i = 1, \dots, k$ ;  $j = 1, \dots, n_i$  where  $\alpha_i \sim N(0, \sigma_A^2)$  and  $e_{ij} \sim N(0, \sigma^2)$ . To test the hypothesis  $H_0: \sigma_A^2 = 0$  vs.  $H_1: \sigma_A^2 > 0$ ,

- (1) Compute the test statistic  $F = \frac{\text{Between MS}}{\text{Within MS}}$ , which follows an  $F_{k-1, N-k}$  distribution under  $H_0$  where

$$\text{Between MS} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{\bar{y}})^2 / (k-1)$$

$$\text{Within MS} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / (N-k)$$

- (2) If  $F > F_{k-1, N-k, 1-\alpha}$ , then reject  $H_0$ .  
If  $F \leq F_{k-1, N-k, 1-\alpha}$ , then accept  $H_0$ .
- (3) The exact  $p$ -value is given by the area to the right of  $F$  under an  $F_{k-1, N-k}$  distribution.
- (4) The within-group variance component ( $\sigma^2$ ) is estimated by the Within MS.
- (5a) If we have a balanced design (i.e.,  $n_1 = n_2 = \dots = n_k = n$ ), then the between-group variance component ( $\sigma_A^2$ ) is estimated by

$$\hat{\sigma}_A^2 = \max \left[ \left( \frac{\text{Between MS} - \text{Within MS}}{n} \right), 0 \right]$$

(5b) If we have an unbalanced design (i.e., at least two of the  $n_i$  are unequal), then the between-group variance component ( $\sigma_A^2$ ) is estimated by

$$\hat{\sigma}_A^2 = \max \left[ \left( \frac{\text{Between MS} - \text{Within MS}}{n_0} \right), 0 \right]$$

where

$$n_0 = \left( \sum_{i=1}^k n_i - \sum_{i=1}^k n_i^2 \Big/ \sum_{i=1}^k n_i \right) \Big/ (k-1)$$

### Example 12.28

**Endocrinology** Test whether the underlying mean ln plasma estradiol is the same for different subjects using the data in Table 12.20.

#### Solution

We have used the SAS GLM (general linear model) procedure to perform the significance test in Equation 12.33 based on the ln estradiol values in Table 12.20. These results are given in Table 12.21.

**Table 12.21**

**Analysis of the plasma-estradiol data (ln scale) in Table 12.20 using the SAS GLM procedure**

The GLM Procedure						
Dependent Variable: LESTRADL						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	4	2.65774661	0.66443665	22.15	0.0022	
Error	5	0.15001218	0.03000244			
Corrected Total	9	2.80775879				
R-Square		Coeff Var	Root MSE	LESTRADL Mean		
0.946572		6.744243	0.173212	2.568296		

In Table 12.20 we have 5 subjects and 2 replicates per subject. The  $F$  statistic (given under  $F$  Value) = 22.15  $\sim F_{4,5}$  under  $H_0$ . The  $p$ -value =  $Pr(F_{4,5} > 22.15) = .0022$  (given under  $Pr > F$ ). Thus there are significant differences among the underlying mean ln(plasma estradiol) values for different subjects.

### Example 12.29

**Endocrinology** Estimate the between-person and within-person variance components for ln(plasma estradiol) using the data in Table 12.20.

#### Solution

From Equation 12.33 we have the within-person variance component estimated by the Within Mean Square. In Table 12.21, this is called the Error Mean Square (or MSE) = 0.030. To estimate the between-person variance, we refer to Equation 12.33. Because this is a balanced design, we have

$$\begin{aligned}\hat{\sigma}_A^2 &= \left[ \max \left( \frac{\text{Between MS} - \text{Within MS}}{2} \right), 0 \right] \\ &= \frac{0.6644 - 0.0300}{2} = \frac{0.6344}{2} = 0.317\end{aligned}$$

Thus the between-person variance is about 10 times as large as the within-person variance, which indicates good reproducibility.

Alternatively, we could refer to the SAS GLM procedure output given in Table 12.22.

**Table 12.22 Representation of the Expected Mean Squares in terms of sources of variance using the data in Table 12.20**

<b>The GLM Procedure</b>	
<b>Source</b>	<b>Type III Expected Mean Square</b>
<b>person</b>	<b>Var(Error) + 2 Var(person)</b>

We see that the person (or Between Mean Square) gives an unbiased estimate of within-person variance + 2 between-person variance [which is denoted by  $\text{Var}(\text{Error}) + 2 \text{Var}(\text{person}) = \sigma^2 + 2\sigma_A^2$ . Thus because we already have an estimate of  $\text{Var}(\text{Error})$  we can compute  $\text{Var}(\text{person})$  by subtraction; that is,  $(\text{Model Mean Square} - \text{Error Mean Square})/2 = 0.317$ . The representation in Table 12.22 is most useful for unbalanced designs because it obviates the need for the user to compute  $n_0$  in step 5b of Equation 12.33. In this case, the expected value of the person Mean Square would be  $\text{Var}(\text{Error}) + n_0 \text{Var}(\text{person})$ .

Another parameter that is often of interest in reproducibility studies is the coefficient of variation (CV). Generally speaking, CVs of <20% are desirable, whereas CVs of >30% are undesirable. The CV in reproducibility studies is defined as

$$CV = 100\% \times \frac{\text{within-person standard deviation}}{\text{within-person mean}}$$

We could estimate the mean and standard deviation for each of the five subjects in Table 12.20 based on the raw plasma-estradiol values and average the individual CV estimates. However, if the standard deviation appears to increase as the mean increases, then a better estimate of the CV is given as follows [7].

#### Equation 12.34

##### Estimation of the CV in Reproducibility Studies

Suppose we have  $k$  subjects enrolled in a reproducibility study where there are  $n_i$  replicates for the  $i$ th subject,  $i = 1, \dots, k$ . To estimate the CV,

- (1) Apply the ln transformation to each of the values.
- (2) Estimate the between- and within-subject variance components using a one-way random-effects model ANOVA as shown in Equation 12.33.
- (3) The CV in the *original scale* is estimated by

$$100\% \times \sqrt{\text{Within MS}} \text{ from step 2}$$

#### Example 12.30

**Endocrinology** Estimate the CV for plasma estradiol given the data in Table 12.21.

#### Solution

We have from Table 12.21 that the Within-person Mean Square based on ln-transformed plasma-estradiol values is given by 0.0300. Thus

$$CV = 100\% \times \sqrt{0.0300} = 17.3\%$$

Alternatively, we could have used  $100\% \times \text{Root MSE} = 100\% \times \sqrt{\text{Error Mean Square}} = 17.3\%$ . Note that the CV given in Table 12.21 (6.74%) is *not* appropriate in this case because it is simply based on  $100\% \times \sqrt{\text{MSE}}/\text{estradiol mean}$ , which would give us the

CV for  $\ln(\text{plasma estradiol})$  rather than plasma estradiol itself. Similarly, if we had run the GLM procedure using the raw scale rather than the  $\ln$  scale (see Table 12.23), the CV given (16.4%) would also not be appropriate because it is based on the assumption that the standard deviation is independent of the mean value, which is not true in the raw scale.

**Table 12.23** Analysis of the plasma-estradiol data (original scale) in Table 12.20 using the SAS GLM procedure

The GLM Procedure						
Dependent Variable: estradiol						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	4	593.3140000	148.3285000	24.55	0.0017	
Error	5	30.2150000	6.0430000			
Corrected Total	9	623.5290000				
R-Square		Coeff Var	Root MSE	estradiol Mean		
0.951542		16.39928	2.458251	14.99000		

In some examples, there are more than two sources of variation.

### Example 12.31

**Hypertension** Suppose we obtain blood-pressure recordings from each of  $k$  subjects. We ask each subject to return to the clinic at  $n_1$  visits. At each of the  $n_1$  visits, we obtain  $n_2$  blood-pressure readings. In this setting, we would be interested in three components of blood-pressure variation: (1) between persons, (2) between different visits for the same person, and (3) between different readings for the same person at the same visit.

This is called a random-effects ANOVA model with more than one level of nesting. This type of problem is beyond the scope of this book. See Snedecor and Cochran [8] for a lucid description of this problem.

In this section, we have examined the **one-way ANOVA random-effects model**. The random-effects model differs from the fixed-effects model in several important ways. First, with a random-effects model we are not interested in comparing mean levels of the outcome variable (e.g., estradiol) among specific levels of the grouping (or categorical variable). Thus, in Example 12.27, we were not interested in comparing mean estradiol levels among different specific women. Instead, the women in the study are considered a random sample of all women who could have participated in the study. It is usually a foregone conclusion that individual women will have different estradiol levels. Instead, what is of interest is estimating what proportion of the total variability of estradiol is attributable to between-person vs. within-person variation. Conversely, in the fixed-effects ANOVA (e.g., Example 12.1) we were interested specifically in comparing mean FEF levels for nonsmokers vs. passive smokers. In the fixed-effects case, the levels of the categorical variable have inherent meaning and the primary goal is to compare mean levels of the outcome variable (FEF) among different levels of the grouping variable.

### REVIEW QUESTIONS 12E

- 1 What is the difference between a fixed-effects ANOVA model and a random-effects ANOVA model? When do we use each?
- 2 Consider the activated-protein-C (APC) resistance split-sample data in Table 2.17.

- (a) Fit a one-way random-effects ANOVA model to these data.
- (b) Estimate the between-person and within-person components of variation.
- (c) Do you think this is a reproducible assay? Why or why not?

## 12.9 The Intraclass Correlation Coefficient

In Sections 11.7–11.8, we were concerned with Pearson correlation coefficients between two distinct variables denoted by  $x$  and  $y$ . For example, in Example 11.22 we were concerned with the correlation between cholesterol levels of a wife ( $x$ ) and a husband ( $y$ ). In Example 11.29 we were concerned with the correlation of body weight between a father ( $x$ ) and his first-born son ( $y$ ). In some instances, we are interested in the correlation between variables that are not readily distinguishable from each other.

### Example 12.32

**Endocrinology** In Example 12.27, we were concerned with the reproducibility of two replicate measures obtained from split samples of plasma estradiol from 5 women. In this instance, the replicate-sample determinations are indistinguishable from each other, since each plasma sample was split into two halves at random. Thus it is impossible to specifically identify an  $x$  or  $y$  variable.

A more fundamental issue is that in Equation 11.17 the sample correlation coefficient was written in the form

$$s_{xy}/(s_x s_y)$$

Thus from Equation 11.17 an alternative definition of the correlation coefficient  $r$  is the ratio of the sample covariance between  $x$  and  $y$  divided by the product of the standard deviation of  $x$  multiplied by the standard deviation of  $y$ . The implicit assumption in Equation 11.17 is that  $x$  and  $y$  are distinct variables and thus the sample mean and standard deviation are computed separately for  $x$  and  $y$ . In Example 12.27, we could, at random, assign one of the two replicates to be  $x$  and the other  $y$  for each woman and compute  $r$  in Equation 11.17 based on separate estimates of the sample mean and standard deviation for  $x$  and  $y$ . However, if  $x$  and  $y$  are indistinguishable from each other, a more efficient estimate of the mean and standard deviation can be obtained by using all available replicates for each woman. Thus a special type of correlation is needed between repeated measures on the same subject, called an intraclass correlation coefficient.

### Definition 12.13

---

Suppose we have  $k$  subjects and obtain  $n_i$  replicates from the  $i$ th subject,  $i = 1, \dots, k$ . Let  $y_{ij}$  represent the  $j$ th replicate from the  $i$ th subject. The correlation between two replicates from the same subject—i.e., between  $y_{ij}$  and  $y_{il}$  where  $j \neq l$  and  $1 \leq j \leq n_i$ ,  $1 \leq l \leq n_i$  is called an **intraclass correlation coefficient**, denoted by  $\rho_i$ .

---

If  $y_{ij}$  follows a one-way random-effects ANOVA model, where

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \alpha_i \sim N(0, \sigma_A^2), \quad e_{ij} \sim N(0, \sigma^2)$$

then it can be shown that  $\rho_i = \sigma_A^2 / (\sigma_A^2 + \sigma^2)$ ; i.e.,  $\rho_i$  is the ratio of the between-person variance divided by the sum of the between-person and the within-person variance. The intraclass correlation coefficient is a measure of reproducibility of replicate measures from the same subject. It ranges between 0 and 1, with  $\rho_i = 0$  indicating no

reproducibility at all (i.e., large within-person variability and 0 between-person variability) and  $\rho_I = 1$  indicating perfect reproducibility (i.e., 0 within-person variability and large between-person variability). According to Fleiss [9],

### Equation 12.35

#### Interpretation of Intraclass Correlation

$\rho_I < 0.4$  indicates poor reproducibility

$0.4 \leq \rho_I < 0.75$  indicates fair to good reproducibility

$\rho_I \geq 0.75$  indicates excellent reproducibility

There are several methods of estimation for the intraclass correlation coefficient. The simplest and perhaps most widely used method is based on the one-way random-effects model ANOVA, which we discussed in Section 12.8.

### Equation 12.36

#### Point and Interval Estimation of the Intraclass Correlation Coefficient

Suppose we have a one-way random-effects model ANOVA, where

$$\gamma_{ij} = \mu + \alpha_i + e_{ij}, e_{ij} \sim N(0, \sigma^2), \alpha_i \sim N(0, \sigma_A^2), i = 1, \dots, k; j = 1, \dots, n_i$$

The intraclass correlation coefficient  $\rho_I = \sigma_A^2 / (\sigma_A^2 + \sigma^2)$ . A point estimate of  $\rho_I$  is given by

$$\hat{\rho}_I = \max \left[ \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}^2}, 0 \right]$$

where  $\hat{\sigma}_A^2$  and  $\hat{\sigma}^2$  are the estimates of the between-subject and within-subject variance components from a one-way random-effects model ANOVA given in Equation 12.33. This estimator is sometimes referred to as the **analysis-of-variance estimator**.

An approximate two-sided  $100\% \times (1 - \alpha)$  CI for  $\rho_I$  is given by  $(c_1, c_2)$  where

$$c_1 = \max \left\{ \frac{F/F_{k-1, N-k, 1-\alpha/2} - 1}{n_0 + F/F_{k-1, N-k, 1-\alpha/2} - 1}, 0 \right\}$$

$$c_2 = \max \left\{ \frac{F/F_{k-1, N-k, \alpha/2} - 1}{n_0 + F/F_{k-1, N-k, \alpha/2} - 1}, 0 \right\}$$

$F$  is the  $F$  statistic from the significance test of the hypothesis  $H_0: \sigma_A^2 = 0$  vs.  $H_1: \sigma_A^2 > 0$  given in Equation 12.33,  $N = \sum_{i=1}^k n_i$  and  $n_0 = \left( \sum_{i=1}^k n_i - \sum_{i=1}^k n_i^2 / \sum_{i=1}^k n_i \right) / (k-1)$  as given in Equation 12.33.

If all subjects have the same number of replicates ( $n$ ), then  $n_0 = n$ .

### Example 12.33

**Endocrinology** Estimate the intraclass correlation coefficient between replicate plasma-estradiol samples based on the data in Table 12.20.

#### Solution

We will use the  $\ln(\text{plasma estradiol})$  values as we did in Table 12.21 because we found that the within-person variance was relatively constant on the log scale but depended on the mean value in the original (raw) scale. From Example 12.29, we have  $\hat{\sigma}_A^2 = 0.317$ ,  $\hat{\sigma}^2 = 0.030$ . Therefore, our point estimate of the intraclass correlation is given by

$$\begin{aligned}\rho_1 &= \frac{0.317}{0.317 + 0.030} \\ &= \frac{0.317}{0.347} = 0.914\end{aligned}$$

Thus there is excellent reproducibility for  $\ln(\text{plasma estradiol})$ . To obtain confidence limits about this estimate, we refer to Equation 12.36. From Table 12.21, we see that the  $F$  statistic from the SAS General Linear Model procedure based on a one-way random-effects model ANOVA is 22.15. Also, because this is a balanced design (i.e., we have the same number of replicates for each subject), we have  $n_0 = n = 2$ . We also need the critical values (which we obtain from Table 9 in the Appendix or from Excel) given by  $F_{4,5,.975} = 7.39$  and  $F_{4,5,.025} = 1/F_{5,4,.975} = 1/9.36 = 0.107$ . Therefore, the 95% CI for  $\rho_1$  is given by  $(c_1, c_2)$ , where

$$\begin{aligned}c_1 &= \max \left\{ \frac{\frac{22.15}{F_{4,5,.975}} - 1}{2 + \frac{22.15}{F_{4,5,.975}} - 1}, 0 \right\} \\ &= \max \left\{ \frac{\frac{22.15}{7.39} - 1}{1 + \frac{22.15}{7.39}}, 0 \right\} \\ &= \frac{1.997}{3.997} = .500 \\ c_2 &= \max \left\{ \frac{\frac{22.15}{F_{4,5,.025}} - 1}{2 + \frac{22.15}{F_{4,5,.025}} - 1}, 0 \right\} \\ &= \max \left\{ \frac{\frac{22.15}{0.107} - 1}{1 + \frac{22.15}{0.107}}, 0 \right\} \\ &= \frac{207.29}{209.29} = .990\end{aligned}$$

Therefore, the 95% CI for  $\rho_1 = (.500, .990)$ , which is quite wide.

Another interpretation for the intraclass correlation is based on *reliability* rather than *reproducibility*.

### Example 12.34

**Hypertension** Suppose we wish to characterize blood pressure for an individual over a short time period (e.g., a 1-month period). Blood pressure is an imprecise measure with a lot of within-person variation. Therefore, the ideal way to characterize a subject's blood pressure is to take many replicate measures over a short period of time and use the average ( $T$ ) as an estimate of the true level of blood pressure. We then might ask to what extent does a single blood pressure ( $X$ ) relate to  $T$ ? The answer to this question is given by the intraclass correlation coefficient.

**Equation 12.37****Alternative Interpretation of the Intraclass Correlation Coefficient as a Measure of Reliability**

Suppose we have a one-way random-effects model ANOVA, where

$$y_{ij} = \mu + \alpha_i + e_{ij}, e_{ij} \sim N(0, \sigma^2), \alpha_i \sim N(0, \sigma_A^2)$$

where  $y_{ij}$  denotes the  $j$ th replicate from the  $i$ th subject. The average of an infinite number of replicates for the  $i$ th subject is denoted by  $Y_i = \mu + \alpha_i$ . The square of the correlation between  $Y_i$  and a single replicate measure  $y_{ij}$  is given by the intraclass correlation coefficient. Therefore, the intraclass correlation coefficient can also be interpreted as a measure of reliability and is sometimes called the **reliability coefficient**.

**Solution to Example 12.34**

It has been shown that the intraclass correlation coefficient for diastolic blood-pressure (DBP) measurements for 30- to 49-year-olds based on a single visit is .79 [10]. Thus the correlation between DBP obtained at one visit and the “true” DBP is  $\sqrt{.79} = .89$ . To increase the reliability of blood-pressure measurements, clinicians often take an average blood pressure over several visits. The rationale for this practice is that the reliability of a mean DBP over, say, three visits is higher (.96) than for a single visit (.89) in the sense that its correlation with the true DBP is higher. In each instance the average of three readings was used to characterize DBP at any one visit. This is the rationale for the screening design of the Trial of Hypertension Prevention (TOHP), in which approximately 20,000 people were screened in order to identify approximately 2000 subjects with high-normal DBP (defined as having a mean DBP of 80–89 mm Hg based on an average of nine readings over three visits with three measurements per visit) [11].

In this section, we have been introduced to the intraclass correlation coefficient. Suppose individuals (e.g., children) are categorized into groups by a grouping variable (e.g., families). The intraclass correlation coefficient is used to estimate the correlation between two separate members of the same group (e.g., two children in the same family). It is defined differently from an ordinary Pearson correlation coefficient (see Section 11.7). For a Pearson correlation, there are two distinct variables being compared (e.g., cholesterol levels for a husband vs. cholesterol levels for a wife). The mean and variance of each variable is estimated separately. For an intraclass correlation coefficient, it is arbitrary which child is denoted as the  $x$  variable and which is denoted as the  $y$  variable. Thus the estimated mean and variance of  $x$  and  $y$  are the same and are obtained by a pooled estimate over all children over all families. The intraclass correlation coefficient can also be interpreted as a measure of the percentage of total variation that is attributable to between-group (e.g., between-family) variation.

**REVIEW QUESTIONS 12F**

- 1** What is an intraclass correlation coefficient? How does it differ from an ordinary Pearson correlation coefficient?
- 2** Refer to Table 2.17.
  - (a)** Compute the intraclass correlation coefficient between replicate APC-resistance values.
  - (b)** Interpret what it means.
  - (c)** What is the correlation coefficient between a single APC-resistance value and the mean APC-resistance value over a large number of values for an individual subject?

## 12.10 Mixed Models

In some instances we wish to classify units of analysis by two categorical variables at the same time (say  $x_1$ ,  $x_2$ ) and look at the effects of  $x_1$  and  $x_2$  simultaneously on a continuous outcome variable  $y$ , where one of the variables (say  $x_1$ ) can be considered random and the other variable ( $x_2$ ) can be considered fixed.

### Example 12.35

**Ophthalmology** A study was performed among dry eye patients comparing an active drug with placebo. The subjects were either given active drug or placebo and then were put into a “dry eye room,” which is a room with very low humidity (<10%), to exacerbate their symptoms, hopefully to a greater extent with placebo than with active drug. An important physiologic measure used with dry eye patients is the tear break-up time (TBUT), which is the time it takes for a tear to dissolve; a smaller number indicates worse symptoms (more tearing). The TBUT was measured at baseline (bas), immediately after drop instillation (im), and at 5 minutes (pst5), 10 minutes (pst10), and 15 minutes (pst15) after instillation. Data were obtained from both the right (od) and left (os) eyes, and two replicates were obtained for each eye. The data were obtained on the same subjects under three different experimental conditions prior to drop instillation (3 second nonblink period, 6 second nonblink period, 10 second nonblink period).

The results for the right eye among placebo subjects ( $n=14$ ) after the 3 second nonblink period are given in Table 12.24. We wish to test whether there are differences in TBUT among subjects and over time. In this design, the two factors to consider are the subject (which is a random effect) and the time point (which is a fixed effect). This type of design is called a *mixed model*. How should we analyze the data in this setting?

**Table 12.24 Tear break-up time over 15 minutes among 14 dry eye patients in the placebo group of a clinical trial after a 3 second nonblink period**

	od bas 1	od bas 2	od im 1	od im 2	od pst5 1	od pst5 2	od pst10 1	od pst10 2	od pst15 1	od pst15 2
1	5.44	4.59	4.34	5.31	4.81	6.53	6.00	4.63	6.47	7.03
2	3.28	3.00	10.87	19.06	13.34	12.31	10.34	9.71	5.81	7.25
3	3.18	2.43	14.78	16.28	12.53	16.84	5.53	6.68	6.78	6.43
4	2.47	1.40	11.12	6.44	8.46	3.84	1.93	2.46	2.62	3.21
5	4.40	4.90	12.93	14.84	7.43	9.78	5.93	6.81	6.28	7.65
6	4.93	4.87	10.56	11.71	4.53	5.50	5.37	3.75	4.56	3.31
7	7.21	6.15	14.34	15.50	10.56	10.87	10.43	8.57	3.15	2.78
8	4.93	4.15	15.31	14.00	7.51	6.59	4.01	3.59	3.90	3.62
9	3.18	2.84	6.90	5.75	8.09	7.03	5.31	6.81	4.09	4.06
10	3.47	1.37	6.91	3.57	7.06	5.09	1.53	1.06	2.96	1.34
11	9.46	8.50	12.03	10.75	17.03	14.93	12.31	14.62	13.06	15.09
12	3.03	3.12	7.12	7.19	5.68	4.32	2.41	3.10	4.47	4.25
13	2.47	2.62	18.97	10.60	2.06	2.66	1.87	2.91	2.22	2.40
14	2.66	2.32	12.97	14.81	5.03	3.03	2.35	1.31	1.32	1.19

We will use a two-way mixed-effects ANOVA model for this purpose given as follows.

**Equation 12.38****Two-way ANOVA Model-Balanced Design with One Fixed Effect and One Random Effect**

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, i = 1, \dots, r; j = 1, \dots, c; k = 1, \dots, n$$

where

$y_{ijk}$  =  $k$ th replicate for subject  $i$  at time point  $j$

$\alpha_i$  = row classification (in this case, subject)

$\beta_j$  = column classification (in this case, time)

$(\alpha\beta)_{ij}$  = interaction effect between row and column classifications

$e_{ijk}$  = error term assumed to follow an  $N(0, \sigma^2)$  distribution

where  $\alpha_i$  is a random effect distributed as  $N(0, \sigma_A^2)$ ,  $\beta_j$  is a fixed effect with 5 levels,  $(\alpha\beta)_{ij}$  is an interaction effect, and  $e_{ijk}$  is an error term distributed as  $N(0, \sigma^2)$ .

The results can be analyzed in terms of an ANOVA table as follows.

**Table 12.25 Analysis of two-way ANOVA table with one fixed effect and one random effect**

Effect	SS	df	MS	F Stat	df
Row (random)	$RSS = \sum_{i=1}^r y_{i..}^2 / (nc) - y_{...}^2 / (nrc)$	$r-1$	$RMS = RSS / (r-1)$	$RMS / EMS$	$r-1, rc(n-1)$
Column (fixed)	$CSS = \sum_{j=1}^c y_{.j}^2 / (nr) - y_{...}^2 / (nrc)$	$c-1$	$CMS = CSS / (c-1)$	$CMS / IMS$	$c-1, (r-1)(c-1)$
Interaction	$ISS = \sum_{i=1}^r \sum_{j=1}^c y_{ij}^2 / n - RSS - CSS - y_{...}^2 / (nrc)$	$(r-1)(c-1)$	$IMS = ISS / [(r-1)(c-1)]$	$IMS / EMS$	$(r-1)(c-1), rc(n-1)$
Error	$ESS = \sum_{i=1}^r \sum_{j=1}^c (y_{ijk} - \bar{y}_{ij.})^2$	$rc(n-1)$	$EMS = ESS / [rc(n-1)]$		

In general, with two-way ANOVA we might have designs where

- (a) both factors are fixed (Model I)
- (b) both factors are random (Model II)
- (c) one factor is fixed and one is random (Model III)

Methods for computation of  $F$  statistics for significance tests for these three designs are given in Table 12.26.

**Table 12.26 Computation of the  $F$  statistics for tests of significance in a two-factor ANOVA with replication**

Hypothesized effect	Model I (factors $A$ and $B$ both fixed)	Model II (factors $A$ and $B$ both random)	Model III (factor $A$ random; factor $B$ fixed)
Factor $A$	$\frac{\text{factor } A \text{ MS}}{\text{error MS}}$	$\frac{\text{factor } A \text{ MS}}{A \times B \text{ MS}}$	$\frac{\text{factor } A \text{ MS}}{\text{error MS}}$
Factor $B$	$\frac{\text{factor } B \text{ MS}}{\text{error MS}}$	$\frac{\text{factor } B \text{ MS}}{A \times B \text{ MS}}$	$\frac{\text{factor } B \text{ MS}}{A \times B \text{ MS}}$
$A \times B$ interaction	$\frac{A \times B \text{ MS}}{\text{error MS}}$	$\frac{A \times B \text{ MS}}{\text{error MS}}$	$\frac{A \times B \text{ MS}}{\text{error MS}}$

**Example 12.36** **Ophthalmology** Analyze the dry eye data in Table 12.24 using mixed-model methods.

**Solution** We use the general linear model procedure of MINITAB treating id as random and time as fixed with results given in Table 12.27.

**Table 12.27** Use of the general linear model to analyze the dry eye data in Table 12.24

General Linear Model: tbut vs. id, time						
Factor	Type	Levels	Values			
id	random	14	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14			
time	fixed	5	1=bas, 2=im, 3=5 min pst, 4=10 min pst, 5=15 min pst			
Analysis of Variance for tbut, using adjusted SS for tests						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
id	13	897.582	897.582	69.045	5.48	0.000
time	4	964.637	964.637	241.159	19.13	0.000
id*time	52	655.523	655.523	12.606	5.89	0.000
Error	70	149.769	149.769	2.140		
Total	139	2667.511				
$S = 1.46272 \quad R\text{-sq} = 94.39\% \quad R\text{-Sq(adj)} = 88.85\%$						
Least Squares Means for tbut						
time	Mean					
1=bas	4.013					
2=im	11.249					
3=5 min pst	7.980					
4=10 min pst	5.405					
5=15 min pst	4.904					

We see that TBUT increases immediately after drop instillation but then drops towards baseline over the next 15 minutes. We see that there are significant effects of both subject and time ( $p < .001$ ). Note that MINITAB computes the  $F$  statistic for the random effect (id) differently than in Table 12.26. The appropriate  $F$  statistic for id should be  $69.045 / 2.140 = 32.26 \sim F_{13,70}$  under  $H_0$  with  $p\text{-value} = Pr(F_{13,70} > 32.26) < .001$ . In addition, we can compare means of groups corresponding to specific levels of the fixed effect (time) using the Tukey approach. To compare mean TBUT at time  $i_1$  with mean TBUT at time  $i_2$ , we compute the studentized range statistic  $q$ , where

$$q = \frac{\bar{y}_{i_1} - \bar{y}_{i_2}}{\sqrt{s_{\text{INT}}^2 / (rn)}} \sim q_{c,(r-1)(c-1)}$$

except that for a mixed model  $s_{\text{INT}}^2$  is estimated from the Interaction Mean Square (IMS) rather than the Error Mean Square (EMS). The  $q$  statistic is compared with  $q_{c,(r-1)(c-1),.05}$  for significance. Percentiles of the studentized range statistic are given in Appendix Table 16. Finally, 95% confidence limits are obtained from

$$\bar{y}_{i_1} - \bar{y}_{i_2} \pm q_{c,(r-1)(c-1),.05} \sqrt{s_{\text{INT}}^2 / (rn)}$$

The results for the TBUT data in Table 12.24 are shown in Table 12.28 and Figure 12.15.

**Table 12.28 Comparison of specific groups using the Tukey approach in the dry eye data**

Tukey Simultaneous Tests

Response Variable tbut

All Pairwise Comparisons among Levels of time

time = 1 = bas subtracted from:

time	Difference of Means	SE Difference	t-value	Adjusted p-value
1=im	7.2354	0.9489	7.6249	0.0000
3=5 min pst	3.9668	0.9489	4.1803	0.0010
4=10 min pst	1.3914	0.9489	1.4663	0.5885
5=15 min pst	0.8904	0.9489	0.9383	0.8806

time = 2=im subtracted from:

time	Difference of Means	SE Difference	t-value	Adjusted p-value
3=5 min pst	-3.269	0.9489	-3.445	0.0096
4=10 min pst	-5.844	0.9489	-6.159	0.0000
5=15 min pst	-6.345	0.9489	-6.687	0.0000

time = 3=5 min pst subtracted from:

time	Difference of Means	SE Difference	t-value	Adjusted p-value
4=10 min pst	-2.575	0.9489	-2.714	0.0654
5=15 min pst	-3.076	0.9489	-3.242	0.0170

time = 4=10 min pst subtracted from:

time	Difference of Means	SE Difference	t-value	Adjusted p-value
5=15 min pst	-0.5011	0.9489	-0.5280	0.9841

We see that there are significant effects on TBUT after drop instillation despite the fact that a placebo drop was used. Compared with baseline, there is a significant increase (i.e., improvement) in TBUT immediately after drop instillation ( $p < .001$ ) and 5 minutes post drop instillation ( $p = .001$ ), which wears off after 10 minutes. In addition, mean TBUT is significantly lower 5, 10, and 15 minutes post drop instillation compared with immediately after drop instillation ( $p < .01$ ). Furthermore, mean TBUT 5 minutes post drop instillation is significantly different from mean TBUT 15 minutes post drop instillation ( $p = .017$ ) and nearly significantly different from mean TBUT 10 minutes post drop instillation ( $p = .065$ ). Clearly, if an active eye drop is to be tested vs. a placebo eye drop, patients must be followed for longer than 5 minutes to have the potential to show significant differences vs. placebo. Finally, 95% CI for differences between pairs of means using the Tukey approach are given in Figure 12.15.

We can also use ANOVA to assess more than two factors at the same time (including both fixed and random effects) (i.e., multi-way ANOVA) using the mixed-model approach, but this is beyond the scope of this text. In addition, we can use mixed-model methods if there are an unequal number of replicates per subject at each time point.

**Figure 12.15** 95% Confidence intervals comparing pairs of groups using the dry eye data in Table 12.24

Tukey 95% Simultaneous Confidence Intervals

Response Variable tbtr

All Pairwise Comparisons among Levels of time

time = 1 = bas subtracted from:

time	Lower	Center	Upper	
2 = im	4.551	7.2354	9.919	(----*----)
3 = 5 min pst	1.283	3.9668	6.651	(----*----)
4 = 10 min pst	-1.293	1.3914	4.075	(----*----)
5 = 15 min pst	-1.794	0.8904	3.574	(----*----)
				-----+-----+-----+
				-6.0    0.0    6.0    12.0

time = 2 = im subtracted from:

time	Lower	Center	Upper	
3 = 5 min pst	-5.953	-3.269	-0.585	(----*----)
4 = 10 min pst	-8.528	-5.844	-3.160	(----*----)
5 = 15 min pst	-9.029	-6.345	-3.661	(----*----)
				-----+-----+-----+
				-6.0    0.0    6.0    12.0

time = 3 = 5 min pst subtracted from:

time	Lower	Center	Upper	
4 = 10 min pst	-5.259	-2.575	0.1086	(----*----)
5 = 15 min pst	-5.760	-3.076	-0.3925	(----*----)
				-----+-----+-----+
				-6.0    0.0    6.0    12.0

time = 4 = 10 min pst subtracted from:

time	Lower	Center	Upper	
5 = 15 min pst	-3.185	-0.5011	2.183	(----*----)
				-----+-----+-----+

## 12.11 Summary

One-way ANOVA methods enable us to relate a normally distributed outcome variable to the levels of a single categorical independent variable. Two different ANOVA models based on a fixed- or random-effects model, respectively, were considered. In a fixed-effects model, the levels of the categorical variable are determined in advance. A major objective of this design is to test the hypothesis that the mean level of the dependent variable is different for different groups defined by the categorical variable. The specific groups that are different can also be identified, using *t* tests based on the LSD procedure if the comparisons have been planned in advance or using multiple-comparisons methods if they have not. More complex comparisons such as testing for dose-response relationships involving more than two levels of the categorical variable can also be accomplished using linear-contrast methods. Fixed-effects ANOVA methods can be thought of as a generalization of two-sample inference based on *t* tests presented in Chapter 8.

In addition, nonparametric methods for fixed-effects ANOVA based on the Kruskal-Wallis test were also discussed for situations when the assumption of

normality is questionable. This test can be thought of as a generalization of the Wilcoxon rank-sum test presented in Chapter 9 if more than two groups are being compared.

Under a random-effects model, the levels of the categorical variable are determined at random, with the levels being drawn from an underlying normal distribution whose variance characterizes between-subject variation in the study population. For a given subject, there is, in addition, within-subject variation around the underlying mean for that subject. A major objective of a random-effects design is to estimate the between- and within-subject variance components. Random-effects models are also useful in computing coefficients of variation in reproducibility studies.

We also discussed the two-way ANOVA, in which we are interested in jointly comparing the mean levels of an outcome variable according to the levels of two categorical variables (e.g., mean level of blood pressure by both sex and ethnic group). With two-way ANOVA, we can simultaneously estimate the main effects of sex (e.g., the effect of sex on blood pressure after controlling for ethnic group), the main effects of ethnic group (e.g., the effect of ethnic group on blood pressure after controlling for sex), and the interaction effects between sex and ethnic group (e.g., estimates of differences of ethnic-group effects on blood pressure between males and females). We also saw that both one-way and two-way fixed-effects ANOVA models can be considered as special cases of multiple-regression models by using dummy-variable coding for the effects of the categorical variable(s).

Furthermore, we examined both one-way and two-way ANCOVA. In one-way ANCOVA, we are interested primarily in relating a continuous outcome variable to a categorical variable but want to control for other covariates. Similarly, in two-way ANCOVA we are interested primarily in relating a continuous outcome variable simultaneously to two categorical variables but want to control for other covariates. We also saw that both one-way and two-way ANCOVA models can be represented as special cases of multiple-regression models. Finally, we investigated mixed-effects models where one or more factors is fixed by design and one or more factors is random.

## PROBLEMS

### Nutrition

Researchers compared protein intake among three groups of postmenopausal women: (1) women eating a standard American diet (STD), (2) women eating a lacto-ovo-vegetarian diet (LAC), and (3) women eating a strict vegetarian diet (VEG). The mean  $\pm$   $sd$  for protein intake (mg) is presented in Table 12.29.

**Table 12.29 Protein intake (mg) among three dietary groups of postmenopausal women**

Group	Mean	sd	n
STD	75	9	10
LAC	57	13	10
VEG	47	17	6

\***12.1** Perform a statistical procedure to compare the means of the three groups using the critical-value method.

\***12.2** What is the  $p$ -value from the test performed in Problem 12.1?

\***12.3** Compare the means of each specific pair of groups using the LSD methodology.

\***12.4** Suppose that in the general population, 70% of vegetarians are lacto-ovo-vegetarians, whereas 30% are strict vegetarians. Perform a statistical procedure to test whether the contrast  $L = 0.7\bar{y}_2 + 0.3\bar{y}_3 - \bar{y}_1$  is significantly different from 0. What does the contrast mean?

**12.5** Using the data in Table 12.29, perform a multiple-comparisons procedure to identify which specific underlying means are different.

### Pulmonary Disease

Twenty-two young asthmatic volunteers were studied to assess the short-term effects of sulfur dioxide ( $SO_2$ ) exposure under various conditions [12]. The baseline data in Table 12.30 were presented regarding bronchial reactivity to  $SO_2$ , stratified by lung function (as defined by forced expiratory volume / forced vital capacity [ $FEV_1/FVC$ ]) at screening.

**Table 12.30 Relationship of bronchial reactivity to SO<sub>2</sub> (cm H<sub>2</sub>O/s) grouped by lung function at screening among 22 asthmatic volunteers**

Lung-function group		
Group A FEV <sub>1</sub> /FVC ≤ 74%	Group B FEV <sub>1</sub> /FVC 75–84%	Group C FEV <sub>1</sub> /FVC ≥ 85%
20.8	7.5	9.2
4.1	7.5	2.0
30.0	11.9	2.5
24.7	4.5	6.1
13.8	3.1	7.5
	8.0	
	4.7	
	28.1	
	10.3	
	10.0	
	5.1	
	2.2	

Source: Reprinted with permission of the *American Review of Respiratory Disease*, 131(2), 221–225, 1985.

\***12.6** Test the hypothesis that there is an overall mean difference in bronchial reactivity among the three lung-function groups.

\***12.7** Compare the means of each pair of groups using the LSD method.

\***12.8** Compare the means of each pair of groups using the Bonferroni method.

### Hypertension

Automated blood-pressure measuring devices have appeared in many banks, drugstores, and other public places. A study was conducted to assess the comparability of machine readings vs. readings using the standard cuff [13]. Readings were taken using both the machine and the standard cuff at four separate locations. The results are given in Table 12.31. Suppose we want to test whether the mean difference between

machine and standard cuff readings is consistent over the four locations (i.e., if the bias is comparable over all four locations).

**12.9** Is a fixed-effects or a random-effects ANOVA appropriate here?

**12.10** Test whether the mean difference is consistent over all four locations.

**12.11** Estimate the proportion of the variance attributable to between-machine vs. within-machine variability.

### Mental Health

For the purpose of identifying older nondemented people with early signs of senile dementia, a Mental Function Index was constructed based on three short tests of cognitive function. In Table 12.32, data relating the Mental Function Index at baseline to clinical status determined independently at baseline and follow-up, with a median follow-up period of 959 days, are presented [14].

**Table 12.32 Relationship between clinical status at baseline and follow-up (median follow-up period of 959 days) to mean Mental Function Index at baseline**

Clinical Status		Mean	sd	n
Baseline	Follow-up			
Normal	Unchanged	0.04	0.11	27
Normal	Questionably or mildly affected	0.22	0.17	9
Questionably affected	Progressed	0.43	0.35	7
Definitely affected	Progressed	0.76	0.58	10

Source: Reprinted with permission of the *American Journal of Epidemiology*, 120(6), 922–935, 1984.

**12.12** What test procedure can be used to test for significant differences among the groups?

**12.13** Perform the test mentioned in Problem 12.12, and report appropriate p-values identifying differences between specific groups.

**Table 12.31 Mean SBP and difference between machine and human readings at four locations**

Location	SBP machine (mm Hg)			SBP standard cuff (mm Hg)			SBP machine – SBP standard cuff (mm Hg)		
	Mean	sd	n	Mean	sd	n	Mean	sd	n
A	142.5	21.0	98	142.0	18.1	98	0.5	11.2	98
B	134.1	22.5	84	133.6	23.2	84	0.5	12.1	84
C	147.9	20.3	98	133.9	18.3	98	14.0	11.7	98
D	135.4	16.7	62	128.5	19.0	62	6.9	13.6	62

Source: Reprinted with permission of the *American Heart Association, Hypertension*, 2(2), 221–227, 1980.

## Hypertension

Some common strategies for treating hypertensive patients by nonpharmacologic methods include (1) weight reduction and (2) trying to get the patient to relax more by meditational or other techniques. Suppose these strategies are evaluated by randomizing hypertensive patients to four groups who receive the following types of nonpharmacologic therapy:

- Group 1:** Patients receive counseling for both weight reduction and meditation.
- Group 2:** Patients receive counseling for weight reduction but not for meditation.
- Group 3:** Patients receive counseling for meditation but not for weight reduction.
- Group 4:** Patients receive no counseling at all.

Suppose 20 hypertensive patients are assigned at random to each of the four groups, and the change in diastolic blood pressure (DBP) is noted in these patients after a 1-month period. The results are given in Table 12.33.

**Table 12.33 Change in DBP among hypertensive patients who receive different kinds of nonpharmacologic therapy**

Group	Mean change in DBP (baseline – follow-up) (mm Hg)	sd change	n
1	8.6	6.2	20
2	5.3	5.4	20
3	4.9	7.0	20
4	1.1	6.5	20

**12.14** Test the hypothesis that mean change in DBP is the same among the four groups.

**12.15** Analyze whether counseling for weight reduction has a significant effect on reducing blood pressure.

**12.16** Analyze whether meditation instruction has a significant effect on reducing blood pressure.

**12.17** Is there any relationship between the effects of weight-reduction counseling and meditation counseling on blood-pressure reduction? That is, does weight-reduction counseling work better for people who receive meditational counseling or for people who do not receive meditational counseling, or is there no difference in effect between these two subgroups?

## Hypertension

An instructor in health education wants to familiarize her students with the measurement of blood pressure. Each student is given a portable blood-pressure machine to take home and is told to take two readings on each of 10 consecutive days. The data for one student are given in Table 12.34.

**Table 12.34 Systolic blood-pressure (SBP) recordings on one participant for 10 consecutive days with two readings per day**

Day	Reading	
	1	2
1	98	99
2	102	93
3	100	98
4	99	100
5	96	100
6	95	100
7	90	98
8	102	93
9	91	92
10	90	94

**\*12.18** Estimate the between-day and within-day components of variance for this participant.

**\*12.19** Is there a difference in underlying mean blood pressure by day for this participant?

## Bioavailability

Intake of high doses of beta-carotene in food has been associated in some observational studies with a decreased incidence of cancer. A clinical trial was planned comparing the incidence of cancer in a group taking beta-carotene in capsule form compared with a group taking beta-carotene placebo capsules. One issue in planning such a study is which preparation to use for the beta-carotene capsules. Four preparations were considered: (1) Solatene (30-mg capsules), (2) Roche (60-mg capsules), (3) Badische Anilin und Soda Fabrik (BASF) (30-mg capsules), and (4) BASF (60-mg capsules). To test efficacy of the four agents in raising plasma-carotene levels, a small bioavailability study was conducted. After two consecutive-day fasting blood samples, 23 volunteers were randomized to one of the four preparations, taking 1 pill every other day for 12 weeks. The primary endpoint was level of plasma carotene attained after moderately prolonged steady ingestion. For this purpose, blood samples were drawn at 6, 8, 10, and 12 weeks, with results given in Data Set BETACAR.DAT, on the Companion Website. The format of the data is given in Table 12.35.

**12.20** Use ANOVA methods to estimate the CV for plasma beta-carotene for the 23 participants, based on the two baseline measurements.

**12.21** Is there a significant difference in bioavailability of the four different preparations? Use ANOVA methods to assess this based on the 6-week data in comparison with baseline.

**Table 12.35 Format of BETACAR.DAT**

Variable	Column	Code
Preparation	1	1 = SOL; 2 = ROCHE; 3 = BASF-30; 4 = BASF-60
Subject number	3–4	
First baseline level	6–8	
Second baseline level	10–12	
Week 6 level	14–16	
Week 8 level	18–20	
Week 10 level	22–24	
Week 12 level	26–28	

**12.22** Use the methods in Problem 12.21 to compare the bioavailability of the four preparations at 8 weeks vs. baseline.

**12.23** Use the methods in Problem 12.21 to compare the bioavailability of the four preparations at 10 weeks vs. baseline.

**12.24** Use the methods in Problem 12.21 to compare the bioavailability of the four preparations at 12 weeks vs. baseline.

**12.25** Use the methods in Problem 12.21 to compare the bioavailability of the four preparations based on the average plasma beta-carotene at (6, 8, 10, and 12 weeks) vs. baseline.

**12.26** Perform a two-way ANOVA to jointly assess the effects of preparation and follow-up time on plasma-carotene levels. Is there any evidence that the effect of preparation differs by time period?

### Hepatic Disease

Refer to Data Set HORMONE.DAT, on the Companion Website (see p. 315 for a description of the data set).

**12.27** Use ANOVA methods to test whether the change in biliary-secretion levels is comparable for the five hormone groups. Identify and test for any specific group differences.

**12.28** Answer Problem 12.27 for changes in pancreatic-secretion levels.

**12.29** Answer Problem 12.27 for changes in biliary-pH levels.

**12.30** Answer Problem 12.27 for changes in pancreatic-pH levels.

### Endocrinology

A study was conducted [15] concerning the effect of calcium supplementation on bone loss among postmenopausal women. Women were randomized to (1) estrogen cream and calcium placebo ( $n = 15$ ), (2) placebo estrogen cream and 2000 mg/day of calcium ( $n = 15$ ), or (3) placebo estrogen cream and calcium placebo ( $n = 13$ ). Subjects were seen every 3 months for a 2-year period. The rate of bone loss was computed for each woman and expressed as a percentage of the initial bone mass. The results are shown in Table 12.36.

**Table 12.36 Mean ( $\pm 1$  sd) slope of total-body bone mass (percentage per year) in the three treatment groups**

Treatment group		
(1) Estrogen ( $n = 15$ )	(2) Calcium ( $n = 15$ )	(3) Placebo ( $n = 13$ )
$-0.43 \pm 1.60$	$-2.62 \pm 2.68$	$-3.98 \pm 1.63$

Source: Reprinted with permission from *The New England Journal of Medicine*, 316(4), 173–177, 1987.

**12.31** What test can be used to compare the mean rate of bone loss in the three groups?

**12.32** Implement the test in Problem 12.31, and report a *p*-value.

**12.33** Identify which pairs of groups are different from each other, using both *t* tests and the method of multiple comparisons. Report a *p*-value for each pair of treatment groups.

**12.34** Which methodology do you think is more appropriate in Problem 12.33?

### Endocrinology

Refer to Data Set ENDOCRIN.DAT, on the Companion Website. The data set consists of split-sample plasma determinations of four hormones for each of 5 subjects from one laboratory. The format of the data is given in Table 12.37.

**Table 12.37 Format of ENDOCRIN.DAT**

	Column	Units
Subject number	1	
Replicate number	3	
Plasma estrone	5–8	pg/mL
Plasma estradiol	10–14	pg/mL
Plasma androstanedione	16–19	ng/dL
Plasma testosterone	21–24	ng/dL

**12.35** Estimate the between-subject and within-subject variation for plasma estrone, plasma androstenedione, and plasma testosterone.

**12.36** Estimate the CV for each of the hormones in Problem 12.35.

### Environmental Health

A student wants to find out whether specific locations within a house show heat loss. To assess this she records temperatures at 20 sites within her house for each of 30 days. In addition, she records the outside temperature. The data are given in Data Set TEMPERAT.DAT, on the Companion Website. The format of the data is given in Table 12.38.

**Table 12.38 Format of TEMPERAT.DAT**

	Column	Comment
1. Date	1–6	(mo/da/yr)
2. Outside temperature	8–9	(°F)
3. Location within house	11–12	(1–20)
4. Inside temperature	14–17	(°F)

Note: The data were collected by Sarah Rosner.

**12.37** Assume a random-effects model. Estimate the between-day vs. within-day variation in temperature within this house.

**12.38** Are there significant differences in temperature between different locations in the house?

**12.39** Assume a fixed-effects model. Use the method of multiple comparisons to assess which specific locations in the house are different in mean temperature.

### Environmental Health, Pediatrics

Refer to Data Set LEAD.DAT, on the Companion Website.

**12.40** Use ANOVA methods to assess whether there are any overall differences between the control group, the currently exposed group, and the previously exposed group in mean full-scale IQ. Also, compare each pair of groups and report a *p*-value.

**12.41** Determine a 95% CI for the underlying mean difference in full-scale IQ for each pair of groups considered.

### Gastroenterology

In Table 12.39, we present data relating protein concentration to pancreatic function as measured by trypsin secretion among patients with cystic fibrosis [16].

**12.42** If we do not want to assume normality for these distributions, then what statistical procedure can be used to compare the three groups?

**12.43** Perform the test mentioned in Problem 12.42, and report a *p*-value. How do your results compare with a parametric analysis of the data?

### Environmental Health, Pediatrics

Refer to Data Set LEAD.DAT, on the Companion Website.

**12.44** Use nonparametric methods to compare MAXFWT among the three exposure groups defined by the variable LEAD\_GRP.

**12.45** Answer Problem 12.44 for IQF (full-scale IQ).

**12.46** Compare your results in Problems 12.44 and 12.45 with the corresponding results in Table 12.9 for MAXFWT and Problem 12.40 for IQF where parametric methods were used.

**Table 12.39 Relationship between protein concentration (mg/ml) of duodenal secretions to pancreatic function as measured by trypsin secretion [U/(kg/hr)]**

$\leq 50$		Trypsin secretion [U/(kg/hr)] 51–1000		$>1000$	
Subject number	Protein concentration	Subject number	Protein concentration	Subject number	Protein concentration
1	1.7	1	1.4	1	2.9
2	2.0	2	2.4	2	3.8
3	2.0	3	2.4	3	4.4
4	2.2	4	3.3	4	4.7
5	4.0	5	4.4	5	5.0
6	4.0	6	4.7	6	5.6
7	5.0	7	6.7	7	7.4
8	6.7	8	7.6	8	9.4
9	7.8	9	9.5	9	10.3
		10	11.7		

Source: Reprinted with permission of *The New England Journal of Medicine*, 312(6), 329–334, 1985.

### Renal Disease

Refer to Data Set SWISS.DAT on the Companion Website.

**12.47** Use ANOVA methods to compare the mean change in serum-creatinine values from the baseline visit to the 1978 visit among women in the high-NAPAP group, low-NAPAP group, and the control group.

One issue in Problem 12.47 is that only the first and last visits are used to assess change in serum creatinine over time.

**12.48** Fit a regression line relating serum creatinine to time for each person in each of the high-NAPAP group, the low-NAPAP group, and the control group. How do you interpret the slope and intercept for each person?

**12.49** Use regression analysis or ANOVA methods to compare the slopes of the three groups.

**12.50** Answer the question in Problem 12.49 for the intercepts in the three groups.

**12.51** What are your overall conclusions regarding the comparison of serum creatinine among the three groups?

(Note: One issue in the preceding analyses is that we have considered all subjects as yielding identical information regardless of the number of visits available for analysis. A more precise approach would be to use methods of *longitudinal data analysis* to weight subjects according to the number and spacing of their visits. This is beyond the scope of this text.)

### Bioavailability

Refer to Table 12.35.

**12.52** Compute the intraclass correlation coefficient between replicate plasma beta-carotene blood samples at baseline. Provide a 95% CI about this estimate. Perform the analysis based on the entire data set.

**12.53** Use linear-regression methods to assess whether plasma-carotene levels increase over the 12-week period. Use appropriate data transformations if necessary. Perform separate analyses for each of the four preparations.

**12.54** Do you have any recommendations as to which preparation should be used in the main clinical trial?

### Endocrinology

Refer to Table 12.37.

**12.55** Estimate the intraclass correlation for each of plasma estrone, plasma androstenedione, and plasma testosterone, and provide associated 95% confidence limits. Is the reproducibility of these plasma-hormone levels excellent, good, or poor?

### Environmental Health

Refer to Table 12.38.

**12.56** Use regression methods to assess whether there is any relationship between the indoor and outdoor temperature readings.

### Ophthalmology

The term *retinitis pigmentosa* (RP) refers to a group of hereditary, retinal pigmentary degenerations in which patients report night blindness and loss of visual field, usually between the ages of 10 and 40 years. Some patients lose all useful vision (i.e., become legally blind) by the age of 30 years, while others retain central vision even beyond the age of 60 years. A specific gene has been linked to some types of RP where the mode of genetic transmission is autosomal dominant. The most reliable methods of following the course of RP in humans is by using the electroretinogram (ERG), which is a measure of the electrical activity in the retina. As the disease progresses, the patient's ERG amplitude declines. The ERG amplitude has been strongly related to the patient's ability to perform routine activities, such as driving or walking unaided, especially at night.

One hypothesis is that direct exposure of the retina to sunlight is harmful to RP patients, so many patients wear sunglasses. To test the sunlight hypothesis, researchers introduced this gene into a group of mice and mated them over many generations to produce a group of "RP mice." Then they randomly assigned the mice to lighting conditions from birth that were either (1) light, (2) dim, or (3) dark. A control group of normal mice was also randomized to similar lighting conditions. The mice had their ERG amplitudes (labeled BAMP and AAMP for B-wave amplitude and A-wave amplitude, respectively), which correspond to different frequencies of light, measured at 15, 20, and 35 days of life. In addition, the same protocol was used for a group of normal mice except that only BAMP was measured. The data for both RP mice and normal mice are available in the Data Set MICE.DAT and the documentation in MICE.DOC (both on the Companion Website).

**12.57** Analyze the data regarding the sunlight hypothesis, and summarize your findings. (Hint: Estimate a slope for each ERG amplitude for each lighting-condition group, and compare the slopes among the light, dim, and dark groups using either ANOVA or regression methods. Do separate analyses for AAMP and BAMP. Consider appropriate data transformations to ensure approximate normality for the outcome measures.)

### Hypertension

Refer to Table 12.15. A similar two-way ANOVA was run using PROC GLM of SAS comparing mean diastolic blood pressure (DBP) by study group and sex. The results are given in Table 12.40.

**12.58** Summarize the findings in a few sentences.

**Table 12.40** SAS GLM procedure output illustrating the effects of study group and sex on DBP using the data set in Example 12.20

SAS GENERAL LINEAR MODELS PROCEDURE								
DEPENDENT VARIABLE: MDIAS								
SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.	
MODEL	3	48186.99270094	16062.33090031	134.15	0.0001	0.350741	15.0906	
ERROR	745	89199.44205496	119.73079470			ROOT MSE		MDIAS MEAN
CORRECTED TOTAL	748	137386.43475590			10.94215677		72.50972853	
SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE III SS	F VALUE	PR > F
STUDY	2	45269.88509153	189.05	0.0001	2	46573.92818903	194.49	0.0001
SEX	1	2917.10760942	24.36	0.0001	1	2917.10760942	24.36	0.0001
STUDY								
PROB >  T								
SV LV NOR								
SV		.	0.0001	0.0001				
LV		0.0001	.	0.0001				
NOR		0.0001	0.0001	.				
SEX								
PROB >  T								
MALE FEMALE								
MALE		.	0.0001	0.0001				
FEMALE		0.0001	.	0.0001				
PARAMETER								
ESTIMATE								
INTERCEPT		76.47708914		115.61		PR >  T		STD ERROR OF ESTIMATE
STUDY	SV	-17.30065001		-19.40		0.0001		0.66152912
	LV	-10.65302392		-7.23		0.0001		0.89174716
	NOR	0.00000000		.		0.0001		1.47258921
SEX	MALE	3.98399582		4.94		0.0001		0.80713389
	FEMALE	0.00000000		.		0.0001		.

## Hypertension

Refer to Table 12.16. An ANCOVA was performed using PROC GLM of SAS comparing mean DBP by study group and sex after controlling for effects of age and weight. The results are given in Table 12.41.

**12.59** Summarize the findings in a few sentences, and compare them with the results in Problem 12.58.

## Cardiovascular Disease

A recent randomized trial examined the effects of lipid-modifying therapy (simvastatin plus niacin) and antioxidants (vitamins E and C, beta-carotene, and selenium) on cardiovascular protection in patients with clinical coronary disease, low HDL cholesterol, and normal LDL cholesterol [17]. A total of 160 patients were randomized into four groups: placebo lipid-lowering and placebo antioxidants, active lipid-lowering and placebo antioxidants,

placebo lipid-lowering and active antioxidants, or active lipid-lowering and active antioxidants. All participants had substantial stenoses (blockages) of the coronary arteries quantified by catheterization at baseline, and the primary endpoint was the percent change in a person's stenoses after 3 years of treatment, with a positive change indicating an increased amount of stenosis, as shown in Table 12.42. Because some patients did not complete the study, the primary endpoint was assessed in 146 participants.

**12.60** Perform a one-way ANOVA to assess whether there are significant differences in mean change in percent stenosis among the four groups.

**12.61** Using the LSD method, identify which pairs of groups are significantly different.

**12.62** Are there significant interaction effects between simvastatin-niacin and antioxidants? What does an *interaction*

**Table 12.41** SAS GLM procedure output illustrating the effects of study group and sex on DBP after controlling for age and weight using the data set in Example 12.20

SAS GENERAL LINEAR MODELS PROCEDURE								
DEPENDENT VARIABLE: MDIAS								
SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.	
MODEL	5	57521.19225928	11504.23845186	107.08	0.0001	0.419457	14.2973	
ERROR	741	79611.39165542	107.43777551			ROOT MSE		MDIAS MEAN
CORRECTED TOTAL	746	137132.58391470				10.36521951		72.49770638
SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE III SS	F VALUE	PR > F
STUDY	2	45234.31356527	210.51	0.0001	2	12349.74237359	57.47	0.0001
SEX	1	2912.79699312	27.11	0.0001	1	624.47946004	5.81	0.0162
AGE	1	5237.52247659	48.75	0.0001	1	4245.67574526	39.52	0.0001
WGT	1	4136.55922429	38.50	0.0001	1	4136.55922429	38.50	0.0001
STUDY								
PROB >  T								
SV LV NOR								
SV	.	0.0186	0.0001					
LV	0.0186	.	0.0001					
NOR	0.0001	0.0001	.					
SEX								
PROB >  T								
MALE FEMALE								
MALE	.	0.0162						
FEMALE	0.0162	.						
PARAMETER								
ESTIMATE								
T FOR H0: PARAMETER=0								
INTERCEPT	52.96724415	20.41				0.0001		2.59544038
STUDY	SV -11.18295628	-10.45				0.0001		1.06990647
	LV -7.58825363	-5.29				0.0001		1.43486888
	NOR 0.00000000	.				.		.
SEX	MALE 2.10948458	2.41				0.0162		0.87497527
	FEMALE 0.00000000	.				.		.
AGE	0.30162065	6.29				0.0001		0.04798065
WGT	0.08278540	6.20				0.0001		0.01334174

**Table 12.42** Mean changes ( $\pm$  sd), per patient, in the percentage of stenosis by treatment group

	Placebo (n = 34)	Simvastatin-niacin (n = 33)	Antioxidants (n = 39)	Simvastatin-niacin plus antioxidants (n = 40)
Mean change in stenosis (percentage of diameter)	3.9 ( $\pm$ 5.2)	-0.4 ( $\pm$ 2.8)	1.8 ( $\pm$ 4.2)	0.7 ( $\pm$ 3.2)

effect mean in the context of this trial? (Hint: Use the linear contrasty  $\bar{Y}_4 - \bar{Y}_2 - \bar{Y}_3 + \bar{Y}_1$ , where the group numbers are in the same order as in Table 12.42.)

### Ophthalmology

Refer to Data Set TEAR.DAT on the Companion Website. (See p. 265 for a description of the data set.)

**12.63** Use two-way ANOVA methods to compare the mean change in TBUT immediately post-instillation against baseline according to participant (1, 2, ..., 14) and non-blink time period (3 sec vs. 6 sec vs. 10 sec). For this analysis, average the first and second replicates and right and left eyes for each participant. Thus each participant is represented by an average of four values. (Hint: Represent

the participant effect by a set of 13 dummy variables and the nonblink time-period effect by a set of 2 dummy variables. If there are overall significant differences for the nonblink time period, then identify which nonblink time periods are significantly different.

**12.64** Repeat the analyses in Problem 12.63 for the mean change in TBUT at 5 minutes post-instillation vs. baseline.

**12.65** Repeat the analyses in Problem 12.63 for the mean change in TBUT at 10 minutes post-instillation vs. baseline.

**12.66** Repeat the analyses in Problem 12.63 for the mean change in TBUT at 15 minutes post-instillation vs. baseline.

**12.67** What are your overall conclusions concerning the effects of nonblink time period on change in TBUT?

### Genetics, Diabetes

Suppose we have separately analyzed the effects of 10 SNPs comparing people with type I diabetes vs. controls. The  $p$ -values from these separate analyses are given in Table 12.43.

**Table 12.43 Effects of 10 SNPs on type I diabetes**

SNP	$p$ -value	SNP	$p$ -value
1	.04	6	.62
2	.10	7	.001
3	.40	8	.01
4	.55	9	.80
5	.34	10	.005

**12.68** Use the Bonferroni method to correct for multiple comparisons. Which SNPs show statistically significant effects?

**12.69** Use the FDR method to correct for multiple comparisons using an  $FDR = .05$ . Which SNPs show statistically significant effects? How do the results compare with those in Problem 12.68?

### Ophthalmology

Data were collected on TBUT, a measure of how long it takes for tears to form. This is a commonly used measure to assess dry eye patients, with a smaller number indicating more severe disease. One issue is how reproducible this measure is. For this purpose, replicate values of this measure were obtained from the right eye of 14 dry eye patients 5 minutes after instillation of a placebo drop. The data for the first 7 patients are given in Table 12.44.

**Table 12.44  $\log_e$  (TBUT) obtained 5 minutes after instillation of a placebo drop (seconds)**

Replicate 1	Replicate 2	Mean ( $\bar{y}_i$ )
1.57	1.87	1.72
2.59	2.51	2.55
2.53	2.82	2.675
2.14	1.34	1.74
2.00	2.28	2.14
1.50	1.70	1.60
2.36	2.39	2.365
overall mean ( $\bar{y}$ )		2.113

Assume overall that the  $\ln_e$  TBUT is approximately normally distributed. The  $sd$  ( $\ln_e$  TBUT over the 14 observations) is 0.459.

**12.70** What is a reasonable type of model to fit these data? Write out the model, and explain what the terms mean.

**12.71** Fit this model to the above data and report the  $p$ -value for the overall  $F$  test.

**12.72** What is the intraclass correlation coefficient for (TBUT) for these data? What does it mean?

**12.73** What is the coefficient of variation for TBUT as estimated from these data?

### Nutrition, Endocrinology

A study was performed to relate aspects of childhood diet to measurements of bone density in middle age (50- to 70-year-old) women [18]. The data in Table 12.45 were reported from a Norwegian study relating cod liver oil supplementation in childhood to bone mineral density (BMD) in the distal forearm.

**Table 12.45 Mean BMD by cod liver oil intake during childhood**

Cod liver oil intake during childhood	Mean distal forearm BMD ( $g/cm^2$ ) ( $\bar{y}_i$ )	sd	N
Never	0.435	0.058	267
Irregularly	0.424	0.061	695
Fall and winter	0.428	0.062	1655
Whole year	0.420	0.067	237
Overall mean ( $\bar{y}$ )	0.427		2854

**12.74** We wish to use a one-way ANOVA model to compare the means of the four groups. Should we use a fixed-effects or random-effects model, and why?

**12.75** Table 12.46 is the ANOVA table.

**Table 12.46 ANOVA table of effect of cod liver oil intake during childhood on BMD**

	SS	df	MS
Between	0.0366	3	0.0122
Within	10.8946	2850	0.0038

Assess whether the mean BMD varies significantly among the four groups at the 5% level.

**12.76** Suppose we specifically wish to compare women who used cod liver oil for the entire year during childhood vs. women who never used it.

- (i) Perform this comparison using the LSD method, and report a *p*-value.
- (ii) Perform this comparison using the Bonferroni method. Are the results significant at the 5% level after Bonferroni correction? Why or why not?

(Hint: Assume that a  $t_d$  distribution is the same as an  $N(0,1)$  distribution of  $d \geq 200$ .)

**12.77** A reasonable summary of the data might be provided by relating the score  $S_i$  given by  $S_i = 1$  if never,  $S_i = 2$  if irregular,  $S_i = 3$  if fall and winter, and  $S_i = 4$  if whole year. What method can be used to relate mean BMD to this score variable?

### Ornithology

Data are available from four source populations of stonechat birds:

- (1) An African group that are resident year-round in equatorial Africa
- (2) A European group from Austria that migrate a comparatively short distance to Northern Africa
- (3) An Irish group that winters along the coast of Britain and Ireland
- (4) A Siberian group from Kazakhstan that migrates a long distance to India, China, and Northern Africa

The data for this problem were supplied by Maude Baldwin, a graduate student in the Biology Department at Harvard University. The data in Table 12.47 on wing length were presented from male birds of four different populations.

**Table 12.47 Wing length (cm) from males of four different populations of stonechat birds**

	Mean	sd	N
African	72.73	1.42	22
European	66.26	1.33	43
Irish	68.45	1.15	30
Siberian	69.11	1.22	12
Overall	68.52		107

Suppose we assume within each subspecies that wing length is normally distributed and that the underlying standard deviation is the same.

**12.78** What test can be used to compare the mean wing length in the four groups?

**12.79** Perform the test, and report a *p*-value (two-tailed).

**12.80** Identify which pairs of subspecies are significantly different in mean wing length using a two-sided Bonferroni-adjusted  $\alpha$  level of 0.05. (Hint:  $t_{103,9958} \approx 2.68$ .)

The full data set consists of both male and female birds with results as shown in Table 12.48.

**Table 12.48 Wing length (cm) from four different species of stonechats, by gender**

	Male			Female		
	Mean	sd	N	Mean	sd	N
African	72.73	1.42	22	70.64	1.27	23
European	66.26	1.33	43	64.25	1.93	45
Irish	68.45	1.15	30	66.52	1.57	31
Siberian	69.11	1.22	12	66.49	0.62	7

**12.81** Write down a model to assess the effects of species and gender in the same model, allowing for the possibility that differences between mean wing lengths of different species vary by gender. What is the name of the type of model you specify? (Do not implement this model.) Explain what each of the terms in the model mean.

### Ophthalmology, Infectious Disease

Fluoroquinolones are antibiotics which are used for treating certain types of bacterial infections. These drugs are FDA approved to be taken systemically. However, some phase IV (post-approval) studies have shown that use of these drugs can be a risk factor for the development of peripheral neuropathy (i.e., neurologic symptoms), and the safety labeling has been changed accordingly.

A small clinical trial was set up to study the safety and effectiveness of two ophthalmic solutions (i.e., eye drops) in this drug class that are intended to be used to treat bacterial ocular infections. Two active drugs in this class (drug M and drug G) together with a placebo (drug P) were studied according to the following design. Ninety three normal subjects were randomized to one of three groups in approximately equal numbers, as shown in Table 12.49.

**Table 12.49 Treatment assignments in ocular infection clinical trial**

Group	Eye 1	Eye 2
A	G	P
B	M	P
C	G	M

Thus, for each subject in group A, drug G was administered in a randomly selected eye and drug P in the fellow eye. Groups B and C are defined similarly. Each person was told to administer the two assigned drugs four times per day for 10 days. The principal outcome (or response) measure in the study was corneal sensitivity, measured in millimeters, which has values in the range of 40–60 mm. High values of corneal sensitivity indicate greater sensitivity (i.e., normal), whereas low values indicate poor sensitivity (i.e., abnormal).

Corneal sensitivity was measured at baseline, at 7 days, and at 14 days. Note that each person was still taking study drug at day 7 but not at day 14.

The sensitivity of the central cornea as well as each corneal quadrant (superior, inferior, temporal, and nasal) was measured for each eye at day 0, 7, and 14. We will refer to these as the five regions of the cornea. The data are available in CORNEAL.DAT with documentation in CORNEAL.DOC on the Companion Website.

## REFERENCES

- [1] White, J. R., & Froeb, H. R. (1980). Small-airway dysfunction in nonsmokers chronically exposed to tobacco smoke. *New England Journal of Medicine*, 302(13), 720–723.
- [2] Michels, K. B., & Rosner, B. (1996). Data trawling: To fish or not to fish. *Lancet*, 348, 1152–1153.
- [3] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1): 289–300.
- [4] Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariable methods* (2nd ed.). Boston: Duxbury.
- [5] Abelson, M. B., Kliman, G. H., Butrus, S. I., & Weston, J. H. (1983). Modulation of arachidonic acid in the rabbit conjunctiva: Predominance of the cyclo-oxygenase pathway. Presented at the Annual Spring Meeting of the Association for Research in Vision and Ophthalmology, Sarasota, Florida, May 2–6, 1983.
- [6] Hankinson, S. E., Manson, J. E., Spiegelman, D., Willett, W. C., Longcope, C., & Speizer, R. E. (1995). Reproducibility of plasma hormone levels in postmenopausal women over a 2–3 year period. *Cancer Epidemiology, Biomarkers and Prevention*, 4(6), 649–654.
- [7] Chinn, S. (1990). The assessment of methods of measurement. *Statistics in Medicine*, 9, 351–362.
- [8] Snedecor, G., & Cochran, W. G. (1988). *Statistical methods*. Ames: Iowa State University Press.
- [9] Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.
- [10] Cook, N. R., & Rosner, B. (1993). Screening rules for determining blood pressure status in clinical trials: Application to the Trials of Hypertension Prevention. *American Journal of Epidemiology*, 137(12), 1341–1352.
- [11] Satterfield, S., Cutler, J. A., Langford, H. G., et al. (1991). Trials of Hypertension Prevention: Phase I design. *Annals of Epidemiology*, 1, 455–457.
- [12] Linn, W. S., Shamoo, D. A., Anderson, K. R., Whynot, J. D., Avol, E. L., & Hackney, J. D. (1985). Effects of heat and humidity on the responses of exercising asthmatics to sulfur dioxide exposure. *American Review of Respiratory Disease*, 131(2), 221–225.
- [13] Polk, B. F., Rosner, B., Feudo, R., & Van Denburgh, M. (1980). An evaluation of the Vita Stat automatic blood pressure measuring device. *Hypertension*, 2(2), 221–227.
- [14] Pfeffer, R. I., Kurosaki, T. T., Chance, J. M., Filos, S., & Bates, D. (1984). Use of the mental function index in older adults: Reliability, validity and measurement of change over time. *American Journal of Epidemiology*, 120(6), 922–935.
- [15] Riis, B., Thomsen, K., & Christiansen, C. (1987). Does calcium supplementation prevent post-menopausal bone loss? A double-blind controlled study. *New England Journal of Medicine*, 316(4), 173–177.
- [16] Kopelman, H., Durie, P., Gaskin, K., Weizman, Z., & Forstner, G. (1985). Pancreatic fluid secretion and protein hyperconcentration in cystic fibrosis. *New England Journal of Medicine*, 312(6), 329–334.
- [17] Brown, B. G., Zhao, X.-Q., Chait, A., Fisher, L. D., Cheung, M. C., Morse, J. S., Dowdy, A. A., Marino, E. K., Bolson, E. L., Alaupovic, P., Frohlich, J., Serafini, L., Huss-Frechette, E., Wang, S., DeAngelis, D., Dodek, A., & Albers, J. J. (2001). Simvastatin and niacin, antioxidant vitamins, or the combination for the prevention of coronary disease. *New England Journal of Medicine*, 345, 1583–1592.
- [18] Forsmo, S., Fjeldbo, S. K., & Langhammer, A. (2008). Childhood cod liver oil consumption and bone mineral density in a population-based cohort of peri- and postmenopausal women. *American Journal of Epidemiology*, 167(4), 406–411.

# 13

## Design and Analysis Techniques for Epidemiologic Studies

### 13.1 Introduction

In Chapter 10 we discussed methods of analysis for categorical data. The data were displayed in a single  $2 \times 2$  contingency table or more generally in a single  $R \times C$  contingency table. In epidemiologic applications, the rows of the table refer to disease categories and the columns to exposure categories (or vice versa). It is natural in such applications to define a measure of effect based on the counts in the contingency table (such as the relative risk) and to obtain confidence limits for such measures. An important issue is whether a disease–exposure relationship is influenced by other variables (called *confounders*). In this chapter, we discuss

- (1) Some common study designs used in epidemiologic work
- (2) Several measures of effect that are commonly used for categorical data
- (3) Techniques for assessing a primary disease–exposure relationship while controlling for confounding variable(s), including
  - (a) Mantel-Haenszel methodology
  - (b) Logistic regression
  - (c) Extensions to logistic regression
- (4) Meta-analysis, a popular methodology for combining results over more than one study
- (5) Several alternative study designs, including
  - (a) Active-control designs
  - (b) Cross-over designs
- (6) Some newer data-analysis techniques in epidemiology when the assumptions of standard methods are not satisfied, including
  - (a) Methods for clustered binary data
  - (b) Methods for handling data with substantial measurement error
- (7) Methods for handling missing data
- (8) Introduction to longitudinal data analysis

### 13.2 Study Design

Let's look at Table 10.2. In this table, we examined the association between use of oral contraceptives (OC) at baseline and development of myocardial infarction (MI)

over a 3-year follow-up period. In this setting, OC use is sometimes called an *exposure variable*, and the occurrence of MI, a *disease variable*. Researchers often are interested in exposure–disease relationships, as shown in Table 13.1.

**Table 13.1 Hypothetical exposure–disease relationship**

		Disease		
		Yes	No	
Exposure	Yes	a	b	$a + b = n_1$
	No	c	d	$c + d = n_2$
		$a + c = m_1$	$b + d = m_2$	

There are a total of  $n_1 = a + b$  exposed subjects, of whom  $a$  have disease, and a total of  $n_2 = c + d$  unexposed subjects, of whom  $c$  have disease. Three main study designs are used to explore such relationships: a prospective study design, a retrospective study design, and a cross-sectional study design.

#### Definition 13.1

A **prospective study** is a study in which a group of disease-free individuals is identified at one point in time and are followed over a period of time until some of them develop the disease. The development of disease over time is then related to other variables measured at baseline, generally called *exposure variables*. The study population in a prospective study is often called a **cohort**. Thus another name for this type of study is a **cohort study**.

#### Definition 13.2

A **retrospective study** is a study in which two groups of individuals are initially identified: (1) a group that has the disease under study (the cases) and (2) a group that does not have the disease under study (the controls). An attempt is then made to relate their *prior* health habits to their current disease status. This type of study is also sometimes called a **case-control study**.

#### Definition 13.3

A **cross-sectional study** is one in which a study population is ascertained at a single point in time. All participants in the study population are asked about their current disease status and their current or past exposure status. This type of study is sometimes called a **prevalence study** because the prevalence of disease at one point in time is compared between exposed and unexposed individuals. This contrasts with a prospective study, in which one is interested in the incidence rather than the prevalence of disease.

#### Example 13.1

**Cardiovascular Disease** What type of study design was used in Table 10.2?

#### Solution

The study presented in Table 10.2 is an example of a prospective design. All participants were disease free at baseline and had their exposure (OC use) measured at that time. They were followed for 3 years, during which some developed disease while others remained disease free.

**Example 13.2**

**Cancer** What type of study design was used in the international breast-cancer study in Example 10.4?

**Solution**

This is an example of a retrospective study. Breast-cancer cases were identified together with controls who were of comparable age and in the hospital at the same time as the cases but who did not have breast cancer. Pregnancy history (age at first birth) of cases and controls was compared.

What are the advantages of the two types of studies? A prospective study is usually more definitive because the patients' knowledge of their current health habits is more precise than their (or related individuals') recall of their past health habits. Second, a retrospective study has a greater chance of bias for two reasons. First, it is much more difficult to obtain a representative sample of people who already have the disease in question. For example, some of the diseased individuals may have already died and only the mildest cases (or if it is a study of deceased people, the most severe cases) may be included. This type of bias is called **selection bias**. Second, the diseased individuals, if still alive, or their surrogates may tend to give biased answers about prior health habits if they *believe* there is a relationship between these prior health habits and the disease; this type of bias is called **recall bias**. However, a retrospective study is much less expensive to perform and can be completed in much less time than a prospective study. For example, if the study in Example 10.4 were done as a prospective study, it would require a very large study population followed for many years before 3000 cases of breast cancer would occur. Thus an inexpensive retrospective study may be done initially as a justification for doing the ultimate, definitive prospective study.

**Example 13.3**

**Hypertension** Suppose a study is performed concerning infant blood pressure. All infants born in a specific hospital are ascertained within the first week of life while in the hospital and have their blood pressure measured in the newborn nursery. The infants are divided into two groups: a high-blood-pressure group, if their blood pressure is in the top 10% of infant blood pressure based on national norms; and a normal-blood-pressure group otherwise. The infants' blood-pressure group is then related to their birthweight (low if  $\leq 88$  oz and normal otherwise). This is an example of a cross-sectional study because the blood pressures and birthweights are measured at approximately the same point in time.

Not all studies fit neatly into the characterizations given in Definitions 13.1–13.3. Indeed, some case-control studies are based on exposure variables that are collected prospectively.

**Example 13.4**

**Cardiovascular Disease** The Physicians' Health Study was a large, randomized clinical trial. Participants were approximately 22,000 male physicians ages 40–84 who were initially (in 1982) free of coronary heart disease (CHD) and cancer (except for nonmelanoma skin cancer). The principal aims of the study were to investigate the effect of aspirin use on CHD and the effect of beta-carotene use on cancer incidence. Accordingly, participants were randomized to one of four treatment groups (group 1 received aspirin placebo and beta-carotene placebo capsules, group 2 received active aspirin and beta-carotene placebo capsules, group 3 received aspirin placebo and active beta-carotene capsules, and group 4 received active aspirin and active beta-carotene capsules). The aspirin arm of the study was stopped in 1990 when it became clear aspirin had an important protective effect in preventing the development of

CHD. The beta-carotene arm of the study was discontinued in 1997 when it became clear that beta-carotene had no effect on cancer incidence. As a secondary aim of the study, blood samples were collected at baseline from all physicians in the cohort. The goal of this part of the study was to relate lipid abnormalities identified in the blood samples to the occurrence of CHD. However, it would have been prohibitively expensive to analyze all the blood samples that were collected. Instead, all men who developed CHD (the case group) ( $n \approx 300$ ) and a random sample of physicians who did not develop CHD, but who had approximately the same age distribution as the case group (the control group) ( $n \approx 600$ ), were identified and their blood samples analyzed. The type of study is a case-control study nested within a prospective study that does not fit neatly into the characterizations in Definitions 13.1 and 13.2. Specifically, the issue of biased ascertainment of exposure in retrospective and case-control studies is not an issue here because blood samples were obtained at baseline. However, the methods of analysis described hereafter for case-control studies are also applicable to this type of study.

In this section, we have discussed the principal designs used in observational epidemiologic studies. In a prospective study, a cohort of disease-free participants is ascertained at baseline and followed over time until some members of the cohort develop disease. It is generally considered the gold standard of designs for observational studies. However, it is relatively expensive because for a meaningful number of events to occur over time, a large number of participants must often be followed for a long time. In a case-control design, a group of participants with disease (the cases) and a group of participants without disease (the controls) are recruited. Usually a retrospective history of health habits prior to getting disease is obtained. This design is relatively inexpensive because we don't have to wait until participants develop disease, which for rare diseases can often take a long time. However, the results from using this study design are sometimes problematic to interpret because of

- (1) Recall bias of previous exposures by people who already have disease
- (2) Potential selection bias of
  - (a) The case group if, for example, a milder series of case participants who are still alive is used
  - (b) The control group if, for example, control selection is related, often unexpectedly, to the exposure

Thus case-control studies are often used as preliminary steps to justify the ultimate, definitive prospective study. Cross-sectional studies are conducted at one point in time and have many of the same problems as case-control studies, except that the relative number of cases and controls is not fixed in advance.

### 13.3 Measures of Effect for Categorical Data

We would like to compare the frequency of disease between exposed and unexposed subjects. Doing so is most straightforward in the context of prospective studies in which we compare incidence rates, or in cross-sectional studies in which we compare prevalence rates between exposed and unexposed individuals. We will discuss these issues for prospective studies in terms of comparing incidence rates, but the same measures of effect can be used for cross-sectional studies in terms of prevalence.

**Definition 13.4**

Let

$p_1$  = probability of developing disease for exposed individuals

$p_2$  = probability of developing disease for unexposed individuals

The **risk difference** (RD) is defined as  $p_1 - p_2$ . The **risk ratio** (RR) or **relative risk** is defined as  $p_1/p_2$ .

## The Risk Difference

Suppose that  $\hat{p}_1$  and  $\hat{p}_2$  are the sample proportions with disease for exposed and unexposed subjects based on sample sizes of  $n_1$  and  $n_2$ , respectively. An unbiased point estimate of  $p_1 - p_2$  is given by  $\hat{p}_1 - \hat{p}_2$ . To obtain an interval estimate, we assume the normal approximation to the binomial distribution holds, whereby from Chapter 6,  $\hat{p}_1 \sim N(p_1, p_1 q_1 / n_1)$ ,  $\hat{p}_2 \sim N(p_2, p_2 q_2 / n_2)$ . Because these are two independent samples, from Equation 5.10,

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

Therefore, if  $p_1 q_1 / n_1 + p_2 q_2 / n_2$  is approximated by  $\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2$  then

$$Pr\left(p_1 - p_2 - z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq \hat{p}_1 - \hat{p}_2 \leq p_1 - p_2 + z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}\right) = 1 - \alpha$$

This can be rewritten as two inequalities:

$$p_1 - p_2 - z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq \hat{p}_1 - \hat{p}_2$$

$$\text{and } \hat{p}_1 - \hat{p}_2 \leq p_1 - p_2 + z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

If  $z_{1-\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}$  is added to both sides of the first inequality and subtracted from both sides of the second inequality, then we obtain

$$p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$\text{and } \hat{p}_1 - \hat{p}_2 - z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \leq p_1 - p_2$$

This leads to the following method for point and interval estimation of the risk difference:

**Equation 13.1**

### Point and Interval Estimation for the Risk Difference

Let  $\hat{p}_1, \hat{p}_2$  represent the sample proportion who develop disease in a prospective study, based on sample sizes of  $n_1$  exposed subjects and  $n_2$  unexposed subjects, respectively. A point estimate of the risk difference is given by  $\hat{p}_1 - \hat{p}_2$ . A  $100\% \times (1 - \alpha)$  CI for the risk difference is given by

$$\hat{p}_1 - \hat{p}_2 - [1 / (2n_1) + 1 / (2n_2)] \pm z_{1-\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2} \text{ if } \hat{p}_1 > \hat{p}_2$$

$$\hat{p}_1 - \hat{p}_2 + [1 / (2n_1) + 1 / (2n_2)] \pm z_{1-\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2} \text{ if } \hat{p}_1 \leq \hat{p}_2$$

Use these expressions for the CI only if  $n_1 \hat{p}_1 \hat{q}_1 \geq 5$  and  $n_2 \hat{p}_2 \hat{q}_2 \geq 5$ .

The term  $1/(2n_1) + 1/(2n_2)$  serves as a continuity correction in Equation 13.1.

### Example 13.5

**Cardiovascular Disease** Referring to the OC-MI data in Table 10.2, provide a point estimate and a 95% CI for the difference between the proportion of women who develop MI among OC users and the comparable proportion among non-OC users.

#### Solution

We have that  $n_1 = 5000$ ,  $\hat{p}_1 = 13/5000 = .0026$ ,  $n_2 = 10,000$ ,  $\hat{p}_2 = 7/10,000 = .0007$ . Thus a point estimate of the risk difference ( $p_1 - p_2$ ) is given by  $\hat{p}_1 - \hat{p}_2 = .0019$ . Because  $n_1 \hat{p}_1 \hat{q}_1 = 5000(.0026)(.9974) = 13.0 \geq 5$ ,  $n_2 \hat{p}_2 \hat{q}_2 = 10,000(.0007)(.9993) = 7.0 \geq 5$ , the large-sample CI in Equation 13.1 can be used. The 95% CI is given by

$$\begin{aligned} .0026 - .0007 - \left[ \frac{1}{2(5000)} + \frac{1}{2(10,000)} \right] &\pm 1.96 \sqrt{\frac{.0026(.9974)}{5000} + \frac{.0007(.9993)}{10,000}} \\ &= .00175 \pm 1.96(.00077) \\ &= .00175 \pm .00150 = (.0002, .0033) \end{aligned}$$

Thus the true risk difference is significantly greater than zero.

## The Risk Ratio

A point estimate of the risk ratio ( $RR = p_1/p_2$ ) is given by

### Equation 13.2

$$\hat{RR} = \hat{p}_1 / \hat{p}_2$$

To obtain an interval estimate, we assume the normal approximation to the binomial distribution is valid. Under this assumption, it can be shown that the sampling distribution of  $\ln(\hat{RR})$  more closely follows a normal distribution than  $\hat{RR}$  itself.

We note that

$$\begin{aligned} Var[\ln(\hat{RR})] &= Var[\ln(\hat{p}_1) - \ln(\hat{p}_2)] \\ &= Var[\ln(\hat{p}_1)] + Var[\ln(\hat{p}_2)] \end{aligned}$$

To obtain  $Var[\ln(\hat{p}_1)]$ , we employ a principle known as the *delta method*.

### Equation 13.3

#### Delta Method

The variance of a function of a random variable  $f(X)$  is approximated by

$$Var[f(X)] \approx [f'(X)]^2 Var(X)$$

**Example 13.6****Solution**

Use the delta method to find the variance of  $\ln(\hat{p}_1)$ ,  $\ln(\hat{p}_2)$ , and  $\ln(\hat{RR})$ .

In this case  $f(X) = \ln(X)$ . Because  $f'(X) = \frac{1}{X}$ , we obtain

$$\text{Var}[\ln(\hat{p}_1)] \approx \frac{1}{\hat{p}_1^2} \text{Var}(\hat{p}_1) = \frac{1}{\hat{p}_1^2} \left( \frac{\hat{p}_1 \hat{q}_1}{n_1} \right) = \frac{\hat{q}_1}{\hat{p}_1 n_1}$$

However, from Table 13.1,  $\hat{p}_1 = a / n_1$ ,  $\hat{q}_1 = b / n_1$ . Therefore,

$$\text{Var}[\ln(\hat{p}_1)] = \frac{b}{an_1}$$

Also, using similar methods,

$$\text{Var}[\ln(\hat{p}_2)] = \frac{\hat{q}_2}{\hat{p}_2 n_2} = \frac{d}{cn_2}$$

It follows that

$$\begin{aligned} \text{Var}[\ln(\hat{RR})] &= \frac{b}{an_1} + \frac{d}{cn_2} \\ \text{or } \text{se}[\ln(\hat{RR})] &= \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}} \end{aligned}$$

Therefore, an approximate two-sided  $100\% \times (1 - \alpha)$  CI for  $\ln(RR)$  is given by

**Equation 13.4**

$$\left[ \ln(\hat{RR}) - z_{1-\alpha/2} \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}}, \ln(\hat{RR}) + z_{1-\alpha/2} \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}} \right]$$

The antilog of each end of the interval in Equation 13.4 then provides a two-sided  $100\% \times (1 - \alpha)$  CI for  $RR$  itself, given by

**Equation 13.5**

$$\left[ e^{\ln(\hat{RR}) - z_{1-\alpha/2} \sqrt{b/(an_1) + d/(cn_2)}}, e^{\ln(\hat{RR}) + z_{1-\alpha/2} \sqrt{b/(an_1) + d/(cn_2)}} \right]$$

The estimation procedures for the  $RR$  are summarized as follows.

**Equation 13.6****Point and Interval Estimation for the Risk Ratio (RR)**

Let  $\hat{p}_1, \hat{p}_2$  represent the sample proportions of exposed and unexposed individuals who develop disease in a prospective study, based on samples of size  $n_1$  and  $n_2$ , respectively. A point estimate of the  $RR$  (or relative risk) is given by  $\hat{p}_1 / \hat{p}_2$ . A  $100\% \times (1 - \alpha)$  CI for the  $RR$  is given by  $[\exp(c_1), \exp(c_2)]$ , where

$$c_1 = \ln(\hat{RR}) - z_{1-\alpha/2} \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}}$$

$$c_2 = \ln(\hat{RR}) + z_{1-\alpha/2} \sqrt{\frac{b}{an_1} + \frac{d}{cn_2}}$$

where  $a, b$  = number of exposed subjects who do and do not develop disease, respectively, and  $c, d$  = number of unexposed subjects who do and do not develop disease, respectively. This method of interval estimation is only valid if  $n_1\hat{p}_1\hat{q}_1 \geq 5$  and  $n_2\hat{p}_2\hat{q}_2 \geq 5$ .

**Example 13.7**

**Cardiovascular Disease** Referring to Table 10.2, provide a point estimate and a 95% CI for the relative risk of MI among OC users compared with non-OC users.

**Solution**

We have from Example 13.5 that  $\hat{p}_1 = 13/5000 = .0026$ ,  $n_1 = 5000$ ,  $\hat{p}_2 = 7/10,000 = .0007$ ,  $n_2 = 10,000$ . Thus our point estimate of RR is  $\hat{RR} = \hat{p}_1/\hat{p}_2 = .0026/.0007 = 3.71$ . To compute a 95% CI, we obtain  $c_1, c_2$  in Equation 13.6. We have  $a = 13$ ,  $b = 4987$ ,  $c = 7$ , and  $d = 9993$ . Thus

$$\begin{aligned} c_1 &= \ln\left(\frac{.0026}{.0007}\right) - 1.96 \sqrt{\frac{4987}{13(5000)} + \frac{9993}{7(10,000)}} \\ &= 1.312 - 1.96(0.4685) \\ &= 1.312 - 0.918 = 0.394 \\ c_2 &= 1.312 + 0.918 = 2.230 \end{aligned}$$

Therefore, our 95% CI for  $RR = (e^{0.394}, e^{2.230}) = (1.5, 9.3)$ , which implies that the true RR is significantly greater than 1.

## The Odds Ratio

In the previous section, the  $RR$  (or relative risk) was introduced. The relative risk can be expressed as the ratio of the probability of disease among exposed subjects ( $p_1$ ) divided by the probability of disease among unexposed subjects ( $p_2$ ). Although easily understood, the  $RR$  has the disadvantage of being constrained by the denominator probability ( $p_2$ ). For example, if  $p_2 = .5$ , then the  $RR$  can be no larger than  $1/.5 = 2$ ; if  $p_2 = .8$ , then the  $RR$  can be no larger than  $1/.8 = 1.25$ . To avoid this restriction, another comparative measure relating two proportions is sometimes used, called the odds ratio (OR). The odds in favor of a success are defined as follows.

**Definition 13.5**


---

If the probability of a success =  $p$ , then the **odds in favor of success** =  $p/(1-p)$ .

---

If two proportions  $p_1, p_2$  are considered and the odds in favor of success are computed for each proportion, then the ratio of odds, or  $OR$ , becomes a useful measure for relating the two proportions.

**Definition 13.6**


---

Let  $p_1, p_2$  be the underlying probability of success for two groups. The  $OR$  is defined as

$$OR = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1q_2}{p_2q_1} \text{ and estimated by } \hat{OR} = \frac{\hat{p}_1\hat{q}_2}{\hat{p}_2\hat{q}_1}$$

Equivalently, if the four cells of the  $2 \times 2$  contingency table are labeled by  $a, b, c, d$ , as they are in Table 13.1, then

$$\hat{OR} = \frac{[a / (a + b)] \times [d / (c + d)]}{[c / (c + d)] \times [b / (a + b)]} = \frac{ad}{bc}$$

In the context of a prospective study, the  $OR$  can be interpreted as the odds in favor of disease for exposed subjects divided by the odds in favor of disease for unexposed subjects. This is sometimes referred to as the *disease-odds ratio*.

**Definition 13.7**

The **disease-odds ratio** is the odds in favor of disease for the exposed group divided by the odds in favor of disease for the unexposed group.

**Example 13.8**

**Cardiovascular Disease** Using the OC-MI data in Table 10.2, estimate the  $OR$  in favor of MI for an OC user compared with a non-OC user (i.e., the disease-odds ratio).

**Solution**

We have  $\hat{p}_1 = .0026$ ,  $\hat{q}_1 = .9974$ ,  $\hat{p}_2 = .0007$ ,  $\hat{q}_2 = .9993$ . Thus

$$OR = \frac{.0026(.9993)}{.0007(.9974)} = 3.72$$

Thus the odds in favor of an MI for an OC user is 3.7 times the odds in favor of an MI for a non-OC user.  $\hat{OR}$  could also have been computed from the contingency table in Table 10.2, whereby

$$\hat{OR} = \frac{13 \times 9993}{7 \times 4987} = 3.72$$

If the probability of disease is the same for exposed and unexposed subjects, then  $OR = 1$ . Conversely,  $OR$ 's greater than 1 indicate a greater likelihood of disease among the exposed than among the unexposed, whereas  $OR$ 's less than 1 indicate a greater likelihood of disease among the unexposed than among the exposed. Notice that there is no restriction on the  $OR$  as there was for the  $RR$ . Specifically, as the probability of disease among the exposed ( $p_1$ ) approaches 0,  $OR$  approaches 0, whereas as  $p_1$  approaches 1,  $OR$  approaches  $\infty$ , regardless of the value of the probability of disease among the unexposed ( $p_2$ ). This property is particularly advantageous when combining results over several  $2 \times 2$  tables, as discussed in Section 13.6. Finally, if the probabilities of success are low (i.e.,  $p_1, p_2$  are small), then  $1 - p_1$  and  $1 - p_2$  will each be close to 1, and the  $OR$  will be approximately the same as the relative risk. Thus the  $OR$  is often used as an approximation to the  $RR$  for rare diseases.

In Example 13.8, we computed the  $OR$  as a disease-odds ratio. However, another way to express the  $OR$  is as an exposure-odds ratio.

**Definition 13.8**

The **exposure-odds ratio** is the odds in favor of being exposed for diseased subjects divided by the odds in favor of being exposed for nondiseased subjects.

From Table 13.1, this is given by

**Equation 13.7**

$$\begin{aligned}\text{Exposure-odds ratio} &= \frac{[a / (a + c)] / [c / (a + c)]}{[b / (b + d)] / [d / (b + d)]} \\ &= \frac{ad}{bc} = \text{disease-odds ratio}\end{aligned}$$

Therefore the exposure-odds ratio is the same as the disease-odds ratio. This relationship is particularly useful for case-control studies. For prospective studies, we have seen that we can estimate the risk difference, the *RR*, or the *OR*. For case-control studies, we cannot directly estimate either the risk difference or the *RR*. To see why, let *A*, *B*, *C*, and *D* represent the true number of subjects in the reference population, corresponding to cells *a*, *b*, *c*, and *d* in our sample as shown in Table 13.2.

**Table 13.2 Hypothetical exposure-disease relationships in a sample and a reference population**

		Sample		Population	
		Disease		Disease	
		Yes	No	Yes	No
Exposed	Yes	<i>a</i>	<i>b</i>	Exposed	<i>A</i>
	No	<i>c</i>	<i>d</i>		<i>C</i>

In a case-control study, we assume that a random fraction  $f_1$  of subjects with disease and a random fraction  $f_2$  of subjects without disease in the reference population are included in our study sample. We also assume there is no sampling bias, so  $f_1$  is the sampling fraction for both exposed and unexposed subjects with disease and  $f_2$  is the sampling fraction for both exposed and unexposed subjects without disease. Therefore  $a = f_1A$ ,  $c = f_1C$ ,  $b = f_2B$ ,  $d = f_2D$ . If we estimate the *RR* from our study sample, we obtain

**Equation 13.8**

$$\begin{aligned}\hat{RR} &= \frac{a / (a + b)}{c / (c + d)} \\ &= \frac{f_1A / (f_1A + f_2B)}{f_1C / (f_1C + f_2D)} \\ &= \frac{A / (f_1A + f_2B)}{C / (f_1C + f_2D)}\end{aligned}$$

However, from Table 13.2, the true *RR* in the reference population is

**Equation 13.9**

$$RR = \frac{A / (A + B)}{C / (C + D)}$$

The expressions on the right-hand side of Equations 13.8 and 13.9 are only the same if  $f_1 = f_2$ —that is, if the sampling fraction of subjects with disease and without disease

are the same. However, this is very unlikely in a case-control study because the usual sampling strategy is to oversample subjects with disease.

**Example 13.9**

**Cancer** Consider a case-control study of the relationship between dietary factors and colon cancer. Suppose 100 colon-cancer cases are selected from a tumor registry, and 100 controls are chosen who live in the same census tract as the cases and have approximately the same age and sex distribution. Thus an equal number of cases and controls are in the sample, even though the fraction of people with colon cancer in the census tract may be very low. Therefore  $f_1$  will be much larger than  $f_2$ . Thus  $\hat{RR}$  will provide a biased estimate of  $RR$  in most case-control studies. This is also true for the risk difference. However, we can estimate the  $OR$  from our sample in Table 13.2 given by

$$\begin{aligned}\hat{OR} &= \frac{ad}{bc} \\ &= \frac{f_1 A(f_2 D)}{f_2 B(f_1 C)} \\ &= \frac{AD}{BC} = OR\end{aligned}$$

Thus the  $OR$  estimated from our sample provides an unbiased estimate of the  $OR$  from our reference population. We saw in Equation 13.7 that the exposure- and disease-odds ratios are the same for any  $2 \times 2$  table relating exposure to disease, regardless of sampling strategy. Therefore, the  $OR$  from a case-control study provides an unbiased estimate of the true disease-odds ratio. However, if the disease under study is rare, then the disease-odds ratio is approximately the same as the  $RR$ . This lets us indirectly estimate the  $RR$  for case-control studies.

The general method of estimation of the  $RR$  in case-control studies is summarized as follows.

**Equation 13.10****Estimation of the Risk Ratio for Case-Control Studies**

Suppose we have a  $2 \times 2$  table relating exposure to disease, as in Table 13.1. If the data are collected using a case-control study design, and the disease under study is rare (i.e., disease incidence  $< .10$ ), then we can estimate the  $RR$  approximately by

$$\hat{RR} \approx \hat{OR} = \frac{ad}{bc}$$

**Example 13.10**

**Cancer** Estimate the  $RR$  for breast cancer for women with a late age at first birth ( $\geq 30$ ) compared with women with an early age at first birth ( $\leq 29$ ) based on the data in Table 10.1.

**Solution**

The estimated  $OR$  is given by

$$\begin{aligned}\hat{OR} &= \frac{ad}{bc} \\ &= \frac{683(8747)}{2537(1498)} = 1.57\end{aligned}$$

This is an estimate of the *RR* because, although breast cancer is one of the most common cancers in women, its incidence in the general population of women is relatively low, unless very old women are considered.

## Interval Estimation for the Odds Ratio

In the previous section, we discussed how to estimate the *OR*. We saw that, in a case-control study with a rare disease outcome, the *OR* provides an approximate estimate of the *RR*. The issue remains as to how to obtain interval estimates for the *OR*. Several methods exist for this purpose. One of the most popular approaches is the Woolf method [1]. Woolf showed that, approximately,

$$\text{Var}[\ln(\hat{O}R)] \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

where  $a$ ,  $b$ ,  $c$ , and  $d$  are the four cells of our  $2 \times 2$  contingency table. To see this, suppose we have a prospective design. From Definition 13.6 we can represent the estimated *OR* as a disease-odds ratio of the form  $(\hat{p}_1 / \hat{q}_1) / (\hat{p}_2 / \hat{q}_2)$

where  $\hat{p}_1 = a / (a+b)$ ,  $\hat{p}_2 = c / (c+d)$ ,  $\hat{q}_1 = 1 - \hat{p}_1$ ,  $\hat{q}_2 = 1 - \hat{p}_2$

Furthermore,

$$\begin{aligned}\text{Var}[\ln(\hat{O}R)] &= \text{Var}[\ln((\hat{p}_1 / \hat{q}_1) / (\hat{p}_2 / \hat{q}_2))] \\ &= \text{Var}[\ln(\hat{p}_1 / \hat{q}_1) - \ln(\hat{p}_2 / \hat{q}_2)] \\ &= \text{Var}[\ln(\hat{p}_1 / \hat{q}_1)] + \text{Var}[\ln(\hat{p}_2 / \hat{q}_2)]\end{aligned}$$

To obtain  $\text{Var}[\ln(\hat{p}_1 / \hat{q}_1)]$ , we use the delta method. We have

$$\frac{d[\ln(\hat{p}_1 / \hat{q}_1)]}{d\hat{p}_1} = \frac{1}{\hat{p}_1 \hat{q}_1}$$

Furthermore,  $\text{Var}(\hat{p}_1) = \hat{p}_1 \hat{q}_1 / (a+b)$

Hence

$$\begin{aligned}\text{Var}[\ln(\hat{p}_1 / \hat{q}_1)] &= \left( \frac{1}{\hat{p}_1 \hat{q}_1} \right)^2 \frac{\hat{p}_1 \hat{q}_1}{a+b} \\ &= \frac{1}{(a+b) \hat{p}_1 \hat{q}_1} \\ &= \frac{1}{(a+b) \left( \frac{a}{a+b} \right) \left( \frac{b}{a+b} \right)} \\ &= \frac{a+b}{ab} = \frac{1}{a} + \frac{1}{b}\end{aligned}$$

Similarly,  $\text{Var}[\ln(\hat{p}_2 / \hat{q}_2)] = \frac{1}{c} + \frac{1}{d}$ . Because  $\hat{p}_1$  and  $\hat{p}_2$  are independent random variables, it follows that

$$\text{Var}[\ln(\hat{O}R)] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

A similar result can be obtained if we have a case-control design instead of a prospective design.

If we assume approximate normality of  $\ln(\hat{OR})$ , then a  $100\% \times (1 - \alpha)$  CI for  $\ln(OR)$  is given by

$$\ln(\hat{OR}) \pm z_{1-\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

If we take the antilog of each end of the CI, then it follows that a  $100\% \times (1 - \alpha)$  CI for  $OR$  is given by

$$e^{\ln(\hat{OR}) \pm z_{1-\alpha/2} \sqrt{1/a+1/b+1/c+1/d}} = (\hat{OR} e^{-z_{1-\alpha/2} \sqrt{1/a+1/b+1/c+1/d}}, \hat{OR} e^{z_{1-\alpha/2} \sqrt{1/a+1/b+1/c+1/d}})$$

This approach is summarized as follows.

### Equation 13.11

#### Point and Interval Estimation for the Odds Ratio (Woolf Procedure)

Suppose we have a  $2 \times 2$  contingency table relating exposure to disease, with cell counts  $a, b, c, d$  as given in Table 13.1.

- (1) A point estimate of the true  $OR$  is given by  $\hat{OR} = ad/bc$
- (2) An approximate two-sided  $100\% \times (1 - \alpha)$  CI for  $OR$  is given by  $(e^{c_1}, e^{c_2})$ , where

$$c_1 = \ln(\hat{OR}) - z_{1-\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$c_2 = \ln(\hat{OR}) + z_{1-\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

- (3) In a prospective or a cross-sectional study, the CI in (2) should only be used if  $n_1 \hat{p}_1 \hat{q}_1 \geq 5$  and  $n_2 \hat{p}_2 \hat{q}_2 \geq 5$  where  
 $n_1$  = the number of exposed individuals  
 $\hat{p}_1$  = the sample proportion with disease among exposed individuals, and  
 $\hat{q}_1 = 1 - \hat{p}_1$   
 $n_2$  = the number of unexposed individuals  
 $\hat{p}_2$  = the sample proportion with disease among unexposed individuals, and  
 $\hat{q}_2 = 1 - \hat{p}_2$
- (4) In a case-control study, the CI should only be used if  $m_1 \hat{p}_1^* \hat{q}_1^* \geq 5$  and  $m_2 \hat{p}_2^* \hat{q}_2^* \geq 5$  where  
 $m_1$  = the number of cases  
 $\hat{p}_1^*$  = the proportion of cases that are exposed, and  $\hat{q}_1^* = 1 - \hat{p}_1^*$   
 $m_2$  = the number of controls  
 $\hat{p}_2^*$  = the proportion of controls that are exposed, and  $\hat{q}_2^* = 1 - \hat{p}_2^*$
- (5) If the disease under study is rare, then  $\hat{OR}$  and its associated  $100\% \times (1 - \alpha)$  CI can be interpreted as approximate point and interval estimates of the  $RR$ . This is particularly important in case-control studies in which no direct estimate of the  $RR$  is available.

### Example 13.11

**Cancer** Compute a point estimate and a 95% CI for the  $OR$  relating age at first birth to breast-cancer incidence based on the data in Table 10.1.

### Solution

From Example 13.10 we see that the point estimate of the  $OR = 1.57$ . To obtain an interval estimate, we first compute a 95% CI for  $\ln(OR)$  as follows:

$$\begin{aligned}
 &= \ln(\hat{OR}) \pm z_{.975} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \\
 &= \ln(1.572) \pm 1.96 \sqrt{\frac{1}{683} + \frac{1}{2537} + \frac{1}{1498} + \frac{1}{8747}} \\
 &= 0.452 \pm 1.96(0.0514) \\
 &= 0.452 \pm 0.101 = (0.352, 0.553)
 \end{aligned}$$

A 95% CI for *OR* is thus given by

$$(e^{0.352}, e^{0.553}) = (1.42, 1.74)$$

Because the 95% CI excludes 1, we can conclude that the true *OR* is significantly greater than 1. Also, this is a relatively rare disease, so we can also interpret this interval as an approximate 95% CI for the *RR*.

In this section, we have examined the *RD*, *RR*, and *OR*, the main effect measures used in epidemiologic studies. The *RD* and *RR* can be estimated directly from prospective studies but not from case-control studies. The *OR* is estimable from both prospective and case-control studies. In case-control studies with a rare disease outcome, the *OR* provides an indirect estimate of the *RR*. We also discussed large-sample methods for obtaining confidence limits for the preceding effect measures. To obtain confidence limits for the *RR* and *OR*, we introduced a general technique called the *delta method* to obtain the variance of a function of a random variable, such as  $\ln(X)$ , if the variance of the random variable (*X*) is already known.

### REVIEW QUESTIONS 13A

- 1 Refer to the OC-MI data in Table 10.2
  - (a) Compute the *RR* and a 95% CI about this estimate.
  - (b) Compute the *OR* and a 95% CI about this estimate.
- 2 Suppose 200 obese children and 500 normal-weight children are identified in a school-based screening for hypertension. Eighteen of the obese children and 10 of the normal-weight children are hypertensive.
  - (a) What is the estimated risk of hypertension in each group?
  - (b) Provide an estimate of the risk difference and a 95% CI about this estimate.
- 3 Suppose 100 lung-cancer cases and 200 age- and sex-matched controls are identified and a smoking history is obtained. Fifty of the lung-cancer cases and 20 of the controls are current smokers.
  - (a) Is it possible to estimate the difference in lung-cancer incidence between current smokers vs. noncurrent smokers from these data? If so, estimate it and provide a 95% CI.
  - (b) Is it possible to approximately estimate the *RR* between current smokers and noncurrent smokers for lung cancer from these data? If so, then estimate it and provide a 95% CI.

## 13.4 Attributable Risk

In some cases, a risk factor may have a large *RR*. However, if the risk factor is relatively rare, only a small proportion of cases may be attributable to this risk factor.

Conversely, if a risk factor is common, then even a moderate  $RR$  may translate to a large number of cases attributable to this risk factor. The concept of attributable risk ( $AR$ ) is useful in these circumstances.

**Definition 13.9** Suppose a risk factor is dichotomous with probability of occurrence =  $p$  and the relative risk of disease for persons with the risk factor compared with persons without the risk factor =  $RR$ . The  $AR$  percent for this risk factor is:

$$AR = 100\% \frac{(RR - 1)p}{[(RR - 1)p + 1]}.$$

Suppose the probability of disease =  $d$  for persons without the risk factor and  $RRd$  for persons with the risk factor. The overall probability of disease ( $p_D$ ) is then

$$p_D = RRdp + d(1 - p).$$

If all persons with the risk factor become risk factor free then  $(RR - 1)pd$  cases of disease could be prevented. If we express the number of cases prevented as a proportion of the total number of cases and multiply by 100%, we obtain the definition of  $AR$  given in Definition 13.9.

**Example 13.12** **Cancer** A group of 10,000 current smoking women and 40,000 never smoking women ages 50–69 are followed for 5 years. Fifty of the current smokers and 10 of the never smokers develop lung cancer over 5 years. What is the  $RR$  of current smoking vs. never smoking for lung cancer? Suppose that 20% of the general population of women in this age group are current smokers and 80% are never smokers (we ignore ex-smokers for simplicity). What proportion of lung cancer is attributable to current smoking?

**Solution** From the study results, the  $RR$  for lung cancer =  $(50/10,000)/(10/40,000) = 20$ . In this case,  $p = .2$  and  $RR = 20$ . From Definition 13.9, we have

$$\begin{aligned} AR &= 100\% \times 19(.2)/[19(.2) + 1] \\ &= 100\% \times 3.8/4.8 = 79.2\%. \end{aligned}$$

We now will consider interval estimation for  $AR$ . Based on Definition 13.9,

$$AR/100 = (RR - 1)p / [(RR - 1)p + 1]$$

and

$$1 - AR/100 = 1 / [(RR - 1)p + 1].$$

Hence,

$$\left( \frac{AR/100}{1 - AR/100} \right) = (RR - 1)p.$$

If we take logs of both sides of the equation, we obtain:

$$y = \ln \left( \frac{AR/100}{1 - AR/100} \right) = \ln [AR/(100 - AR)] = \ln p + \ln(RR - 1)$$

We will use the delta method to obtain  $\text{var}(y)$ . We have

$$\frac{dy}{d(RR)} = \frac{1}{(RR - 1)}$$

Hence,

$$\text{var}(y) = \frac{1}{(RR - 1)^2} \text{var}(RR)$$

Also, by the delta method

$$\text{var}(\ln RR) = \text{var}(RR)/RR^2$$

or

$$\text{var}(RR) = RR^2 \text{var}(\ln RR).$$

Thus,

$$\text{var}(y) = (RR/|RR - 1|)^2 \text{var}(\ln RR)$$

Referring to Example 13.6, we have

$$\text{var}(\ln RR) = b/(an_1) + d/(cn_2)$$

where  $a, b, c, d, n_1$ , and  $n_2$  are given in Table 13.1. If we assume normality of  $y$ , then a two-sided  $100\% \times (1 - \alpha)$  CI for AR is given by  $[100\% \times e^{c_1}/(1 + e^{c_1}), 100\% \times e^{c_2}/(1 + e^{c_2})]$ , where  $(c_1, c_2) = \gamma \pm z_{1-\alpha/2} (RR/|RR - 1|)[b/(an_1) + d/(cn_2)]^{1/2}$ .

This is summarized as follows.

### Equation 13.12

#### Interval Estimation for the Attributable Risk

Suppose we have a dichotomous risk factor with known prevalence =  $p$  and estimated relative risk =  $\hat{RR}$ .

(1) A point estimate of the AR is given by

$$\hat{AR} = 100\% \times (\hat{RR} - 1)p / [(\hat{RR} - 1)p + 1]$$

(2) An approximate two-sided  $100\% \times (1 - \alpha)$  CI for AR is given by  $[100\% \times e^{c_1}/(1 + e^{c_1}), 100\% \times e^{c_2}/(1 + e^{c_2})]$ , where  $(c_1, c_2) = \gamma \pm z_{1-\alpha/2} (RR/|RR - 1|)[b/(an_1) + d/(cn_2)]^{1/2}$ ,  $a, b, c, d, n_1$ , and  $n_2$  are defined in Table 13.1 and  $\gamma = \ln[\hat{AR}/(100 - \hat{AR})]$ .

### Example 13.13

**Cancer** Provide a 95% CI for the AR of current smoking for lung cancer using the data in Example 13.12.

### Solution

We refer to Equation 13.12. We have the following  $2 \times 2$  table relating current smoking to lung cancer in this study. The results are shown in Table 13.3.

**Table 13.3**

**Association between current smoking status and lung cancer risk**

		Lung cancer		Total
		Yes	No	
Current	Yes	50	9950	10,000
	Never	10	39,990	40,000
		60	49,940	50,000

Hence,  $a = 50$ ,  $b = 9950$ ,  $c = 10$ ,  $d = 39,990$ ,  $n_1 = 10,000$ ,  $n_2 = 40,000$ . Referring to Equation 13.12, the 95% CI for AR is given by

$$[100\% \times e^{c_1} / (1 + e^{c_1}), 100\% \times e^{c_2} / (1 + e^{c_2})]$$

where

$$(c_1, c_2) = \gamma \pm z_{.975}(RR / |RR - 1|)[b / (an_1) + d / (cn_2)]^{1/2}$$

$$\gamma = \ln[\hat{AR}/(100 - \hat{AR})].$$

In this case, from Example 13.12, we have  $\gamma = \ln[79.2 / (100 - 79.2)] = 1.335$  and  $RR = 20$ .

Hence,

$$\begin{aligned} (c_1, c_2) &= 1.335 \pm 1.96(20 / 19)[9950 / [50(10,000)] + 39,990 / [10(40,000)]]^{1/2} \\ &= 1.335 \pm 2.063(0.0199 + 0.0998)^{1/2} \\ &= 1.335 \pm 0.714 \\ &= (0.621, 2.049) \end{aligned}$$

Thus,

$$e^{c_1} / (1 + e^{c_1}) = e^{0.621} / (1 + e^{0.621}) = 0.650$$

$$e^{c_2} / (1 + e^{c_2}) = e^{2.049} / (1 + e^{2.049}) = 0.886$$

and the 95% CI for  $AR$  is (65%, 88.6%).

In Definition 13.9, we considered  $AR$  for the case of a categorical exposure with two categories. This definition can be extended to the case of a categorical exposure with more than two categories.

#### Example 13.14

**Cancer** Consider the data set in Example 13.12. Suppose in addition to the 10,000 current smoking women and 40,000 never smoking women we have 50,000 ex-smoking women who previously smoked but do not currently smoke, of whom 25 developed lung cancer over a 5-year period. What percent of lung cancer is attributable to smoking (i.e., either current or past smoking)?

#### Solution

In this case, we have 2 smoking groups and one group of never smokers. We wish to determine the percent of lung cancer cases that could be prevented if none of the women had ever started smoking (i.e., were never smokers). In general, suppose we have  $k$  groups where group 1 is an unexposed group and group  $2, \dots, k$  are exposed groups with possibly different amounts of exposure. Let  $p_i$  be the probability of being in the  $i$ th exposure group,  $i = 1, \dots, k$ , and let  $RR_i$  = relative risk of disease for persons in the  $i$ th exposure group, compared with the unexposed group (group 1),  $i = 1, \dots, k$ . Let the probability of disease among the unexposed group =  $d$ . It follows that the overall probability of disease ( $p_D$ ) is

$$p_D = \sum_{i=1}^k p_i d RR_i = d \left[ 1 + \sum_{i=2}^k p_i (RR_i - 1) \right]$$

If all persons in the  $i$ th exposure group were unexposed, then  $(RR_i - 1)p_i d$  cases of disease could be prevented,  $i = 2, \dots, k$ . If we express the number of cases prevented as a proportion of the total number of cases and multiply by 100%, we obtain the following definition of  $AR$  for an exposure with multiple categories:

$$AR = 100\% \sum_{i=2}^k (RR_i - 1)p_i / [1 + \sum_{i=2}^k (RR_i - 1)p_i]$$

To obtain confidence limits for  $AR$  in this setting, one can use a multivariate extension of the delta method. The result is given as follows.

**Equation 13.13****Estimation of AR with Multiple Exposed Groups**

Suppose we have a prospective study with one unexposed group (denoted by group 1) and  $k - 1$  exposed group denoted by group  $2, \dots, k$ . Suppose we have a  $2 \times k$  table relating exposure to disease of the form:

		Exposure group		
		1	2	$k$
Disease	+	$a_1$	$a_2$	$\dots$
	-	$n_1 - a_1$	$n_2 - a_2$	$\dots$
		$n_1$	$n_2$	$n_k$

and let the  $RR$  in exposure group  $i$  be denoted by  $RR_i$  and estimated by  $\hat{RR}_i = (a_i / n_i) / (a_1 / n_1)$ ,  $i = 2, \dots, k$ . Assume the proportion of subjects in the  $i$ th exposure group in the reference population is denoted by  $p_i$  and is assumed to be known without error. The  $AR$  if all persons in the  $i$ th exposure group were unexposed is given by

$$AR = 100\% \times \sum_{i=2}^k (RR_i - 1)p_i / [1 + \sum_{i=2}^k (RR_i - 1)p_i]$$

If  $RR_i$  is estimated by  $\hat{RR}_i$ , then

- (1) Compute a point estimate of  $AR$  given by  $\hat{AR} = 100\% \times \sum_{i=2}^k (\hat{RR}_i - 1)p_i / [1 + \sum_{i=2}^k (\hat{RR}_i - 1)p_i]$ .
- (2) To obtain a  $100\% \times (1 - \alpha)$  CI for  $AR$ , compute  $y = \ln[(\hat{AR}/100) / (1 - \hat{AR}/100)]$ .
- (3) Compute  $var(y) = \sum_{i=2}^k (p_i \hat{RR}_i)^2 \left( \frac{n_1 - a_1}{a_1 n_1} + \frac{n_i - a_i}{a_i n_i} \right) / [\sum_{i=2}^k (\hat{RR}_i - 1)p_i]^2$   
 $- 2 \sum_{i=2}^k \sum_{j=2, j \neq i}^k p_{i1} p_{j2} \frac{\hat{RR}_{i1} \hat{RR}_{j2} (n_1 - a_1)}{a_1 n_1} / [\sum_{i=2}^k (\hat{RR}_i - 1)p_i]^2$
- (4) Compute  $(c_1, c_2) = y \pm z_{1-\alpha/2} [var(y)]^{1/2}$
- (5) A  $100\% \times (1 - \alpha)$  CI for  $AR$  is  $100\% \times [e^{c_1} / (1 + e^{c_1}), e^{c_2} / (1 + e^{c_2})]$ .

**Example 13.15**

**Cancer** Obtain the estimated  $AR$  of smoking for lung cancer with the associated 95% CI based on the data in Examples 13.12 and 13.14.

**Solution**

Table 13.4 is a  $2 \times 3$  table.

**Table 13.4****Association between smoking status and lung cancer risk**

		Smoking			Total
		Never	Ex	Current	
Lung cancer	Yes	10	25	50	85
	No	39,990	49,975	9950	99,915
		40,000	50,000	10,000	100,000

We assume the proportion of never, ex, and current smoking women in the general population is the same as in the study population, that is, 40%, 50%, and 10%, respectively. From Table 13.4,  $\hat{RR}_2 = (25 / 50,000) / (10 / 40,000) = 2.0$ ,  $\hat{RR}_3 = (50 / 10,000) / (10 / 40,000) = 20.0$ . Hence, we estimate the AR from Equation 13.13, Step 1 as follows:

$$\begin{aligned}\hat{AR} &= 100\% \times [(2-1)(.50) + (20-1)(.10)] / [1 + (2-1)(.50) + (20-1)(.10)] \\ &= 100\% \times 2.4 / 3.4 = 70.6\%.\end{aligned}$$

Thus, about 71% of lung cancer is attributable to smoking.

To obtain a 95% CI for AR, we compute  $y$  from Equation 13.13, Step 2 as follows:

$$y = \ln(.706/.294) = 0.875.$$

To obtain the variance of  $y$  we follow Equation 13.13, Step 3, where:

$$\begin{aligned}\text{var}(y) &= \frac{(p_2 \hat{RR}_2)^2}{\left[ \sum_{i=2}^3 (\hat{RR}_i - 1)p_i \right]^2} \left( \frac{n_1 - a_1}{a_1 n_1} + \frac{n_2 - a_2}{a_2 n_2} \right) \\ &\quad + \frac{(p_3 \hat{RR}_3)^2}{\left[ \sum_{i=2}^3 (\hat{RR}_i - 1)p_i \right]^2} \left( \frac{n_1 - a_1}{a_1 n_1} + \frac{n_3 - a_3}{a_3 n_3} \right) - 2 \frac{p_2 p_3}{\left[ \sum_{i=2}^3 (\hat{RR}_i - 1)p_i \right]^2} \frac{\hat{RR}_2 \hat{RR}_3 (n_1 - a_1)}{a_1 n_1} \\ &\equiv A + B - 2C\end{aligned}$$

We have:

$$A = \frac{[.50(2)]^2}{2.4^2} \left[ \frac{39,990}{10(40,000)} + \frac{49,975}{25(50,000)} \right] = \frac{1}{5.76} (0.100 + 0.040) = 0.0243$$

$$B = \frac{[.10(20)]^2}{2.4^2} \left[ \frac{39,990}{10(40,000)} + \frac{9950}{50(10,000)} \right] = 0.6944(0.100 + 0.020) = 0.0832$$

$$C = \frac{.50(.10)(2)(20)(39,990)}{2.4^2(10)(40,000)} = 0.0347$$

Hence,

$$\text{var}(y) = 0.0243 + 0.0832 - 2(0.0347) = 0.0381$$

$$se(y) = 0.0381^{1/2} = 0.1952.$$

A 95% CI for  $\ln[AR/(100 - AR)]$  is given in Equation 13.13, Step 4, by:

$$(c_1, c_2) = 0.875 \pm 1.96(0.1952) = (0.493, 1.258)$$

The corresponding 95% CI for AR is given in Equation 13.13, Step 5, by:

$$\begin{aligned}&[100(e^{0.493}) / (1 + e^{0.493}), 100(e^{1.258}) / (1 + e^{1.258})] \\ &= [100(1.637 / 2.637), 100(3.519) / 4.519] \\ &= (62.1\%, 77.9\%) \end{aligned}$$

## 13.5 Confounding and Standardization

### Confounding

When looking at the relationship between a disease and an exposure variable, it is often important to control for the effect of some other variable that is associated with both the disease and the exposure variable.

#### Definition 13.10

A **confounding variable** is a variable that is associated with both the disease and the exposure variable. Such a variable must usually be controlled for before looking at a disease–exposure relationship.

#### Example 13.16

**Cancer** Suppose we are interested in the relationship between lung-cancer incidence and heavy drinking (defined as  $\geq 2$  drinks per day). We conduct a prospective study in which drinking status is determined at baseline and the cohort is followed for 10 years to determine cancer endpoints. Table 13.5 is a  $2 \times 2$  table relating lung-cancer incidence to initial drinking status, in which we compare heavy drinkers ( $\geq 2$  drinks per day) with nondrinkers.

**Table 13.5**

**Crude relationship between lung-cancer incidence and drinking status**

		Lung cancer		1700
		Yes	No	
Drinking status	Heavy drinker	33	1667	
	Nondrinker	27	2273	2300
		60	3940	4000

Because lung cancer is relatively rare, we estimate the *RR* by the  $OR = (33 \times 2273)/(27 \times 1667) = 1.67$ . Thus it appears heavy drinking is a risk factor for lung cancer.

We refer to Table 13.5 as expressing the “crude” relationship between lung-cancer incidence and drinking status. The adjective “crude” means the relationship is presented without any adjustment for possible confounding variables. One such confounding variable is smoking because smoking is related to both drinking and lung-cancer incidence. Specifically, one explanation for the crude association found in Table 13.5 between lung-cancer incidence and drinking may be that heavy drinkers are more likely than nondrinkers to be smokers, and smokers are more likely to develop lung cancer than nonsmokers. To investigate this hypothesis, we look at the relationship between lung-cancer incidence and drinking status after controlling for smoking (i.e., separately for smokers and nonsmokers at baseline). These data are given in Table 13.6.

We see that smoking is related to drinking status. Specifically, 800 of the 1000 smokers (80%) vs. 900 of the 3000 nonsmokers (30%) are heavy drinkers. Also, smoking is related to lung cancer. Specifically, 30 of the 1000 smokers (3%) vs. 30 of the 3000 nonsmokers (1%) developed lung cancer.

**Example 13.17**

**Cancer** Investigate the relationship between lung-cancer incidence and drinking status, while controlling for smoking.

**Solution**

To investigate the relationship between lung-cancer incidence and drinking status, while controlling for smoking status, we can compute separate *ORs* for Table 13.6a and b. The *OR* relating lung cancer to drinking status among smokers is  $OR = (24 \times 194)/(6 \times 776) = 1.0$ , whereas the comparable *OR* among nonsmokers is  $OR = (9 \times 2079)/(21 \times 891) = 1.0$ . Thus, after controlling for the confounding variable smoking, we find *no* relationship between lung cancer and drinking status.

**Table 13.6 Relationship between lung-cancer incidence and drinking status while controlling for smoking status at baseline**

		(a) Smokers at baseline		(b) Nonsmokers at baseline	
		Lung cancer		Lung cancer	
		Yes	No	Yes	No
Drinking status	Heavy drinker	24	776	800	9
	Nondrinker	6	194	200	
		30	970	1000	30
					2970
					3000

**Definition 13.11**

The analysis of disease–exposure relationships in separate subgroups of the data, in which the subgroups are defined by one or more potential confounders, is called **stratification**. The subgroups themselves are called **strata**.

We refer to smoking as a *positive confounder* because it is related in the same direction to both lung-cancer incidence and heavy drinking.

**Definition 13.12**

A **positive confounder** is a confounder that either

- (1) is positively related to both exposure and disease, or
- (2) is negatively related to both exposure and disease

After adjusting for a positive confounder, an adjusted *RR* or *OR* is lower than the crude *RR* or *OR*.

**Definition 13.13**

A **negative confounder** is a confounder that either

- (1) is positively related to disease and negatively related to exposure, or
- (2) is negatively related to disease and positively related to exposure

After adjusting for a negative confounder, an adjusted *RR* or *OR* is greater than the crude *RR* or *OR*.

**Example 13.18**

**Cancer** In Table 13.6, what type of confounder is smoking?

**Solution**

Smoking is a positive confounder because it is positively related to both heavy drinking (the exposure) and lung cancer (the disease). Indeed, the crude positive association between lung cancer and drinking status ( $OR = 1.67$ ) was reduced to no association at all ( $OR = 1.0$ ) once smoking was controlled for.

**Table 13.7****Association between MI and OC use by age**

Age	Recent OC use	Cases (MI)	Controls	$\hat{OR}$	Proportion OC user	Proportion MI
25–29	Yes	4	62	7.2	23	2
	No	2	224			
30–34	Yes	9	33	8.9	9	5
	No	12	390			
35–39	Yes	4	26	1.5	8	9
	No	33	330			
40–44	Yes	6	9	3.7	3	16
	No	65	362			
45–49	Yes	6	5	3.9	3	24
	No	93	301			
Total	Yes	29	135	1.7		
	No	205	1607			

**Example 13.19**

**Cardiovascular Disease** The relationship between OC use and MI after stratification by age was considered by Shapiro et al. [2]. The data are given in Table 13.7.

We see that the age-specific  $ORs$  tend to be higher than the crude  $OR$  (1.7) when age was not controlled for. Age is an example of a negative confounder here because it is negatively associated with OC use (older women use OCs less frequently than younger women do) and positively associated with disease (older women are more likely to be cases than are younger women). Thus the age-specific  $ORs$  tend to be higher than the crude  $OR$ .

An often asked question is, When is it reasonable to control for a confounder when exploring the relationship between an exposure and disease? This depends on whether or not the confounder is in the “causal pathway” between exposure and disease.

**Definition 13.14**

A confounder is said to be in the **causal pathway** between exposure and disease if (1) the exposure is causally related to the confounder and (2) the confounder is causally related to disease.

**Example 13.20**

**Cardiovascular Disease** Suppose we are interested in the possible association between obesity and the development of coronary heart disease (CHD). If we examine the crude relationship between obesity and CHD, we usually find that obese people have higher rates of CHD than do people of normal weight. However, obesity is positively related to both hypertension and diabetes. In some studies, once hypertension and/or diabetes is controlled for as a confounder, there is a much weaker relationship or even no relationship between obesity and the development of CHD. Does this mean there is no real association between obesity and CHD?

**Solution**

No, it does not. If obesity is an important cause of both hypertension and diabetes and both are causally related to the development of CHD, then hypertension and diabetes are in the causal pathway between obesity and the development of CHD. It is inappropriate to include hypertension or diabetes as confounders of the relationship between obesity and CHD because they are in the causal pathway.

In other words, the decision as to which confounders are in the causal pathway should be made on the basis of biological rather than purely statistical considerations.

## Standardization

Age is often an important confounder influencing both exposure and disease rates. For this reason, it is often routine to control for age when assessing disease-exposure relationships. A first step is sometimes to compute rates for the exposed and unexposed groups that have been “age standardized.” The term “age standardized” means the expected disease rates in the exposed and unexposed groups are each based on an age distribution from a standard reference population. If the same standard is used for both the exposed and unexposed groups, then a comparison can be made between the two standardized rates that is not confounded by possible age differences between the two populations.

**Example 13.21**

**Infectious Disease** The presence of bacteria in the urine (bacteriuria) has been associated with kidney disease. Conflicting results have been reported from several studies concerning the possible role of OCs in bacteriuria. The following data were collected in a population-based study of nonpregnant premenopausal women younger than age 50 [3]. The data are presented on an age-specific basis in Table 13.8.

**Table 13.8****Risk of bacteriuria among OC users and nonusers**

Age group	% with bacteriuria			
	OC users		Non-OC users	
	%	n	%	n
16–19	1.2	84	3.2	281
20–29	5.6	284	4.0	552
30–39	6.3	96	5.5	623
40–49	22.2	18	2.7	482

Source: Reprinted with permission of *The New England Journal of Medicine*, 299, 536–537, 1978.

The prevalence of bacteriuria generally increases with age. In addition, the age distribution of OC users and non-OC users differs considerably, with OC use more common among younger women. Thus, for descriptive purposes we would like to compute age-standardized rates of bacteriuria separately for OC users and non-OC users and compare them using an *RR*.

**Definition 13.15**

Suppose people in a study population are stratified into  $k$  age groups. Let the risk of disease among the exposed in the  $i$ th age group =  $\hat{p}_{i1} = x_{i1} / n_{i1}$  where  $x_{i1}$  = number of exposed subjects with disease in the  $i$ th age group and  $n_{i1}$  = total number of exposed subjects in the  $i$ th age group,  $i = 1, \dots, k$ . Let the risk of disease among the unexposed in the  $i$ th age group =  $\hat{p}_{i2} = x_{i2} / n_{i2}$ , where  $x_{i2}$  = number of unexposed subjects

with disease in the  $i$ th age group and  $n_{i2}$  = total number of unexposed subjects in the  $i$ th age group,  $i = 1, \dots, k$ . Let  $n_i$  = number of subjects in the  $i$ th age group in a *standard* population,  $i = 1, \dots, k$ .

$$\text{Age-standardized risk of disease among the exposed} = \hat{p}_1^* = \sum_{i=1}^k n_i \hat{p}_{i1} / \sum_{i=1}^k n_i$$

$$\text{Age-standardized risk of disease among the unexposed} = \hat{p}_2^* = \sum_{i=1}^k n_i \hat{p}_{i2} / \sum_{i=1}^k n_i$$

$$\text{Standardized RR} = \hat{p}_1^* / \hat{p}_2^*$$

**Example 13.22**

**Infectious Disease** Using the data in Table 13.8, compute the age-standardized risk of bacteriuria separately for OC users and non-OC users, using the total study population as the standard, and compute the standardized *RR* for bacteriuria for OC users vs. non-OC users.

**Solution**

The age distribution of the total study population is shown in Table 13.9.

**Table 13.9****Age distribution of total study population**

Age group	<i>n</i>
16–19	365
20–29	836
30–39	719
40–49	500
Total	2420

The age-standardized risk of bacteriuria for OC users (the exposed) is

$$\begin{aligned}\hat{p}_1^* &= \frac{365(.012) + 836(.056) + 719(.063) + 500(.222)}{2420} \\ &= \frac{207.493}{2420} = .086\end{aligned}$$

The age-standardized risk of bacteriuria for non-OC users (the unexposed) is

$$\begin{aligned}\hat{p}_2^* &= \frac{365(.032) + 836(.040) + 719(.055) + 500(.027)}{2420} \\ &= \frac{98.165}{2420} = .041\end{aligned}$$

The standardized *RR* =  $.086/.041 = 2.1$ .

This method of standardization is sometimes referred to as **direct standardization**. Using age-standardized risks is somewhat controversial because results may differ depending on which standard is used. However, space limitations often make it impossible to present age-specific results in a paper, and the reader can get a quick summary of the overall results from the age-standardized risks.

Using standardized risks is a good descriptive tool for controlling for confounding. In the next section, we discuss how to control for confounding in assessing

disease–exposure relationships in a hypothesis-testing framework using the Mantel-Haenszel test. Finally, standardization can be based on stratification by factors other than age. For example, standardization by both age and sex is common. Similar methods can be used to obtain age–sex standardized risks and standardized *RRs* as given in Definition 13.15.

In this section, we have introduced the concept of a confounding variable (*C*), a variable related to both the disease (*D*) and exposure (*E*) variables. Furthermore, we classified confounding variables as positive confounders if the associations between *C* and *D* and *C* and *E*, respectively, are in the same direction and as negative confounders if the associations between *C* and *D* and *C* and *E* are in opposite directions. We also discussed when it is or is not appropriate to control for a confounder, according to whether *C* is or is not in the causal pathway between *E* and *D*. Finally, because age is often an important confounding variable, it is reasonable to consider descriptive measures of proportions and relative risk that control for age. Age-standardized proportions and *RRs* are such measures.

### REVIEW QUESTIONS 13B

- 1 Suppose we are interested in the association between smoking and bone density in women.
  - (a) If body-mass index (BMI) is inversely associated with smoking and positively associated with bone density, then is BMI a positive confounder, a negative confounder, or neither?
  - (b) If alcohol intake is positively associated with smoking and is unrelated to bone density, then is alcohol intake a positive confounder, a negative confounder, or neither?
- 2 Suppose the age-specific risks of hypertension in adults with left ventricular hypertrophy (LVH) and controls are as shown in Table 13.10.

**Table 13.10** Age-specific hypertension risks among patients with LVH and controls

	LVH		Control	
	Risk	N	Risk	N
40–49	.16	20	.14	50
50–59	.20	40	.18	40
60–69	.28	30	.20	36
70–79	.36	35	.25	29

- (a) Calculate the age-standardized risk of hypertension in each group using the total population in the two groups as the standard.
- (b) Calculate the age-standardized *RR* of hypertension for LVH patients vs. controls.

## 13.6 Methods of Inference for Stratified Categorical Data—The Mantel-Haenszel Test

### Example 13.23

**Cancer** A 1985 study identified a group of 518 cancer cases ages 15–59 and a group of 518 age- and sex-matched controls by mail questionnaire [4]. The main purpose of the study was to look at the effect of passive smoking on cancer risk. The study

defined passive smoking as exposure to the cigarette smoke of a spouse who smoked at least one cigarette per day for at least 6 months. One potential confounding variable was smoking by the participants themselves (i.e., personal smoking) because personal smoking is related to both cancer risk and spouse smoking. Therefore, it was important to control for personal smoking before looking at the relationship between passive smoking and cancer risk.

To display the data, a  $2 \times 2$  table relating case-control status to passive smoking can be constructed for both nonsmokers and smokers. The data are given in Table 13.11 for nonsmokers and Table 13.12 for smokers.

**Table 13.11 Relationship of passive smoking to cancer risk among nonsmokers**

Case-control status	Passive smoker		Total
	Yes	No	
Case	120	111	231
Control	80	155	235
Total	200	266	466

Source: From Sandler et al., "Passive Smoking in Adulthood and Cancer Risk," *American Journal of Epidemiology*, 1985 121: 37–48. Reprinted by permission of Oxford University Press.

**Table 13.12 Relationship of passive smoking to cancer risk among smokers**

Case-control status	Passive smoker		Total
	Yes	No	
Case	161	117	278
Control	130	124	254
Total	291	241	532

Source: From Sandler et al., "Passive Smoking in Adulthood and Cancer Risk," *American Journal of Epidemiology*, 1985 121: 37–48. Reprinted by permission of Oxford University Press.

The passive-smoking effect can be assessed separately for nonsmokers and smokers. Indeed, we notice from Tables 13.11 and 13.12 that the *OR* in favor of a case being exposed to cigarette smoke from a spouse who smokes vs. a control is  $(120 \times 155)/(80 \times 111) = 2.1$  for nonsmokers, whereas the corresponding *OR* for smokers is  $(161 \times 124)/(130 \times 117) = 1.3$ . Thus for both subgroups the trend is in the direction of more passive smoking among cases than among controls. The key question is how to combine the results from the two tables to obtain an overall estimated *OR* and test of significance for the passive-smoking effect.

In general, the data are stratified into  $k$  subgroups according to one or more confounding variables to make the units within a stratum as homogeneous as possible. The data for each stratum consist of a  $2 \times 2$  contingency table relating exposure to disease, as shown in Table 13.13 for the  $i$ th stratum.

**Table 13.13 Relationship of disease to exposure in the  $i$ th stratum**

Disease	Exposure		Total	
	Yes	Yes		
		$a_i$	$b_i$	
Yes	No	$c_i$	$d_i$	$a_i + b_i$
		$a_i + c_i$	$b_i + d_i$	
No				$n_i$

Based on our work on Fisher's exact test, the distribution of  $a_i$  follows a **hypergeometric distribution**. The test procedure is based on a comparison of the observed number of units in the (1, 1) cell of each stratum (denoted by  $O_i = a_i$ ) with the

expected number of units in that cell (denoted by  $E_i$ ). The test procedure is the same regardless of order of the rows and columns; that is, which row (or column) is designated as the first row (or column) is arbitrary. Based on the hypergeometric distribution (Equation 10.9), the expected number of units in the (1, 1) cell of the  $i$ th stratum is given by

**Equation 13.14**

$$E_i = \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

The observed and expected numbers of units in the (1, 1) cell are then summed over all strata, yielding  $O = \sum_{i=1}^k O_i$ ,  $E = \sum_{i=1}^k E_i$ , and the test is based on  $O - E$ . Based on the hypergeometric distribution (Equation 10.9), the variance of  $O_i$  is given by

**Equation 13.15**

$$V_i = \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}$$

Furthermore, the variance of  $O$  is denoted by  $V = \sum_{i=1}^k V_i$ . The test statistic is given by  $X_{MH}^2 = (|O - E| - .5)^2 / V$ , which should follow a chi-square distribution with 1 degree of freedom ( $df$ ) under the null hypothesis of no association between disease and exposure.  $H_0$  is rejected if  $X_{MH}^2$  is large. The abbreviation *MH* refers to Mantel-Haenszel; this procedure is known as the Mantel-Haenszel test and is summarized as follows.

**Equation 13.16**

#### Mantel-Haenszel Test

To assess the association between a dichotomous disease and a dichotomous exposure variable after controlling for one or more confounding variables, use the following procedure:

- (1) Form  $k$  strata, based on the level of the confounding variable(s), and construct a  $2 \times 2$  table relating disease and exposure within each stratum, as shown in Table 13.13.
- (2) Compute the total observed number of units ( $O$ ) in the (1, 1) cell over all strata, where

$$O = \sum_{i=1}^k O_i = \sum_{i=1}^k a_i$$

- (3) Compute the total expected number of units ( $E$ ) in the (1, 1) cell over all strata, where

$$E = \sum_{i=1}^k E_i = \sum_{i=1}^k \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

- (4) Compute the variance ( $V$ ) of  $O$  under  $H_0$ , where

$$V = \sum_{i=1}^k V_i = \sum_{i=1}^k \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}$$

- (5) The test statistic is then given by

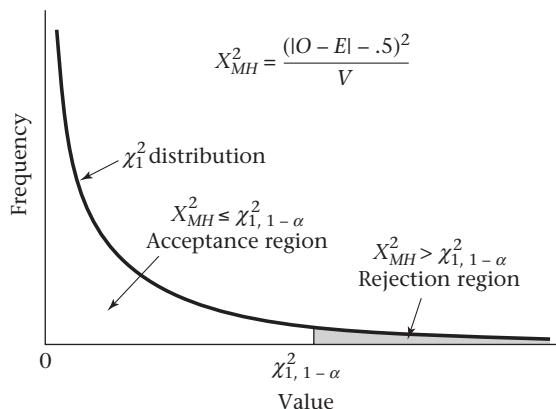
$$X_{MH}^2 = \frac{(|O - E| - .5)^2}{V}$$

which under  $H_0$  follows a chi-square distribution with 1  $df$ .

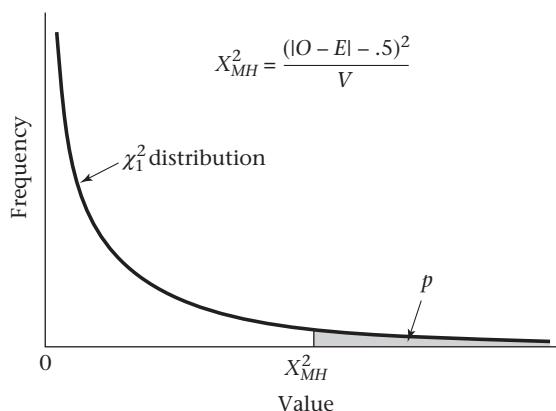
- (6) For a two-sided test with significance level  $\alpha$ ,
- if  $X_{MH}^2 > \chi_{1,1-\alpha}^2$  then reject  $H_0$ .
  - if  $X_{MH}^2 \leq \chi_{1,1-\alpha}^2$  then accept  $H_0$ .
- (7) The exact  $p$ -value for this test is given by
- $$p = Pr(\chi_1^2 > X_{MH}^2)$$
- (8) Use this test only if the variance  $V$  is  $\geq 5$ .
- (9) Which row or column is designated as first is arbitrary. The test statistic  $X_{MH}^2$  and the assessment of significance are the same regardless of the order of the rows and columns.

The acceptance and rejection regions for the Mantel-Haenszel test are shown in Figure 13.1. The computation of the  $p$ -value for the Mantel-Haenszel test is illustrated in Figure 13.2.

**Figure 13.1 Acceptance and rejection regions for the Mantel-Haenszel test**



**Figure 13.2 Computation of the  $p$ -value for the Mantel-Haenszel test**



**Example 13.24**

**Cancer** Assess the relationship between passive smoking and cancer risk using the data stratified by personal smoking status in Tables 13.11 and 13.12.

**Solution**

Denote the nonsmokers as stratum 1 and the smokers as stratum 2.

$$O_1 = \text{observed number of nonsmoking cases who are passive smokers} = 120$$

$$O_2 = \text{observed number of smoking cases who are passive smokers} = 161$$

Furthermore,

$$E_1 = \frac{231 \times 200}{466} = 99.1$$

$$E_2 = \frac{278 \times 291}{532} = 152.1$$

Thus the total observed and expected numbers of cases who are passive smokers are, respectively,

$$O = O_1 + O_2 = 120 + 161 = 281$$

$$E = E_1 + E_2 = 99.1 + 152.1 = 251.2$$

Therefore, more cases are passive smokers than would be expected based on their personal smoking habits. Now compute the variance to assess whether this difference is statistically significant.

$$V_1 = \frac{231 \times 235 \times 200 \times 266}{466^2 \times 465} = 28.60$$

$$V_2 = \frac{278 \times 254 \times 291 \times 241}{532^2 \times 531} = 32.95$$

$$\text{Therefore } V = V_1 + V_2 = 28.60 + 32.95 = 61.55$$

Thus the test statistic  $X_{MH}^2$  is given by

$$X_{MH}^2 = \frac{(|281 - 251.2| - .5)^2}{61.55} = \frac{858.17}{61.55} = 13.94 \sim \chi_1^2 \text{ under } H_0$$

Because  $\chi_{1,999}^2 = 10.83 < 13.94 = X_{MH}^2$ , it follows that  $p < .001$ . Thus there is a highly significant positive association between case-control status and passive-smoking exposure, even after controlling for personal cigarette-smoking habit.

### Estimation of the Odds Ratio for Stratified Data

The Mantel-Haenszel method tests significance of the relationship between disease and exposure. However, it does not measure the strength of the association. Ideally, we would like a measure similar to the *OR* presented for a single  $2 \times 2$  contingency table in Definition 13.6. Assuming that the underlying *OR* is the same for each stratum, an estimate of the common underlying *OR* is provided by the Mantel-Haenszel estimator as follows.

**Equation 13.17****Mantel-Haenszel Estimator of the Common Odds Ratio for Stratified Data**

In a collection of  $k$   $2 \times 2$  contingency tables, where the table corresponding to the  $i$ th stratum is denoted as in Table 13.13, the Mantel-Haenszel estimator of the common *OR* is given by

$$\hat{OR}_{MH} = \frac{\sum_{i=1}^k a_i d_i / n_i}{\sum_{i=1}^k b_i c_i / n_i}$$

**Example 13.25**

**Cancer** Estimate the OR in favor of being a passive smoker for cancer cases vs. controls after controlling for personal smoking habit.

**Solution**

From Equation 13.17, Table 13.11, and Table 13.12,

$$\hat{OR}_{MH} = \frac{(120 \times 155 / 466) + (161 \times 124 / 532)}{(80 \times 111 / 466) + (130 \times 117 / 532)} = \frac{77.44}{47.65} = 1.63$$

Thus the odds in favor of being a passive smoker for a cancer case is 1.6 times as large as that for a control. Because cancer is relatively rare, we can also interpret these results as indicating that risk of cancer for a passive smoker is 1.6 times as great as for a nonpassive smoker, even after controlling for personal smoking habit.

We are also interested in estimating confidence limits for the OR in Equation 13.17. A variance estimate of  $\ln(\hat{OR}_{MH})$  has been provided by Robins et al. [5], which is accurate under a wide range of conditions, particularly if there are many strata with small numbers of subjects in each stratum. This variance estimate can be used to obtain confidence limits for  $\ln(OR)$ . We can then take the antilog of each of the confidence limits for  $\ln(OR)$  to obtain confidence limits for OR. This procedure is summarized as follows.

**Equation 13.18****Interval Estimate for the Common Odds Ratio from a Collection of  $k$   $2 \times 2$  Contingency Tables**

A two-sided  $100\% \times (1 - \alpha)$  CI for the common OR from a collection of  $k$   $2 \times 2$  tables is given by

$$\exp\left[\ln \hat{OR}_{MH} \pm z_{1-\alpha/2} \sqrt{Var(\ln \hat{OR}_{MH})}\right]$$

where

$$Var(\ln \hat{OR}_{MH}) = \frac{\sum_{i=1}^k P_i R_i}{2 \left( \sum_{i=1}^k R_i \right)^2} + \frac{\sum_{i=1}^k (P_i S_i + Q_i R_i)}{2 \left( \sum_{i=1}^k R_i \right) \left( \sum_{i=1}^k S_i \right)} + \frac{\sum_{i=1}^k Q_i S_i}{2 \left( \sum_{i=1}^k S_i \right)^2} \equiv A + B + C$$

where  $A$ ,  $B$ , and  $C$  correspond to the first, second, and third terms on the right-hand side of  $Var(\ln \hat{OR}_{MH})$ , and

$$P_i = \frac{a_i + d_i}{n_i}, Q_i = \frac{b_i + c_i}{n_i}, R_i = \frac{a_i d_i}{n_i}, S_i = \frac{b_i c_i}{n_i}$$

**Example 13.26**

**Cancer** Estimate 95% confidence limits for the common OR using the data in Tables 13.11 and 13.12.

**Solution**

Note from Example 13.25 that the point estimate of the  $OR = \hat{OR}_{MH} = 1.63$ . To obtain confidence limits, we first compute  $P_i$ ,  $Q_i$ ,  $R_i$ , and  $S_i$  as follows:

$$P_1 = \frac{120+155}{466} = .590, \quad Q_1 = 1 - P_1 = .410$$

$$R_1 = \frac{120(155)}{466} = 39.91, \quad S_1 = \frac{80(111)}{466} = 19.06$$

$$P_2 = \frac{161+124}{532} = .536, \quad Q_2 = 1 - P_2 = .464$$

$$R_2 = \frac{161(124)}{532} = 37.53, \quad S_2 = \frac{130(117)}{532} = 28.59$$

Thus

$$\text{Var}(\ln \hat{OR}_{MH}) = A + B + C$$

where

$$A = \frac{.590(39.91) + .536(37.53)}{2(39.91 + 37.53)^2} = 0.00364$$

$$B = \frac{.590(19.06) + .410(39.91) + .536(28.59) + .464(37.53)}{2(39.91 + 37.53)(19.06 + 28.59)} = 0.00818$$

$$C = \frac{.410(19.06) + .464(28.59)}{2(19.06 + 28.59)^2} = 0.00464$$

Thus  $\text{Var}(\ln \hat{OR}_{MH}) = 0.00364 + 0.00818 + 0.00464 = 0.01646$ . The 95% CI for  $\ln(OR)$  is

$$\ln(1.63) \pm 1.96\sqrt{0.01646} = (0.234, 0.737)$$

The 95% CI for  $OR$  is

$$(e^{0.234}, e^{0.737}) = (1.26, 2.09)$$

## Effect Modification

One assumption made in the estimation of a common  $OR$  in Equation 13.17 is that the strength of association is the same in each stratum. If the underlying  $OR$  is different in the various strata, then it makes little sense to estimate a common  $OR$ .

### Definition 13.16

Suppose we are interested in studying the association between a disease variable  $D$  and an exposure variable  $E$  but are concerned about the possible confounding effect of another variable  $C$ . We stratify the study population into  $g$  strata according to the variable  $C$  and compute the  $OR$  relating disease to exposure in each stratum. If the underlying (true)  $OR$  is different across the  $g$  strata, then there is said to be **interaction** or **effect modification** between  $E$  and  $C$ , and the variable  $C$  is called an **effect modifier**.

In other words, if  $C$  is an effect modifier, then the relationship between disease and exposure differs for different levels of  $C$ .

**Example 13.27**

**Cancer** Consider the data in Tables 13.11 and 13.12. We estimated that the *OR* relating cancer and passive smoking is 2.1 for nonsmokers and 1.3 for smokers. If these were the underlying *ORs* in these strata, then personal smoking would be an effect modifier. Specifically, the relationship between passive smoking and cancer is much stronger for nonsmokers than for smokers. The rationale for this is that the home environment of active smokers already contains cigarette smoke and the extra degradation of the environment by spousal smoking may not be that meaningful.

The issue remains, how can we detect whether another variable *C* is an effect modifier? We use a generalization of the Woolf procedure for obtaining confidence limits for a single *OR* given in Equation 13.11. Specifically, we want to test the hypothesis  $H_0: OR_1 = \dots = OR_k$  vs.  $H_1$ : at least two of the  $OR_i$  differ from each other.

We base our test on the test statistic  $X^2 = \sum_{i=1}^k w_i (\ln \hat{OR}_i - \bar{\ln OR})^2$  where  $\ln \hat{OR}_i$  = the estimated log *OR* relating disease to exposure in the *i*th stratum of the potential effect modifier, *C*,  $\bar{\ln OR}$  = the estimated “weighted average” log *OR* over all strata, and *w* is a weight that is inversely proportional to the variance of  $\ln \hat{OR}_i$ . The purpose of the weighting is to weight strata with lower variance (which usually correspond to strata with more subjects) more heavily. If  $H_0$  is true, then  $X^2$  will be small because each of the stratum-specific log *ORs* will be relatively close to each other and to the “average” log *OR*. Conversely, if  $H_1$  is true, then  $X^2$  will be large. Under  $H_0$ , it can be shown that  $X^2$  follows a chi-square distribution with  $k - 1$  *df*. Thus we will reject  $H_0$  if  $X^2 > \chi^2_{k-1,1-\alpha}$  and accept  $H_0$  otherwise. This procedure is summarized as follows.

**Equation 13.19****Chi-Square Test for Homogeneity of *ORs* over Different Strata (Woolf Method)**

Suppose we have a dichotomous disease variable *D* and exposure variable *E*. We stratify our study population into *k* strata according to a confounding variable *C*. Let  $OR_i$  = underlying *OR* in the *i*th stratum. To test the hypothesis  $H_0: OR_1 = \dots = OR_k$  vs.  $H_1$ : at least two of the  $OR_i$  are different with a significance level  $\alpha$ , use the following procedure:

(1) Compute the test statistic  $X_{\text{HOM}}^2 = \sum_{i=1}^k w_i (\ln \hat{OR}_i - \bar{\ln OR})^2 \sim \chi^2_{k-1}$  under  $H_0$

where  $\ln \hat{OR}_i$  = estimated  $\ln OR$  in the *i*th stratum =  $\ln [a_i d_i / (b_i c_i)]$  and  $a_i, b_i, c_i, d_i$  are the cells of the  $2 \times 2$  table relating disease to exposure in the *i*th stratum as shown in Table 13.13.

$$w_i = \left( \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right)^{-1}$$

(1a) An alternative computational form of the test statistic is

$$X_{\text{HOM}}^2 = \sum_{i=1}^k w_i (\ln \hat{OR}_i)^2 - \left( \sum_{i=1}^k w_i \ln \hat{OR}_i \right)^2 \Bigg/ \sum_{i=1}^k w_i$$

(2) If  $X_{\text{HOM}}^2 > \chi^2_{k-1,1-\alpha}$ , then reject  $H_0$ ,

If  $X_{\text{HOM}}^2 \leq \chi^2_{k-1,1-\alpha}$ , then accept  $H_0$ .

(3) The exact *p*-value =  $Pr(\chi^2_{k-1} > X^2)$

**Example 13.28**

**Cancer** Assess whether the *ORs* relating passive smoking to cancer are different for smokers vs. nonsmokers, using the data in Tables 13.11 and 13.12.

**Solution**

Let stratum 1 refer to nonsmokers and stratum 2 to smokers. Referring to Tables 13.11 and 13.12, we see that

$$\begin{aligned}\ln \hat{OR}_1 &= \ln \left( \frac{120 \times 155}{80 \times 111} \right) = \ln(2.095) = 0.739 \\ w_1 &= \left( \frac{1}{120} + \frac{1}{111} + \frac{1}{80} + \frac{1}{155} \right)^{-1} = (0.036)^{-1} = 27.55 \\ \ln \hat{OR}_2 &= \ln \left( \frac{161 \times 124}{130 \times 117} \right) = \ln(1.313) = 0.272 \\ w_2 &= \left( \frac{1}{161} + \frac{1}{117} + \frac{1}{130} + \frac{1}{124} \right)^{-1} = (0.031)^{-1} = 32.77\end{aligned}$$

Thus, based on step 1a in Equation 13.19, the test statistic is given by

$$\begin{aligned}X_{\text{HOM}}^2 &= 27.55(0.739)^2 + 32.77(0.272)^2 - [27.55(0.739) + 32.77(0.272)]^2 / (27.55 + 32.77) \\ &= 17.486 - (29.284)^2 / 60.32 \\ &= 17.486 - 14.216 = 3.27 \sim \chi^2_1 \text{ under } H_0\end{aligned}$$

Referring to Table 6 in the Appendix, we note that  $\chi^2_{1,90} = 2.71$ ,  $\chi^2_{1,95} = 3.84$ . Because  $2.71 < 3.27 < 3.84$ , it follows that  $.05 < p < .10$ . Thus there is no significant effect modification; that is, the *ORs* in the two strata are not significantly different.

In general, it is important to test for homogeneity of the stratum-specific *ORs*. If the true *ORs* are significantly different, then it makes no sense to obtain a pooled-*OR* estimate such as given by the Mantel-Haenszel estimator in Equation 13.17. Instead, separate *ORs* should be reported.

### Estimation of the *OR* in Matched-Pair Studies

There is a close connection between McNemar's test for matched-pair data in Equation 10.12 and the Mantel-Haenszel test procedure for stratified categorical data in Equation 13.16. Matched pairs are a special case of stratification in which each matched pair corresponds to a separate stratum of size 2. It can be shown that McNemar's test is a special case of the Mantel-Haenszel test for strata of size 2. Furthermore, the Mantel-Haenszel *OR* estimator in Equation 13.17 reduces to  $\hat{OR} = \frac{n_A}{n_B}$  for matched-pair data, where  $n_A$  = number of discordant pairs of type A and  $n_B$  = number of discordant pairs of type B. Also, it can be shown that the variance of  $\ln(\text{OR})$  for a matched-pair study is given by  $\text{Var}[\ln(\hat{OR})] = \frac{1}{np\hat{q}}$ , where  $n$  = total number of discordant pairs =  $n_A + n_B$ ,  $\hat{p}$  = proportion of discordant pairs of type A =  $n_A/(n_A + n_B)$ ,  $\hat{q} = 1 - \hat{p}$ . This leads to the following technique for estimating the disease-exposure *OR* in matched-pair studies.

**Equation 13.20**

#### Estimation of the *OR* in Matched-Pair Studies

Suppose we want to study the relationship between a dichotomous disease and exposure variable, in a case-control design. We control for confounding by

forming matched pairs of subjects with disease (cases) and subjects without disease (controls), where the two subjects in a matched pair are the same or similar on one or more confounding variables.

- (1) The *OR* relating disease to exposure is estimated by

$$\hat{OR} = n_A/n_B$$

where

$n_A$  = number of matched pairs in which the case is exposed and the control is not exposed

$n_B$  = number of matched pairs in which the case is not exposed and the control is exposed

- (2) A two-sided  $100\% \times (1 - \alpha)$  CI for *OR* is given by  $(e^{c_1}, e^{c_2})$ , where

$$c_1 = \ln(\hat{OR}) - z_{1-\alpha/2} \sqrt{\frac{1}{n\hat{p}\hat{q}}}$$

$$c_2 = \ln(\hat{OR}) + z_{1-\alpha/2} \sqrt{\frac{1}{n\hat{p}\hat{q}}}$$

$$n = n_A + n_B$$

$$\hat{p} = \frac{n_A}{n_A + n_B}, \quad \hat{q} = 1 - \hat{p}$$

- (3) The same methodology can be used for prospective or cross-sectional studies in which exposed and unexposed individuals are matched on one or more confounding variables and disease outcomes are compared between exposed and unexposed individuals. In this setting,

$n_A$  = number of matched pairs in which the exposed subject has disease and the unexposed subject does not

$n_B$  = number of matched pairs in which the exposed subject does not have disease and the unexposed subject does and steps 1 and 2 are as just indicated

- (4) This method should only be used if  $n$  = number of discordant pairs is  $\geq 20$ .

### Example 13.29

**Cancer** Estimate the *OR* relating type of treatment to 5-year mortality using the matched-pair data in Table 10.14.

#### Solution

We have from Table 10.14 that

$n_A$  = number of matched pairs in which the treatment A patient dies within 5 years and the treatment B patient survives for 5 years = 5

$n_B$  = number of matched pairs in which the treatment B patient dies within 5 years and the treatment A patient survives for 5 years = 16

Thus  $(\hat{OR}) = 5/16 = 0.31$ . To obtain a 95% CI we see that  $n = 21$ ,  $\hat{p} = 5/21 = .238$ ,  $\hat{q} = .762$ , and  $n\hat{p}\hat{q} = 3.81$ . Thus  $\ln(\hat{OR}) = -1.163$ ,  $Var[\ln(\hat{OR})] = 1/3.81 = 0.263$ , and a 95% CI for  $\ln(OR)$  is  $(-1.163 - 1.96\sqrt{0.263}, -1.163 + 1.96\sqrt{0.263}) = (-2.167, -0.159)$ . The corresponding 95% CI for *OR* is  $(e^{-2.167}, e^{-0.159}) = (0.11, 0.85)$ .

## Testing for Trend in the Presence of Confounding— Mantel-Extension Test

**Example 13.30**

**Sleep Disorders** Sleep-disordered breathing is very common among adults. To estimate the prevalence of this disorder, a questionnaire concerning sleep habits was mailed to 3513 individuals 30–60 years of age who worked for three large state agencies in Wisconsin [6]. Subjects were classified as habitual snorers if they reported either (1) snoring, snorting, or breathing pauses every night or almost every night or (2) extremely loud snoring. The results are given by age and sex group in Table 13.14.

**Table 13.14****Prevalence of habitual snoring by age and sex group**

Age	Women			Men		
	Yes	No	Total	Yes	No	Total
30–39	196	603	799	188	348	536
40–49	223	486	709	313	383	696
50–60	103	232	335	232	206	438
Total	522	1321	1843	733	937	1670

We would like to assess whether the prevalence of habitual snoring increases with age.

In this study, we want to assess whether there is a trend in the prevalence rates with age after controlling for sex. To address this issue, we need to generalize the chi-square test for trend given in Equation 10.24 to allow for stratification of our study sample by relevant confounding variables. We can also describe this problem as a generalization of the Mantel-Haenszel test given in Equation 13.16 in which we are combining results from several  $2 \times k$  tables (rather than just  $2 \times 2$  tables). Suppose we have  $s$  strata and  $k$  ordered categories for the exposure variable. Consider the  $2 \times k$  table relating the dichotomous disease variable  $D$  to the ordered categorical exposure variable  $E$  for subjects in the  $i$ th stratum (see Table 13.15). We assume there is a score for the  $j$ th exposure category denoted by  $x_j$ ,  $j = 1, \dots, k$ .

**Table 13.15****Relationship of disease to exposure in the  $i$ th stratum,  
 $i = 1, \dots, s$** 

		Exposure				$n_i$
		1	2	...	$k$	
Disease	+	$n_{i1}$	$n_{i2}$	...	$n_{ik}$	$m_i$
	-	$m_{i1}$	$m_{i2}$	...	$m_{ik}$	$N_i$
		$t_{i1}$	$t_{i2}$	...	$t_{ik}$	
Score		$x_1$	$x_2$	...	$x_k$	

The total observed score among subjects with disease in the  $i$ th stratum =  $O_i = \sum_{j=1}^k n_{ij}x_j$ . The expected score among diseased subjects in the  $i$ th stratum under the null hypothesis that the average score for subjects with and without disease in a stratum is the

same =  $E_i = \left( \sum_{j=1}^k t_{ij}x_j \right) \frac{n_i}{N_i}$ . If diseased subjects tend to have higher exposure scores on average than nondiseased subjects, then  $O_i$  will be greater than  $E_i$  for most strata. If diseased subjects tend to have lower exposure scores than nondiseased subjects, then  $O_i$  will be less than  $E_i$  for most strata. Therefore, we will base our test on  $O - E$  where  $O = \sum_{i=1}^s O_i$ ,  $E = \sum_{i=1}^s E_i$ . The test procedure is given as follows.

**Equation 13.21****Chi-Square Test for Trend-Multiple Strata (Mantel Extension Test)**

- (1) Suppose we have  $s$  strata. In each stratum, we have a  $2 \times k$  table relating disease (2 categories) to exposure ( $k$  ordered categories) with score for the  $j$ th category =  $x_j$  as shown in Table 13.15.
- (2) Let  $p_{ij} =$  proportion of subjects with disease among subjects in the  $i$ th stratum and  $j$ th exposure category

To test the hypothesis  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ , where

$$p_{ij} = \alpha_i + \beta x_j$$

We compute the test statistic

$$X_{TR}^2 = (|O - E| - 0.5)^2 / V \sim \chi_1^2 \text{ under } H_0$$

where

$$\begin{aligned} O &= \sum_{i=1}^s O_i = \sum_{i=1}^s \sum_{j=1}^k n_{ij}x_j \\ E &= \sum_{i=1}^s E_i = \sum_{i=1}^s \left[ \left( \sum_{j=1}^k t_{ij}x_j \right) \frac{n_i}{N_i} \right] \\ V &= \sum_{i=1}^s V_i = \sum_{i=1}^s \frac{n_i m_i (N_i s_{2i} - s_{1i}^2)}{N_i^2 (N_i - 1)} \\ s_{1i} &= \sum_{j=1}^k t_{ij}x_j, i = 1, \dots, s \\ s_{2i} &= \sum_{j=1}^k t_{ij}x_j^2, i = 1, \dots, s \end{aligned}$$

- (3) If  $X_{TR}^2 > \chi_{1,1-\alpha}^2$  we reject  $H_0$ .  
If  $X_{TR}^2 \leq \chi_{1,1-\alpha}^2$  we accept  $H_0$ .
- (4) The exact  $p$ -value =  $Pr(\chi_1^2 > X_{TR}^2)$ .
- (5) This test should only be used if  $V \geq 5$ .

**Example 13.31**

Use the data in Table 13.14 to assess whether the prevalence of habitual snoring increases with age, after controlling for sex.

**Solution**

In this example, we have two strata, corresponding to women ( $i = 1$ ) and men ( $i = 2$ ), respectively. We will use scores of 1, 2, 3 for the three age groups. We have

$$\begin{aligned}
 O_1 &= 196(1) + 223(2) + 103(3) = 951 \\
 O_2 &= 188(1) + 313(2) + 232(3) = 1510 \\
 O &= 951 + 1510 = 2461 \\
 E_1 &= [799(1) + 709(2) + 335(3)]522/1843 = 912.6 \\
 E_2 &= [536(1) + 696(2) + 438(3)]733/1670 = 1423.0 \\
 E &= 912.6 + 1423.0 = 2335.6 \\
 s_{11} &= 799(1) + 709(2) + 335(3) = 3222 \\
 s_{21} &= 799(1^2) + 709(2^2) + 335(3^2) = 6650 \\
 s_{12} &= 536(1) + 696(2) + 438(3) = 3242 \\
 s_{22} &= 536(1^2) + 696(2^2) + 438(3^2) = 7262 \\
 V_1 &= \frac{522(1321)[1843(6650) - 3222^2]}{1843^2(1842)} = 206.61 \\
 V_2 &= \frac{733(937)[1670(7262) - 3242^2]}{1670^2(1669)} = 238.59 \\
 V &= 206.61 + 238.59 = 445.21
 \end{aligned}$$

Thus the test statistic is given by

$$X_{TR}^2 = \frac{(|2461 - 2335.6| - .5)^2}{445.21} = \frac{124.9^2}{445.21} = 35.06 \sim \chi_1^2$$

Because  $\chi_{1,999}^2 = 10.83$  and  $X_{TR}^2 = 35.06 > 10.83$ , it follows that  $p < .001$ . Therefore, there is a significant association between the prevalence of habitual snoring and age, with older subjects snoring more frequently. This analysis was performed while controlling for the possible confounding effects of sex.

In this section, we have learned about analytic techniques for controlling for confounding in epidemiologic studies. If we have a dichotomous disease variable ( $D$ ), a dichotomous exposure variable ( $E$ ), and a categorical confounder ( $C$ ), then we can use the Mantel-Haenszel test to assess the association between  $D$  and  $E$  while controlling for  $C$ . On the master flowchart in the back of the book (p. 846), starting at ⑥, we answer yes to (1)  $2 \times 2$  contingency table? and at ④ arrive at the box labeled “Use two-sample test for binomial proportions, or  $2 \times 2$  contingency-table methods if no confounding is present, or the Mantel-Haenszel test if confounding is present.”

If  $E$  is categorical but has more than two categories, then we can use the Mantel Extension test for this purpose. Referring to the master flowchart again, we answer no to (1)  $2 \times 2$  contingency table? yes to (2)  $2 \times k$  contingency table? and yes to (3) interested in trend over  $k$  proportions? This leads us to the box labeled “Use chi-square test for trend if no confounding is present, or the Mantel Extension test if confounding is present.”

### REVIEW QUESTIONS 13C

- 1 What is the purpose of the Mantel-Haenszel test? How does it differ from the ordinary chi-square test for  $2 \times 2$  tables?
- 2 A case-control study was performed relating environmental arsenic exposure to nonmelanoma skin cancer. Distance of a residence from a power station was

considered as an indirect measure of arsenic exposure [7]). The data, presented by gender, are shown in Table 13.16.

- (a) Use the Mantel-Haenszel test to assess the association between case-control status and distance to the power station.
  - (b) Estimate the *OR* between case-control status and distance to the power station, and provide a 95% CI around this estimate.
  - (c) Assess the homogeneity of the preceding *ORs* for males vs. females.
- 3 What is the difference between the Mantel Extension test and the Mantel-Haenszel test?

**Table 13.16** Association between nonmelanoma skin cancer and distance of residence from a power station

	Males		Females		
	Distance to power station		Distance to power station		
	< 5 km	> 10 km	< 5 km	> 10 km	
Cases	15	30		21	50
Controls	23	38		19	52

- 4 The complete distribution of distance to the power station  $\times$  case-control status for the study mentioned in Review Question 13C.2 is given in Table 13.17.

**Table 13.17** Association between nonmelanoma skin cancer and distance of residence from a power station

	Males			Females			
	Distance to power station			Distance to power station			
	< 5 km	5–10 km	> 10 km	< 5 km	5–10 km	> 10 km	
Cases	15	84	30		21	64	50
Controls	23	81	38		19	73	52

Use the Mantel Extension test to assess whether distance to the power station is associated with case-control status after controlling for gender.

## 13.7 Power and Sample-Size Estimation for Stratified Categorical Data

### Example 13.32

**Cancer** A study was performed [8] based on a sample of 106,330 women enrolled in the Nurses' Health Study (NHS) relating ever use of OCs at baseline (in 1976) to breast-cancer incidence from 1976 to 1980. Because both OC use and breast cancer are related to age, the data were stratified by 5-year age groups and the Mantel-Haenszel test was employed to test for this association. The results supported the null hypothesis. The estimated *OR* ( $\hat{OR}_{MH}$ ) was 1.0 with 95% CI = (0.8, 1.3). What power did the study have to detect a significant difference if the underlying *OR* = 1.3?

The power formulas given in Section 10.5 are not applicable because a stratified analysis was used rather than a simple comparison of binomial proportions. However, an approximate power formula is available [9]. To use this formula, we need to know (1) the proportion of exposed subjects in each stratum, (2) the proportion of diseased subjects in each stratum, (3) the proportion of subjects in each stratum of the total study population, and (4) the size of the total study population. The power formula is given as follows.

**Equation 13.22**
**Power Estimation for a Collection of  $2 \times 2$  Tables Based on the Mantel-Haenszel Test**

Suppose we wish to relate a dichotomous disease variable  $D$  to a dichotomous exposure variable  $E$  and want to control for a categorical confounding variable  $C$ . We subdivide the study population into  $k$  strata, where the  $2 \times 2$  table in the  $i$ th stratum is given by

		Exposure		$N_i$
		+	-	
Disease	+	$a_i$	$b_i$	$N_i$
	-	$c_i$	$d_i$	
		$M_{1i}$	$M_{2i}$	$N_i$

We wish to test the hypothesis  $H_0: OR = 1$  vs.  $H_1: OR = \exp(\gamma)$  for  $\gamma \neq 0$ , where  $OR$  is the underlying stratum-specific  $OR$  relating disease to exposure that is assumed to be the same in each stratum. Let

$N$  = size of the total study population

$r_i$  = proportion of exposed subjects in stratum  $i$

$s_i$  = proportion of diseased subjects in stratum  $i$

$t_i$  = proportion of total study population in stratum  $i$

If we use the Mantel-Haenszel test with a significance level of  $\alpha$ , then the power is given by

$$\text{Power} = \Phi \left[ \frac{\sqrt{N} \left( \gamma B_1 + \frac{\gamma^2}{2} B_2 \right) - z_{1-\alpha/2} \sqrt{B_1}}{(B_1 + \gamma B_2)^{1/2}} \right]$$

where

$$B_1 = \sum_{i=1}^k B_{1i}$$

$$B_{1i} = r_i s_i t_i (1 - r_i)(1 - s_i)$$

$$B_2 = \sum_{i=1}^k B_{2i}$$

$$B_{2i} = B_{1i}(1 - 2r_i)(1 - 2s_i)$$

**Example 13.33** | **Cancer** Estimate the power of the study described in Example 13.32 for the alternative hypothesis that  $OR = 1.3$ .

**Solution**

The data were stratified by 5-year age groups. The age-specific proportion of ever OC users ( $r_i$ ), the age-specific 4-year incidence of breast cancer ( $s_i$ ), and the age distribution of the total study population ( $t_i$ ) are given in Table 13.18 together with  $B_1$  and  $B_2$ .

We see that the proportion of ever OC users goes down sharply with age and breast-cancer incidence rises sharply with age. Thus there is evidence of strong negative confounding and a stratified analysis is essential. To compute power, we note that  $N = 106,330$ ,  $\gamma = \ln(1.3) = 0.262$ ,  $z_{1-\alpha/2} = z_{.975} = 1.96$ . Thus

$$\text{Power} = \Phi \left\{ \frac{\sqrt{106,330} [0.262 \times 1.06 \times 10^{-3} + 0.262^2 (2.32 \times 10^{-4}/2)] - 1.96 \sqrt{1.06 \times 10^{-3}}}{[1.06 \times 10^{-3} + 0.262 (2.32 \times 10^{-4})]^{1/2}} \right\}$$

$$= \Phi \left( \frac{0.0296}{0.0335} \right) = \Phi(0.882) = .81$$

Thus the study had 81% power to detect a true OR of 1.3.

An alternative (and simpler) method for computing power would be to pool data over all strata and compute crude power based on the overall  $2 \times 2$  table relating disease to exposure. Generally, if there is positive confounding, then the true power (i.e., the power based on the Mantel-Haenszel test—Equation 13.16) is lower than the crude power; if there is negative confounding (as was the case in Example 13.33), then the true power is greater than the crude power.

**Table 13.18** Power calculation for studying the association between breast-cancer incidence and OC use based on NHS data

Age group	Proportion ever OC use ( $r$ )	4-year incidence of breast cancer <sup>a</sup> ( $s$ )	Proportion of total study population ( $t$ )	$B_{1i}$	$B_{2i}$
30–34	.771	160	.188	$5.30 \times 10^{-5}$ <sup>b</sup>	$-2.86 \times 10^{-5}$ <sup>b</sup>
35–39	.629	350	.195	$1.59 \times 10^{-4}$	$-4.07 \times 10^{-5}$
40–44	.465	530	.209	$2.74 \times 10^{-4}$	$1.90 \times 10^{-5}$
45–49	.308	770	.199	$3.24 \times 10^{-4}$	$1.23 \times 10^{-4}$
50–55	.178	830	.209	$2.52 \times 10^{-4}$	$1.59 \times 10^{-4}$
Total				$1.06 \times 10^{-3}$	$2.32 \times 10^{-4}$
				$(B_1)$	$(B_2)$

<sup>a</sup> $\times 10^{-5}$ .

<sup>b</sup>e.g.,  $B_{11} = .771 \times 160 \times 10^{-5} \times .188 \times (1 - .771) \times (1 - 160 \times 10^{-5}) = 5.30 \times 10^{-5}$   
 $B_{21} = 5.30 \times 10^{-5} \times [1 - 2(.771)] \times [1 - 2(160 \times 10^{-5})] = -2.86 \times 10^{-5}$

Alternatively, before a study begins we may want to specify the power and compute the size of the total study population needed to achieve that level of power, given that we know the distribution of the study population by stratum and the

overall exposure and disease rates within each stratum. The appropriate sample-size formula is given by

**Equation 13.23****Sample-Size Estimation for a Collection of  $2 \times 2$  Tables  
Based on the Mantel-Haenszel Test**

$N$  = total number of subjects in the entire study needed for a stratified design using the Mantel-Haenszel test as the method of analysis

$$= \left( z_{1-\alpha/2} \sqrt{B_1} + z_{1-\beta} \sqrt{B_1 + \gamma B_2} \right)^2 \left/ \left( \gamma B_1 + \frac{\gamma^2}{2} B_2 \right) \right.^2$$

where  $\alpha$  = type I error,  $1 - \beta$  = power,  $\gamma = \ln(OR)$  under  $H_1$ , and  $B_1$  and  $B_2$  are defined in Equation 13.22.

**REVIEW QUESTION 13D**

- 1 Consider the study in Review Question 13.C2. How much power did the study have to detect an *OR* of 1.5 for the association between distance to power station and nonmelanoma skin cancer, assuming that the exposure prevalence, disease prevalence, and gender distribution are the same as in Table 13.16? (*Hint:* Use Equation 13.22.)

## 13.8 Multiple Logistic Regression

### Introduction

In Section 13.6, we learned about the Mantel-Haenszel test and the Mantel Extension test, which are techniques for controlling for a single categorical covariate  $C$  while assessing the association between a dichotomous disease variable  $D$  and a categorical exposure variable  $E$ . If

- (1)  $E$  is continuous
- (2) or  $C$  is continuous
- (3) or there are several confounding variables  $C_1, C_2, \dots$ , each of which may be either categorical or continuous

then it is either difficult or impossible to use the preceding methods to control for confounding. In this section, we will learn about the technique of multiple logistic regression, which can handle all the situations in Section 13.6 as well as those in (1), (2), and (3) above. Multiple logistic regression can be thought of as an analog to multiple linear regression, discussed in Chapter 11, where the outcome (or dependent) variable is binary as opposed to normally distributed.

### General Model

**Example 13.34**

**Infectious Disease** *Chlamydia trachomatis* is a microorganism that has been established as an important cause of nongonococcal urethritis, pelvic inflammatory disease, and other infectious diseases. A study of risk factors for *C. trachomatis* was conducted in a population of 431 female college students [10]. Because multiple risk

factors may be involved, several risk factors must be controlled for simultaneously in analyzing variables associated with *C. trachomatis*.

A model of the following form might be considered.

**Equation 13.24**

$$p = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

where  $p$  = probability of disease. However, because the right-hand side of Equation 13.24 could be less than 0 or greater than 1 for certain values of  $x_1, \dots, x_k$ , predicted probabilities that are either less than 0 or greater than 1 could be obtained, which is impossible. Instead, the logit (logistic) transformation of  $p$  is often used as the dependent variable.

**Definition 13.17**

The **logit transformation**  $\text{logit}(p)$  is defined as

$$\text{logit}(p) = \ln[p / (1 - p)]$$

Unlike  $p$ , the logit transformation can take on any value from  $-\infty$  to  $+\infty$ .

**Example 13.35**

Compute  $\text{logit}(.1)$ ,  $\text{logit}(.95)$ .

**Solution**

$$\text{logit}(.1) = \ln(.1 / .9) = \ln(1 / 9) = -\ln(9) = -2.20$$

$$\text{logit}(.95) = \ln(.95 / .05) = \ln(19) = 2.94$$

If  $\text{logit}(p)$  is modeled as a linear function of the independent variables  $x_1, \dots, x_k$ , then the following multiple logistic-regression model is obtained.

**Equation 13.25****Multiple Logistic-Regression Model**

If  $x_1, \dots, x_k$  are a collection of independent variables and  $y$  is a binomial-outcome variable with probability of success =  $p$ , then the multiple logistic-regression model is given by

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

or, equivalently, if we solve for  $p$ , then the model can be expressed in the form

$$p = \frac{e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

In the second form of the model, we see that  $p$  must always lie between 0 and 1 regardless of the values of  $x_1, \dots, x_k$ . Complex numeric algorithms are generally required to fit the parameters of the model in Equation 13.25. The best-fitting model relating the prevalence of *C. trachomatis* to the risk factors (1) race and (2) the lifetime number of sexual partners is presented in Table 13.19.

**Interpretation of Regression Parameters**

How can the regression coefficients in Table 13.19 be interpreted? The regression coefficients in Table 13.19 play a role similar to that played by partial-regression

coefficients in multiple linear regression (See Definition 11.17). Specifically, suppose we consider two individuals with different values of the independent variables as shown in Table 13.20, where the  $j$ th independent variable is a binary variable.

If we refer to the independent variables as exposure variables, then individuals A and B are the same on all risk factors in the model except for the  $j$ th exposure variable, where individual A is exposed (coded as 1) and individual B is not exposed (coded as 0). According to Equation 13.25, the logit of the probability of success for individuals A and B, denoted by  $\text{logit}(p_A)$ , and  $\text{logit}(p_B)$ , are given by

**Equation 13.26**

$$\begin{aligned}\text{logit}(p_A) &= \alpha + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_j(1) + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k \\ \text{logit}(p_B) &= \alpha + \beta_1 x_1 + \cdots + \beta_{j-1} x_{j-1} + \beta_j(0) + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k\end{aligned}$$

**Table 13.19** Multiple logistic-regression model relating prevalence of *C. trachomatis* to race and number of lifetime sexual partners

Risk factor	Regression coefficient $(\hat{\beta}_j)$	Standard error $se(\hat{\beta}_j)$	$[z]$ $[\hat{\beta}_j / se(\hat{\beta}_j)]$
Constant	-1.637		
Black race	+2.242	0.529	+4.24
Lifetime number of sexual partners among users of nonbarrier <sup>a</sup> methods of contraception <sup>b</sup>	+0.102	0.040	+2.55

<sup>a</sup>Barrier methods of contraception include diaphragm, diaphragm and foam, and condom; nonbarrier methods include all other forms of contraception or no contraception.

<sup>b</sup>This variable is defined as 0 for users of barrier methods of contraception.

Source: From McCormack, et al., "Infection with Chlamydia Trachomatis in Female College Students," *American Journal of Epidemiology*, 1985 121: 107-115. Reprinted by permission of Oxford University Press.

**Table 13.20** Two hypothetical subjects with different values for a binary independent variable ( $x$ ) and the same values for all other variables in a multiple logistic-regression model

Individual	Independent variable					
	1	2	...	$j - 1$	$j$	$j + 1 \dots k$
A	$x_1$	$x_2$	...	$x_{j-1}$	1	$x_{j+1} \dots x_k$
B	$x_1$	$x_2$	...	$x_{j-1}$	0	$x_{j+1} \dots x_k$

If we subtract  $\text{logit}(p_B)$  from  $\text{logit}(p_A)$  in Equation 13.26, we obtain

**Equation 13.27**

$$\text{logit}(p_A) - \text{logit}(p_B) = \beta_j$$

However, from Definition 13.17,  $\text{logit}(p_A) = \ln[p_A/(1 - p_A)]$ ,  $\text{logit}(p_B) = \ln[p_B/(1 - p_B)]$ . Therefore, on substituting into Equation 13.27, we obtain

$$\ln[p_A / (1 - p_A)] - \ln[p_B / (1 - p_B)] = \beta_j$$

or

**Equation 13.28**

$$\ln \left[ \frac{p_A / (1 - p_A)}{p_B / (1 - p_B)} \right] = \beta_j$$

If we take the antilog of each side of Equation 13.28, then we have

**Equation 13.29**

$$\frac{p_A / (1 - p_A)}{p_B / (1 - p_B)} = e^{\beta_j}$$

However, from the definition of an *OR* (Definition 13.6), we know the odds in favor of success for subject A (denoted by  $\text{Odds}_A$ ) is given by  $\text{Odds}_A = p_A / (1 - p_A)$ . Similarly,  $\text{Odds}_B = p_B / (1 - p_B)$ . Therefore, we can rewrite Equation 13.29 as follows.

**Equation 13.30**

$$\frac{\text{Odds}_A}{\text{Odds}_B} = e^{\beta_j}$$

Thus, in words, the odds in favor of disease for subject A divided by the odds in favor of disease for subject B =  $e^{\beta_j}$ . However, we can also think of  $\text{Odds}_A / \text{Odds}_B$  as the *OR* relating disease to the  $j$ th exposure variable for two hypothetical individuals, one of whom is exposed for the  $j$ th exposure variable (subject A) and the other of whom is not exposed for the  $j$ th exposure variable (subject B), where the individuals are the same for all other risk factors considered in the model. Thus this *OR* is an *OR* relating disease to the  $j$ th exposure variable, adjusted for the levels of all other risk factors in our model. This is summarized as follows.

**Equation 13.31**

#### **Estimation of ORs in Multiple Logistic Regression for Dichotomous Independent Variables**

Suppose there is a dichotomous exposure variable ( $x_j$ ), which is coded as 1 if present and 0 if absent. For the multiple logistic-regression model in Equation 13.25, the *OR* relating this exposure variable to the dependent variable is estimated by

$$\hat{OR} = e^{\hat{\beta}_j}$$

This relationship expresses the odds in favor of success if  $x_j = 1$  divided by the odds in favor of success if  $x_j = 0$  (i.e., the disease-exposure *OR*) *after controlling for all other variables in the logistic-regression model*. Furthermore, a two-sided 100%  $\times$  (1 -  $\alpha$ ) CI for the true *OR* is given by

$$\left[ e^{\hat{\beta}_j - z_{1-\alpha/2} \text{se}(\hat{\beta}_j)}, e^{\hat{\beta}_j + z_{1-\alpha/2} \text{se}(\hat{\beta}_j)} \right]$$

**Example 13.36**

**Infectious Disease** Estimate the odds in favor of infection with *C. trachomatis* for black women compared with white women after controlling for previous sexual experience, and provide a 95% CI about this estimate.

**Solution**

From Table 13.19,

$$\hat{OR} = e^{2.242} = 9.4$$

Thus the odds in favor of infection for black women are nine times as great as those for white women after controlling for previous sexual experience. Furthermore, because  $z_{1-\alpha/2} = z_{.975} = 1.96$  and  $se(\hat{\beta}_j) = 0.529$ , a 95% CI for  $OR$  is given by

$$[e^{2.242-1.96(0.529)}, e^{2.242+1.96(0.529)}] = (e^{1.205}, e^{3.279}) = (3.3, 26.5)$$

We can also use Equation 13.31 to make a connection between logistic regression and contingency-table analysis for  $2 \times 2$  tables given in Chapter 10. Specifically, suppose there is only one risk factor in the model, which we denote by  $E$  and which takes the value 1 if exposed and 0 if unexposed, and we have a dichotomous disease variable  $D$ . We can relate  $D$  to  $E$  using the logistic-regression model.

**Equation 13.32**

$$\log[p / (1 - p)] = \alpha + \beta E$$

where  $p$  = probability of disease given a specific exposure status  $E$ . Therefore, the probability of disease among the unexposed =  $e^\alpha / (1 + e^\alpha)$  and among the exposed =  $e^{\alpha+\beta} / (1 + e^{\alpha+\beta})$ . Also, from Equation 13.31,  $e^\beta$  represents the  $OR$  relating  $D$  to  $E$  and is the same  $OR$  [ $ad/(bc)$ ] obtained from the  $2 \times 2$  table in Table 10.7 relating  $D$  to  $E$ . We have formulated the models in Equations 13.25 and 13.32 under the assumption that we have conducted either a prospective study or a cross-sectional study (i.e., that our study population is representative of the general population and we have not oversampled cases in our study population, as would be true in a case-control study). However, logistic regression is applicable to data from case-control studies as well. Suppose we have a case-control study in which there is a disease variable  $D$  and an exposure variable  $E$  and no other covariates. If we use the logistic-regression model in Equation 13.32—that is,  $D$  as the outcome variable and  $E$  as the independent (or predictor) variable—then the probability of disease among the unexposed [ $e^\alpha / (1 + e^\alpha)$ ] and the exposed [ $e^{\alpha+\beta} / (1 + e^{\alpha+\beta})$ ] will *not* be generalizable to the reference population because they are derived from a selected sample with a greater proportion of cases than in the reference population. However, the  $OR$   $e^\beta$  will be generalizable to the reference population. Thus we can estimate  $ORs$  from case-control studies, but we cannot estimate probabilities of disease. This statement is also true if there are multiple exposure variables in a logistic-regression model derived from data from a case-control study. The relationships between logistic-regression analysis and contingency-table analysis are summarized as follows.

**Equation 13.33****Relationship Between Logistic-Regression Analysis and Contingency-Table Analysis**

Suppose we have a dichotomous disease variable  $D$  and a single dichotomous exposure variable  $E$ , derived from either a prospective, cross-sectional, or case-control study design, and that the  $2 \times 2$  table relating disease to exposure is given by

		$E$	
		+	-
$D$	+	$a$	$b$
	-	$c$	$d$

(1) We can estimate the *OR* relating *D* to *E* in either of two equivalent ways:

- (a) We can compute the *OR* directly from the  $2 \times 2$  table =  $ad/bc$
- (b) We can set up a logistic-regression model of the form

$$\ln[p/(1-p)] = \alpha + \beta E$$

where  $p$  = probability of disease *D* given exposure status *E* and where we estimate the *OR* by  $e^\beta$ .

(2) For prospective or cross-sectional studies, we can estimate the probability of disease among exposed ( $p_E$ ) subjects and unexposed ( $p_{\bar{E}}$ ) subjects in either of two equivalent ways:

- (a) From the  $2 \times 2$  table, we have

$$p_E = a / (a + c), p_{\bar{E}} = b / (b + d)$$

- (b) From the logistic-regression model,

$$p_E = e^{\hat{\alpha} + \hat{\beta}} / (1 + e^{\hat{\alpha} + \hat{\beta}}), p_{\bar{E}} = e^{\hat{\alpha}} / (1 + e^{\hat{\alpha}})$$

where  $\hat{\alpha}$ ,  $\hat{\beta}$  are the estimated parameters from the logistic-regression model.

(3) For case-control studies, it is impossible to estimate absolute probabilities of disease unless the sampling fraction of cases and controls from the reference population is known, which is almost always *not* the case.

### **Example 13.37**

Assess the relationship between mother's age at first birth and breast-cancer incidence based on the data in Table 10.1 using logistic-regression analysis.

#### **Solution**

We will use the logistic-regression model

$$\ln[p / (1 - p)] = \alpha + \beta \times \text{AGEGE30}$$

where

$p$  = probability of breast cancer

$\text{AGEGE30} = 1$  if age at first birth  $\geq 30$

= 0 otherwise

The results using the SAS PROC LOGISTIC program are given in Table 13.21. We see that the estimated *OR* relating breast-cancer incidence to AGEGE30 =  $e^{0.4526} = 1.57$ . This is identical to the *OR* estimated using contingency-table methods given in Example 13.10. Notice that although there are actually 13,465 subjects in the study, PROC LOGISTIC tells us there are two observations. The reason is that the program allows us to enter data in grouped form with all observations with the same combination of independent variables entered as one record. In this case, there is only one covariate (AGEGE30), which only has two possible values (0 or 1); thus there are two "observations." For each level of AGEGE30, we need to provide the number of cases (i.e., successes) and the number of trials (i.e., observations). Entering the data in grouped form usually reduces the computer time needed to fit a logistic-regression model (in some data sets dramatically, if the number of covariate patterns is small relative to the number of subjects).

**Table 13.21** Association between age at first birth and breast-cancer incidence based on the data in Table 10.1 using the SAS PROC LOGISTIC procedure

Case/Trials Model (recommended instead of freq) Logistic Regression			
The LOGISTIC Procedure			
Model Information			
Data Set			WORK.AFB1
Response Variable (Events)			cases
Response Variable (Trials)			trials
Model			binary logit
Optimization Technique			Fisher's scoring
Number of Observations Read 2			
Number of Observations Used 2			
Sum of Frequencies Read 13465			
Sum of Frequencies Used 13465			
Response Profile			
Ordered Value	Binary Outcome	Total Frequency	
1	Event	3220	
2	Nonevent	10245	
Model Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied.			
Model Fit Statistics			
Criterion		Intercept Only	Intercept and Covariates
AIC		14815.785	14743.181
SC		14823.293	14758.197
-2 Log L		14813.785	14739.181
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	74.6042	1	<.0001
Score	78.3698	1	<.0001
Wald	77.5782	1	<.0001
Analysis of Maximum Likelihood Estimates			
Parameter	DF	Estimate	Standard Error
INTERCEPT	1	-1.2377	0.0225
AGEGE30	1	0.4526	0.0514
Wald Chi-Square			
		3012.7767	<.0001
		77.5782	<.0001
Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	Wald
AGEGE30	1.572	1.422 1.739	Chi-Square
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	18.1	Somers' D	0.066
Percent Discordant	11.5	Gamma	0.222
Percent Tied	70.4	Tau-a	0.024
Pairs	32988900	c	0.533

We are also interested in expressing the strength of association between a continuous independent variable and the dependent variable in terms of an *OR* after controlling for the other independent variables in the model.

Suppose we have two individuals A and B who are the same for all independent variables in the model except for a single continuous risk factor  $x_j$ , where they differ by an amount  $\Delta$  (see Table 13.22).

**Table 13.22** Two hypothetical subjects with different values for a continuous independent variable ( $x_j$ ) in a multiple logistic-regression model and the same values for all other variables

Individual	Independent variable			
	1	$2 \dots j-1$	$j$	$j+1 \dots k$
A	$x_1$	$x_2 \dots x_{j-1}$	$x_j + \Delta$	$x_{j+1} \dots x_k$
B	$x_1$	$x_2 \dots x_{j-1}$	$x_j$	$x_{j+1} \dots x_k$

Following the same argument as in Equation 13.26, we have

**Equation 13.34**

$$\begin{aligned}\text{logit}(p_A) &= \alpha + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_j (x_j + \Delta) + \beta_{j+1} x_{j+1} + \dots + \beta_k x_k \\ \text{logit}(p_B) &= \alpha + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \dots + \beta_k x_k\end{aligned}$$

If we subtract  $\text{logit}(p_B)$  from  $\text{logit}(p_A)$  in Equation 13.34, we obtain

$$\text{logit}(p_A) - \text{logit}(p_B) = \beta_j \Delta$$

$$\text{or } \ln\left(\frac{p_A}{1-p_A}\right) - \ln\left(\frac{p_B}{1-p_B}\right) = \beta_j \Delta$$

$$\text{or } \ln\left[\frac{p_A / (1-p_A)}{p_B / (1-p_B)}\right] = \beta_j \Delta$$

$$\text{or } OR = \frac{p_A / (1-p_A)}{p_B / (1-p_B)} = e^{\beta_j \Delta}$$

Thus the odds in favor of disease for subject A vs. subject B =  $e^{\beta_j \Delta}$ . This result is summarized as follows.

**Equation 13.35**

#### Estimation of ORs in Multiple Logistic Regression for Continuous Independent Variables

Suppose there is a continuous independent variable ( $x_j$ ). Consider two individuals who have values of  $x + \Delta$  and  $x$  for  $x_j$ , respectively, and have the same values for all other independent variables in the model. The *OR* in favor of success for the first individual vs. the second individual is estimated by

$$(\hat{OR}) = e^{\hat{\beta}_j \Delta}$$

Furthermore, a two-sided 100%  $\times (1 - \alpha)$  CI for *OR* is given by

$$\left\{ e^{[\hat{\beta}_j - z_{1-\alpha/2} \hat{\sigma}(\hat{\beta}_j)]\Delta}, e^{[\hat{\beta}_j + z_{1-\alpha/2} \hat{\sigma}(\hat{\beta}_j)]\Delta} \right\}$$

Thus  $OR$  represents the odds in favor of success for an individual with level  $x + \Delta$  for  $x_j$  vs. an individual with level  $x$  for  $x_j$ , after controlling for all other variables in the model. The symbol  $\Delta$  is usually chosen to represent a meaningful increment in the continuous variable  $x$ .

### Example 13.38

**Infectious Disease** Given the data in Table 13.19, what is the extra risk of infection for each additional sexual partner for women of a particular race who use nonbarrier methods of contraception? Provide a 95% CI associated with this estimate.

### Solution

We have that  $\Delta = 1$ . From Table 13.19,  $\hat{\beta}_j = 0.102$ ,  $se(\hat{\beta}_j) = 0.040$ . Thus

$$\hat{OR} = e^{0.102 \times 1} = e^{0.102} = 1.11$$

Thus the odds in favor of infection increase an estimated 11% for each additional sexual partner for women of a particular race who use nonbarrier methods of contraception. A 95% CI for  $OR$  is given by

$$\left\{ e^{[0.102 - 1.96(0.040)]}, e^{[0.102 + 1.96(0.040)]} \right\} = (e^{0.0236}, e^{0.1804}) = (1.02, 1.20)$$

## Hypothesis Testing

How can the statistical significance of the risk factors in Table 13.19 be evaluated? The statistical significance of each of the independent variables after controlling for all other independent variables in the model should be assessed. This task can be accomplished by first computing the test statistic  $z = \hat{\beta}_j / se(\hat{\beta}_j)$ , which should follow an  $N(0, 1)$  distribution under the null hypothesis that the  $j$ th independent variable has no association with the dependent variable after controlling for the other variables.  $H_0$  will be rejected for either large positive or large negative values of  $z$ . This procedure is summarized as follows.

### Equation 13.36

#### Hypothesis Testing in Multiple Logistic Regression

To test the hypothesis  $H_0: \beta_j = 0$ , all other  $\beta_i \neq 0$ , vs.  $H_1: \text{all } \beta_j \neq 0$  for the multiple logistic-regression model in Equation 13.25, use the following procedure:

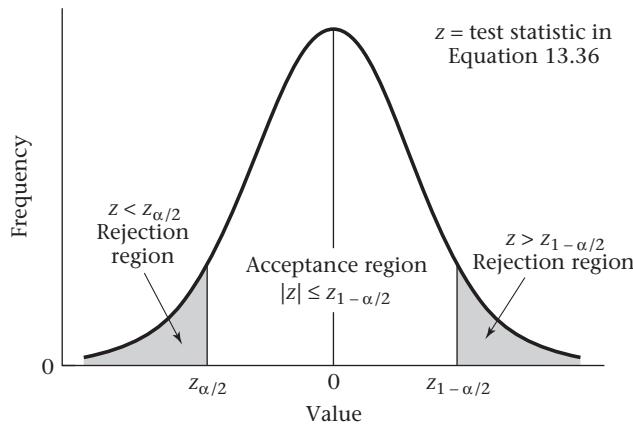
- (1) Compute the test statistic  $z = \hat{\beta}_j / se(\hat{\beta}_j) \sim N(0, 1)$  under  $H_0$ .
- (2) To conduct a two-sided test with significance level  $\alpha$ ,
 

if $z < z_{\alpha/2}$	or	$z > z_{1-\alpha/2}$	then reject $H_0$
if $z_{\alpha/2} \leq z \leq z_{1-\alpha/2}$			then accept $H_0$
- (3) The exact  $p$ -value is given by
 
$$2 \times [1 - \Phi(z)] \quad \text{if } z \geq 0$$

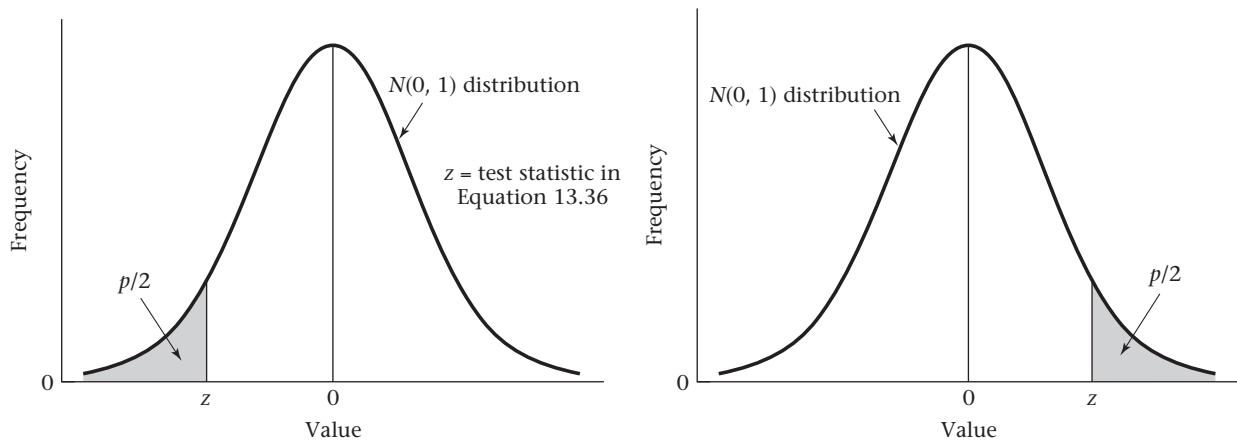
$$2 \times \Phi(z) \quad \text{if } z < 0$$
- (4) This large-sample procedure should only be used if there are at least 20 successes and 20 failures, respectively, in the data set.

The acceptance and rejection regions for this test are shown in Figure 13.3. Computation of the exact  $p$ -value is illustrated in Figure 13.4.

**Figure 13.3** Acceptance and rejection regions for the test of the hypothesis  $H_0: \beta_j = 0$ , all other  $\beta_j \neq 0$ , vs.  $H_1: \text{all } \beta_j \neq 0$  in multiple logistic regression



**Figure 13.4** Computation of the  $p$ -value for the test of the hypothesis  $H_0: \beta_j = 0$ , all other  $\beta_j \neq 0$ , vs.  $H_1: \text{all } \beta_j \neq 0$  in multiple logistic regression



(a) If  $z = \hat{\beta}_j/se(\hat{\beta}_j) < 0$ , then  $p = 2 \times (\text{area to the left of } z \text{ under an } N(0, 1) \text{ distribution})$ .

(b) If  $z = \hat{\beta}_j/se(\hat{\beta}_j) \geq 0$ , then  $p = 2 \times (\text{area to the right of } z \text{ under an } N(0, 1) \text{ distribution})$ .

### Example 13.39

**Infectious Disease** Assess the significance of the independent variables in the multiple logistic-regression model presented in Table 13.19.

### Solution

First compute the test statistic  $z = \hat{\beta}_j/se(\hat{\beta}_j)$  for each of the independent variables, as shown in Table 13.19. For an  $\alpha$  level of .05, compare  $|z|$  with  $z_{.975} = 1.96$  to assess statistical significance. Because both independent variables satisfy this criterion, they are both significant at the 5% level. The exact  $p$ -values are given by

$$\begin{aligned} p(\text{race}) &= 2 \times [1 - \Phi(4.24)] < .001 \\ p(\text{number of sexual partners}) &= 2 \times [1 - \Phi(2.55)] = .011 \end{aligned}$$

Thus both variables are significantly associated with *C. trachomatis*. Specifically, after controlling for the other variable in the model, there is an increased probability of infection for black women vs. white women and for women with more previous sexual experience vs. women with less previous sexual experience.

We also can quantify the magnitude of the association. From Example 13.36, the odds in favor of *C. trachomatis* are 9.4 times as high for black women than for white women where both women either used barrier methods of contraception (in which case  $x_2 = 0$ ) or used nonbarrier methods of contraception and had the same lifetime number of sexual partners ( $x_2 > 0$ ). From Example 13.38, for women of the same race who used nonbarrier methods of contraception, the odds in favor of *C. trachomatis* increase by a factor of 1.11 for each added sexual partner during her lifetime.

Finally, the 95% CIs in Equations 13.31 and 13.35 will contain 1 only if there is a nonsignificant association between  $x_j$  and the dependent variable. Similarly, these intervals will not contain 1 only if there is a significant association between  $x_j$  and the dependent variable. Thus, because both independent variables in Table 13.19 are statistically significant, the CIs in Examples 13.36 and 13.38 both exclude 1.

#### Example 13.40

**Cardiovascular Disease** The Framingham Heart Study began in 1950 by enrolling 2282 men and 2845 women ages 30–59 years, who have been followed up to the present [11]. Coronary risk-factor information about cohort members has been obtained at biannual examinations. The relationship between the incidence of CHD and selected risk factors, including age, sex, serum cholesterol, serum glucose, BMI, systolic blood pressure (SBP), and cigarette smoking, was assessed [12]. For the analysis, men were chosen who were free of CHD (either nonfatal MI or fatal CHD) at examination 4 and for whom all risk-factor information from examination 4 was available. In this analysis, the cohort was assessed for the development of CHD over the next 10 years (examinations 5–9). There were 1731 men who satisfied these criteria and constituted the study population, of whom 163 developed CHD over the 10-year study period. The baseline characteristics of the study population are presented in Table 13.23. Subjects were 50 years of age on average and showed a wide range for each of the CHD risk factors.

In Equation 13.31, we studied how to assess logistic-regression coefficients for dichotomous exposure variables. In Equation 13.35, we studied how to assess logistic-regression coefficients for continuous exposure variables. In some instances, we wish to use categorical variables in a multiple logistic-regression model that have more than two categories. In this case, we can represent such variables by a collection of  $k - 1$  dummy variables in a similar manner to that used for multiple regression in Equation 12.18.

In the analysis of the data in this example, age was treated as a categorical variable (with categories 35–44, 45–54, 55–64, and 65–69), whereas all other risk factors were treated as continuous variables. The reason for treating age as a categorical variable is that in other studies, the increase in incidence of CHD with age was reported to be nonlinear with age; for example, the *OR* relating incidence for 50- to 54-year-olds vs. incidence for 45- to 49-year-olds is different from that for 60- to 64-year-olds vs. 55- to 59-year-olds. If age is entered as a continuous variable, then the increase in risk for every 5-year increase in age (as measured by the *OR*) is assumed to be the same. Therefore, we choose one category (35–44) to be the reference

**Table 13.23** Baseline characteristics of study population, Framingham Heart Study

Risk factor	Mean	Standard deviation	No.	%
Serum cholesterol (mg/dL) <sup>a</sup>	234.8	40.6		
Serum glucose (mg/dL) <sup>b</sup>	81.8	27.4		
BMI (kg/m <sup>2</sup> )	26.5	3.4		
SBP (mm Hg) <sup>c</sup>	132.1	20.1		
Age (years)	49.6	8.5		
35–39			228	13
40–49			670	39
50–59			542	31
60–69			291	17
Current smoking (cigarettes/day) <sup>d</sup>	13.1	13.5		
0			697	40
1–10			183	11
11–20			510	29
≥21			341	20

Note: The subjects were 1731 men who were seen at examination 4 and were free of CHD at or before examination 4.

<sup>a</sup>Based on the Abell-Kendall method.

<sup>b</sup>Based on a casual specimen of the subject's whole blood, using the Nelson method.

<sup>c</sup>Average of two replicate measurements at examination 4, using a standard mercury sphygmomanometer.

<sup>d</sup>A current smoker is defined as a person who smoked within the past year.

category and create three dummy variables to represent group membership in age groups 45–54, 55–64, and 65–69, respectively. Which category is assigned to be the reference category is determined arbitrarily. In some instances, a particular category is a natural reference category based on scientific considerations. Also, all other risk factors except number of cigarettes currently smoked were converted to the ln scale to reduce the positive skewness. The resulting model is

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{AGE4554} + \beta_2 \text{AGE5564} + \beta_3 \text{AGE6569} + \beta_4 \text{LCHLD235} \\ + \beta_5 \text{LSUGRD82} + \beta_6 \text{SMOKEM13} + \beta_7 \text{LBMID26} + \beta_8 \text{LMSYD132}$$

where

$p$  = probability of developing CHD over a 10-year period

AGE4554 = 1 if a subject is age 45–54, = 0 otherwise

AGE5564 = 1 if a subject is age 55–64, = 0 otherwise

AGE6569 = 1 if a subject is age 65–69, = 0 otherwise

LCHLD235 = ln(serum cholesterol/235)

LSUGRD82 = ln(serum glucose/82)

SMOKEM13 = number of cigarettes currently smoked – 13

LBMID26 = ln(BMI/26)

LMSYD132 = ln(SBP/132)

Each of the risk factors (except for the age variables) have been **mean-centered**; that is, the approximate mean has either been subtracted from each value (for cigarettes per day, which is in the original scale) or each value has been divided by the approximate mean (for all other risk factors, which are in the ln scale). The reason for doing this is to make the constant ( $\alpha$ ) more meaningful. In this analysis, the constant  $\alpha$  represents logit( $p$ ) for an “average” subject in the reference group (35–44 years of age); that is, where all other risk factors are 0, which means that serum cholesterol = 235, serum glucose = 82, number of cigarettes per day = 13, BMI = 26, SBP = 132. The model was fitted using SAS PROC LOGISTIC, and the results are given in Table 13.24.

**Table 13.24** Multiple logistic-regression model for predicting the cumulative incidence of CHD over 10 years based on 1731 men in the Framingham Heart Study who were disease free at baseline using SAS PROC LOGISTIC

Logistic Regression			
The LOGISTIC Procedure			
Model Information			
Data Set			WORK.FRAME
Response Variable			cmbmichd
Number of Response Levels			2
Model			binary logit
Optimization Technique			Fisher's scoring
Number of Observations Read			1731
Number of Observations Used			1731
Response Profile			
Ordered Value	cmbmichd	Frequency	Total
1	1	163	163
2	1	1568	1568
Probability modeled is cmbmichd=1			
Model Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied			
Model Fit Statistics			
Criterion	Intercept		Intercept
	Only	Covariates	and
AIC	1082.387	994.458	
SC	1087.843	1043.566	
-2 Log L	1080.387	976.458	
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	103.9293	8	<.0001
Score	104.3489	8	<.0001
Wald	91.4205	8	<.0001

(continued)

**Table 13.24** Multiple logistic-regression model for predicting the cumulative incidence of CHD over 10 years based on 1731 men in the Framingham Heart Study who were disease free at baseline using SAS PROC LOGISTIC (Continued)

Analysis of Maximum Likelihood Estimates					
PARAMETER	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
INTERCEPT	1	-3.1232	0.2090	223.2970	<.0001
AGEGE4554	1	0.7199	0.2474	8.4684	0.0036
AGEGE5564	1	1.1661	0.2551	20.9034	<.0001
AGEGE6569	1	1.4582	0.3762	15.0242	0.0001
LCHLD235	1	1.8303	0.5085	12.9537	0.0003
LSUGRD82	1	0.5728	0.3262	3.0847	0.0790
SMOKEML3	1	0.0177	0.00637	7.6909	0.0055
LBMID26	1	1.4818	0.7012	4.4662	0.0346
LMSYD132	1	2.7968	0.5737	23.7665	<.0001
Odds Ratio Estimates					
Effect		Point Estimate	95% Wald Confidence Limits		
AGE4554		2.054	1.265	3.336	
AGE5564		3.210	1.947	5.291	
AGE6569		4.298	2.056	8.985	
LCHLD235		6.236	2.302	16.896	
LSUGRD82		1.773	0.936	3.360	
SMOKEM13		1.018	1.005	1.031	
LBMID26		4.401	1.114	17.394	
LMSYD132		16.392	5.325	50.461	
Association of Predicted Probabilities and Observed Responses					
Percent Concordant		72.6	Somers' D	0.459	
Percent Discordant		26.7	Gamma	0.462	
Percent Tied		0.7	Tau-a	0.078	
Pairs		255584	c	0.729	

We see that each of the risk factors is significantly related to the incidence of CHD, with the exception of serum glucose. The OR gives us information as to the magnitude of the associations. The OR for each of the age variables gives an estimate of the odds in favor of CHD for that age group as compared with the reference group (ages 35–44). The odds in favor of CHD for 45- to 54-year-olds are 2.1 ( $e^{0.7199}$ ) times as great as for the reference group, holding all other variables constant. Similarly, the odds in favor of CHD are 3.2 and 4.3 times as great for 55-to 64-year-olds and 65- to 69-year-olds vs. the reference group. The OR for cholesterol ( $e^{1.8303} = 6.2$ ) indicates that for two men who are 1 ln unit apart on ln serum cholesterol and are comparable on all other risk factors, the odds in favor of CHD for the man with the higher cholesterol is 6.2 times as great as for the man with the lower cholesterol. Recall that 1 ln unit apart is equivalent to a cholesterol ratio of  $e^1 = 2.7$ . If we want to compare men with a different cholesterol ratio (e.g., twice as great), we convert 2 to the ln scale and compute the OR as  $e^{1.8303 \ln 2} = e^{1.8303 \times 0.6931} = e^{1.27} = 3.6$ . Thus, if man A has a cholesterol level twice as great as man B and is the same on all other risk factors, then the odds in favor of CHD over a 10-year period are 3.6 times as great for man A vs. man B. The other continuous variables that were converted to the ln scale (glucose, BMI, and SBP) are interpreted similarly. Cigarette smoking was left in the original scale, so the OR of 1.018 provides a comparison of two men 1 cigarette

per day apart. Because this is a trivial difference, a more meaningful comparison is obtained if we compare a smoker of 1 pack per day (i.e., 20 cigarettes per day—man A) vs. a nonsmoker (i.e., 0 cigarettes per day—man B). The odds in favor of CHD for man A vs. man B =  $e^{20(0.0177)} = e^{0.354} = 1.42$ . Thus the smoker of 1 pack per day is 1.4 times as likely to develop CHD over a 10-year period as is the nonsmoker, given that they are the same for all other risk factors. Because we mean-centered all risk factors except for age, the intercept allows us to estimate the 10-year cumulative incidence of CHD for an “average” man in the reference group (age 35–44). Specifically, the 10-year cumulative incidence of CHD for a 35- to 44-year-old man with cholesterol = 235, glucose = 82, number of cigarettes per day = 13, BMI = 26, and SBP = 132 is  $e^{-3.1232}/(1 + e^{-3.1232}) = 0.0440/1.0440 = .042$ , or 4.2%.

### Prediction with Multiple Logistic Regression

We can use a multiple logistic-regression model to predict the probability of disease for an individual subject with covariate values  $x_1, \dots, x_k$ . If the regression parameters were known, then the probability of disease would be estimated using Equation 13.25 by

$$p = \frac{e^L}{1 + e^L}$$

where  $L = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$ .  $L$  is sometimes called a **linear predictor**. Because the parameters are unknown, we substitute estimates of them to obtain the predicted probability

**Equation 13.37**

$$\hat{p} = \frac{e^{\hat{L}}}{1 + e^{\hat{L}}}$$

$$\text{where } \hat{L} = \hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

To obtain two-sided  $100\% \times (1 - \alpha)$  confidence limits for the true probability  $p$ , we must first obtain confidence limits for the linear predictor  $L$  given by

$$\hat{L} \pm z_{1-\alpha/2} se(\hat{L}) = (L_1, L_2)$$

and then transform back to the probability scale to obtain the CI =  $(p_1, p_2)$  where

$$p_1 = e^{L_1}/(1 + e^{L_1}), p_2 = e^{L_2}/(1 + e^{L_2})$$

The actual expression for  $se(\hat{L})$  is complex, requiring matrix algebra, and is beyond the scope of this text, although it can be easily evaluated on the computer. This approach is summarized as follows.

**Equation 13.38**

#### Point and Interval Estimation of Predicted Probabilities Using Logistic Regression

Suppose we wish to estimate the predicted probability of disease ( $p$ ) for a subject with covariate values  $x_1, \dots, x_k$  and obtain confidence limits about this prediction.

- (1) We compute the linear predictor

$$\hat{L} = \hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

where  $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$  are the estimated regression coefficients from the logistic-regression model.

- (2) The point estimate of  $p$  is given by  $e^{\hat{L}}/(1 + e^{\hat{L}})$   
 (3) A two-sided 100%  $\times (1 - \alpha)$  CI for  $p$  is given by  $(p_1, p_2)$ , where

$$p_1 = e^{L_1}/(1 + e^{L_1}), p_2 = e^{L_2}/(1 + e^{L_2})$$

and

$$L_1 = \hat{L} - z_{1-\alpha/2} se(\hat{L})$$

$$L_2 = \hat{L} + z_{1-\alpha/2} se(\hat{L})$$

To obtain  $se(\hat{L})$  requires matrix algebra and is always done by computer. Also,  $se(\hat{L})$  will vary, depending on the covariate values  $x_1, \dots, x_k$ .

- (4) These estimates are only valid for prospective or cross-sectional studies.

### Example 13.41

**Cardiovascular Disease** Obtain a point estimate and 95% confidence limits for the predicted probability of CHD for participants in the Framingham Heart Study data set in Example 13.40.

#### Solution

We refer to Table 13.25, which provides the raw data, the predicted probability of CHD (labeled phat), and the lower and upper 95% confidence limits about the point estimates (labeled lcl and ucl) for subjects 945–1002. For example, for subject 969 the outcome variable is given in the tenth column (cmbmichd) and = 2 (which indicates no event). The values of the independent variables are given in columns 2–9. For example, we know that the subject is 35–44 years old because all the age dummy variables (columns 2–4) are 0. Also, the subject is a nonsmoker because SMOKE13 = −13. His serum cholesterol =  $235 \times e^{-0.00855} = 233$ , etc. From Table 13.24, the linear predictor is

$$\begin{aligned}\hat{L} = & -3.1232 + 0.7199(0) + 1.1661(0) + 1.4582(0) + 1.8303(-0.00855) \\ & + \dots + 2.7968(0.04445) = -3.1192\end{aligned}$$

The predicted probability is

$$\hat{p}_i = e^{-3.1192}/(1 + e^{-3.1192}) = .0423 \text{ (see PHAT column)}$$

The lower and upper 95% confidence limits are lcl = .0269 and ucl = .0660.

## Assessing Goodness of Fit of Logistic-Regression Models

We can use the predicted probabilities for individual subjects to define residuals and assess goodness of fit of logistic-regression models.

### Equation 13.39

#### Residuals in Logistic Regression

If our data are in ungrouped form—that is, each subject has a unique set of covariate values (as in Table 13.25)—then we can define the **Pearson residual** for the  $i$ th observation by

$$r_i = \frac{y_i - \hat{p}_i}{se(\hat{p}_i)}$$

**Table 13.25 Raw data, predicted probabilities, and 95% confidence limits for a subset of the Framingham Heart Study data based on the logistic-regression model in Table 13.24**

obs	age4554	age5564	age6569	Predicted Probabilities and 95% Confidence Limits										
				lchild235	1sgtrgd82	smokhem13	1bnid26	cmbmichd	—level—	phat	lcl	ucl		
945	0	0	0	0.08168	-0.07599	7	-0.02242	-0.04652	1	0.04492	0.03035	0.06602		
946	1	0	0	-0.11248	-0.03727	7	-0.07708	-0.19145	1	0.04083	0.02733	0.06059		
947	0	0	0	0.25084	-0.05919	-13	0.31715	-0.08701	2	0.06702	0.03678	0.11905		
948	0	0	0	-0.04246	-0.02459	-13	0.03620	-0.02299	2	0.03275	0.02082	0.05116		
949	0	1	0	0.05559	0.03593	17	0.11339	0.11441	2	0.20764	0.20019	0.35488		
950	0	0	0	-0.11725	-0.02469	30	-0.07708	-0.19145	2	0.03014	0.01740	0.05169		
951	1	0	0	0	0.05787	-0.08923	-4	0.17157	0.03352	1	0.11192	0.08182	0.15126	
952	0	0	1	-0.41616	-0.17261	-10	-0.10609	0.02247	2	0.0750	0.02768	0.11563		
953	1	0	0	0	-0.11248	0.02410	17	-0.05058	-0.16330	2	0.0575	0.03713	0.08292	
954	0	0	0	0	0.02935	0.04763	30	0.01165	-0.01527	2	0.07328	0.04603	0.11471	
955	1	0	0	0	-0.07509	0.09369	-13	-0.04094	-0.17337	2	0.03688	0.02400	0.05628	
956	0	1	0	0	0.28342	0.01212	17	0.29208	0.43532	1	0.62700	0.46706	0.76327	
957	0	0	0	-0.17132	-0.11626	30	-0.09393	-0.09975	2	0.0426	0.03103	0.09320		
958	1	0	0	0	-0.13658	-0.02469	-13	0.27706	0.34294	1	0.17835	0.10977	0.27647	
959	1	1	0	0	0	0.04349	0.25672	17	0.12755	0.01130	1	0.14006	0.09710	0.19785
960	1	1	0	0	-0.22848	-0.10414	7	0.04349	-0.05151	1	0.0224	0.03487	0.07758	
961	0	0	0	0	-0.15719	-0.13005	7	-0.00107	0.00755	2	0.05915	0.03976	0.08712	
962	0	0	0	-0.10300	-0.10265	7	-0.16215	0.25300	2	0.18767	0.13242	0.25908		
963	1	0	0	-0.13171	-0.10265	-13	-0.09975	-0.20331	2	0.02521	0.01556	0.04060		
964	1	0	0	0	0.07380	0.01212	-13	0.02766	0.05884	2	0.09233	0.06695	0.12602	
965	0	0	0	0	0.03158	-0.34628	-13	0.09027	-0.10368	2	0.05611	0.01533	0.04250	
966	0	0	0	0	0.05787	-0.07599	17	-0.12016	-0.16330	2	0.03236	0.02017	0.05152	
967	0	0	0	0	0.18939	-0.12227	-13	0.03847	0.09393	2	0.17835	0.13293	0.23508	
968	0	0	0	-0.21256	0.24724	-13	0.04545	-0.0878	2	0.07084	0.04637	0.10677		
969	0	0	0	-0.09855	0.09369	-13	0.04847	0.04445	2	0.04232	0.02690	0.06598		
970	0	0	0	-0.07051	-0.06291	2	-0.20067	-0.20441	2	0.0083	0.01312	0.03292		
971	0	0	0	0	0.09337	0.56309	7	0.18137	-0.04652	2	0.08569	0.04953	0.14426	
972	1	1	0	0	-0.08331	-0.05001	-13	0.17018	0.08701	2	0.08943	0.06178	0.12778	
973	1	1	0	0	-0.02105	-0.08923	-13	0.21605	0.19337	2	0.14337	0.10469	0.20469	
974	0	0	0	0	0.08487	-0.17261	-13	0.18028	0.10080	2	0.05278	0.03258	0.08442	
975	0	0	0	-0.09829	0.37240	-13	-0.02096	0.03352	2	0.11000	0.07608	0.15647		
976	1	0	0	0	0.59542	-0.09309	7	-0.15609	0.01130	1	0.20513	0.11702	0.33447	
977	0	0	0	0	-0.08855	-0.10265	17	0.13038	-0.05459	2	0.05435	0.03555	0.08230	
978	1	0	0	0	-0.08331	-0.05001	-13	0.17018	0.08701	2	0.08943	0.06178	0.12778	
979	1	0	0	0	-0.34159	0.21825	7	0.13331	0.01504	2	0.19337	0.14337	0.23508	
980	0	0	0	-0.11626	-0.11626	-13	0.02966	-0.06556	2	0.01504	0.07308	0.11739		
981	0	0	0	0	0.15586	-0.20203	-13	-0.13666	-0.15541	2	0.02434	0.01512	0.04342	
982	0	0	0	0	-0.05956	-0.03727	7	-0.05885	-0.00760	2	0.06917	0.04325	0.10885	
983	0	0	0	0	-0.08855	-0.10265	17	0.13038	-0.05459	2	0.05732	0.02454	0.05639	
984	0	0	0	0	-0.09241	0.84030	7	0.10293	0.15415	2	0.0763	0.0478	0.12558	
985	0	0	0	0	-0.34159	-0.1626	7	0.13331	0.01504	2	0.19337	0.14337	0.23508	
986	0	0	0	0	-0.14989	-0.13005	7	-0.10193	0.07303	2	0.07308	0.04645	0.11739	
987	0	0	0	0	-0.04349	-0.06291	7	-0.12436	-0.18322	1	0.07303	0.04342	0.13447	
988	0	0	0	0	-0.22399	-0.21706	7	-0.07469	-0.25783	1	0.03732	0.02454	0.05639	
989	0	0	0	0	-0.13171	0.08192	7	-0.16689	0.01130	2	0.0763	0.0478	0.12558	
990	0	0	0	0	-0.15822	-0.15822	-13	-0.01445	0.06596	2	0.0763	0.0478	0.12558	
991	0	0	0	0	-0.43570	-0.21706	7	-0.10975	-0.10368	2	0.0763	0.0478	0.12558	
992	0	0	0	0	-0.27699	0.07062	17	-0.00224	0.05884	2	0.10733	0.06959	0.16053	
993	0	0	0	0	-0.11248	-0.27952	7	-0.12436	-0.18322	1	0.02168	0.01342	0.03485	
994	0	0	1	0	-0.19290	-0.06291	2	-0.07469	-0.25783	1	0.05595	0.03475	0.08889	
995	0	0	0	0	-0.04794	0.16799	-13	0.09985	-0.26274	2	0.0763	0.0478	0.12558	
996	0	1	0	0	-0.10110	-0.25672	-13	0.17120	0.09393	2	0.03200	0.01912	0.05310	
997	0	0	0	0	0.00000	-0.18721	7	-0.13501	-0.08289	2	0.03245	0.01840	0.05661	
998	1	0	0	0	-0.14253	-0.13005	-13	-0.07303	0.07303	1	0.03772	0.02352	0.05996	
999	0	0	0	0	-0.29480	-0.00000	-13	-0.02247	0.05884	2	0.10733	0.06959	0.16311	
1000	0	0	0	0	-0.41616	0.02410	-13	0.01742	-0.16062	1	0.03662	0.02355	0.05651	
1001	1	0	0	0	-0.43757	-0.11626	17	0.08095	-0.13178	2	0.01653	0.00578	0.01974	
1002	0	0	0	0	-0.11725	0.09309	-13	0.02401	-0.10080	2	0.01602	0.00867	0.03130	

where

$\gamma_i = 1$  if the  $i$ th observation is a success and = 0 if it is a failure

$$\hat{p}_i = \frac{e^{\hat{L}_i}}{1 + e^{\hat{L}_i}}$$

$\hat{L}_i$  = linear predictor for the  $i$ th subject =  $\hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$

$$se(\hat{p}_i) = \sqrt{\hat{p}_i(1 - \hat{p}_i)}$$

If our data are in grouped form—that is, if the subjects with the same covariate values have been grouped together (as in Table 13.21)—then the Pearson residual for the  $i$ th group of observations is defined by

$$r_i = \frac{\gamma_i - \hat{p}_i}{se(\hat{p}_i)}$$

where

$\gamma_i$  = proportion of successes among the  $i$ th group of observations

$$\hat{p}_i = \frac{e^{\hat{L}_i}}{1 + e^{\hat{L}_i}} \text{ as defined for ungrouped data}$$

$$se(\hat{p}_i) = \sqrt{\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i}}$$

$n_i$  = number of observations in the  $i$ th group

Thus the Pearson residual is similar to the Studentized residual in linear regression as defined in Equation 11.14. As was the case in linear regression, the residuals do not have the same standard error. The standard error is computed based on the binomial distribution, where the probability of success is estimated by  $\hat{p}_i$ . Thus, for ungrouped data,  $se(\hat{p}_i) = \sqrt{\hat{p}_i(1 - \hat{p}_i)}$  because each observation constitutes a sample size of 1. For grouped data,  $se(\hat{p}_i) = \sqrt{\hat{p}_i(1 - \hat{p}_i)/n_i}$  where  $n_i$  = number of observations in the  $i$ th group. The standard error decreases as  $\hat{p}_i$  approaches either 0 or 1; for grouped data, the standard error decreases as  $n_i$  increases.

### Example 13.42

**Cardiovascular Disease** Compute the Pearson residual for the 969th observation in the Framingham Heart Study data set for the logistic-regression model fitted in Table 13.24.

#### Solution

From Example 13.41 we have the predicted probability  $\hat{p}_i = .0423$ . The standard error of  $\hat{p}_i$  is  $se(\hat{p}_i) = \sqrt{.0423(.9577)} = .2013$ . Also, from Table 13.25 we saw that the subject did *not* have an event; thus  $\gamma_i = 0$ . Therefore, the Pearson residual is

$$r_i = \frac{0 - .0423}{.2013} = -.2102$$

In Table 13.26, we display the Pearson residuals for a subset of the subjects in the Framingham Heart Study data set (listed under the column labeled Value). Thus, for

**Table 13.26** Display of Pearson residuals for a subset of the Framingham Heart Study data based on the logistic-regression model fitted in Table 13.24

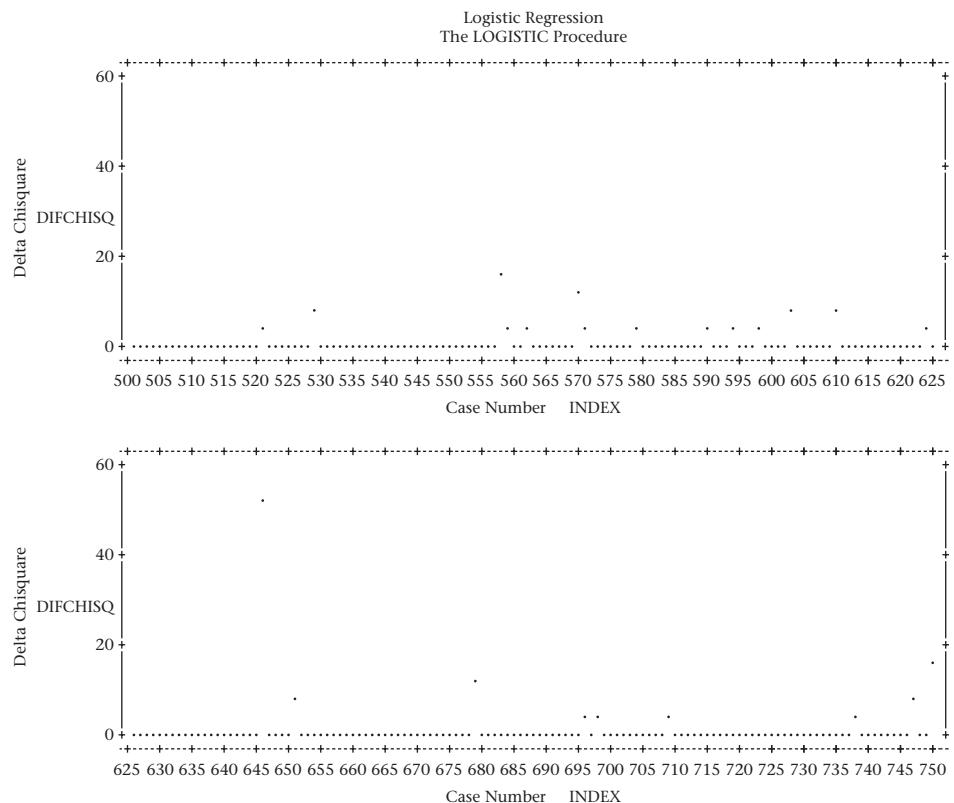
Case Number	Covariates										Pearson Residual (1 unit = 0.9)
	age4554	age5564	age6569	1chld235	1sugr82	smokem13	1bmdd26	1msydi32	Value		
947	0	0	0	0.2508	0.0592	-13.0000	0.3171	-0.0870	-0.2680	*	*
948	0	0	0	-0.00426	-0.0247	-13.0000	0.0362	-0.0230	-0.1840	*	*
949	0	1.0000	0	0.0856	0.0359	17.0000	0.1134	0.1144	-0.6091	*	*
950	0	0	0	-0.1173	-0.0247	30.0000	-0.0771	-0.1915	-0.1763	*	*
951	1.0000	0	0	0.0579	-0.0892	-4.0000	0.1716	0.0335	-0.3550	*	*
952	0	0	1.0000	-0.4162	-0.1726	-10.0000	-0.1061	0.0225	-0.2470	*	*
953	1.0000	0	0	0	-0.1125	0.0241	17.0000	-0.0506	-0.1643	-0.2430	*
954	0	0	0	0.0294	0.0476	30.0000	0.0117	-0.0153	-0.2812	*	*
955	1.0000	0	0	-0.0751	0.0931	-13.0000	-0.0409	-0.1733	-0.1957	*	*
956	0	1.0000	0	0	0.2834	0.0121	17.0000	0.2921	0.4353	0.7713	*
957	0	0	0	-0.1713	-0.1163	30.0000	-0.0997	0.0939	-0.2395	*	*
958	1.0000	0	0	-0.1366	-0.0247	-13.0000	0.2771	0.3429	-0.4659	*	*
959	1.0000	0	0	-0.0435	0.2567	17.0000	0.1276	0.0113	2.4779	*	*
960	1.0000	0	0	-0.2285	0.1041	7.0000	0.00439	-0.0953	4.2594	*	*
961	0	0	0	0	0.1572	-0.1301	7.0000	-0.00107	0.00755	-0.2507	*
962	0	1.0000	0	-0.1030	-0.1027	7.0000	-0.0622	0.2530	-0.4806	*	*
963	1.0000	0	0	-0.1317	-0.1027	-13.0000	-0.0997	-0.2053	-0.1608	*	*
964	1.0000	0	0	0.0738	0.0121	-13.0000	0.0277	0.0588	-0.3189	*	*
965	0	0	0	0.0376	-0.3463	-13.0000	0.0903	-0.1037	-0.1621	*	*
966	0	0	0	0.0579	-0.0760	17.0000	-0.1202	-0.1643	-0.1829	*	*
967	0	1.0000	0	0.1894	-0.0123	-13.0000	0.0385	0.0939	-0.4659	*	*
968	0	1.0000	0	-0.2126	0.2472	-13.0000	0.0545	-0.0788	-0.2761	*	*
969	0	0	0	-0.00855	0.0931	-13.0000	0.0485	0.0445	-0.2102	*	*
970	0	0	0	-0.0705	-0.0629	2.0000	-0.0244	-0.2007	-0.1458	*	*
971	0	0	0	0.0934	0.5631	7.0000	0.1814	-0.0465	-0.3061	*	*
972	1.0000	0	0	-0.0843	-0.0500	-13.0000	0.1702	0.0870	-0.3134	*	*
973	1.0000	0	0	0.0211	-0.0892	-13.0000	0.2161	0.1924	-0.4091	*	*
974	0	0	0	0	0.00847	-0.1726	-13.0000	0.1803	0.1008	-0.2361	*
975	0	1.0000	0	-0.0983	0.3724	-13.0000	-0.0210	0.0335	-0.3516	*	*
976	1.0000	0	0	0.5854	0.0931	7.0000	-0.1561	0.0113	1.9685	*	*
977	0	0	0	-0.00855	-0.1027	17.0000	0.1304	-0.0545	-0.2397	*	*
978	1.0000	0	0	-0.0524	0.8403	7.0000	0.1029	0.1542	-0.5193	*	*
979	1.0000	0	0	-0.3416	0.2183	7.0000	0.1333	0.0150	-0.2808	*	*
980	0	0	0	-0.0705	-0.1163	-13.0000	0.0297	-0.0666	-0.1580	*	*
981	0	1.0000	0	0.1859	-0.2020	-13.0000	-0.1366	-0.1554	-0.2726	*	*
982	0	0	0	-0.0660	-0.0373	7.0000	-0.0588	-0.00760	-0.1969	*	*
983	0	0	0	-0.0479	-0.1582	7.0000	-0.0727	-0.2627	-0.1340	*	*
984	0	0	0	-0.1866	-0.0247	-13.0000	-0.0144	0.0660	-0.1699	*	*
985	0	0	0	-0.1357	-0.2171	7.0000	-0.1098	-0.1037	-0.1123	*	*
986	1.0000	0	0	0.1499	-0.1301	7.0000	-0.1019	0.0730	-0.3630	*	*
987	0	0	0	-0.0435	-0.0629	7.0000	-0.1244	-0.1823	6.7172	*	*
988	1.0000	0	0	0.2240	-0.2171	7.0000	-0.0747	-0.2578	4.1078	*	*
989	0	0	-0.1317	0.0819	7.0000	-0.1669	0.0113	-0.1818	-0.1818	*	*

observation 969, Value =  $-0.2102$ . The Pearson residuals are also displayed in graphic form at the right in Table 13.26.

We can use the Pearson residuals to identify outlying values. However, the utility of individual residuals is more limited for logistic regression than for linear regression, particularly if the data are in ungrouped form. Nevertheless, Pearson residuals with large absolute values are worth further checking to be certain that the values of the dependent and independent variables are correctly entered and possibly to identify patterns in covariate values that consistently lead to outlying values. For ease of observation, the square of the Pearson residuals (referred to as DIFCHISQ) are displayed in Figure 13.5 for a subset of the data. The largest Pearson residual in this data set is for observation 646, which corresponds to a young smoker with no other risk factors who had a predicted probability of CHD of approximately 2% and developed CHD during the 10 years. He had a Pearson residual of 7.1, corresponding to a DIFCHISQ of  $7.1^2 \approx 50$ . If there are several other young smokers with no other risk factors who had events, we may want to modify our model to indicate interaction effects between smoking and age, that is, to allow the effect of smoking to be different for younger vs. older men.

As in linear regression, another aspect of assessing goodness of fit is to determine how influential particular observations are in estimating the regression coefficients. Suppose the  $j$ th regression coefficient when estimated from the full data set is denoted by  $\hat{\beta}_j$  and from the reduced data set obtained by deleting the  $i$ th individual by  $\hat{\beta}_j^{(i)}$ . A measure of influence of the  $i$ th observation on the estimation of  $\hat{\beta}_j$  is given by

**Figure 13.5** **Display of DIFCHISQ (the square of the Pearson residual) for a subset of the Framingham Heart Study data**



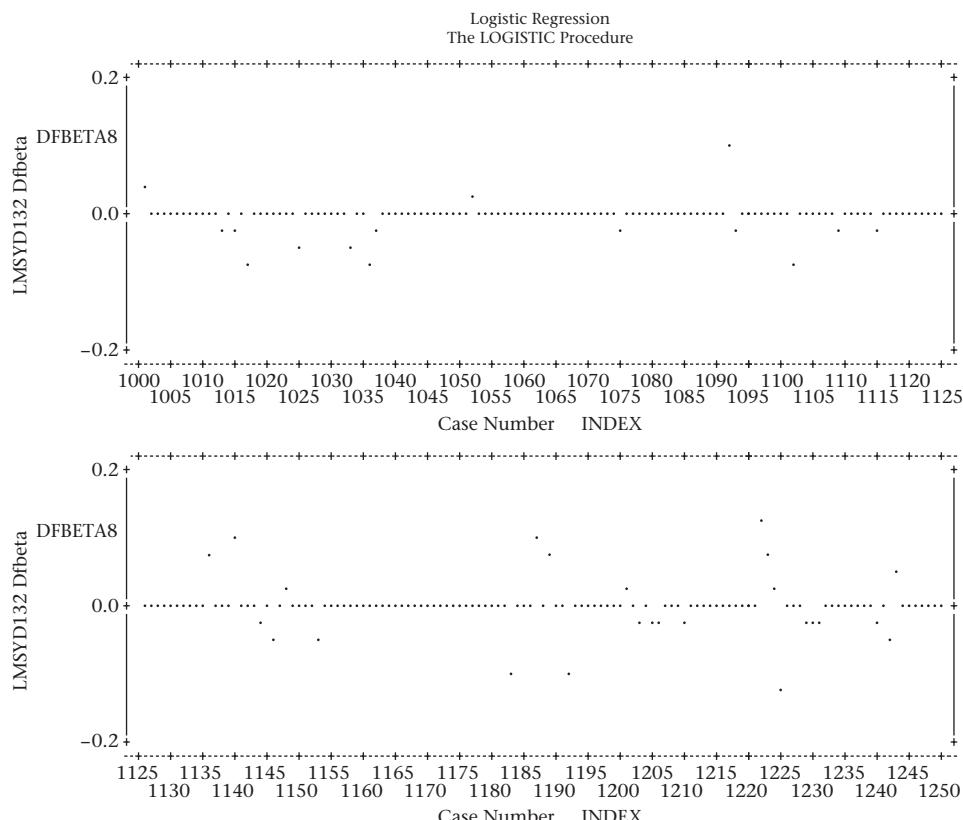
**Equation 13.40**

$$\Delta\beta_j^{(i)} = \frac{\hat{\beta}_j - \hat{\beta}_j^{(i)}}{se(\hat{\beta}_j)}$$

= influence of the  $i$ th observation on estimation of the  $j$ th regression coefficient

where  $se(\hat{\beta}_j)$  is the standard error of  $\hat{\beta}_j$  from the full data set. If  $|\Delta\beta_j^{(i)}|$  is large, the  $i$ th observation will have a great influence on the estimation of  $\beta_j$ .

**Figure 13.6** Influence of individual observations on the estimation of the regression coefficient for SBP for the Framingham Heart Study data fitted in Table 13.24

**Example 13.43**

**Cardiovascular Disease** Assess the influence of individual observations on the estimation of the regression coefficient for SBP for the model in Table 13.24.

**Solution**

SBP is represented as  $\ln(SBP/132)$  and is denoted by LMSYD132 in Table 13.24. The influence measure in Equation 13.40 is displayed in Figure 13.6 for a subset of the data and is denoted by DFBETA8 (because this is the eighth regression coefficient in the model, other than the intercept). We see that none of the observations has a great influence in this subset of the data. The maximum value of  $|\Delta\beta_j^{(i)}| \approx 0.2$ , which is an actual difference of  $0.2 \times se(\hat{\beta}_8) = 0.2 \times 0.5737 \approx 0.11$  (see Table 13.24). Because the actual value of  $\hat{\beta}_8$  in the full model = 2.80, this is a relatively minor change ( $\approx 4\%$ ). This is also true for the remaining observations in the data set as well as for the other regression coefficients. However, influential observations are more likely to appear in smaller data sets, particularly for individual subjects with predictor variables that are far from the mean value.

In this section, we have learned about multiple logistic regression. This is an extremely important technique used to control for one or more continuous or categorical covariates (independent variables) when the outcome (dependent) variable is binary. It can be viewed as an extension of Mantel-Haenszel techniques and usually enables more complete control of confounding effects of several risk factors simultaneously. Also, it is analogous to multiple linear regression for normally distributed outcome variables.

On the master flowchart (p. 844), we start at ④ and generally answer no to (1) interested in relationships between two variables? This leads us to the box labeled “more than two variables of interest.” We then answer “binary” to (2) outcome variable continuous or binary? We then answer no to (3) time of events important? which leads us to the box labeled “multiple logistic-regression methods.”

It is, of course, also possible to use multiple logistic regression if only two variables are of interest, where the dependent variable is binary and the single independent variable is continuous. If the dependent variable is binary and we have a single independent variable that is categorical, then we can use either multiple logistic-regression or contingency-table methods, which should give identical results. The latter are probably preferable, for simplicity.

### REVIEW QUESTIONS 13E

- 1** What is the principal difference between multiple logistic regression and multiple linear regression?
- 2** A study was performed relating incidence of the common cold to intake of different alcoholic beverages in people employed at five Spanish universities [13]. One analysis included users of red wine only or nondrinkers. A logistic regression was performed of incidence of the common cold related to red-wine consumption, which was categorized as follows in drinks per week (0, 1 to 7, 8 to 14, and >14). Other covariates controlled for in the analysis were age (continuous), sex, and faculty/staff status. The results were as shown in Table 13.27.

**Table 13.27** Common cold vs. average red-wine consumption (data)

Red-wine consumption (drinks per week)	Beta	se
0	(ref. <sup>a</sup> )	
1–7	−0.416	0.132
8–14	−0.527	0.238
>14	−0.892	0.341

<sup>a</sup>Reference group.

- (a)** What is the *OR* for red-wine consumption of 1–7 drinks per week vs. 0 drinks per week?
- (b)** What does it mean?
- (c)** Provide a 95% CI for this *OR*.

## 13.9 Extensions to Logistic Regression

### Matched Logistic Regression

In some instances, we have a categorical outcome, but the units of analysis have been selected using a *matched design*, and thus are not independent. Ordinary logistic regression methods are not appropriate here, but an extension of logistic regression called *conditional logistic regression* can be employed.

**Example 13.44**

**Cancer** In 1989–1990, 32,826 NHS participants provided a blood sample for research purposes. The blood was frozen and stored for future analyses. In general, it is too expensive to analyze the blood for all participants. Instead, a nested case-control design is typically used. For example, estradiol is a hormone that has been related to breast cancer in several other studies. To study this question using NHS data, 235 women with breast cancer occurring between 1990 and 2000 and after blood collection were identified. One or two controls were selected per case, yielding a total of 346 controls. The controls were matched on age, time of day of blood draw, fasting status of blood draw, and previous use of post-menopausal hormones. All cases and controls were postmenopausal at the time of the blood draw (1989–1990). Because of possible lab drift, the matched sets (case and 1 or 2 controls) were analyzed at the same time for a number of analytes, including plasma estradiol. How should the association between plasma estradiol and breast cancer be assessed?

Ideally we would like to use logistic regression to predict breast cancer as a function of plasma estradiol and other breast cancer risk factors. However, we need to account for the dependence between women in the same matched set. Conditional logistic regression can be used for this purpose.

**Equation 13.41****Conditional Logistic Regression**

Suppose we wish to assess the association between the incidence of breast cancer ( $D$ ) and plasma estradiol ( $x$ ) but wish to control for other covariates ( $z_1, z_2, \dots, z_k$ ), denoted in summary by  $z$ . Examples of other covariates include age, parity (i.e., number of children), family history of breast cancer, and others. Suppose we subdivide the data into  $S$  matched sets ( $i = 1, \dots, S$ ). The  $i$ th matched set consists of a single case and  $n_i$  controls, where  $n_i \geq 1$  and  $n_i$  may vary among matched sets. Let  $D_{ij}$  = case status of the  $j$ th subject in the  $i$ th matched set. We use a logistic model of the form

$$\text{logit}[Pr(D_{ij} = 1)] = \alpha_i + \beta x_{ij} + \gamma_1 z_{1,ij} + \dots + \gamma_k z_{k,ij} \equiv \alpha_i + \beta x_{ij} + \gamma z_{ij}$$

where  $\alpha_i$  = indicator variable for being in the  $i$ th matched set, which = 1 if a subject is in the  $i$ th matched set and = 0 otherwise.

The problem is that we cannot determine  $\alpha_i$  because the matched sets are small and purposely selected in such a way as to have 1 case and 1 or more controls. Thus, we cannot use logistic regression to determine the absolute probability of disease because of the way the samples are selected. However, we can determine the conditional probability that the  $j$ th member of a matched set is a case given that there is exactly one case in the matched set, denoted by  $p_{ij}$ , or

$$\begin{aligned} p_{ij} &= Pr(D_{ij} = 1 | \sum_{k=1}^{n_i} D_{ik} = 1) = \frac{Pr(D_{ij} = 1) \prod_{k=1, k \neq j}^{n_i} Pr(D_{ik} = 0)}{\sum_{l=1}^{n_i} Pr(D_{il} = 1) \prod_{k \neq l}^{n_i} Pr(D_{ik} = 0)} \\ &= \frac{\exp(\alpha_i + \beta x_{ij} + \gamma z_{ij}) / \prod_{k=1}^{n_i} [1 + \exp(\alpha_i + \beta x_{ik} + \gamma z_{ik})]}{\sum_{l=1}^{n_i} \exp(\alpha_i + \beta x_{il} + \gamma z_{il}) / \prod_{k=1}^{n_i} [1 + \exp(\alpha_i + \beta x_{ik} + \gamma z_{ik})]} \\ &= \exp(\beta x_{ij} + \gamma z_{ij}) / \sum_{l=1}^{n_i} \exp(\beta x_{il} + \gamma z_{il}) \end{aligned}$$

If the  $j$ th subject of the  $i$ th matched set is the case, then the expression in equation 13.41 is referred to as the contribution to the conditional likelihood for the  $i$ th matched set. We can use maximum likelihood methods to find estimates of  $\beta$  and  $\gamma$  which maximize  $L = \prod_{i=1}^S p_{ij_i}$ , where  $j_i$  = case in the  $i$ th matched set. These are called conditional likelihood methods and the model is referred to as a *conditional logistic regression model*.

**Equation 13.42****Interpretation of Parameters in a Conditional Logistic Regression Model**

To interpret the parameters of the conditional logistic regression model in equation 13.41, we consider two subjects  $j$  and  $l$  in the  $i$ th matched set, one of whom is a case and the other a control. We assume these subjects have the same value for all other covariates, that is,  $z_{ij} = z_{il}$ , but differ by one unit on the primary exposure variable, that is,  $x_{ij} = x_{il} + 1$ .

The relative risk that the subject with the higher exposure is the case, is given by

$$RR = Pr(D_{ij} = 1) / Pr(D_{il} = 1) = \exp(\beta)$$

A similar interpretation holds for the other regression coefficients.

**Example 13.45**

**Cancer** Estimate the association between breast cancer incidence and plasma estradiol using the matched design in Example 13.44.

**Solution**

For this example, we had a total of 235 breast cancer cases and 346 controls. A conditional logistic regression model was fit to the data with a single primary exposure variable ln estradiol ( $x$ ) and several other breast cancer risk predictors using the SAS procedure PROC PHREG. This is the same algorithm used to fit proportional hazards survival models in SAS, a topic we will discuss in Chapter 14. The results are given in Table 13.28.

**Table 13.28** Use of SAS PROC PHREG to perform conditional logistic regression on the breast cancer data

```
proc phreg;
model cscn*case(0)=tmtl b4a b4b x5 tmtbm bbd b21 b22 b23
dur3 dur4 dur8
curpmh pstpmh sumbmi2a sumbmi3a sumhgt2a sumhgt3a
sumalcl sumalc2 sumalc3 famhx lestrad1;
strata matchid;
```

**Convergence Status**  
Convergence criterion (GCONV=1E-8) satisfied.

Criterion	Model Fit Statistics	
	Without Covariates	With Covariates
-2 LOG L	407.529	361.237
AIC	407.529	407.237
SBC	407.529	486.808

(continued)

**Table 13.28** Use of SAS PROC PHREG to perform conditional logistic regression on the breast cancer data (Continued)

Testing Global Null Hypothesis: BETA=0						
Test		Chi-Square	DF	Pr > ChiSq		
Likelihood Ratio			23	0.0028		
Score			23	0.0065		
Wald			23	0.0356		
Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
tmt1	1	0.00617	0.08160	0.0057	0.9398	1.006
b4a	1	-0.00661	0.08770	0.0057	0.9399	0.993
b4b	1	-0.03637	0.08894	0.1672	0.6826	0.964
x5	1	-0.0006087	0.02125	0.0008	0.9771	0.999
tmtbm	1	0.0007292	0.00281	0.0672	0.7954	1.001
bbd	1	-0.07291	3.28786	0.0005	0.9823	0.930
b21	1	0.09160	0.13953	0.4310	0.5115	1.096
b22	1	-0.01387	0.05787	0.0575	0.8105	0.986
b23	1	0.0006671	0.04040	0.0003	0.9868	1.001
The PHREG Procedure						
Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
dur3	1	-0.01543	0.06945	0.0494	0.8242	0.985
dur4	1	0.20479	0.11699	3.0640	0.0800	1.227
dur8	1	0.08135	0.07618	1.1403	0.2856	1.085
curpmh	1	-0.15969	0.38707	0.1702	0.6799	0.852
pstpmh	1	-0.06442	0.25656	0.0630	0.8017	0.938
sumbmi2a	1	-0.00105	0.00129	0.6628	0.4156	0.999
sumbmi3a	1	0.0009756	0.00215	0.2065	0.6495	1.001
sumhgt2a	1	-0.00325	0.00232	1.9627	0.1612	0.997
sumhgt3a	1	0.00212	0.00638	0.1105	0.7396	1.002
sumalcl1	1	0.0006666	0.0004087	2.6600	0.1029	1.001
sumalcl2	1	0.00676	0.00380	3.1621	0.0754	1.007
sumalcl3	1	-0.00130	0.0008548	2.3242	0.1274	0.999
famhx	1	0.60465	0.25490	5.6270	0.0177	1.831
lestrad1	1	0.73944	0.21913	11.3866	0.0007	2.095

The actual model fit was as follows.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \beta x_{ij} + \sum_{k=1}^{22} \gamma_k Z_{ijk}$$

where  $i$  = matched pair,  $j$  = subject within the matched pair, and the variables are defined as follows:

$$x_{ij} = \ln(\text{estradiol}) - 2.36 (\ln \text{ pg/mL})$$

$$Z_{ij1} = \text{tmt1} = \text{age at menopause} - \text{age at menarche} = \text{premenopausal time}$$

$$Z_{ij2} = b4a = \text{age} - \text{age at menopause if natural menopause}, = 0 \text{ otherwise}$$

$$Z_{ij3} = b4b = \text{age} - \text{age at menopause if bilateral oophorectomy}, = 0 \text{ otherwise}$$

$$Z_{ij4} = x5 = \text{age at first birth} - \text{age at menarche if parous}, = 0 \text{ if nulliparous}$$

$$Z_{ij5} = \text{tmtbm} = \text{birth index} = \sum_{i=1}^b (\text{age at menopause} - \text{age at } i\text{th birth}) \text{ if parous}, = 0 \text{ if nulliparous}$$

where  $b$  = parity (number of children)

$z_{ij6} = bbd = 1$  if benign breast disease, = 0 otherwise  
 $z_{ij7} = b21 = bbd \times \text{age at menarche}$   
 $z_{ij8} = b22 = bbd \times \text{tmt1}$   
 $z_{ij9} = b23 = bbd \times (\text{age} - \text{age at menopause})$   
 $z_{ij10} = \text{dur3} = \text{duration of estrogen use (yrs)}$   
 $z_{ij11} = \text{dur4} = \text{duration of estrogen + progesterone use (yrs)}$   
 $z_{ij12} = \text{dur8} = \text{duration of use of other postmenopausal hormones (PMH) (yrs)}$   
 $z_{ij13} = \text{curpmh} = \text{current use of PMH} = 1$  if yes/= 0 if no  
 $z_{ij14} = \text{pstpmh} = \text{past use of PMH} = 1$  if yes/= 0 if no  
 $z_{ij15} = \text{sumbmi2a} = \text{average BMI pre-menopause} \times \text{tmt1}$   
 $z_{ij16} = \text{sumbmi3a} = \text{average BMI post-menopause} \times (\text{age} - \text{age at menopause})$   
 $z_{ij17} = \text{sumhgt2a} = \text{height} \times \text{tmt1}$   
 $z_{ij18} = \text{sumhgt3a} = \text{height} \times (\text{age} - \text{age at menopause})$   
 $z_{ij19} = \text{sumalc1} = \text{average alcohol intake pre-menopause (g)} \times \text{tmt1}$   
 $z_{ij20} = \text{sumalc2} = \text{average alcohol intake post-menopause while on PMH} \times \text{duration PMH use}$   
 $z_{ij21} = \text{sumalc3} = \text{average alcohol intake post-menopause while not on PMH} \times (\text{age} - \text{age at menopause} - \text{duration PMH use})$   
 $z_{ij22} = \text{famhx} = \text{family history of breast cancer (1 = yes/0 = no)}$

We see that there is a significant association between breast cancer incidence and  $\ln(\text{estradiol})$  ( $\text{Beta} = 0.739 \pm 0.219, p < .001$ ). The relative risk (listed under the Hazard Ratio column) is  $2.1 (= e^{0.73944})$ . This implies that if we have two women in a matched pair, one with breast cancer and the other without breast cancer, whose  $\ln(\text{estradiol})$  levels differ by 1 ln unit and who are the same for all other breast cancer risk factors, the woman with the higher estradiol is 2.1 times as likely to be the case than the woman with the lower estradiol. Note that a difference of 1 ln unit is equivalent to a ratio of  $e^1 = 2.7$ . Thus, the woman with higher estradiol has a plasma estradiol level that is 2.7 times as high as that of the lower estradiol woman. To obtain 95% confidence limits for  $\beta$ , we compute  $(e^{c_1}, e^{c_2})$ , where  $(c_1, c_2) = \hat{\beta} \pm 1.96\text{se}(\hat{\beta})$ . In this case, we have  $(c_1, c_2) = 0.739 \pm 1.96(0.219) = (0.310, 1.168)$  and the 95% CI =  $(e^{0.310}, e^{1.168}) = (1.4, 3.2)$ . We will discuss hazard ratios in more detail when we study proportional hazards regression models in Chapter 14. In the context of conditional logistic regression, we can consider the hazard ratio as a ratio of incidence rates for the higher vs. lower estradiol woman.

Note that there is only one other predictor that is a significant risk factor for breast cancer in this example, that is, family history of breast cancer (famHx), with an RR of 1.8 and a *p*-value of .018. Several other breast cancer risk factors (dur4 = duration of estrogen + progesterone use, *p* = .08 and sumalc2 = total alcohol consumption while using PMH, *p* = .075) show a trend toward statistical significance. Actually, all the risk factors in the model have been shown to be significant risk factors for breast cancer in large data sets (see Colditz et al. [14] for more details about the variables used for breast cancer modeling). Conditional logistic regression can also be extended to allow for more than one case and/or more than one control in a matched pair (see Breslow and Day [15] for more details about conditional logistic regression).

## Polychotomous Logistic Regression

In some cases, we have a categorical outcome variable with more than two categories. Often we might have a single control group to be compared with multiple case groups or a single case group to be compared with multiple control groups.

**Example 13.46**

**Cancer** Breast cancers are commonly typed using a biochemical assay to determine estrogen receptor (ER) and progesterone receptor (PR) status. Tumors can be jointly classified as having ER positive (ER+) vs. ER negative (ER-) status and PR positive (PR+) vs. PR negative (PR-) status. This distinction is important because different treatments are used according to ER/PR status. A study was performed within the NHS to determine risk factor profiles for specific types of breast cancer according to ER/PR status [14]. There were 2096 incident cases of breast cancer from 1980–2000, of which 1281 were ER+/PR+, 417 were ER-/PR-, 318 were ER+/PR-, and 80 were ER-/PR+. There was a common control group for all types of breast cancers. How should the data be analyzed?

It is tempting to perform separate logistic regression analyses of each case group vs. the control group. This is a valid approach but will not allow us to compare regression coefficients for the same risk factor between different types of breast cancer. Instead, we analyze all the data simultaneously using *polychotomous logistic regression* (PLR). We can generalize logistic regression in this setting as follows. Suppose there are  $Q$  outcome categories, where group 1 is a control group and groups  $2, \dots, Q$  are different case groups. Suppose also that there are  $k$  exposure variables. The PLR model is given as follows.

**Equation 13.43****Polychotomous Logistic Regression**

$$\begin{aligned} Pr(\text{1st outcome category}) &= \frac{1}{1 + \sum_{r=2}^Q \exp\left(\alpha_r + \sum_{k=1}^K \beta_{rk} x_k\right)} \\ Pr(q\text{th outcome category}) &= \frac{\exp\left(\alpha_q + \sum_{k=1}^K \beta_{qk} x_k\right)}{1 + \sum_{r=2}^Q \exp\left(\alpha_r + \sum_{k=1}^K \beta_{rk} x_k\right)}, \quad q = 2, \dots, Q \end{aligned}$$

**Equation 13.44****Interpretation of Parameters in PLR**

Suppose we have 2 individuals who differ by 1 unit on the  $k$ th exposure variable and are the same on all other exposure variables. We will call the individual with the higher exposure ( $x_k + 1$ ) subject A and the subject with the lower exposure ( $x_k$ ) subject B. Based on equation 13.43,

$$\begin{aligned} \text{odds}_{q,A} &= \frac{Pr(\text{subject A is in the } q\text{th outcome category})}{Pr(\text{subject A is in the 1st outcome category (control group)})} \\ &= \exp\left[\alpha_q + \sum_{l=1}^{k-1} \beta_{ql} x_l + \beta_{qk}(x_k + 1) + \sum_{l=k+1}^K \beta_{ql} x_l\right] \\ \text{odds}_{q,B} &= \frac{Pr(\text{subject B is in the } q\text{th outcome category})}{Pr(\text{subject B is in the 1st outcome category})} \\ &= \exp\left[\alpha_q + \sum_{l=1}^{k-1} \beta_{ql} x_l + \beta_{qk}(x_k) + \sum_{l=k+1}^K \beta_{ql} x_l\right] \end{aligned}$$

Hence,

$$\begin{aligned} \text{the OR for being in category } q \text{ vs. category 1 for subject A vs. subject B} \\ = \frac{\text{odds}_{q,A}}{\text{odds}_{q,B}} = \exp(\beta_{qk}) \equiv OR_{qk} \end{aligned}$$

A  $100\% \times (1 - \alpha)$  CI for  $OR_{qk}$  is given by  $(e^{c_1}, e^{c_2})$ , where

$$(c_1, c_2) = \hat{\beta}_{q,k} \pm z_{1-\alpha/2} se(\hat{\beta}_{q,k})$$

Note that a special case of PLR is when  $Q = 2$ , in which case there is one control group and one case group and PLR is the same as ordinary logistic regression.

Another capability of PLR is to compare the strength of association of the same variable for 2 different case categories. For example, we might be interested in whether a breast cancer risk factor had the same *OR* for 2 different types of breast cancer.

In general, the *OR* for being in outcome category  $q_1$  vs. outcome category  $q_2$  for subject A compared with subject B is given by  $\exp(\hat{\beta}_{q_1,k} - \hat{\beta}_{q_2,k})$  with 95% confidence limits given by  $(e^{c_1}, e^{c_2}) = \hat{\beta}_{q_1,k} - \hat{\beta}_{q_2,k} \pm z_{1-\alpha/2} se(\hat{\beta}_{q_1,k} - \hat{\beta}_{q_2,k})$ . In general,  $\hat{\beta}_{q_1,k}$  and  $\hat{\beta}_{q_2,k}$  will be correlated because a common control group is used to estimate each *OR*. Hence,

$$se(\hat{\beta}_{q_1,k} - \hat{\beta}_{q_2,k}) = \left[ \text{var}(\hat{\beta}_{q_1,k}) + \text{var}(\hat{\beta}_{q_2,k}) - 2\text{cov}(\hat{\beta}_{q_1,k}, \hat{\beta}_{q_2,k}) \right]^{1/2}$$

The covariance between estimated regression parameters is available in most computer packages that implement PLR.

#### Example 13.47

Assess the effect of alcohol use before menopause on different types of breast cancer based on the data set described in Example 13.46.

#### Solution

We have fitted the PLR model in Equation 13.43 to the breast cancer data described in Example 13.46. There were a total of 5 groups (one control group and 4 case groups). There were a total of 22 variables in the model. Hence, if the control group is the reference group, there are a total of 88 regression parameters to be estimated plus 4 separate intercept terms. The results for alcohol consumption before menopause (adjusted for the other 21 variables in the model) are given in Table 13.29.

**Table 13.29** Effect of alcohol consumption before menopause<sup>a</sup> on different types of breast cancer, NHS data, 1980–2000

Group	Beta	se	p-value	RR <sup>b</sup> (95% CI)	Number of cases
no breast cancer	(ref)			1.0	
ER+/PR+	0.00029	0.00009	0.001	1.12 (1.04–1.20)	1281
ER+/PR-	0.00022	0.00017	0.20	1.09 (0.96–1.24)	318
ER-/PR+	0.00015	0.00037	0.68	1.06 (0.80–1.40)	80
ER-/PR-	-0.00003	0.00017	0.86	0.99 (0.87–1.12)	417

<sup>a</sup>Cumulative grams of alcohol before menopause (g/day × years).

<sup>b</sup>The relative risk for 1 drink per day of alcohol from age 18 to age 50  $\approx 12 \text{ grams alcohol/drink} \times 32 \text{ years} = 384 \text{ gram-years} \times \text{Beta after controlling for 21 other breast cancer risk factors.}$

We see that there is a significant effect of alcohol for ER+/PR+ breast cancer but not for any other type of breast cancer. The *RR* for 1 g/day = 384 g-years/day =  $\exp[0.00029(384)] = 1.12$ . The 95% CI =  $(e^{c_1}, e^{c_2})$ , where  $(c_1, c_2) = 384[0.00029 \pm 1.96(0.00009)]$ . Thus, the 95% CI = (1.04 – 1.20). In addition, results for ER-/PR- breast cancer are almost completely null.

The model in equation 13.43 forces the regression parameters for all variables for different case groups to be different (i.e., a total of 92 parameters to be estimated

in Example 13.47). It is possible to extend PLR to allow for parameter estimates for some variables to be the same over all case groups and parameter estimates for other variables to be different. This allows one to test whether the effects of a risk factor are significantly heterogeneous among all case groups. Details for this approach are given in Marshall and Chisholm [16].

## Ordinal Logistic Regression

### Example 13.48

**Sports Medicine** In the data set TENNIS1.DAT (on the Companion Website) we have data from an observational study among about 400 members of several tennis clubs in the Boston area. The objective of the study was to examine risk factors for tennis elbow. Subjects were asked how many current or previous episodes of tennis elbow they had. The distribution ranged from 0 to 8 and was very skewed. Hence, we elected to categorize the number of episodes into 3 categories (0/1/2+). We could treat these 3 categories as nominal categorical data and use PLR, but this type of analysis would lose the ordering of the categories in the above scale. Instead, we used a technique called *ordinal logistic regression* to relate the number of episodes of tennis elbow to age, sex, and the material of the racquet.

### Equation 13.45

#### Ordinal Logistic Regression

Suppose an outcome variable  $y$  has  $c$  ordered categories ( $c \geq 2$ ), which we arbitrarily refer to as  $1, \dots, c$ . Suppose also there are  $k$  covariates  $x_1, \dots, x_k$ . An ordinal logistic regression model is defined by

$$\log[Pr(y \leq j) / Pr(y \geq j+1)] = \alpha_j + \beta_1 x_1 + \dots + \beta_k x_k, \quad j = 1, \dots, c-1.$$

The regression coefficients  $\beta_q$  have a similar interpretation as for ordinary logistic regression. Specifically,

$$\begin{aligned} e^{\beta_q} &= (\text{odds that } y \leq j | x_q = x) / (\text{odds that } y \leq j | x_q = x-1), \\ q &= 1, \dots, k; \\ j &= 2, \dots, c \\ &\equiv \text{odds ratio for } y \leq j \text{ given } x_q = x \text{ vs. } x_q = x-1 \\ &\quad \text{holding all other variables constant} \end{aligned}$$

Note that if  $c = 2$ , then the ordinal logistic regression model reduces to ordinary logistic regression.

Note also that in ordinal regression,  $e^{\beta_q}$  is assumed to be the same for each value of  $j$ . This type of ordinal regression model is called a *cumulative odds* or *proportional odds ordinal logistic regression model*.

### Example 13.49

**Sports Medicine** Apply the ordinal logistic regression model in Equation 13.45 to the tennis elbow data in Example 13.48.

#### Solution

We applied the ordinal regression model to the tennis elbow data. For this purpose we categorized the outcome variable ( $y$ ) in terms of the number of episodes of tennis elbow (0/1/2+), which was coded as (0/1/2). Note that the program will work equally well with any numeric values for  $y$ ; the computer will identify the number of unique values of  $y$  and will order these values into ordered categories before performing the analysis. The predictor variables were age, sex (1 = M / 2 = F), and the material of the racquet [1 = wood (reference)/2 = metal (i.e., either aluminum or steel)/3 = fiberglass, graphite, or composite].

The results are given in Table 13.30.

**Table 13.30 Application of the MINITAB ordinal logistic regression program to the tennis elbow data****Ordinal Logistic Regression: num\_episodes(0/1/2+) vs. Age, Sex, material\_current**

Link Function: Logit

Response Information

Variable	Value	Count
num_episodes (0/1/2+)	0	167
	1	150
	2	116
	Total	433

\* NOTE \* 433 cases were used

\* NOTE \* 9 cases contained missing values

**Logistic Regression Table**

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
					Lower	Upper	
Const(1)	2.85057	0.592981	4.81	0.000			
Const(2)	4.42467	0.615680	7.19	0.000			
Age	-0.0580569	0.0107481	-5.40	0.000	0.94	0.92	0.96
Sex*	-0.393689	0.186264	-2.11	0.035	0.67	0.47	0.97
material_current							
2 <sup>+</sup>	-0.319725	0.228137	-1.40	0.161	0.73	0.46	1.14
3 <sup>+</sup>	-0.589181	0.246255	-2.39	0.017	0.55	0.34	0.90

Log-Likelihood = -451.974

Test that all slopes are zero: G = 37.877, DF = 4, p-value = 0.000

\*1 = M/2 = F.

†category 1 = wood (reference)/category 2 = metal (aluminum or steel)/category 3 = fiberglass, graphite, or composite.

We see that there are significant effects of age ( $OR = 0.94$ ,  $p < .001$ ) and gender ( $1 = M / 2 = F$ ), (females vs. males) ( $OR = 0.67$ ,  $p = .035$ ). Thus, older players and females are likely to have more episodes of tennis elbow.

In addition, we categorized the material of the current racquet into 3 categories (1 = wood = reference/2 = metal = aluminum or steel/3 = composite, fiberglass, or graphite).

We see that the  $OR$  comparing metal to wood is 0.73 (95% CI = 0.46, 1.14) but is not statistically significant ( $p = .16$ ).

However, the  $OR$  comparing fiberglass, graphite, or composite with wood is 0.55 (95% CI = 0.34, 0.90), which is statistically significant ( $p = .017$ ) even after controlling for age and sex. Furthermore,  $\exp[\text{const}(1)]$  is an estimate of the odds of 0 episodes vs. 1+ episodes and  $\exp[\text{const}(2)]$  is an estimate of the odds of  $\leq 1$  episode vs. 2+ episodes, both for subjects with all x's equal to zero.

In general, users of wood racquets have the least number of episodes of tennis elbow, and users of composite racquets have the greatest; users of metal racquets are in between.

## 13.10 Meta-Analysis

In the previous sections of this chapter, and in all previous chapters, we have examined methods of analysis for a single study. However, often more than one investigation is performed to study a particular research question, often by different research groups. In some instances, results are seemingly contradictory, with some research groups reporting significant differences for a particular finding and other research groups reporting no significant differences.

### Example 13.50

**Renal Disease** In Data Set NEPHRO.DAT (on the Companion Website), we present data from a literature search comparing the nephrotoxicity (development of abnormal kidney function) of several different aminoglycosides [17]. In Table 13.31, we focus on a subset of eight studies that compared two of the aminoglycosides—gentamicin and tobramycin. In seven of eight studies, the *OR* for tobramycin in comparison with gentamicin is less than 1, implying that there are fewer nephrotoxic side effects for tobramycin than for gentamicin. However, many of the studies are small and individually are likely to yield nonsignificant results. The question is, What is the appropriate way to combine evidence across all the studies so as to reduce sampling error and increase the power of the investigation and, in some instances, to resolve the inconsistencies among the study results? The technique for accomplishing this is called *meta-analysis*. In this section, we will present the methods of DerSimonian and Laird [18] for addressing this problem.

**Table 13.31 Comparison of nephrotoxicity of gentamicin vs. tobramycin in NEPHRO.DAT**

Study	Gentamicin		Tobramycin		Odds ratio <sup>b</sup>	$y_i^c$	$w_i^d$	$w_i^{*e}$
	No. of subjects	No. of positives <sup>a</sup>	No. of subjects	No. of positives <sup>a</sup>				
1. Walker	40	7	40	2	0.25	-1.394	1.430	1.191
2. Wade	43	13	47	11	0.71	-0.349	4.367	2.709
3. Greene	11	2	15	2	0.69	-0.368	0.842	0.753
4. Smith	72	19	74	9	0.39	-0.951	5.051	2.957
5. Fong	102	18	103	15	0.80	-0.229	6.873	3.500
6. Brown	103	5	96	2	0.42	-0.875	1.387	1.161
7. Feig	25	10	29	8	0.57	-0.560	2.947	2.086
8. Matzke	99	9	97	17	2.13	+0.754	5.167	2.996

<sup>a</sup>Number who developed nephrotoxicity.

<sup>b</sup>Odds in favor of nephrotoxicity for tobramycin patients / odds in favor of nephrotoxicity for gentamicin patients.

<sup>c</sup> $y_i = \ln(OR_i)$ .

<sup>d</sup> $w_i = (1/a_i + 1/b_i + 1/c_i + 1/d_i)^{-1}$

<sup>e</sup> $w_i^* = [(1/w_i + \hat{\Delta}^2)^{-1}]$ .

Suppose there is an underlying log odds ratio  $\theta_i$  for the  $i$ th study, which is estimated by  $y_i = \ln(\hat{OR}_i)$   $i = 1, \dots, 8$ , where the estimated  $OR_i$  are given in Table 13.31 in the Odds ratio column. We assume there is *within-study variation* of  $y_i$  about  $\theta_i$ , where the variance of  $y_i$  is

$$s_i^2 = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \equiv \frac{1}{w_i}$$

and  $a_i$ ,  $b_i$ ,  $c_i$ , and  $d_i$  are the cell counts in the  $2 \times 2$  table for the  $i$ th study. We also assume that there is *between-study variation* of  $\Delta_i$  about an average true log OR  $\mu$  over all studies so that

$$\theta_i = \mu + \delta_i$$

$$\text{and } \text{Var}(\Delta_i) = \Delta^2$$

This is similar to the random-effects analysis of variance (ANOVA) model presented in Section 12.8. To estimate  $\mu$ , we calculate a weighted average of the study-specific log ORs given by

$$\hat{\mu} = \sum_{i=1}^k w_i^* \gamma_i / \sum_{i=1}^k w_i^*$$

where

$$w_i^* = (s_i^2 + \hat{\Delta}^2)^{-1}$$

i.e., the weight for the  $i$ th study is inversely proportional to the total variance for that study (which equals  $s_i^2 + \Delta^2$ ), and

$$se(\hat{\mu}) = 1 / \sqrt{\left( \sum_{i=1}^k w_i^* \right)^{1/2}}$$

It can be shown that the best estimate of  $\Delta^2$  is given by

$$\hat{\Delta}^2 = \max \left\{ 0, [Q_w - (k-1)] / \left[ \left( \sum_{i=1}^k w_i - \left( \sum_{i=1}^k w_i^2 / \sum_{i=1}^k w_i \right) \right) \right] \right\}$$

where

$$Q_w = \sum_{i=1}^k w_i (y_i - \bar{y}_w)^2$$

and

$$\bar{y}_w = \sum_{i=1}^k w_i y_i / \sum_{i=1}^k w_i$$

This procedure is summarized as follows.

#### Equation 13.46

#### Meta-Analysis, Random-Effects Model

Suppose we have  $k$  studies, each with the goal of estimating an  $OR = \exp(\mu)$  defined as the odds of disease in a treated group compared with the odds of disease in a control group.

- (1) The best estimate of the average study-specific log OR from the  $k$  studies is given by

$$\hat{\mu} = \sum_{i=1}^k w_i^* \gamma_i / \sum_{i=1}^k w_i^*$$

where  $\gamma_i$  = estimated log OR for the  $i$ th study

$$w_i^* = (s_i^2 + \hat{\Delta}^2)^{-1}$$

$$1/w_i = s_i^2 = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} = \text{within-study variance}$$

and  $a_i$ ,  $b_i$ ,  $c_i$ , and  $d_i$  are the cell counts for the  $2 \times 2$  table for the  $i$ th study.

$$\hat{\Delta}^2 = \max \left\{ 0, [Q_w - (k-1)] / \left[ \sum_{i=1}^k w_i - \left( \sum_{i=1}^k w_i^2 / \sum_{i=1}^k w_i \right) \right] \right\}$$

$$Q_w = \sum_{i=1}^k w_i (\gamma_i - \bar{\gamma}_w)^2 = \sum_{i=1}^k w_i \gamma_i^2 - \left( \sum_{i=1}^k w_i \gamma_i \right)^2 / \sum_{i=1}^k w_i$$

and

$$\bar{\gamma}_w = \sum_{i=1}^k w_i \gamma_i / \sum_{i=1}^k w_i$$

The corresponding point estimate of the  $OR = \exp(\hat{\mu})$ .

(2) The standard error of  $\hat{\mu}$  is given by  $se(\hat{\mu}) = \left( 1 / \sum_{i=1}^k w_i^* \right)^{1/2}$

(3) A  $100\% \times (1 - \alpha)$  CI for  $\mu$  is given by

$$\hat{\mu} \pm z_{1-\alpha/2} se(\hat{\mu}) = (\mu_1, \mu_2)$$

The corresponding  $100\% \times (1 - \alpha)$  CI for  $OR$  is  $[\exp(\mu_1), \exp(\mu_2)]$ .

(4) To test the hypothesis  $H_0: \mu = 0$  vs.  $H_1: \mu \neq 0$  (or equivalently to test  $H_0: OR = 1$  vs.  $H_1: OR \neq 1$ ), we use the test statistic

$$z = \hat{\mu} / se(\hat{\mu})$$

which under  $H_0$  follows an  $N(0, 1)$  distribution. The two-sided  $p$ -value is given by  $2 \times [1 - \Phi(|z|)]$ .

### Example 13.51

**Renal Disease** Estimate the combined nephrotoxicity  $OR$  comparing tobramycin with gentamicin based on the data in Table 13.31. Obtain a 95% CI, and provide a two-sided  $p$ -value for the hypothesis that the two treatments have equal rates of nephrotoxicity.

### Solution

We first compute the study-specific log  $OR$  ( $\gamma_i$ ) and weight  $w_i = (1/a_i + 1/b_i + 1/c_i + 1/d_i)^{-1}$ , which are shown in Table 13.31.

Next we compute the estimated between-study variance  $\hat{\Delta}^2$ . We have

$$\sum_{i=1}^8 w_i = 1.430 + \dots + 5.167 = 28.0646$$

$$\sum_{i=1}^8 w_i \gamma_i = 1.430(-1.394) + \dots + 5.167(0.754) = -9.1740$$

$$\sum_{i=1}^8 w_i \gamma_i^2 = 1.430(-1.394)^2 + \dots + 5.167(0.754)^2 = 13.2750$$

Hence,

$$Q_w = 13.2750 - (-9.1740)^2 / 28.0646 = 10.276$$

Furthermore,

$$\sum_{i=1}^8 w_i^2 = 1.430^2 + \dots + 5.167^2 = 131.889$$

$$\begin{aligned} \hat{\Delta}^2 &= (10.276 - 7) / (28.0646 - 131.889 / 28.0646) \\ &= 0.140 = \text{between-study variance} \end{aligned}$$

Hence,

$$w_i^* = (1 / w_i + \hat{\Delta}^2)^{-1}$$

as shown in Table 13.31. Finally,

$$\begin{aligned}\sum_{i=1}^8 w_i^* \gamma_i &= 1.191(-1.394) + \dots + 2.996(0.754) = -6.421 \\ \sum_{i=1}^8 w_i^* &= 1.191 + \dots + 2.996 = 17.3526\end{aligned}$$

and

$$\hat{\mu} = -6.421 / 17.3526 = -0.370$$

with standard error given by

$$se(\hat{\mu}) = (1 / 17.3526)^{1/2} = 0.240$$

Hence, the point estimate of the overall  $OR = \exp(\mu)$  is given by  $\exp(-0.370) = 0.69$ . A 95% CI for  $\exp(\mu)$  is given by  $\exp[-0.370 \pm 1.96(0.240)] = (0.43, 1.11)$ .

To test the hypothesis  $H_0: \mu = 0$  vs.  $H_1: \mu \neq 0$ , we use the test statistic

$$\begin{aligned}z &= \hat{\mu} / se(\hat{\mu}) \\ &= -0.370 / 0.240 = -1.542\end{aligned}$$

with corresponding two-sided  $p$ -value given by  $p = 2 \times [1 - \Phi(1.542)] = .123$ . Hence the  $OR$ , although less than 1, does not significantly differ from 1.

## Test of Homogeneity of Odds Ratios

Some investigators feel the procedure in Equation 13.46 should only be used if there is not significant heterogeneity among the  $k$  study-specific  $ORs$ . To test the hypothesis  $H_0: \theta_1 = \dots = \theta_k$  vs.  $H_1: \text{at least two } \theta_i's \text{ are different}$ , we use Equation 13.47.

### Equation 13.47

#### Test of Homogeneity of Study-Specific ORs in Meta-Analysis

To test the hypothesis  $H_0: \theta_1 = \dots = \theta_k$  vs.  $H_1: \text{at least two } \theta_i's \text{ are different}$ , where  $\theta_i = \text{estimated log } OR$  in the  $i$ th study, we use the test statistic

$$Q_w = \sum_{i=1}^k w_i (y_i - \bar{y}_w)^2$$

as defined in Equation 13.46. It can be shown under  $H_0$  that  $Q_w \sim \chi_{k-1}^2$ . Hence, to obtain the  $p$ -value, we compute

$$p\text{-value} = Pr(\chi_{k-1}^2 > Q_w)$$

### Example 13.52

**Renal Disease** Test for the homogeneity of the  $ORs$  in Table 13.31.

#### Solution

From the solution to Example 13.51, we have  $Q_w = 10.276 \sim \chi_7^2$  under  $H_0$ . Because  $\chi_{7,.75}^2 = 9.04$  and  $\chi_{7,.90}^2 = 12.02$  and  $9.04 < 10.276 < 12.02$ , it follows that  $1 - .90 < p < 1 - .75$  or  $.10 < p < .25$ . Hence there is not significant heterogeneity among the study-specific  $ORs$ .

It is controversial among research workers whether a fixed- or random-effects model should be used when performing a meta-analysis. Under a fixed-effects model, the between-study variance ( $\Delta^2$ ) is ignored in computing the study weights and only the within-study variance is considered. Hence one uses  $w_i$  for the weights in Equation 13.46 instead of  $w_i^*$ . Some statisticians argue that if there is substantial variation among the study-specific *ORs*, then one should investigate the source of the heterogeneity (different study designs, etc.) and not report an overall pooled estimate of the *OR* as given in Equation 13.46. Others feel that between-study variation should always be considered in meta-analyses. Generally speaking, using a fixed-effects model results in tighter confidence limits and more significant results. However, note that the fixed-effects model and the random-effects model give different relative weights to the individual studies. A fixed-effects model only considers within-study variation. A random-effects model considers both between- and within-study variation. If the between-study variation is substantial relative to the within-study variation (as is sometimes the case), then larger studies will get proportionally more weight under a fixed-effects model than under a random-effects model. Hence the summary *ORs* under these two models may also differ. This is indeed the case in Table 13.31, where the relative weight of larger studies compared with smaller studies is greater for the fixed-effects model weights ( $w_i$ ) than for the random-effects model weights ( $w_i^*$ ). For example, for the data in Table 13.31, if one uses the  $w_i$  for weights instead of  $w_i^*$ , one obtains a point estimate for the overall *OR* [ $\exp(\hat{\mu})$ ] of 0.72 with 95% confidence limits of (0.50, 1.04) with *p*-value = .083 for testing  $H_0$ :  $OR = 1$  vs.  $H_1$ :  $OR \neq 1$ , compared with an *OR* of 0.69 with 95% confidence limits of (0.43, 1.11) for the random-effects model.

One drawback to the random-effects approach is that one cannot use studies with zero events in either treatment group. Under a fixed-effects model such studies get 0 weight. However, under a random-effects model such studies get nonzero weight if the between-study variance is greater than 0. This is problematic because the log *OR* is either  $+\infty$  or  $-\infty$  unless both groups have no events. We excluded one small study in our survey in Table 13.31 for this reason [19], where there were 11 gentamicin patients who experienced 0 events and 11 tobramycin patients who experienced two events. A reasonable compromise might be to check for significant heterogeneity among the study-specific *ORs* using Equation 13.47 and use the decision rule in Table 13.32.

**Table 13.32** Models used for meta-analysis

<i>p</i> -Value for heterogeneity	Type of model used
$\geq .5$	Use fixed-effects model
$.05 \leq p < .5$	Use random-effects model
$<.05$	Do not report pooled <i>OR</i> ; assess sources of heterogeneity

Meta-analysis methods can also be performed based on other effect measures (e.g., mean differences between treatment groups instead of *ORs*). However, this is beyond the scope of this text. For a complete description of meta-analysis, see Hedges and Olkin [20].

In this section, we have studied meta-analysis, a technique for formally combining results over more than one study to maximize precision in estimating parameters and to maximize power for testing hypotheses for particular research questions. We studied both a fixed-effects model, where the weight received by individual studies

is determined only by within-study variation, and a random-effects model, where both between- and within-study variation determine the weight. We also discussed a possible strategy for determining which of the two models, if either, to use in specific situations.

### REVIEW QUESTIONS 13F

- 1 What is the purpose of meta-analysis?
- 2 What is the difference between a fixed-effects model and a random-effects model for meta-analysis?
  - (a) How do the weights compare under these two models?

## 13.11 Equivalence Studies

### Introduction

In Chapter 10, we considered the estimation of sample size for studies in which the null hypothesis is that two treatments are equally effective vs. the alternative hypothesis that the effects of the two treatments are different from each other and effectiveness in each treatment group is expressed as a binomial proportion. These types of studies, which constitute the majority of clinical trials, are referred to as *superiority studies*. However, a newer type of study design has emerged in recent years in which the major goal is to show that two treatments are equivalent rather than that one is superior to the other. Consider the following example, presented by Makuch and Simon [21].

#### Example 13.53

**Cancer** Suppose we want to design a clinical trial to compare two surgical treatments for early-stage breast cancer. The treatments are simple mastectomy and a more conservative tumor resection. In this setting, it would be unethical to compare the experimental treatment with a placebo. Instead, two active treatments are compared with each other. The former treatment is the standard and yields a 5-year survival rate of 80%. The latter is an experimental treatment that is less debilitating than the standard. However, it will only be considered acceptable if it can be shown in some statistical sense to be no more than 10% inferior to the standard treatment in terms of 5-year survival. How can we test whether the experimental treatment is acceptable, and how can we estimate sample size for such a study?

#### Definition 13.18

---

The type of study in Example 13.53, where the goal is to show approximate equivalence of two experimental treatments, is called an **equivalence study**.

---

### Inference Based on Confidence-Interval Estimation

Suppose  $p_1$  is the survival rate for the standard treatment and  $p_2$  is the survival rate for the experimental treatment. The approach we will take is to determine a lower one-sided  $100\% \times (1 - \alpha)$  CI for  $p_1 - p_2$ . In Equation 13.1, we provide a two-sided CI for  $p_1 - p_2$ . The corresponding lower one-sided interval is given by

#### Equation 13.48

$$p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{1-\alpha} \sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}$$

where we have ignored the generally small continuity correction.

We will consider the treatments as equivalent if the upper bound of this one-sided CI does not exceed  $\delta$ , where  $\delta$  is a prespecified threshold.

**Example 13.54**

**Cancer** Suppose we have a clinical trial with 100 patients each on the standard treatment and on the experimental treatment. We find that 80% of patients on the standard treatment and 75% of the patients on the experimental treatment survive for 5 years. Can the treatments be considered equivalent if the threshold for equivalence is that the underlying survival rate for the experimental treatment is no more than 10% worse than for the standard treatment based on a one-sided 95% CI approach?

**Solution**

We construct a lower one-sided 95% CI for  $p_1 - p_2$ . From Equation 13.48, this is given by

$$\begin{aligned} p_1 - p_2 &< .80 - .75 + z_{.95} \sqrt{.80(.20) / 100 + .75(.25) / 100} \\ &= .05 + 1.645(.0589) \\ &= .05 + .097 = .147 \end{aligned}$$

The upper bound of the lower 95% CI exceeds 10%, so the treatments *cannot* be considered equivalent. Thus, although the observed survival rates are only 5% apart, the underlying rates may differ by as much as 15%, which implies the treatments cannot be considered equivalent.

### Sample-Size Estimation for Equivalence Studies

It seems clear from Example 13.54 that large sample sizes are needed to demonstrate equivalence. In some cases, depending on the threshold  $\delta$  specified for equivalence, the sample size needed may be considerably larger than for typical superiority studies. The approach we will take is to require a sample size large enough so that with high probability ( $1 - \beta$ ) the upper confidence limit in Equation 13.48 does not exceed  $\delta$ . Hence we want

**Equation 13.49**

$$\Pr \left[ \hat{p}_1 - \hat{p}_2 + z_{1-\alpha} \sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2} \leq \delta \right] = 1 - \beta$$

However, if we subtract  $p_1 - p_2$  from both sides of Equation 13.49, divide by  $\sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}$ , and then subtract  $z_{1-\alpha}$  from both sides, we obtain

$$\Pr \left[ \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}} \leq \frac{\delta - (p_1 - p_2)}{\sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}} - z_{1-\alpha} \right] = 1 - \beta$$

Under the hypothesis that the true difference in survival rates between treatment groups is  $p_1 - p_2$ , the random variable on the left side is approximately a standard normal deviate. Therefore, to satisfy this equation we have

**Equation 13.50**

$$\frac{\delta - (p_1 - p_2)}{\sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}} - z_{1-\alpha} = z_{1-\beta}$$

We now add  $z_{1-\alpha}$  to both sides of Equation 13.50 and divide by  $\delta - (p_1 - p_2)$  to obtain

$$\frac{1}{\sqrt{\hat{p}_1 \hat{q}_1 / n_1 + \hat{p}_2 \hat{q}_2 / n_2}} = \frac{z_{1-\alpha} + z_{1-\beta}}{\delta - (p_1 - p_2)}$$

If we assume the experimental treatment sample size ( $n_2$ ) is  $k$  times as large as the standard treatment sample size ( $n_1$ ), we obtain

$$\frac{\sqrt{n_1}}{\sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2 / k}} = \frac{z_{1-\alpha} + z_{1-\beta}}{\delta - (p_1 - p_2)}$$

Solving for  $n_1$  yields

**Equation 13.51**

$$n_1 = \frac{(\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2 / k)(z_{1-\alpha} + z_{1-\beta})^2}{[\delta - (p_1 - p_2)]^2}, n_2 = kn_1$$

We summarize these results as follows.

**Equation 13.52**
**Sample-Size Estimation for Equivalence Studies**

Suppose we want to establish equivalence between a standard treatment (treatment 1) and an experimental treatment (treatment 2), where  $p_1$  and  $p_2$  are treatment success rates in groups 1 and 2, respectively. The treatments are considered equivalent (in the sense that the experimental treatment is not substantially worse than the standard treatment) if the upper bound of a lower 100%  $\times$  (1 -  $\alpha$ ) CI for  $p_1 - p_2$  is  $\leq \delta$ . If we want to establish equivalence with a probability of  $1 - \beta$ , we require

$$n_1 = \frac{(p_1 q_1 + p_2 q_2 / k)(z_{1-\alpha} + z_{1-\beta})^2}{[\delta - (p_1 - p_2)]^2} \text{ subjects in group 1}$$

$n_2 = kn_1$  subjects in group 2 where  $k$  is specified in advance

**Example 13.55**

**Cancer** Estimate the required sample size for the study described in Example 13.54 if (1) we want a probability of 80% for establishing equivalence, (2) the sample sizes are the same in the two groups, (3) the underlying 5-year survival rate in both groups is 80%, (4) the threshold for equivalence is 10%, and (5) we establish equivalence based on an upper bound of a lower 95% CI.

**Solution**

We have  $p_1 = p_2 = .80$ ,  $q_1 = q_2 = .20$ ,  $k = 1$ ,  $\alpha = .05$ ,  $\beta = .20$ ,  $\delta = .10$ . Therefore,

$$\begin{aligned} n_1 &= \frac{.80(.20)(2)(z_{.95} + z_{.80})^2}{(.10)^2} \\ &= \frac{.32(1.645 + .84)^2}{.01} = 197.6 = n_2 \end{aligned}$$

Therefore, we require 198 subjects in each group to have an 80% probability of establishing equivalence under this design. This is larger than the sample size in Example 13.54, where we were unable to demonstrate equivalence.

In this section, we have considered methods of analysis and methods of sample-size estimation for equivalence studies (sometimes called *active-control studies*). An equivalence study is one in which we want to establish with high probability that the difference in effect between two treatment groups does not exceed some prespecified threshold with high probability ( $1 - \beta$ ). The threshold  $\delta$ , the probability  $1 - \beta$ , and the underlying success rates in each group need to be specified in advance.

When is it reasonable to consider an equivalence vs. a superiority study? Some people feel that a superiority study based on placebo control is always the design of choice to establish the efficacy of a treatment [22]. Others feel that if a standard therapy has already proven its effectiveness, then it would be unethical to withhold treatment from patients (e.g., by using a placebo as one of the treatment groups in a clinical trial to establish the efficacy of a new treatment for schizophrenia). These issues are discussed in more detail by Rothman and Michels [23].

### REVIEW QUESTIONS 13G

- 1 What is the difference between an equivalence study and a superiority study?
- 2 The drug ibuprofen is often used by patients with osteoarthritis to reduce inflammation and pain. Suppose ibuprofen is effective in 90% of patients. One possible side effect of the drug when taken for a long period of time is gastric bleeding. A new drug is proposed for patients with osteoarthritis. The goal is that the drug will be equivalent to ibuprofen in efficacy but with fewer side effects.
  - (a) To assess equivalence, a study is performed with 100 patients receiving each drug. Ninety of the ibuprofen patients and 86 of the new drug patients show efficacy from the treatment. Can the new drug be considered equivalent to ibuprofen if the criterion for equivalence is that ibuprofen is no more than 5% higher in efficacy than the new drug and a one-sided 90% CI is used to establish equivalence?
  - (b) Suppose a larger equivalence study is planned. How many patients need to be enrolled in each group if the assumptions in Review Question 13G.2a hold and we want a 90% chance of demonstrating equivalence?

## 13.12 The Cross-Over Design

### Example 13.56

**Sports Medicine** In Problem 8.89 we were introduced to Data Set TENNIS2.DAT, on the Companion Website. This was a clinical trial comparing Motrin vs. placebo for the treatment of tennis elbow. Each participant was randomized to receive either Motrin (group A) or placebo (group B) for a 3-week period. All participants then had a 2-week washout period during which they received no study medication. All participants were then “crossed-over” for a second 3-week period to receive the opposite study medication from that initially received. Participants in group A received 3 weeks of placebo while participants in group B received 3 weeks of Motrin. This type of design is called a *cross-over design*. How should we compare the efficacy of Motrin vs. placebo using this design?

### Definition 13.19

A **cross-over design** is a type of randomized clinical trial. In this design, each participant is randomized to either group A or group B. All participants in group A receive drug 1 in the first treatment period and drug 2 in the second treatment period. All

participants in group B receive drug 2 in the first treatment period and drug 1 in the second treatment period. Often there is a *washout* period between the two active drug periods during which they receive no study medication. The purpose of the washout period is to reduce the likelihood that study medication taken in the first period will have an effect that carries over to the next period.

**Definition 13.20**

A **washout period** in a cross-over design is a period between active drug periods, during which subjects receive no study medication.

**Definition 13.21**

A **carry-over effect** in a cross-over design is when the effects of one or both study medications taken during the first active drug period have a residual biological effect during the second active drug period.

The cross-over design described in Definition 13.19 is actually a two-period cross-over design. There are also cross-over designs with more than two periods and/or more than two treatments being compared. These latter designs are beyond the scope of this text. See Fleiss [24] for a discussion of these designs.

### Assessment of Treatment Effects

A 6-point scale for pain relief was used in the study. At the end of each active treatment period, the participants were asked to rate their degree of pain relative to baseline—that is, the beginning of the study before either active treatment period. The rating scale was 1 if worse, 2 if unchanged, 3 if slightly improved (25%), 4 if moderately improved (50%), 5 if mostly improved (75%), and 6 if completely improved (100%). We want to compare the degree of pain relief for participants while on Motrin with the degree of pain relief while on placebo. Let  $x_{ijk}$  = the pain relief rating for the  $j$ th subject in the  $i$ th group during the  $k$ th period, where  $i$  = group (1 = group A, 2 = group B),  $j$  = subject ( $j = 1, \dots, n_1$  if subject is in group A,  $j = 1, \dots, n_2$  if subject is in group B),  $k$  = period (1 = first period, 2 = second period). For the  $j$ th patient in group A, the measure of drug efficacy is  $d_{1j} = x_{1j1} - x_{1j2}$ , whereas for the  $j$ th patient in group B, the measure of drug efficacy is  $d_{2j} = x_{2j2} - x_{2j1}$ . In each case, a large number indicates that the patient experiences less pain while on Motrin than on placebo. The summary measure of efficacy for patients in group A is, therefore,

$$\bar{d}_1 = \sum_{j=1}^{n_1} d_{1j} / n_1$$

and for patients in group B it is

$$\bar{d}_2 = \sum_{j=1}^{n_2} d_{2j} / n_2$$

The overall measure of drug efficacy is

**Equation 13.53**

$$\bar{d} = \frac{1}{2}(\bar{d}_1 + \bar{d}_2)$$

To compute the standard error of  $\bar{d}$ , we assume that the underlying variance of the within-subject differences in group A and group B are the same and estimate this variance ( $\sigma_d^2$ ) by the pooled estimate

**Equation 13.54**

$$s_{d,\text{pooled}}^2 = \frac{\sum_{j=1}^{n_1} (d_{1j} - \bar{d}_1)^2 + \sum_{j=1}^{n_2} (d_{2j} - \bar{d}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_{d_1}^2 + (n_2 - 1)s_{d_2}^2}{n_1 + n_2 - 2}$$

Therefore,

$$\begin{aligned} \text{Var}(\bar{d}) &= \frac{1}{4} [\text{Var}(\bar{d}_1) + \text{Var}(\bar{d}_2)] \\ &= \frac{1}{4} \left( \frac{\sigma_d^2}{n_1} + \frac{\sigma_d^2}{n_2} \right) \\ &= \frac{\sigma_d^2}{4} \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \end{aligned}$$

which is estimated by

$$\frac{s_{d,\text{pooled}}^2}{4} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

with  $n_1 + n_2 - 2$  df. The standard error of  $\bar{d}$  is thus

$$se(\bar{d}) = \sqrt{\frac{s_{d,\text{pooled}}^2}{4} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = \frac{s_{d,\text{pooled}}}{2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

This leads to the following test procedure for assessing the overall treatment effect in cross-over designs.

**Equation 13.55**

#### The Cross-Over Design—Assessment of Overall Treatment Effects

Let  $x_{ijk}$  represent the score for the  $j$ th patient in the  $i$ th group during the  $k$ th period for patients entered into a study using a cross-over design,  $i = 1, 2$ ;  $j = 1, \dots, n_i$ ;  $k = 1, 2$ . Suppose patients in group A receive treatment 1 in period 1 and treatment 2 in period 2, and patients in group B receive treatment 2 in period 1 and treatment 1 in period 2. If we assume that no carry-over effect is present, then we use the following procedure to assess overall treatment efficacy:

(1) We compute

$$\bar{d} = \text{overall estimate of treatment efficacy} = \frac{1}{2} (\bar{d}_1 + \bar{d}_2)$$

$$\text{where } \bar{d}_1 = \sum_{j=1}^{n_1} d_{1j} / n_1$$

$$\bar{d}_2 = \sum_{j=1}^{n_2} d_{2j} / n_2$$

$$d_{1j} = x_{1j1} - x_{1j2}, j = 1, \dots, n_1$$

$$d_{2j} = x_{2j2} - x_{2j1}, j = 1, \dots, n_2$$

(2) The standard error of  $\bar{d}$  is estimated by

$$\sqrt{\frac{s_{d,\text{pooled}}^2}{4} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = \frac{s_{d,\text{pooled}}}{2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$\begin{aligned}s_{d,\text{pooled}}^2 &= \frac{(n_1 - 1)s_{d_1}^2 + (n_2 - 1)s_{d_2}^2}{n_1 + n_2 - 2} \\ s_{d_1}^2 &= \sum_{j=1}^{n_1} (d_{1j} - \bar{d}_1)^2 / (n_1 - 1) \\ s_{d_2}^2 &= \sum_{j=1}^{n_2} (d_{2j} - \bar{d}_2)^2 / (n_2 - 1)\end{aligned}$$

(3) If  $\Delta = \text{underlying mean treatment efficacy}$ , then to test the hypothesis  $H_0: \Delta = 0$  vs.  $H_1: \Delta \neq 0$  using a two-sided level  $\alpha$  significance test, compute the test statistic

$$t = \frac{\bar{d}}{\sqrt{\frac{s_{d,\text{pooled}}^2}{4} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

(4) If  $t > t_{n_1+n_2-2, 1-\alpha/2}$  or  $t < t_{n_1+n_2-2, \alpha/2}$ , then reject  $H_0$ .

If  $t_{n_1+n_2-2, \alpha/2} \leq t \leq t_{n_1+n_2-2, 1-\alpha/2}$ , then accept  $H_0$ .

(5) The exact  $p$ -value is given by

$2 \times \text{area to the left of } t \text{ under a } t_{n_1+n_2-2} \text{ distribution if } t \leq 0$

or

$2 \times \text{area to the right of } t \text{ under a } t_{n_1+n_2-2} \text{ distribution if } t > 0$

(6) A  $100\% \times (1 - \alpha)$  CI for the underlying treatment effect  $\Delta$  is given by

$$\bar{d} \pm t_{n_1+n_2-2, 1-\alpha/2} \sqrt{\frac{s_{d,\text{pooled}}^2}{4} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

### Example 13.57

**Sports Medicine** Test for whether overall degree of pain as compared with baseline is different for patients while on Motrin than while on placebo. Estimate a 95% CI for improvement in degree of pain while on Motrin vs. placebo.

### Solution

There are 88 participants in the data set, 44 in group A and 44 in group B. However, 2 participants in each group had missing pain scores in one or both periods. Hence, 42 participants are available for analysis in each group. We first present the mean pain score vs. baseline for patients in each group during each period as well as the mean difference in pain scores (Motrin – placebo) and the average pain relief score over the two periods (see Table 13.33).

The overall measure of drug efficacy is

$$\bar{d} = \frac{0.071 + 1.357}{2} = 0.714$$

**Table 13.33 Summary statistics of overall impression of drug efficacy compared with baseline ( $n = 84$ )**

	Group							
	A				B			
	Motrin	Placebo	Difference <sup>a</sup>	Average <sup>b</sup>	Motrin	Placebo	Difference <sup>a</sup>	Average <sup>b</sup>
Mean	3.833	3.762	0.071	3.798	4.214	2.857	1.357	3.536
sd	1.188	1.574	1.813	1.060	1.353	1.160	1.376	1.056
n	42	42	42	42	42	42	42	42

Note: The data are obtained from Data Set TENNIS2.DAT (on the Companion Website) using variable 22 for overall impression of drug efficacy during period 1 and variable 43 for overall impression of drug efficacy during period 2.

<sup>a</sup>Pain score on Motrin – pain score on placebo

<sup>b</sup>Average of (pain score on Motrin, and pain score on placebo)

To compute the standard error of  $\bar{d}$ , we first compute the pooled variance estimate given by

$$\begin{aligned}s_{d,\text{pooled}}^2 &= \left[ (n_1 - 1)s_{d_1}^2 + (n_2 - 1)s_{d_2}^2 \right] / (n_1 + n_2 - 2) \\ &= \frac{41(1.813)^2 + 41(1.376)^2}{82} = 2.590\end{aligned}$$

The standard error of  $\bar{d}$  is

$$se(\bar{d}) = \sqrt{\frac{2.590}{4} \left( \frac{1}{42} + \frac{1}{42} \right)} = 0.176$$

The test statistic is

$$t = \frac{0.714}{0.176} = 4.07 \sim t_{82} \text{ under } H_0$$

The exact  $p$ -value =  $2 \times Pr(t_{82} > 4.07)$ . Because  $4.07 > t_{60,9995} = 3.460 > t_{82,9995}$ , it follows that  $p < 2 \times (1 - .9995)$  or  $p < .001$ . Thus there is a highly significant difference in the mean pain score on Motrin vs. the mean pain score on placebo, with patients experiencing less pain when on Motrin.

A 95% CI for the treatment benefit  $\Delta$  is

$$\begin{aligned}\bar{d} \pm t_{n_1+n_2-2,975} se(\bar{d}) \\ = 0.714 \pm t_{82,975} (0.176)\end{aligned}$$

Using Excel, we estimate  $t_{82,975} = 1.989$ . Therefore the 95% CI for  $\Delta = 0.714 \pm 1.989(0.176) = (0.365, 1.063)$ . Thus the treatment benefit is likely to be between 1/3 of a unit and 1 unit on the pain scale.

## Assessment of Carry-Over Effects

In the preceding section, when we computed the overall estimate of the treatment effect in Equation 13.55, we assumed there was no carry-over effect. A carry-over effect is present when the true treatment effect is different for subjects in group A than for subjects in group B.

**Example 13.58**

**Sports Medicine** Suppose Motrin is very effective in relieving pain from tennis elbow and the pain relief is long-lasting (relief continues even after the patients stop taking the medication, whereas placebo has no effect on pain). In this case, the difference between Motrin- and placebo-treated patients is greater during the first treatment period than during the second treatment period. Another way of stating this is that the difference between Motrin and placebo is smaller for patients in group A than for patients in group B. This is because of the carry-over effect of Motrin taken in the first period into the second period. How can we identify such carry-over effects?

Notice that in Example 13.58, if there is a carry-over effect, then the average response for patients in group A over the two periods will be greater than for patients in group B. This forms the basis for our test for identifying carry-over effects.

**Equation 13.56****Assessment of Carry-Over Effects in Cross-Over Studies**

Let  $x_{ijk}$  represent the score for the  $j$ th patient in the  $i$ th group during the  $k$ th period. Define  $\bar{x}_{ij} = (x_{ij1} + x_{ij2})/2$  = average response over both periods for the  $j$ th patient in the  $i$ th group and  $\bar{x}_i = \sum_{j=1}^{n_i} \bar{x}_{ij}/n_i$  = average response over all patients in the  $i$ th group over both treatment periods. We assume that  $\bar{x}_{ij} \sim N(\mu_i, \sigma^2)$ ,  $i = 1, 2; j = 1, \dots, n_i$ . To test the hypothesis  $H_0: \mu_1 = \mu_2$  vs.  $H_1: \mu_1 \neq \mu_2$ :

- (1) We compute the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and

$$s_i^2 = \sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_i)^2 / (n_i - 1), i = 1, 2$$

- (2) If  $t > t_{n_1+n_2-2, 1-\alpha/2}$  or  $t < t_{n_1+n_2-2, \alpha/2}$  we reject  $H_0$ .  
If  $t_{n_1+n_2-2, \alpha/2} \leq t \leq t_{n_1+n_2-2, 1-\alpha/2}$ , we accept  $H_0$ .

- (3) The exact  $p$ -value =  $2 \times Pr(t_{n_1+n_2-2} > t)$  if  $t > 0$   
 $= 2 \times Pr(t_{n_1+n_2-2} < t)$  if  $t \leq 0$

**Example 13.59**

Assess whether there are any carry-over effects, using the tennis-elbow data in Example 13.58.

**Solution**

We refer to Table 13.33 and note that

$$\bar{x}_1 = 3.798, s_1 = 1.060, n_1 = 42$$

$$\bar{x}_2 = 3.536, s_2 = 1.056, n_2 = 42$$

$$s^2 = \frac{41(1.060)^2 + 41(1.056)^2}{82} = 1.119$$

Therefore, the test statistic is given by

$$\begin{aligned} t &= \frac{3.798 - 3.536}{\sqrt{1.119 \left( \frac{1}{42} + \frac{1}{42} \right)}} \\ &= \frac{0.262}{0.231} = 1.135 \sim t_{82} \text{ under } H_0 \end{aligned}$$

Because  $t > 1.046 = t_{60,.85} > t_{82,.85}$ , it follows that  $p < 2 \times (1 - .85) = .30$ . Because  $t < 1.289 = t_{120,.90} < t_{82,.90}$ , it follows that  $p > 2 \times (1 - .90) = .20$ . Therefore,  $.20 < p < .30$ , and there is no significant carry-over effect. We can also gain some insight into possible carry-over effects by referring to Table 13.33. We see the treatment benefit during period 1 is  $3.833 - 2.857 = 0.976$ , whereas the treatment benefit during period 2 is  $4.214 - 3.762 = 0.452$ . Thus, there is some treatment benefit during each period. The degree of benefit is larger in period 1 but is not significantly larger. In general, the power of the test to detect carry-over effects is not great. Also, the effect of possible carry-over effects on the ability to identify overall treatment benefit can be large. Therefore, some authors [25] recommend that the  $p$ -value for declaring significant carry-over effects be set at .10 rather than the usual .05. Even with this more relaxed criterion for achieving statistical significance, we still don't declare a significant carry-over effect with the tennis-elbow data.

Another important insight into the data is revealed by looking for period effects. For example, in Table 13.33 the effect of period 2 vs. period 1 is  $4.214 - 3.833 = 0.381$  while subjects were on Motrin and  $3.762 - 2.857 = 0.905$  while subjects were on placebo. Thus subjects are experiencing less pain in period 2 compared with period 1 regardless of which medication they are taking.

What can we do if we identify a significant carry-over effect using Equation 13.56? In this case, the second-period data are not useful to us because they provide a biased estimate of treatment effects, particularly for subjects who were on active drug in the first period and on placebo in the second period, and we must base our comparison of treatment efficacy on first-period data only. We can use an ordinary two-sample  $t$  test for independent samples based on the first-period data. This test usually has less power than the cross-over efficacy test in Equation 13.55, or requires a greater sample size to achieve a given level of power (see Example 13.60).

### Sample-Size Estimation for Cross-Over Studies

A major advantage of cross-over studies is that they usually require many fewer subjects than the usual randomized clinical trials (which have only 1 period), if no carry-over effect is present. The sample-size formula is as follows.

#### Equation 13.57

##### Sample-Size Estimation for Cross-Over Studies

Suppose we want to test the hypothesis  $H_0: \Delta = 0$  vs.  $H_1: \Delta \neq 0$  using a two-sided test with significance level  $\alpha$ , where  $\Delta$  = underlying treatment benefit for treatment 1 vs. treatment 2 using a cross-over study. If we require a power of  $1 - \beta$ , and we expect to randomize an equal number of subjects to each group (group A receives treatment 1 in period 1 and treatment 2 in period 2; group B receives treatment 2 in period 1 and treatment 1 in period 2), then the appropriate sample size per group =  $n$ , where

$$n = \frac{\sigma_d^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{2\Delta^2}$$

and  $\sigma_d^2$  = variance of difference scores = variance of (response on treatment 1 – response on treatment 2).

*This sample-size formula is only applicable if no carry-over effects (as defined in Definition 13.21) are present.*

### Example 13.60

**Hypertension** Suppose we want to study the effect of postmenopausal hormone (PMH) use on level of diastolic blood pressure (DBP). We intend to enroll  $n$  postmenopausal women per group. Women in group A will get PMH pills in period 1 (4 weeks) and placebo pills in period 2 (4 weeks). Women in group B will get placebo pills in period 1 and PMH pills in period 2. There will be a 2-week washout period between each 4-week active-treatment period. Women will have their blood pressure measured at the end of each active-treatment period based on a mean of 3 readings at a single visit. If we anticipate a 2-mm Hg treatment benefit and the within-subject variance of the difference in mean DBP between the two periods is estimated to be 31, based on pilot-study results, and we require 80% power, then how many participants need to be enrolled in each group?

### Solution

We have  $\sigma_d^2 = 31$ ,  $\alpha = .05$ ,  $\beta = .20$ ,  $\Delta = 2$ . Thus, from Equation 13.57, we have  $z_{1-\alpha/2} = z_{.975} = 1.96$ ,  $z_{1-\beta} = z_{.80} = 0.84$  and

$$\begin{aligned} n &= \frac{31(1.96 + 0.84)^2}{2(4)} \\ &= \frac{31(7.84)}{8} = 30.4 \end{aligned}$$

Thus we need to enroll 31 participants per group, or 62 participants overall, to achieve an 80% power using this design, if no carry-over effect is present.

An alternative design for such a study is the so-called parallel-group design, in which we randomize participants to either PMH or placebo, measure their DBP at baseline and at the end of 4 weeks of follow-up, and base the measure of efficacy for an individual patient on (mean DBP at follow-up – mean DBP at baseline). The sample size needed for such a study is given in Equation 8.27 by

$$\begin{aligned} n &= \text{sample size per group} = \frac{2\sigma_d^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} \\ &= 4 \times \text{sample size per group for cross-over study} \\ &= 4(30.4) = 121.5 = 122 \text{ participants per group or 244 participants overall} \end{aligned}$$

$\sigma_d^2$  is the within-subject variance of the difference in mean DBP (i.e., mean DBP follow-up – mean DBP baseline) = 31. Clearly, the cross-over design is much more efficient, *if the assumption of no carry-over effects is viable*. It is important in planning cross-over studies to include a baseline measurement prior to the active-treatment period. Although the baseline measurement is usually not useful in analyzing cross-over studies, it can be very useful if it is subsequently found that a carry-over effect is present. In this case, one could use a parallel-group design based on the difference between period 1 scores and baseline as the outcome measure, rather than simply

the period 1 scores. The difference score generally has less variability than the period 1 score because it represents within-person variability rather than both between-person and within-person variability as represented by the period 1 score.

In this section, we have examined cross-over designs. Under a cross-over design, each subject receives both treatments but at different times. Randomization determines treatment order for individual subjects. A cross-over design can be more efficient (i.e., require fewer subjects) than a traditional parallel-group design if no carry-over effects are present but will be underpowered if unanticipated carry-over effects are present because the second-period data cannot be validly used. In the latter case, the power can be somewhat improved if a baseline score is obtained before subjects receive either treatment.

It is useful to consider the types of studies in which a cross-over design may be appropriate. In particular, studies based on objective endpoints such as blood pressure, in which the anticipated period of drug efficacy occurs over a short time (i.e., weeks rather than years) and is not long-lasting after drug is withdrawn, are best suited for a cross-over design. However, most phase III clinical trials (i.e., definitive studies used by the FDA as a basis for establishing drug efficacy for new pharmaceutical products or existing products being tested for a new indication) are long-term studies that violate one or more of the preceding principles. Thus, in general, phase III clinical trials usually use the more traditional parallel-group design.

### REVIEW QUESTIONS 13H

- 1 What is the difference between a cross-over design and a parallel-group design?
- 2 What is a carry-over effect?
- 3 Suppose there is no carry-over effect. Which design requires a larger sample size, a cross-over design or a parallel-group design?

## 13.13 Clustered Binary Data

### Introduction

The two-sample test for the comparison of binomial proportions, discussed in Section 10.2, is one of the most frequently cited statistical procedures in applied research. An important assumption underlying this methodology is that the observations within the respective samples are statistically independent.

#### Example 13.61

**Infectious Disease, Dermatology** Rowe et al. [26] reported on a clinical trial of topically applied 3% vidarbine vs. placebo in treating recurrent herpes labialis. During the medication phase of the trial, the characteristics of 53 lesions observed on 31 patients receiving vidarbine were compared with the characteristics of 69 lesions observed on 39 patients receiving placebo. A question of interest is whether the proportion of lesions showing significant shrinkage in the two groups is the same after 7 days. This requires development of a test procedure that adjusts for dependencies in response among lesions observed on the same patient.

### Hypothesis Testing

We assume the sample data arise from two groups of individuals,  $n_1$  individuals in group 1 and  $n_2$  individuals in group 2. Suppose that individual  $j$  in group  $i$  ( $i = 1, 2$ )

contributes  $m_{ij}$  observations to the analysis,  $j = 1, 2, \dots, n_i$  and  $M_i = \sum_{j=1}^{n_i} m_{ij}$  denotes the total number of observations in group  $i$ , each classified as either a success or a failure. Let  $a_{ij}$  denote the observed number of successes for individual  $j$  in group  $i$ , and define  $A_i = \sum_{j=1}^{n_i} a_{ij}$ . Then the overall proportion of successes in group  $i$  may be denoted by  $\hat{p}_i = A_i / M_i = \sum_{j=1}^{n_i} m_{ij} \hat{p}_{ij} / M_i$ , where  $\hat{p}_{ij} = a_{ij} / m_{ij}$  denotes the observed success rate for individual  $j$  in group  $i$ . We further denote the total number of individuals as  $N = n_1 + n_2$  and the total number of observations as  $M = M_1 + M_2$ .

Let  $p_i$  denote the underlying success rate among observations in group  $i$ ,  $i = 1, 2$ . Then we want to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ , assuming the samples are large enough that the normal approximation to the binomial distribution is valid.

An estimate of the degree of clustering within individuals is given by the intraclass correlation for clustered binary data. This is computed in a similar manner as for normally distributed data as given in Section 12.9. The mean square errors between and within individuals, respectively, are given in this case by

$$\begin{aligned} MSB &= \sum_{i=1}^2 \sum_{j=1}^{n_i} m_{ij} (\hat{p}_{ij} - \hat{p}_i)^2 / (N - 2) \\ MSW &= \sum_{i=1}^2 \sum_{j=1}^{n_i} a_{ij} (1 - \hat{p}_{ij}) / (M - N) \end{aligned}$$

The resulting estimate of intraclass correlation is given by

$$\hat{\rho} = (MSB - MSW) / [MSB + (m_A - 1)MSW]$$

where

$$m_A = \left[ M - \sum_{i=1}^2 \left( \sum_{j=1}^{n_i} m_{ij}^2 / M_i \right) \right] / (N - 2)$$

The clustering correction factor in group  $i$  may now be defined as  $C_i = \sum_{j=1}^{n_i} m_{ij} C_{ij} / M_i$ , where  $C_{ij} = 1 + (m_{ij} - 1)\hat{\rho}$ .

The clustering correction factor is sometimes called the *design effect*. Notice that if the intraclass correlation coefficient is 0, then no clustering is present and the design effects in the two samples are each 1. If the intraclass correlation coefficient is  $> 0$ , then the design effects are  $> 1$ . The design effects in the two samples ( $C_1, C_2$ ) are used as correction factors to modify the standard test statistic comparing two binomial proportions (Equation 10.3) for clustering effects. We have the following test procedure.

### Equation 13.58

#### Two-Sample Test for Binomial Proportions (Clustered Data Case)

Suppose we have two samples consisting of  $n_1$  and  $n_2$  individuals, respectively, where the  $j$ th individual in the  $i$ th group contributes  $m_{ij}$  observations to the analysis, of which  $a_{ij}$  are successes. To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ ,

- (1) We compute the test statistic

$$z = \left[ |\hat{p}_1 - \hat{p}_2| - \left( \frac{C_1}{2M_1} + \frac{C_2}{2M_2} \right) \right] / \sqrt{\hat{p}\hat{q}(C_1 / M_1 + C_2 / M_2)}$$

where

$$\begin{aligned}
 \hat{p}_{ij} &= a_{ij}/m_{ij} \\
 \hat{p}_i &= \sum_{j=1}^{n_i} a_{ij} / \sum_{j=1}^{n_i} m_{ij} = \sum_{j=1}^{n_i} m_{ij} \hat{p}_{ij} / \sum_{j=1}^{n_i} m_{ij} \\
 \hat{p} &= \sum_{i=1}^2 \sum_{j=1}^{n_i} a_{ij} / \sum_{i=1}^2 \sum_{j=1}^{n_i} m_{ij} = \sum_{i=1}^2 M_i \hat{p}_i / \sum_{i=1}^2 M_i, \hat{q} = 1 - \hat{p} \\
 M_i &= \sum_{j=1}^{n_i} m_{ij} \\
 C_i &= \sum_{j=1}^{n_i} m_{ij} C_{ij} / M_i \\
 C_{ij} &= 1 + (m_{ij} - 1) \hat{\rho} \\
 \hat{\rho} &= (MSB - MSW) / [MSB + (m_A - 1) MSW] \\
 MSB &= \sum_{i=1}^2 \sum_{j=1}^{n_i} m_{ij} (\hat{p}_{ij} - \hat{p}_i)^2 / (N - 2) \\
 MSW &= \sum_{i=1}^2 \sum_{j=1}^{n_i} a_{ij} (1 - \hat{p}_{ij}) / (M - N) \\
 m_A &= \left[ M - \sum_{i=1}^2 \left( \sum_{j=1}^{n_i} m_{ij}^2 / M_i \right) \right] / (N - 2) \\
 N &= \sum_{i=1}^2 n_i
 \end{aligned}$$

- (2) To test for significance, we reject  $H_0$  if  $|z| > z_{1-\alpha/2}$ , where  $z_{1-\alpha/2}$  is the upper  $\alpha/2$  percentile of a standard normal distribution.
- (3) An approximate  $100\% \times (1 - \alpha)$  CI for  $p_1 - p_2$  is given by

$$\begin{aligned}
 \hat{p}_1 - \hat{p}_2 - [C_1/(2M_1) + C_2/(2M_2)] \pm z_{1-\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 C_1/M_1 + \hat{p}_2 \hat{q}_2 C_2/M_2} &\text{ if } \hat{p}_1 > \hat{p}_2 \\
 \hat{p}_1 - \hat{p}_2 + [C_1/(2M_1) + C_2/(2M_2)] \pm z_{1-\alpha/2} \sqrt{\hat{p}_1 \hat{q}_1 C_1/M_1 + \hat{p}_2 \hat{q}_2 C_2/M_2} &\text{ if } \hat{p}_1 \leq \hat{p}_2
 \end{aligned}$$

- (4) This test should only be used if  $M_1 \hat{p} \hat{q} / C_1 \geq 5$  and  $M_2 \hat{p} \hat{q} / C_2 \geq 5$ .

### Example 13.62

**Dentistry** A longitudinal study of caries lesions on the exposed roots of teeth was reported in the literature [27]. Forty chronically ill subjects were followed for development of root lesions over a 1-year period. The data are given in Table 13.34. Assess whether the male patients had a higher incidence of surfaces with root lesions than did female patients over this time period.

### Solution

We note that 6 of 27 (22.2%) surfaces among 11 male patients developed root lesions compared with 6 of 99 (6.1%) surfaces among 29 female patients. The standard normal deviate test statistic (Equation 10.3) for comparing these two proportions is given by

$$\begin{aligned}
z &= \left[ |\hat{p}_1 - \hat{p}_2| - \left( \frac{1}{2M_1} + \frac{1}{2M_2} \right) \right] / \sqrt{\hat{p}\hat{q}(1/M_1 + 1/M_2)} \\
&= \left[ |.2222 - .0606| - \left[ \frac{1}{2(27)} + \frac{1}{2(99)} \right] \right] / \sqrt{(12/126)(114/126)(1/27 + 1/99)} \\
&= .1380 / .0637 = 2.166 \sim N(0, 1) \text{ under } H_0
\end{aligned}$$

which yields a  $p$ -value of  $2 \times [1 - \Phi(2.166)] = .030$ . However, application of this test procedure ignores the dependency of responses on different surfaces within the same patient. To incorporate this dependency, we use the test procedure in Equation 13.58. We must compute the intraclass correlation  $\hat{\rho}$ , which is given as follows:

$$\hat{\rho} = (MSB - MSW) / [MSB + (m_A - 1)MSW]$$

where

$$\begin{aligned}
MSB &= [4(0/4 - .2222)^2 + \dots + 2(0/2 - .2222)^2 + 2(1/2 - .0606)^2 + \dots + 2(0/2 - .0606)^2] / 38 \\
&= 6.170 / 38 = 0.1624 \\
MSW &= [0(1 - 0/4) + \dots + 0(1 - 0/2)] / (27 + 99 - 40) \\
&= 4.133 / 86 = 0.0481 \\
m_A &= [126 - (77/27 + 403/99)] / (40 - 2) \\
&= (126 - 6.923) / 38 = 119.077 / 38 = 3.134 \\
\hat{\rho} &= (0.1624 - 0.0481) / [0.1624 + (3.134 - 1)0.0481] \\
&= 0.1143 / 0.2649 = .431
\end{aligned}$$

To compute the adjusted test statistic, we need to estimate  $C_1$ ,  $C_2$ , where

$$\begin{aligned}
C_1 &= \frac{2.294(4) + \dots + 1.431(2)}{4 + \dots + 2} \\
&= \frac{48.573}{27} = 1.799 \\
C_2 &= \frac{1.431(2) + \dots + 1.431(2)}{2 + \dots + 2} \\
&= \frac{230.166}{99} = 2.325
\end{aligned}$$

Thus we have the adjusted test statistic

$$\begin{aligned}
z &= \frac{|.2222 - .0606| - \left[ \frac{1.799}{2(27)} + \frac{2.325}{2(99)} \right]}{\sqrt{(12/126)(114/126)(1.799/27 + 2.325/99)}} \\
&= \frac{.1166}{\sqrt{.08617(.09011)}} \\
&= \frac{.1166}{.0881} = 1.323
\end{aligned}$$

This yields a two-tailed  $p$ -value of  $2 \times [1 - \Phi(1.323)] = 2 \times (.0930) = .186$ , which is not statistically significant. Thus the significance level attained from an analysis that ignores the dependency among surfaces within the same patient ( $p = .030$ )

**Table 13.34** Longitudinal data on development of caries lesions over a 1-year period

	ID	Age	Sex	Lesions	Surfaces
Males	1	71	M	0	4
	5	70	M	1	1
	6	65	M	2	2
	7	53	M	0	2
	8	71	M	2	4
	11	74	M	0	3
	15	81	M	0	3
	18	64	M	0	3
	30	40	M	0	1
	32	78	M	1	2
	35	79	M	0	2
Total	11			6	27
Females	2	80	F	1	2
	3	83	F	1	6
	4	86	F	0	8
	9	69	F	1	5
	10	59	F	0	4
	12	88	F	0	4
	13	36	F	1	2
	14	60	F	0	4
	16	71	F	0	4
	17	80	F	0	4
	19	59	F	0	6
	20	65	F	0	2
	21	85	F	0	4
	22	72	F	0	4
	23	58	F	0	2
	24	65	F	0	3
	25	59	F	0	2
	26	45	F	0	2
	27	71	F	0	4
	28	82	F	2	2
	29	48	F	0	2
	31	67	F	0	2
	33	80	F	0	2
	34	69	F	0	4
	36	85	F	0	4
	37	77	F	0	4
	38	71	F	0	3
	39	85	F	0	2
	40	52	F	0	2
Total	29			6	99

is considerably lower than the true significance level attained from the procedure in Equation 13.58, which accounts for the dependence. The existence of such dependence is biologically sensible, given the common factors that affect the surfaces within a mouth, such as nutrition, saliva production, and dietary habits [28].

Using Equation 13.58, we can also develop a 95% CI for  $p_1 - p_2$  = true difference in 1-year incidence of root caries between males and females, which is given as follows:

$$\begin{aligned} & .1166 \pm 1.96 \sqrt{\frac{.2222(.7778)(1.799)}{27} + \frac{.0606(.9394)(2.325)}{99}} \\ & = .1166 \pm 1.96(.1134) \\ & = .1166 \pm .2222 = (-.106, .339) \end{aligned}$$

Note that the inference procedure in Equation 13.58 reduces to the standard two-sample inference procedure when  $\hat{p} = 0$  (i.e., Equation 10.3), or when  $m_{ij} = 1, j = 1, 2, \dots, n_i; i = 1, 2$ . Finally, note that if all individuals in each group contribute exactly the same number of sites ( $m$ ), then the test procedure in Equation 13.58 reduces as follows:

### Equation 13.59

#### Two-Sample Test for Binomial Proportions (Equal Number of Sites per Individual)

If each individual in each of two groups contributes  $m$  observations, then to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ , perform the following procedure:

(1) Let

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{[1+(m-1)\hat{p}]}{2} \left( \frac{1}{M_1} + \frac{1}{M_2} \right)}{\sqrt{\hat{p}\hat{q}(1/M_1 + 1/M_2)}} \times \frac{1}{\sqrt{1+(m-1)\hat{p}}}$$

where  $M_i = n_i m$ ,  $i = 1, 2$ , and  $\hat{p}_i$ ,  $i = 1, 2$ , and  $\hat{p}$  are defined in Equation 13.58.

- (2) To test for significance, we reject  $H_0$  if  $|z| > z_{1-\alpha/2}$ , where  $z_{1-\alpha/2}$  is the upper  $\alpha/2$  percentile of a standard normal distribution.
- (3) An approximate 100%  $\times (1 - \alpha)$  CI for  $p_1 - p_2$  is given by

$$\begin{cases} \hat{p}_1 - \hat{p}_2 - \frac{[1+(m-1)\hat{p}]}{2} (1/M_1 + 1/M_2) \pm z_{1-\alpha/2} \sqrt{[1+(m-1)\hat{p}] (\hat{p}_1 \hat{q}_1 / M_1 + \hat{p}_2 \hat{q}_2 / M_2)} \\ \text{if } \hat{p}_1 > \hat{p}_2 \\ \hat{p}_1 - \hat{p}_2 + \frac{[1+(m-1)\hat{p}]}{2} (1/M_1 + 1/M_2) \pm z_{1-\alpha/2} \sqrt{[1+(m-1)\hat{p}] (\hat{p}_1 \hat{q}_1 / M_1 + \hat{p}_2 \hat{q}_2 / M_2)} \\ \text{if } \hat{p}_1 \leq \hat{p}_2 \end{cases}$$

- (4) This test should only be used if  $M_1 \hat{p} \hat{q} / [1+(m-1)\hat{p}] \geq 5$  and  $M_2 \hat{p} \hat{q} / [1+(m-1)\hat{p}] \geq 5$ .

### Power and Sample Size Estimation for Clustered Binary Data

Suppose we want to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ . If we assume there is independence among the observations within an individual and there are  $n_1$  observations in group 1 and  $n_2$  observations in group 2, then the power is given by  $\Phi(z_{1-\beta})$ , where

**Equation 13.60**

$$z_{1-\beta} = |p_1 - p_2| / \sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2} - z_{1-\alpha/2} \sqrt{\bar{p} \bar{q} (1/n_1 + 1/n_2)}$$

(Also see Equation 10.15.) In the case of clustered binary data, we replace  $n_1$  and  $n_2$  by the effective number of independent observations per group, or

**Equation 13.61**

$$n_i = M_i / C_i, i = 1, 2$$

where  $C_i$  is defined in Equation 13.58. To compute sample size, we specify  $1 - \beta$  and solve for  $n_1$  and  $n_2$  as a function of  $1 - \beta$ . The results are summarized as follows.

**Equation 13.62**
**Power and Sample Size Estimation for Comparing Binomial Proportions Obtained from Clustered Binary Data**

Suppose we wish to test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$ . If we intend to use a two-sided test with significance level  $\alpha$  and have available  $n_i$  individuals from the  $i$ th group,  $i = 1, 2$ , where each individual contributes  $m$  observations with intraclass correlation  $=\rho$ , then the power of the study given by  $\Phi(z_{1-\beta})$ , where

$$z_{1-\beta} = |p_1 - p_2| / \sqrt{C(p_1 q_1 / M_1 + p_2 q_2 / M_2)} \\ - z_{1-\alpha/2} \sqrt{\bar{p} \bar{q} (1/M_1 + 1/M_2)} / \sqrt{p_1 q_1 / M_1 + p_2 q_2 / M_2}$$

where  $C = 1 + (m - 1) \rho$ ,  $M_i = n_i m$  and  $\bar{p} = (n_1 p_1 + n_2 p_2) / (n_1 + n_2)$ .

If we require a given level of power  $= 1 - \beta$  and it is anticipated that  $n_2 = kn_1$ , then the required sample size in each group is given by

$$n_1 = C \left[ z_{1-\alpha/2} \sqrt{\bar{p} \bar{q} (1+1/k)} + z_{1-\beta} \sqrt{p_1 q_1 + p_2 q_2 / k} \right]^2 / \left( m |p_1 - p_2|^2 \right), n_2 = kn_1$$

**Example 13.63**

**Dentistry** A clinical trial is planned of a new therapeutic modality for the treatment of periodontal disease. The unit of observation is the surface within the patient's mouth. Two groups of patients, one randomly assigned to the new modality and the other randomly assigned to a standard treatment, will be monitored at 6 months after therapy to compare the percentage of surfaces over all patients that lose attachment of teeth to the gum surface. It is anticipated from previous studies that approximately two-thirds of teeth from surfaces treated with the standard modality will lose attachment, and a reduction of this proportion to half would be considered clinically significant. Suppose each patient is required to contribute an average of 25 surfaces to the analysis. How many participants are required for each treatment group in order to have 80% power to detect this magnitude of effect if a two-sided test is to be used with significance level  $= .05$ ?

**Solution**

Because surfaces within one patient cannot be regarded as independent, an estimate of required sample size depends on level of intrapatient correlation ( $\rho$ ) with respect to occurrence of attachment loss. Referring to Fleiss et al. [29], a reasonable estimate of  $\rho$  is given by  $.50$ . Also,  $p_1 = .667$ ,  $q_1 = .333$ ,  $p_2 = .500$ ,  $q_2 = .500$ ,  $\bar{p} = (.667 + .500)/2 = .584$ ,  $\bar{q} = .416$ ,  $k = 1$ , and  $C = 1 + (25 - 1)0.5 = 13$ . The required number of participants per group is then obtained from Equation 13.62 as follows:

$$\begin{aligned}
 N_1 = N_2 &= \frac{\left[1 + (25 - 1).50\right] \left(z_{.975} \sqrt{2\bar{p}\bar{q}} + z_{.80} \sqrt{p_1 q_1 + p_2 q_2}\right)^2}{25(p_1 - p_2)^2} \\
 &= \frac{13 \left[1.96 \sqrt{2(.584)(.416)} + 0.84 \sqrt{.667(.333) + .50(.50)}\right]^2}{25(.667 - .500)^2} \\
 &= \frac{13(1.3665 + 0.5772)^2}{0.6944} = \frac{49.1137}{0.6944} = 70.7 \text{ or } 71 \text{ participants per group}
 \end{aligned}$$

**Example 13.64**

**Dentistry** Suppose the investigators feel they can recruit 100 participants per group for the study mentioned in Example 13.63. How much power would such a study have with the parameters given in Example 13.63, if a two-sided test is to be used with  $\alpha = .05$ ?

**Solution**

Because there are 25 surfaces per participant, we have  $M_1 = M_2 = 25(100) = 2500$ . Thus, from Equation 13.62 we have power =  $\Phi(z_{1-\beta})$ , where

$$\begin{aligned}
 z_{1-\beta} &= |.667 - .500| / \sqrt{13[.667(.333)/2500 + .500(.500)/2500]} \\
 &\quad - z_{.975} \sqrt{.584(.416)(2/2500)} / \sqrt{[.667(.333) + .500(.500)]/2500} \\
 &= .167/0.0496 - 1.96/(.0139)/.0137 \\
 &= 3.363 - 1.989 = 1.375
 \end{aligned}$$

Thus, the power =  $\Phi(1.375) = .915$  if 100 participants per group were recruited.

## Regression Models for Clustered Binary Data

In the preceding examples, we have considered a comparison of two binomial proportions where the units of observation are not independent. However, we often would like to consider one or more additional covariates in our modeling. For this purpose, we wish to extend logistic regression methods to allow for correlation between subunits within the same cluster. A technique called generalized estimating equations (GEE) can perform this type of analysis [30].

**Equation 13.63****GEE Model**

Suppose we have  $n$  clusters and  $m_i$  observations (subunits) within the  $i$ th cluster,  $i = 1, \dots, n$ .

Let  $y_{ij}$  = outcome for the  $j$ th subunit in the  $i$ th cluster,  $i = 1, \dots, n$ ;  $j = 1, \dots, m_i$ .

$$= \begin{cases} 1 & \text{with probability } p_{ij} \\ 0 & \text{with probability } q_{ij} = 1 - p_{ij} \end{cases}$$

Let  $x_{ijk}, \dots, x_{ijk}$  be a set of covariates for the  $j$ th subunit in the  $i$ th cluster. A GEE model is a logistic regression model that allows for the correlation between outcomes for multiple subunits in the same cluster, which is specified by

$$\ln[p_{ij} / (1 - p_{ij})] = \alpha + \sum_{k=1}^K \beta_k x_{ijk}$$

where  $\text{corr}(p_{ij_1}, p_{ij_2}) = \rho$ .

This is called a *compound symmetry* or *exchangeable* correlation structure because the correlation between outcomes for any two subunits in the same cluster is assumed to be the same.

**Example 13.65** **Dentistry** Reanalyze the dental data in Example 13.62 using GEE methodology.

**Solution** We have fit the model

$$\ln[p_{ij} / (1 - p_{ij})] = \alpha + \beta_1 \text{ gender}$$

where gender = 1 if male, = 0 if female, and  $p_{ij}$  = probability that the  $j$ th surface from the  $i$ th person developed caries lesions over a 1-year period where  $i = 1, \dots, 40$ ,  $j = 1, \dots, m_i$  = number of surfaces available for the  $i$ th person. In addition, the correlation between any two subunits (surfaces) in the same cluster (subject) is specified by  $\text{corr}(p_{ij_1}, p_{ij_2}) = \rho$ .

This is referred to as an *exchangeable correlation structure* since any two surfaces within the same subject are assumed to have the same correlation. We used PROC GENMOD of SAS to fit this model specifying a binomial model for the distribution of the outcome variable and a logit link which is the function of the outcome variables whose expected value is assumed to be a linear function of the covariates (hence, the term logistic link). The results are given in Table 13.35.

**Table 13.35 Use of PROC GENMOD of SAS to analyze the effect of gender on the development of caries lesions**

The SAS System			
The GENMOD Procedure			
Model Information			
Data Set	WORK.DENTAL		
Distribution	Binomial		
Link Function	Logit		
Dependent Variable	Lesion		
Number of Observations Read	126		
Number of Observations Used	126		
Number of Events	12		
Number of Trials	126		
Class Level Information			
Class	Levels	Values	
ID	40	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40	
Response Profile			
Ordered		Total	
Value	Lesion	Frequency	
1	1	12	
2	0	114	

PROC GENMOD is modeling the probability that Lesion = '1'

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	Gender

(continued)

**Table 13.35 Use of PROC GENMOD of SAS to analyze the effect of gender on the development of caries lesions (Continued)**

The SAS System							
The GENMOD Procedure							
Analysis of Initial Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq	
Intercept	1	-2.7408	0.4212	-3.5664 -1.9153	42.34	<.0001	
Gender	1	1.4881	0.6259	0.2614 2.7147	5.65	0.0174	
Scale	0	1.0000	0.0000	1.0000 1.0000			

NOTE: The scale parameter was held fixed.

GEE Model Information							
Correlation Structure						Exchangeable	
Subject Effect						ID (40 levels)	
Number of Clusters						40	
Correlation Matrix Dimension						8	
Maximum Cluster Size						8	
Minimum Cluster Size						1	

Algorithm converged.

Exchangeable Working Correlation							
Correlation							
0.1210160955							

Analysis of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence			
				Limits		Z	Pr >  z
Intercept		-2.6496	0.4947	-3.6191	-1.6801	-5.36	<.0001
Gender		1.4478	0.7561	-0.0342	2.9297	1.91	0.0555

Note that PROC GENMOD first fits a model assuming zero correlation among sub-units, which in this case is simply a logistic model with the single covariate gender. The results are listed under Analysis of Initial Parameter Estimates. It uses this model as an initial estimate of the regression parameters. It then refines the estimates by successively estimating the correlation ( $\rho$ ) between outcomes from surfaces from the same person and then re-estimating the regression parameters conditional on  $\rho$ . The final model is listed under the Analysis of GEE Parameter Estimates. We see that the test statistic for gender =  $z = \hat{\beta}_1 / se(\hat{\beta}_1) = 1.4478/0.7561 = 1.91$  with  $p\text{-value} = .056$ . This is actually similar to the result obtained in Example 13.62 if no continuity correction is used, given by

$$z_{\text{clustered binomial}} = \frac{.2222 - .0606}{.0881} = \frac{.1616}{.0881} = 1.834 \sim N(0, 1),$$

with  $p\text{-value} = 2 \times [1 - \Phi(1.834)] = .067$ .

Note also that the estimate of the correlation between outcomes for two surfaces from the same subject is estimated to be 0.121. If we compare the GEE parameter estimates with the initial parameter estimates, we see that the standard error of  $\beta_1$  is larger for the GEE parameter estimates, and the  $p$ -value is larger reflecting the fact that there is correlation among the subunits that reduces the effective sample size and provides for a more appropriate analysis.

**Example 13.66** **Dentistry** Assess the effect of gender on the incidence of caries lesions while controlling for age using the data in Example 13.62.

**Solution** We use PROC GENMOD of SAS to fit the following GEE model

$$\ln[p_{ij} / (1 - p_{ij})] = \alpha + \beta_1 \text{ gender} + \beta_2 \text{ age}$$

where gender = 1 if male, = 0 if female, and  $p_{ij}$  = probability that the  $j$ th surface from the  $i$ th person developed caries lesions over a 1-year period, where  $i = 1, \dots, 40$ ,  $j = 1, \dots, m_i$  = number of surfaces available for the  $i$ th person and  $\text{corr}(p_{ij1}, p_{ij2}) = \rho$ .

The results are given in Table 13.36. We note that there is a borderline effect of gender ( $z$  value = 1.94,  $p$ -value = .052) after controlling for age. No significant effect of age was found ( $p = .88$ ).

**Table 13.36 Use of PROC GENMOD of SAS to analyze the effect of gender and age on development of caries lesions**

The SAS System			
The GENMOD Procedure			
Model Information			
Data Set	WORK.DENTAL		
Distribution	Binomial		
Link Function	Logit		
Dependent Variable	Lesion		
Number of Observations Read	126		
Number of Observations Used	126		
Number of Events	12		
Number of Trials	126		
Class Level Information			
Class	Levels	Values	
ID	40	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40	
Response Profile			
Ordered		Total	
Value	Lesion	Frequency	
1	1	12	
2	0	114	

(continued)

**Table 13.36 Use of PROC GENMOD of SAS to analyze the effect of gender and age on development of caries lesions (Continued)**

The SAS System The GENMOD Procedure							
PROC GENMOD is modeling the probability that Lesion = '1'							
Parameter Information							
Parameter Effect							
Prm1 Intercept							
Prm2 Gender							
Prm3 Age							
Algorithm converged.							
Analysis of Initial Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr > ChiSq	
Intercept	1	-2.9880	2.0867	-7.0780 1.1019	2.05	0.1522	
Gender	1	1.4943	0.6283	0.2628 2.7258	5.66	0.0174	
Age	1	0.0034	0.0284	-0.0522 0.0591	0.01	0.9034	
Scale	0	1.0000	0.0000	1.0000 1.0000			
NOTE: The scale parameter was held fixed.							
GEE Model Information							
Correlation Structure Exchangeable							
Subject Effect ID (40 levels)							
Number of Clusters 40							
Correlation Matrix Dimension 8							
Maximum Cluster Size 8							
Minimum Cluster Size 1							
Algorithm converged.							
Exchangeable Working Correlation							
Correlation 0.1174787811							
Analysis of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter	Estimate	Standard Error	95% Confidence Limits	z	Pr >  z		
Intercept	-3.0538	2.5527	-8.0571 1.9494	-1.20	0.2316		
Gender	1.4569	0.7505	-0.0140 2.9278	1.94	0.0522		
Age	0.0057	0.0361	-0.0652 0.0765	0.16	0.8757		

We can also interpret this result in terms of *ORs* using similar methods as for logistic regression. From Table 13.36, the odds of dental caries for surfaces from males vs. females of the same age =  $e^{1.4569} = 4.3 \equiv OR_{\text{gender}}$ . The 95% CI for the *OR* is given by  $(e^{c_1}, e^{c_2})$ , where

$$c_1 = \hat{\beta}_1 - 1.96 \text{ se}(\hat{\beta}_1), \quad c_2 = \hat{\beta}_1 + 1.96 \text{ se}(\hat{\beta}_1). \text{ Hence,}$$

$$c_1 = 1.4569 - 1.96(0.7505) = -0.014, \quad c_2 = 1.4569 + 1.96(0.7505) = 2.928,$$

and the 95% CI for  $OR_{\text{gender}}$

$$= (e^{-0.014}, e^{2.928}) \equiv (1.0, 18.7)$$

Similarly, the  $OR$  for 2 persons of the same sex who differ by 10 years of age is given by

$$e^{10(0.0057)} = 1.1 \text{ with } 95\% \text{ CI} = (e^{c_1}, e^{c_2}) \text{ where}$$

$$c_1 = 0.0057(10) - 1.96(0.0361)(10) = -0.651$$

$$c_2 = 0.0057(10) + 1.96(0.0361)(10) = 2.321$$

Hence, the 95% CI =  $(e^{-0.651}, e^{2.321}) = (0.5, 10.2)$ .

We have used GEE to adjust logistic regression models for clustering. GEE can also be used to adjust other types of models (e.g., linear regression models) for clustering. This might be appropriate if we have a normally distributed outcome (e.g., serum cholesterol) and we have either data collected from family members or longitudinal data with repeated measures on the same subject over time. In addition, different types of correlation structures other than compound symmetry can be specified (see PROC GENMOD of SAS version 9.2 [31] for other applications of GEE methods and other possible correlation structures).

In this section, we have examined methods of analysis and sample-size estimation for clustered binary data (sometimes referred to as *correlated binary data*). Clustered binary data occur in clinical trials in which the unit of randomization differs from the unit of analysis. For example, in dental clinical trials randomization is usually performed at the person level, but the actual unit of analysis is usually either the tooth or the tooth surface. Similarly, in group randomized studies, a large group (such as an entire school) is the unit of randomization. For example, five schools may be randomized to an active nutritional intervention whose goal is to reduce dietary-fat intake, and five other schools may be randomized to a control intervention. Suppose the outcome is reported dietary-fat intake <30% of calories after 1 year. The outcome is obtained on individual students within the school. In the former example, the outcomes on tooth surfaces represent correlated binary data because there is lack of independence of responses from different teeth or surfaces within the same mouth. In the latter example, outcomes on students represent correlated binary data because of the expected similarity of dietary habits of students from the same school, due to similarity in, for example, socioeconomic status. Clustered binary data can also occur in observational studies, such as virtually any study in the field of ophthalmology where the eye is the unit of analysis.

### REVIEW QUESTIONS 13I

- 1** What is clustered binary data? How does it differ from ordinary binary data?
- 2** Why can't the chi-square test for  $2 \times 2$  tables (Equation 10.5) be used for clustered binary data?
- 3** What role does the intraclass correlation play in analyzing correlated binary data?

## 13.14 Longitudinal Data Analysis

An important application of clustered data methods is in longitudinal data analysis, where each subject provides repeated measures over time and the goal is to assess the effect of covariates on the rate of change over time.

### Example 13.67

**Ophthalmology** A clinical trial was performed among subjects with retinitis pigmentosa (RP) to compare the rate of decline of ERG (electroretinogram) amplitude over time among 4 treatment groups. The ERG is an objective measure of the electrical activity in the retina. In normals, the average ERG is about 350  $\mu$ V. In RP patients, it declines over time and is often  $<10 \mu$ V and sometimes  $<1 \mu$ V, after which total blindness often occurs. Subjects were randomized to either group 1 = 15,000 IU of vitamin A per day, group 2 = 400 IU of vitamin E per day, group 3 = 15,000 IU of vitamin A and 400 IU of vitamin E per day, or group 4 = placebo and were followed annually for 4–6 years. The primary analysis was based on the 354 subjects with baseline ERG amplitude of  $\geq 0.68 \mu$ V. How should we compare the rate of decline among the 4 treatment groups?

One approach is to compute a slope for each subject and then compute an average slope over all subjects within a group. However, subjects may not be followed for the same length of time and subjects with a longer follow-up should be weighted more heavily. A better option is to use *longitudinal data analysis*. For the above clinical trial we consider the following model:

### Equation 13.64

$$\gamma_{it} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma t + \delta_1 x_{i1}t + \delta_2 x_{i2}t + \delta_3 x_{i3}t + e_{it}$$

where  $\gamma_{it} = \ln(\text{ERG amplitude})$  for the  $i$ th subject at time  $t$ ,  $i = 1, \dots, 354, t = 0, \dots, 6$ .

$$x_{ij} = \begin{cases} 1 & \text{if the } i\text{th subject is in treatment group } j, j = 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}$$

$e_{it}$  = error term which is assumed to be normally distributed

with mean = 0 and variance =  $\sigma^2$

A distinguishing feature of longitudinal data analysis is that the error terms for the same subject over time are *not* assumed to be independent. Longitudinal models can be specified with a variety of correlation structures. The simplest correlation structure is called an *exchangeable* or *compound symmetry* correlation structure, where  $\text{corr}(e_{it_1}, e_{it_2}) = \rho$  for some  $\rho \neq 0$ . In other words, the correlation between the residuals at two times  $t_1, t_2$  is assumed to be the same regardless of how far apart  $t_1$  and  $t_2$  are. This might be reasonable for the above relatively short-term clinical trial, but not for a longer term study.

Another important issue is the interpretation of the parameters in Equation 13.64.

### Equation 13.65

#### Interpretation of Parameters in Longitudinal Data Analysis

In Equation 13.64, the parameter  $\beta_j$  represents the mean difference in  $\ln(\text{ERG amplitude})$  at baseline ( $t = 0$ ) between subjects in the  $j$ th treatment group and subjects in the placebo group (group 4) ( $j = 1, 2, 3$ ). The parameter  $\gamma$  represents the rate of decline in  $\ln(\text{ERG amplitude})$  per year among subjects in the placebo group (group 4).

The rate of decline per year among subjects in the  $j$ th treatment group is given by  $\gamma + \delta_j$ ,  $j = 1, 2, 3$ .

Hence,  $\delta_j$  represents the difference in the rate of decline between subjects in the  $j$ th treatment group and subjects in the placebo group,  $j = 1, 2, 3$ . The parameters  $\delta_j$  are usually of primary interest in a longitudinal study.

**Example 13.68**

**Ophthalmology** Analyze the data from the clinical trial mentioned in Example 13.67 using longitudinal data methods.

**Solution**

We have used PROC MIXED of SAS to analyze the data using the repeated option with a compound symmetry correlation structure. The results are given in Table 13.37.

**Table 13.37** Analysis of longitudinal data in RP Clinical Trial ( $n=354$ )

The Mixed Procedure	
Model Information	
<b>Data Set</b>	WORK.MIXED
<b>Dependent Variable</b>	erou
<b>Covariance Structure</b>	Compound Symmetry
<b>Subject Effect</b>	subj
<b>Estimation Method</b>	REML
<b>Residual Variance Method</b>	Profile
<b>Fixed Effects SE Method</b>	Model-Based
<b>Degrees of Freedom Method</b>	Between-Within

**Class Level Information**

Class	Levels	Values
timecat	7	0 1 2 3 4 5 6
trtgp	4	1 2 3 4

**Dimensions**

<b>Covariance Parameters</b>	2
<b>Columns in X</b>	10
<b>Columns in Z</b>	0
<b>Subjects</b>	354
<b>Max Obs Per Subject</b>	7

**Number of Observations**

<b>Number of Observations Read</b>	2098
<b>Number of Observations Used</b>	2098
<b>Number of Observations Not Used</b>	0

**Iteration History**

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	6688.64082379	
1	2	1937.80338853	0.00000576

(continued)

**Table 13.37** Analysis of longitudinal data in RP Clinical Trial (*n*=354) (Continued)

The Mixed Procedure																		
Convergence criteria met.																		
Estimated R Matrix for Subject 1																		
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7											
1	1.4044	1.3399	1.3399	1.3399	1.3399	1.3399	1.3399											
2	1.3399	1.4044	1.3399	1.3399	1.3399	1.3399	1.3399											
3	1.3399	1.3399	1.4044	1.3399	1.3399	1.3399	1.3399											
4	1.3399	1.3399	1.3399	1.4044	1.3399	1.3399	1.3399											
5	1.3399	1.3399	1.3399	1.3399	1.4044	1.3399	1.3399											
6	1.3399	1.3399	1.3399	1.3399	1.3399	1.4044	1.3399											
7	1.3399	1.3399	1.3399	1.3399	1.3399	1.3399	1.4044											
Estimated R Correlation Matrix for Subject 1																		
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7											
1	1.0000	0.9541	0.9541	0.9541	0.9541	0.9541	0.9541											
2	0.9541	1.0000	0.9541	0.9541	0.9541	0.9541	0.9541											
3	0.9541	0.9541	1.0000	0.9541	0.9541	0.9541	0.9541											
4	0.9541	0.9541	0.9541	1.0000	0.9541	0.9541	0.9541											
5	0.9541	0.9541	0.9541	0.9541	1.0000	0.9541	0.9541											
6	0.9541	0.9541	0.9541	0.9541	0.9541	1.0000	0.9541											
7	0.9541	0.9541	0.9541	0.9541	0.9541	0.9541	1.0000											
Covariance Parameter Estimates																		
Cov Parm	Subject		Estimate															
CS	subj		1.3399															
Residual			0.06448															
Fit Statistics																		
-2 Res Log Likelihood				1937.8														
AIC (smaller is better)				1941.8														
AICC (smaller is better)				1941.8														
BIC (smaller is better)				1949.5														
Null Model Likelihood Ratio Test																		
DF	Chi-Square		Pr > ChiSq															
1	4750.84		<.0001															
Solution for Fixed Effects																		
Standard																		
Effect	trtgp	Estimate	Error	DF	t Value	Pr >  t												
Intercept		1.1340	0.1235	350	9.18	<.0001												
time		-0.1048	0.006116	1740	-17.14	<.0001												
trtgp	1	0.06132	0.1777	350	0.34	0.7303												
trtgp	2	-0.2346	0.1737	350	-1.35	0.1778												

(continued)

**Table 13.37** Analysis of longitudinal data in RP Clinical Trial ( $n=354$ ) (Continued)

The Mixed Procedure									
	Effect	trtgp	Estimate	Error	DF	t Value	Pr >  t		
	trtgp	3	-0.00992	0.1756	350	-0.06	0.9550		
	trtgp	4	0	.	.	.	.		
	time*trtgp	1	0.01846	0.008852	1740	2.09	0.0371		
	time*trtgp	2	-0.02087	0.008755	1740	-2.38	0.0172		
	time*trtgp	3	0.01270	0.008766	1740	1.45	0.1477		
	time*trtgp	4	0	.	.	.	.		
	Covariance Matrix for Fixed Effects								
Row	Effect	trtgp	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	Intercept		0.01525	-0.00009	-0.01525	-0.01525	-0.01525		0.000094
2	time		-0.00009	0.000037	0.000094	0.000094	0.000094		-0.00004
3	trtgp	1	-0.01525	0.000094	0.03159	0.01525	0.01525		-0.00020
4	trtgp	2	-0.01525	0.000094	0.01525	0.03018	0.01525		-0.00009
5	trtgp	3	-0.01525	0.000094	0.01525	0.01525	0.03085		-0.00009
6	trtgp	4							
7	time*trtgp	1	0.000094	-0.00004	-0.00020	-0.00009	-0.00009		0.000078
8	time*trtgp	2	0.000094	-0.00004	-0.00009	-0.00019	-0.00009		0.000037
9	time*trtgp	3	0.000094	-0.00004	-0.00009	-0.00009	-0.00019		0.000037
10	time*trtgp	4							
Covariance Matrix for Fixed Effects									
Row			Col8		Col9		Col10		
			1	0.000094	0.000094				
			2	-0.00004	-0.00004				
			3	-0.00009	-0.00009				
			4	-0.00019	-0.00009				
			5	-0.00009	-0.00019				
			6						
			7	0.000037	0.000037				
			8	0.000077	0.000037				
			9	0.000037	0.000077				
			10						
Correlation Matrix for Fixed Effects									
Row	Effect	trtgp	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	Intercept		1.0000	-0.1249	-0.6948	-0.7109	-0.7031		0.08629
2	time		-0.1249	1.0000	0.08678	0.08879	0.08782		-0.6909
3	trtgp	1	-0.6948	0.08678	1.0000	0.4939	0.4885		-0.1254
4	trtgp	2	-0.7109	0.08879	0.4939	1.0000	0.4999		-0.06135
5	trtgp	3	-0.7031	0.08782	0.4885	0.4999	1.0000		-0.06067
6	trtgp	4						1.0000	
7	time*trtgp	1	0.08629	-0.6909	-0.1254	-0.06135	-0.06067		1.0000
8	time*trtgp	2	0.08725	-0.6986	-0.06062	-0.1253	-0.06135		0.4827
9	time*trtgp	3	0.08714	-0.6977	-0.06054	-0.06195	-0.1244		0.4820
10	time*trtgp	4							

(continued)

**Table 13.37** Analysis of longitudinal data in RP Clinical Trial ( $n=354$ ) (Continued)

The Mixed Procedure			
Correlation Matrix for Fixed Effects			
Row	Col8	Col9	Col10
1	0.08725	0.08714	
2	-0.6986	-0.6977	
3	-0.06062	-0.06054	
4	-0.1253	-0.06195	
5	-0.06135	-0.1244	
6			
7	0.4827	0.4820	
8	1.0000	0.4874	
9	0.4874	1.0000	
10			1.0000

Type 3 Tests of Fixed Effects				
Effect	Num	Den		
	DF	DF	F Value	Pr > F
time	1	1740	1064.73	<.0001
trtgp	3	350	1.10	0.3481
time*trtgp	3	1740	7.65	<.0001

We see that there are 354 subjects in the analysis who were assessed over 2098 visits (number of observations used). Note that not all subjects had 6 years of follow-up (i.e., 7 visits). Hence, 2098 is less than  $354 \times 7 = 2478$ .

The program also provides the estimated correlation ( $\rho$ ) between outcomes for repeated visits (see Estimated R correlation matrix for subject 1), which is 0.9541, and the estimated covariance between outcomes for repeated visits (see Estimated R matrix for subject 1). The regression parameter estimates are given under Solution for Fixed Effects. The estimated rate of decline in the ln scale =  $-0.1048$  per year in the placebo group, which is equivalent to a rate of decline of  $(1 - e^{-0.1048}) \times 100\% = 9.9\%$  per year in the original scale. The estimated rate of decline is  $(1 - e^{-0.1048+0.0185}) \times 100\% = 8.3\%$  per year in the vitamin A group (group 1),  $(1 - e^{-0.1048-0.0209}) \times 100\% = 11.8\%$  per year in the vitamin E group (group 2), and  $(1 - e^{-0.1048+0.0127}) \times 100\% = 8.8\%$  per year in the vitamin A + E group (group 3).

There are significant differences in the rate of decline between the vitamin A group vs. the placebo group ( $t = 2.09, p = .037$ ) and between the vitamin E group vs. the placebo group ( $t = -2.38, p = .017$ ). Hence, vitamin A is beneficial for the patients since it reduces the rate of decline, while vitamin E is deleterious since it increases the rate of decline.

We can also adjust for other covariates while performing longitudinal analyses.

#### Example 13.69

**Ophthalmology** Adjust for the effects of age and sex while comparing rates of decline in different treatment groups in the RP clinical trial described in Example 13.67.

#### Solution

Since this is a randomized clinical trial, we expect age and sex distributions to be comparable in different treatment groups. However, in a medium-size clinical trial, small differences may still exist. To control for age and sex, we enhance the model in Equation 13.64 as follows:

$$\begin{aligned}y_{it} = & \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma t + \delta_1 x_{i1}t + \delta_2 x_{i2}t + \delta_3 x_{i3}t \\& + \theta_1 z_{i1} + \theta_2 z_{i2} + \lambda_1 z_{i1}t + \lambda_2 z_{i2}t + e_{it}\end{aligned}$$

where  $z_{i1} = \text{age}_i - 35 = \text{age of } i\text{th subject} - 35$  (labelled as age35 in Table 13.38)

and  $z_{i2} = 1$  if the  $i$ th subject is a male,  $= 0$  if the  $i$ th subject is a female  
(labelled as SEX in Table 13.38)

Note that in longitudinal models to control for another covariate  $z$ , we need to include terms for both  $z$  (to control for baseline differences between groups on  $z$ ) and  $z \times t$  (to control for effects of  $z$  on rate of change). The results from fitting this model are given in Table 13.38.

**Table 13.38** Analysis of differences in rates of decline in RP Treatment Trial controlling for group differences in age and sex

The Mixed Procedure			
Model Information			
Data Set		WORK.MIXED	
Dependent Variable		erou	
Covariance Structure		Compound Symmetry	
Subject Effect		subj	
Estimation Method		REML	
Residual Variance Method		Profile	
Fixed Effects SE Method		Model-Based	
Degrees of Freedom Method		Between-Within	
Class Level Information			
Class	Levels	Values	
timecat	7	0 1 2 3 4 5 6	
trtgp	4	1 2 3 4	
Dimensions			
Covariance Parameters		2	
Columns in X		14	
Columns in Z		0	
Subjects		354	
Max Obs Per Subject		7	
Number of Observations			
Number of Observations Read		2098	
Number of Observations Used		2098	
Number of Observations Not Used		0	
Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	6697.64649393	
1	2	1957.51715086	0.00013321
2	1	1957.38799564	0.00000037

(continued)

**Table 13.38** Analysis of differences in rates of decline in RP treatment trial controlling for group differences in age and sex (*Continued*)

The Mixed Procedure Convergence criteria met.														
Estimated R Matrix for Subject 1														
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7							
1	1.4012	1.3370	1.3370	1.3370	1.3370	1.3370	1.3370							
2	1.3370	1.4012	1.3370	1.3370	1.3370	1.3370	1.3370							
3	1.3370	1.3370	1.4012	1.3370	1.3370	1.3370	1.3370							
4	1.3370	1.3370	1.3370	1.4012	1.3370	1.3370	1.3370							
5	1.3370	1.3370	1.3370	1.3370	1.4012	1.3370	1.3370							
6	1.3370	1.3370	1.3370	1.3370	1.3370	1.4012	1.3370							
7	1.3370	1.3370	1.3370	1.3370	1.3370	1.3370	1.4012							
Estimated R Correlation Matrix for Subject 1														
Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7							
1	1.0000	0.9542	0.9542	0.9542	0.9542	0.9542	0.9542							
2	0.9542	1.0000	0.9542	0.9542	0.9542	0.9542	0.9542							
3	0.9542	0.9542	1.0000	0.9542	0.9542	0.9542	0.9542							
4	0.9542	0.9542	0.9542	1.0000	0.9542	0.9542	0.9542							
5	0.9542	0.9542	0.9542	0.9542	1.0000	0.9542	0.9542							
6	0.9542	0.9542	0.9542	0.9542	0.9542	1.0000	0.9542							
7	0.9542	0.9542	0.9542	0.9542	0.9542	0.9542	1.0000							
Covariance Parameter Estimates														
Cov Parm	Subject		Estimate											
CS	subj		1.3370											
Residual			0.06419											
Fit Statistics														
-2 Res Log Likelihood	1957.4													
AIC (smaller is better)	1961.4													
AICC (smaller is better)	1961.4													
BIC (smaller is better)	1969.1													
Null Model Likelihood Ratio Test														
DF	Chi-Square		Pr > Chisq											
1	4740.26		<.0001											
Solution for Fixed Effects														
Standard														
Effect	trtgp	Estimate	Error	DF	t Value	Pr >  t								
Intercept		1.2763	0.1397	349	9.14	<.0001								
time		-0.1114	0.01030	1737	-10.82	<.0001								
trtgp	1	0.08876	0.1781	349	0.50	0.6186								
trtgp	2	-0.1934	0.1745	349	-1.11	0.2685								
trtgp	3	0.006696	0.1756	349	0.04	0.9696								
trtgp	4	0	.	.	.	.								
time*trtgp	1	0.01725	0.008906	1737	1.94	0.0529								

(continued)

**Table 13.38** Analysis of differences in rates of decline in RP treatment trial controlling for group differences in age and sex (*Continued*)

The Mixed Procedure									
Solution for Fixed Effects									
Effect	trtgp	Standard							
time*trtgp	2	-0.02322	0.008802	1737	-2.64	0.0084			
time*trtgp	3	0.01128	0.008767	1737	1.29	0.1986			
time*trtgp	4	0	.	.	.	.	.		
age35		0.005703	0.007677	1737	0.74	0.4577			
SEX		-0.2597	0.1280	349	-2.03	0.0433			
time*age35		-0.00095	0.000382	1737	-2.50	0.0127			
time*SEX		0.008247	0.006446	1737	1.28	0.2009			
Covariance Matrix for Fixed Effects									
Row	Effect	trtgp	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	Intercept		0.01952	-0.00019	-0.01436	-0.01406	-0.01474		0.000087
2	time		-0.00019	0.000106	0.000114	0.000026	0.000089		-0.00003
3	trtgp	1	-0.01436	0.000114	0.03173	0.01542	0.01532		-0.00020
4	trtgp	2	-0.01406	0.000026	0.01542	0.03044	0.01534		-0.00010
5	trtgp	3	-0.01474	0.000089	0.01532	0.01534	0.03083		-0.00010
6	trtgp	4							
7	time*trtgp	1	0.000087	-0.00003	-0.00020	-0.00010	-0.00010		0.000079
8	time*trtgp	2	0.000087	-0.00003	-0.00010	-0.00019	-0.00010		0.000038
9	time*trtgp	3	0.000090	-0.00004	-0.00009	-0.00010	-0.00019		0.000038
10	time*trtgp	4							
11	age35		0.000075	-0.00006	-0.00003	0.000060	1.199E-6		3.714E-7
Covariance Matrix for Fixed Effects									
Row	Col18	Col19	Col10	Col11	Col12	Col13	Col14		
1	0.000087	0.000090		0.000075	-0.00806	9.61E-8	0.000050		
2	-0.00003	-0.00004		-0.00006	-0.00008	-2.57E-7	-0.00002		
3	-0.00010	-0.00009		-0.00003	-0.00181	1.954E-7	0.000015		
4	-0.00019	-0.00010		0.000060	-0.00197	-2.58E-7	0.000015		
5	-0.00010	-0.00019		1.199E-6	-0.00092	-8.36E-8	7.669E-6		
6									
7	0.000038	0.000038		3.714E-7	0.000016	-8.14E-8	-7.38E-6		
8	0.000077	0.000038		-2.13E-8	0.000016	1.073E-7	-6.48E-6		
9	0.000038	0.000077		-3.19E-7	7.162E-6	7.942E-8	-3.49E-6		
10									
11	-2.13E-8	-3.19E-7		0.000059	0.000131	-3.56E-7	-7.74E-7		
Covariance Matrix for Fixed Effects									
Row	Effect	trtgp	Col1	Col2	Col3	Col4	Col5	Col6	Col7
12	SEX		-0.00806	-0.00008	-0.00181	-0.00197	-0.00092		0.000016
13	time*age35		9.61E-8	-2.57E-7	1.954E-7	-2.58E-7	-8.36E-8		-8.14E-8
14	time*SEX		0.000050	-0.00002	0.000015	0.000015	7.669E-6		-7.38E-6

(continued)

**Table 13.38 Analysis of differences in rates of decline in RP treatment trial controlling for group differences in age and sex (Continued)**

The Mixed Procedure								
Covariance Matrix for Fixed Effects								
Row		Col18	Col19	Col10	Col11	Col12	Col13	Col14
12		0.000016	7.162E-6		0.000131	0.01639	-7.43E-7	-0.00010
13		1.073E-7	7.942E-8		-3.56E-7	-7.43E-7	1.459E-7	3.052E-7
14		-6.48E-6	-3.49E-6		-7.74E-7	-0.00010	3.052E-7	0.000042
Correlation Matrix for Fixed Effects								
Row	Effect	trtgp	Col1	Col2	Col3	Col4	Col5	Col6
1	Intercept		1.0000	-0.1355	-0.5771	-0.5770	-0.6010	0.06981
2	time		-0.1355	1.0000	0.06196	0.01467	0.04934	-0.3689
3	trtgp	1	-0.5771	0.06196	1.0000	0.4961	0.4897	-0.1253
4	trtgp	2	-0.5770	0.01467	0.4961	1.0000	0.5006	-0.06174
5	trtgp	3	-0.6010	0.04934	0.4897	0.5006	1.0000	-0.06078
6	trtgp	4					1.0000	
7	time*trtgp	1	0.06981	-0.3689	-0.1253	-0.06174	-0.06078	1.0000
8	time*trtgp	2	0.07041	-0.3804	-0.06136	-0.1251	-0.06159	0.4893
9	time*trtgp	3	0.07345	-0.3933	-0.06079	-0.06238	-0.1243	0.4845
10	time*trtgp	4						
11	age35		0.07036	-0.7335	-0.02014	0.04508	0.000889	0.005432
12	SEX		-0.4506	-0.05978	-0.07922	-0.08813	-0.04093	0.01369
13	time*age35		0.001801	-0.06537	0.002873	-0.00387	-0.00125	-0.02394
Correlation Matrix for Fixed Effects								
Row		Col18	Col19	Col10	Col11	Col12	Col13	Col14
1		0.07041	0.07345		0.07036	-0.4506	0.001801	0.05556
2		-0.3804	-0.3933		-0.7335	-0.05978	-0.06537	-0.3001
3		-0.06136	-0.06079		-0.02014	-0.07922	0.002873	0.01327
4		-0.1251	-0.06238		0.04508	-0.08813	-0.00387	0.01347
5		-0.06159	-0.1243		0.000889	-0.04093	-0.00125	0.006776
6								
7		0.4893	0.4845		0.005432	0.01369	-0.02394	-0.1286
8		1.0000	0.4910		-0.00031	0.01396	0.03193	-0.1142
9		0.4910	1.0000		-0.00474	0.006381	0.02372	-0.06174
10				1.0000				
11		-0.00031	-0.00474		1.0000	0.1329	-0.1213	-0.01565
12		0.01396	0.006381		0.1329	1.0000	-0.01519	-0.1251
13		0.03193	0.02372		-0.1213	-0.01519	1.0000	0.1240
Correlation Matrix for Fixed Effects								
Row	Effect	trtgp	Col1	Col2	Col3	Col4	Col5	Col6
14	time*SEX		0.05556	-0.3001	0.01327	0.01347	0.006776	-0.1286

(continued)

**Table 13.38** Analysis of differences in rates of decline in RP treatment trial controlling for group differences in age and sex (Continued)

The Mixed Procedure							
Correlation Matrix for Fixed Effects							
Row	Col18	Col19	Col10	Col11	Col12	Col13	Col14
14	-0.1142	-0.06174		-0.01565	-0.1251	0.1240	1.0000
Type 3 Tests of Fixed Effects							
		Num	Den				
Effect	DF	DF	F Value		Pr > F		
time	1	1737	145.98		<.0001		
trtgp	3	349	0.93		0.4273		
time*trtgp	3	1737	8.05		<.0001		
age35	1	1737	0.55		0.4577		
SEX	1	349	4.11		0.0433		
time*age35	1	1737	6.23		0.0127		
time*SEX	1	1737	1.64		0.2009		

We see from Table 13.38 that gender is related to baseline level of ERG, with males having significantly lower levels ( $p = .043$ ). Also, more importantly, age has a significant effect on rate of decline ( $p = .013$ ), with older patients declining more rapidly. After controlling for age and sex, there remain significant differences in rate of decline between the vitamin A group (group 1) and the placebo group (group 4) ( $p = .05$ ) as well as between the vitamin E group (group 2) and the placebo group ( $p = .008$ ). In summary, vitamin A diminishes the rate of decline and vitamin E accelerates the rate of decline in ERG amplitude among RP patients.

In this section, we have considered marginal models for longitudinal data analysis. More complex longitudinal models exist, including random effects models [32] and conditional models [33], but these are beyond the scope of this text.

Note in both Tables 13.37 and 13.38 it is also possible to compare the rates of change between any 2 groups. For example, to compare the rates of change between groups 1 and 2, we would compute

$$\hat{\beta}_1 - \hat{\beta}_2 \text{ with}$$

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)},$$

Where  $\text{var}(\hat{\beta}_1)$  is found in the (7,7) element (i.e., row 7 column 7),  $\text{var}(\hat{\beta}_2)$  is found in the (8,8) element (i.e., row 8 column 8), and  $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$  is found in the (7,8) element (i.e., row 7, column 8) of the Covariance Matrix for Fixed Effects part of the SAS Proc Mixed output. A  $100\% \times (1-\alpha)$  CI for  $\beta_1 - \beta_2$  is then given by:

$$\hat{\beta}_1 - \hat{\beta}_2 \pm z_{1-\alpha/2} se(\hat{\beta}_1 - \hat{\beta}_2).$$

## 13.15 Measurement-Error Methods

### Introduction

Exposure variables in epidemiology are often measured with error. An interesting question is, How does this measurement error affect the results obtained from standard analyses?

**Example 13.70**

**Cancer, Nutrition** A hypothesis has been proposed linking breast-cancer incidence to saturated-fat intake. To test this hypothesis, a group of 89,538 women, ages 34–59, who were free of breast cancer in 1980, were followed until 1984. During this period, 590 cases of breast cancer occurred. A logistic-regression model [34] of breast-cancer incidence from 1980 to 1984 on age (using dummy variables based on the age groups 34–39, 40–44, 45–49, 50–54, 55–59), saturated-fat intake as a continuous variable as reported on a 1980 food-frequency questionnaire (FFQ), and alcohol intake (using dummy variables based on the groups 0, 0.1–1.4, 1.5–4.9, 5.0–14.9, 15+ grams per day) was fit to the data. The results are given in Table 13.39.

The OR for a 10 g/day increase in calorie-adjusted saturated fat intake (henceforth referred to as saturated fat) was 0.92 (95% CI = 0.80–1.05). The instrument used to assess diet for the analyses in Table 13.39 was the 1980 FFQ, on which participants reported their average consumption of each of 61 foods over the past year. This instrument is known to have a large amount of measurement error. It is sometimes referred to as a *surrogate* for the ideal instrument that could always measure dietary fat without error. What impact does the measurement error have on the results obtained?

**Table 13.39** Association between breast-cancer incidence and calorie-adjusted saturated-fat intake, NHS, 1980–1984, 590 events, 89,538 participants

Variable	$\beta$	se	$z$	$p$	OR (95% CI)
Saturated-fat intake (g) ( $X$ )	−0.0878	0.0712	−1.23	.22	0.92 (0.80–1.05)

Note: Based on a 10-g increase in saturated-fat intake.

### Measurement-Error Correction with Gold-Standard Exposure

The model fit in Table 13.39 was of the form

**Equation 13.66**

$$\ln[p/(1-p)] = \alpha + \beta X + \sum_{j=1}^m \delta_j u_j$$

where  $X$  is saturated-fat intake from the FFQ that is measured with error and  $u_1, \dots, u_m$  are a set of variables measured without error, which in this example represent dummy variables for age and alcohol intake. Alcohol intake is actually also measured with error, but the degree of measurement error is typically much smaller than for saturated fat [34]. For simplicity, we assume alcohol intake is measured without error.

To assess the impact of measurement error on the estimate of  $\beta$  in Equation 13.66, we would have to consider how the estimate of  $\beta$  would change if average daily saturated fat intake could be ascertained with no error. The diet record (DR) is considered a gold standard by some nutritional epidemiologists. With a DR, a person records each food eaten and the corresponding portion size on a real-time basis. The foods and portion sizes are then entered onto a computer, and a computer program is used to calculate nutrients consumed during this period. Ideally, a DR would be filled out for each of 365 days in 1980 by each of the 89,538 nurses in the main study. However, it is very expensive to collect and process DR data. Instead, a *validation study* was performed among 173 nurses who filled out 4 weeks of DR with individual weeks spaced about 3 months apart. They then filled out an additional FFQ in

1981 to refer to the same time period as the DR. The data from the validation study were used to model the relationship between reported DR saturated-fat intake ( $x$ ) and reported FFQ saturated-fat intake ( $X$ ), using a linear-regression model of the form

**Equation 13.67**

$$x = \alpha' + \gamma X + e$$

where  $e$  is assumed to be  $N(0, \sigma^2)$ . The results are given in Table 13.40.

**Table 13.40** Relationship between DR saturated-fat intake ( $x$ ) and FFQ saturated-fat intake ( $X$ ), NHS, 1981, 173 participants

Variable	$\hat{\gamma}$	se	$t$	$p$ -value
Saturated-fat intake FFQ (g) ( $X$ )	0.468	0.048	9.75	<.001

We see, as expected, a highly significant association between  $x$  and  $X$ . Our goal is to estimate the relationship between breast-cancer incidence and DR saturated-fat intake ( $x$ ) after controlling for age and alcohol intake, assuming a logistic model for this relationship of the form

**Equation 13.68**

$$\ln[p/(1-p)] = \alpha^* + \beta^* x + \sum_{j=1}^m \delta_j^* u_j$$

where  $u_1, \dots, u_m$  are a set of other covariates assumed to be measured without error that represent age and alcohol intake. The problem is that we only observe  $x$  directly on 173 of 89,538 women. Therefore, instead of estimating the logistic regression directly from Equation 13.68, we use an indirect approach. Specifically, because we know  $X$  for each woman in the main study, we can estimate the average DR intake for that value of  $X$ , which we denote by  $E(x|X)$ , and use that as an estimate of  $x$  for that woman. From the linear regression in Equation 13.67, we have

**Equation 13.69**

$$E(x|X) = \alpha' + \gamma X$$

Substituting  $E(x|X)$  from Equation 13.69 for  $x$  in Equation 13.68 yields

**Equation 13.70**

$$\ln[p/(1-p)] = (\alpha^* + \beta^* \alpha') + (\beta^* \gamma) X + \sum_{j=1}^m \delta_j^* u_j$$

If we compare Equation 13.70 with Equation 13.66, we see the dependent and independent variables are the same. Thus we can equate the regression coefficients corresponding to  $X$  (that is, FFQ saturated fat), yielding

**Equation 13.71**

$$\beta^* \gamma = \beta$$

If we divide both sides of Equation 13.71 by  $\gamma$ , we obtain

**Equation 13.72**

$$\beta^* = \beta/\gamma$$

Therefore, to estimate the logistic-regression coefficient of breast cancer on “true” saturated-fat intake, we divide the logistic-regression coefficient ( $\beta$ ) of breast cancer on the surrogate saturated fat ( $X$ ) from the main study by the linear-regression coefficient ( $\gamma$ ) of true ( $x$ ) on surrogate ( $X$ ) saturated fat from the validation study. The equating of Equations 13.66 and 13.70 is an approximation because it ignores the impact of the error distribution from the validation-study model in Equation 13.67 on the estimation of  $x$  for subjects in the main study. However, it is approximately valid if the disease under study is rare and the measurement-error variance ( $\sigma^2$  in Equation 13.67) is small [34].

To obtain the standard error of  $\hat{\beta}^*$  and associated confidence limits for  $\beta^*$ , we use a multivariate extension of the delta method introduced in Equation 13.3. This approach for estimating  $\beta^*$  is called the *regression-calibration method* [34] and is summarized as follows.

**Equation 13.73**
**Regression-Calibration Approach for Estimation of Measurement-Error-Corrected OR Relating a Dichotomous Disease Variable ( $D$ ) to a Single Exposure Variable ( $X$ ) Measured with Error, When a Gold-Standard Exposure ( $x$ ) Is Available**

Suppose we have

- (1) A dichotomous disease variable ( $D$ ), where  $D = 1$  if disease is present, 0 if disease is absent
- (2) A single exposure variable ( $X$ ) measured with error (called the *surrogate exposure*)
- (3) A corresponding gold-standard exposure variable ( $x$ ) that represents true exposure (or is at least an unbiased estimate of true exposure with errors that are uncorrelated with that of the surrogate)
- (4) A set of other covariates  $u_1, \dots, u_m$ , which are assumed to be measured without error

We wish to fit the logistic-regression model

$$\ln[p/(1-p)] = \alpha + \beta^* x + \sum_{j=1}^m \delta_j^* u_j$$

where  $p = Pr(D = 1 | x, u_1, \dots, u_m)$

We have available

- (a) A main-study sample of size  $n$  (usually large), where  $D$ ,  $X$ , and  $u_1, \dots, u_m$  are observed
- (b) A validation-study sample of size  $n_1$  (usually small), where  $x$  and  $X$  are observed. Ideally, the validation-study sample should be a representative sample from the main-study sample or a comparable external sample.

Our goal is to estimate  $\beta^*$ . For this purpose,

- (i) We use the main-study sample to fit the logistic-regression model of  $D$  on  $X$  and  $u_1, \dots, u_m$  of the form

$$\ln[p/(1-p)] = \alpha + \beta X + \sum_{j=1}^m \delta_j u_j$$

- (ii) We use the validation-study sample to fit the linear-regression model of  $x$  on  $X$  of the form

$$x = \alpha' + \gamma X + e$$

$$\text{where } e \sim N(0, \sigma^2)$$

- (iii) We use (i) and (ii) to obtain the point estimate of  $\beta^*$ , given by

$$\hat{\beta}^* = \hat{\beta} / \hat{\gamma}$$

The corresponding estimate of the  $OR$  of  $D$  on  $x$  is given by

$$\hat{OR} = \exp(\hat{\beta}^*)$$

- (iv) We obtain the variance of  $\hat{\beta}^*$  by computing

$$Var(\hat{\beta}^*) = (1/\hat{\gamma}^2) Var(\hat{\beta}) + (\hat{\beta}^2/\hat{\gamma}^4) Var(\hat{\gamma})$$

where  $\hat{\beta}$  and  $Var(\hat{\beta})$  are obtained from (i)

and  $\hat{\gamma}$  and  $Var(\hat{\gamma})$  are obtained from (ii)

- (v) We obtain a  $100\% \times (1 - \alpha)$  CI for  $\beta^*$  by computing

$$\hat{\beta}^* \pm z_{1-\alpha/2} se(\hat{\beta}^*) = (\hat{\beta}_1^*, \hat{\beta}_2^*)$$

where  $\hat{\beta}^*$  is obtained from (iii) and

$$se(\hat{\beta}^*) = [Var(\hat{\beta}^*)]^{1/2} \text{ is obtained from (iv)}$$

The corresponding  $100\% \times (1 - \alpha)$  CI for  $OR$  is given by

$$[\exp(\hat{\beta}_1^*), \exp(\hat{\beta}_2^*)]$$

This method should only be used if the disease under study is rare (incidence <10%) and the measurement-error variance [ $\sigma^2$  in (ii)] is small.

### Example 13.71

**Cancer, Nutrition** Estimate the  $OR$  relating breast-cancer incidence from 1980–1984 to DR intake of saturated fat in 1980 using the regression-calibration method based on the data in Tables 13.39 and 13.40.

#### Solution

From Table 13.39 we have  $\hat{\beta} = -0.0878$ ,  $se(\hat{\beta}) = 0.0712$ . From Table 13.40, we have that  $\hat{\gamma} = 0.468$ ,  $se(\hat{\gamma}) = 0.048$ . Thus, from step (iii) of Equation 13.73, the point estimate of  $\beta^*$  is  $\hat{\beta}^* = -0.0878/0.468 = -0.1876$ . The corresponding point estimate of the  $OR$  relating breast-cancer incidence to an increase of 10 g of “true” (DR) saturated-fat intake is  $\exp(-0.1876) = 0.83$ . To obtain  $Var(\hat{\beta}^*)$ , we refer to step (iv) of Equation 13.73. We have

$$\begin{aligned} Var(\hat{\beta}^*) &= (1/0.468^2)(0.0712)^2 + [(-0.0878)^2/(0.468)^4](0.048)^2 \\ &= 0.02315 + 0.00037 = 0.02352 \\ se(\hat{\beta}^*) &= (0.02352)^{1/2} = 0.1533 \end{aligned}$$

Thus, from step (v) of Equation 13.73, a 95% CI for  $\beta^*$  is given by

$$-0.1876 \pm 1.96(0.1533) = -0.1876 \pm 0.3006 = (-0.488, 0.113) = (\hat{\beta}_1^*, \hat{\beta}_2^*)$$

The corresponding 95% CI for  $OR$  is given by

$$[\exp(-0.488), \exp(0.113)] = (0.61, 1.12)$$

Notice that the measurement-error-corrected estimate of  $OR$  (0.83) is farther away from 1 than the crude or uncorrected estimate (0.92) obtained in Table 13.39. The uncorrected estimate of 0.92 is attenuated (i.e., incorrectly moved closer to 1, the null value) under the influence of measurement error. Therefore, the corrected estimate (0.83) is sometimes called a *deattenuated OR* estimate. Notice also that the CI for the corrected  $OR$  (0.61, 1.12) is much wider than the corresponding CI for the uncorrected  $OR$  (0.80, 1.05) in Table 13.39, which often occurs. Finally, the two terms in the expression for  $Var(\hat{\beta}^*)$  (0.02315 and 0.00037) reflect error in the estimated main-study logistic-regression coefficient ( $\hat{\beta}$ ) and the estimated validation-study linear-regression coefficient ( $\hat{\gamma}$ ), respectively. Usually the first term predominates unless the validation-study sample size is very small.

### Measurement-Error Correction Without a Gold-Standard Exposure

Example 13.70 assumed a dietary instrument (the DR) that at least some nutritionists would regard as a gold standard. Technically, to use the regression-calibration method, the gold-standard instrument need only provide an unbiased estimate of “true” exposure rather than actually be “true” exposure with errors that are uncorrelated with that of the surrogate. Given that the DR in Example 13.70 consisted of average intake over 28 days spaced throughout the year, this seemingly would provide an unbiased estimate of intake over all 365 days, provided the DR is filled out accurately. However, for some exposures even a potential gold-standard instrument doesn’t exist.

#### **Example 13.72**

**Cancer, Endocrinology** Among postmenopausal women, a positive association has generally been observed between plasma-estrogen levels and breast-cancer risk. However, most studies have been small, and many have not evaluated specific estrogen fractions. A substudy of the NHS was conducted among 11,169 postmenopausal women who provided a blood sample during the period from 1989 to 1990 and were not using postmenopausal hormones at the time of the blood collection [35]. However, it was too expensive to analyze hormone levels for all 11,000 women. Instead, hormone levels were assayed from 156 women who developed breast cancer after blood collection but before June 1994. Two control women, matched with respect to age, menopausal status, and month and time of day of blood collection, were selected for each breast-cancer case. In this example, we consider the relationship between  $\ln(\text{plasma estradiol})$  and the development of breast cancer. The  $\ln$  transformation was used to better satisfy the linearity assumptions of logistic regression. The results indicated an  $RR$  of breast cancer of 1.77 (95% CI = 1.06–2.93) comparing women in the highest quartile of  $\ln(\text{estradiol})$  (median estradiol level = 14 pg/mL) with women in the lowest quartile of  $\ln(\text{estradiol})$  (median estradiol level = 4 pg/mL) based on the distribution of  $\ln(\text{estradiol})$  among the controls. However, it is known that plasma estradiol has some measurement error, and we would like to obtain a measurement-error-corrected estimate of the  $RR$ . How can we accomplish this?

Unlike the dietary study in Example 13.70, there is no gold-standard instrument for plasma estradiol similar to the DR for nutrient intake. However, it is reasonable to consider the average of a large number of  $\ln(\text{estradiol})$  measurements ( $x$ ) as a gold standard that can be compared with the single  $\ln(\text{estradiol})$  measurement ( $X$ ) obtained in the study. Although  $x$  is not directly measurable, we can consider a random-effects ANOVA model relating  $X$  to  $x$ , of the form

**Equation 13.74**

$$X_i = x_i + e_i$$

where  $X_i$  = a single  $\ln(\text{estradiol})$  measurement for the  $i$ th woman

$x_i$  = underlying mean  $\ln(\text{estradiol})$  level for the  $i$ th woman

$$x_i \sim N(\mu, \sigma_A^2), e_i \sim N(0, \sigma^2)$$

Here  $\sigma_A^2$  represents between-person variation, and  $\sigma^2$  represents within-person variation for  $\ln(\text{estradiol})$  levels.

To implement the regression-calibration method, we need to obtain an estimate of  $\gamma$  as in Equation 13.67—that is, the regression coefficient of  $x$  [true  $\ln(\text{estradiol})$  level] on  $X$  [a single  $\ln(\text{estradiol})$  value]. We know from Equation 12.37 that

**Equation 13.75**

$$\text{Corr}(x, X) = \text{reliability coefficient} = (\rho_I)^{1/2}$$

$$\text{where } \rho_I = \sigma_A^2 / (\sigma_A^2 + \sigma^2) = \text{intraclass correlation coefficient}$$

Furthermore, from our work on the relationship between a regression coefficient and a correlation coefficient (Equation 11.19), we found that

**Equation 13.76**

$$b(x \text{ on } X) = \text{Corr}(x, X) \text{sd}(x) / \text{sd}(X)$$

Also, from Equation 13.74 we have

**Equation 13.77**

$$\text{sd}(x) = \sigma_A$$

**Equation 13.78**

$$\text{sd}(X) = (\sigma_A^2 + \sigma^2)^{1/2}$$

Therefore, on combining Equations 13.75–13.78 we obtain

**Equation 13.79**

$$\begin{aligned} b(x \text{ on } X) &= (\rho_I)^{1/2} \sigma_A / (\sigma_A^2 + \sigma^2)^{1/2} \\ &= (\rho_I)^{1/2} \left[ \sigma_A^2 / (\sigma_A^2 + \sigma^2) \right]^{1/2} \\ &= (\rho_I)^{1/2} (\rho_I)^{1/2} = \rho_I \end{aligned}$$

Thus we can estimate the regression coefficient of  $x$  on  $X$  by the sample intraclass correlation coefficient  $r_p$ . To obtain  $r_p$  we need to conduct a reproducibility study on

a subsample of subjects with at least two replicates per subject. The reproducibility study plays the same role as a validity study, which is used when a gold standard is available. If we substitute  $r_I$  for  $\gamma$  in Equation 13.67, we obtain the following regression-calibration procedure for measurement-error correction when no gold standard is available.

### Equation 13.80

#### Regression-Calibration Approach for Estimation of Measurement-Error-Corrected OR Relating a Dichotomous Disease Variable ( $D$ ) to a Single Exposure Variable ( $X$ ) Measured with Error When a Gold Standard Is Not Available

Suppose we have

- (1) A dichotomous disease variable  $D$  ( $= 1$  if disease is present,  $= 0$  if disease is absent)
- (2) A single continuous exposure variable  $X$  measured with error
- (3) (Optionally) a set of other covariates  $u_1, \dots, u_m$  measured without error

We define  $x$  as the average of  $X$  over many replicates for an individual subject. We wish to fit the logistic-regression model

$$\ln[p/(1-p)] = \alpha^* + \beta^*x + \sum_{j=1}^m \delta_j^*u_j$$

where  $p = Pr(D = 1 | x, u_1, \dots, u_m)$

We have available

- (a) A main-study sample of size  $n$  (usually large), where  $D$ ,  $X$ , and  $u_1, \dots, u_m$  are observed
- (b) A reproducibility-study sample of size  $n_1$  (usually small), where  $k_i$  replicate observations of  $X$  are obtained for the  $i$ th person, from which the estimated intraclass correlation coefficient  $r_I$  can be obtained (see Equation 12.36)

To estimate  $\beta^*$ , we

- (i) Fit the main-study logistic-regression model of  $D$  on  $X$  and  $u_1, \dots, u_m$  of the form

$$\ln[p/(1-p)] = \alpha + \beta X + \sum_{j=1}^m \delta_j u_j$$

- (ii) Use the reproducibility study to estimate  $\rho_i$  by  $r_I$

(iii) Obtain the point estimate of  $\hat{\beta}^*$ , given by  $\hat{\beta}^* = \hat{\beta}/r_I$  with corresponding OR estimate  $= \exp(\hat{\beta}^*)$

- (iv) Obtain the variance of  $\hat{\beta}^*$ , given by

$$Var(\hat{\beta}^*) = (1/r_I^2) Var(\hat{\beta}) + (\hat{\beta}^2/r_I^4) Var(r_I) = A + B$$

where  $Var(r_I)$  is obtained from [36] as follows:

$$Var(r_I) = 2(1 - r_I)^2 [1 + (k_0 - 1)r_I]^2 / [k_0(k_0 - 1)(n_1 - 1)]$$

$$\text{and } k_0 = \left( \sum_{i=1}^{n_1} k_i - \sum_{i=1}^{n_1} k_i^2 / \sum_{i=1}^{n_1} k_i \right) / (n_1 - 1)$$

(Note: If all subjects provide the same number of replicates ( $k$ ), then  $k_0 = k$ .)

(v) Obtain a  $100\% \times (1 - \alpha)$  CI for  $\hat{\beta}^*$  given by

$$\hat{\beta}^* \pm z_{1-\alpha/2} se(\hat{\beta}^*) = (\hat{\beta}_1^*, \hat{\beta}_2^*)$$

where  $\hat{\beta}^*$  is obtained from (iii) and  $se(\hat{\beta}^*) = [Var(\hat{\beta}^*)]^{1/2}$  is obtained from (iv).

The corresponding  $100\% \times (1 - \alpha)$  CI for  $OR$  is given by

$$[\exp(\hat{\beta}_1^*), \exp(\hat{\beta}_2^*)]$$

### Example 13.73

**Cancer, Endocrinology** Estimate the  $OR$  relating breast-cancer incidence to plasma estradiol after correcting for measurement error for a woman with true plasma-estradiol level of 14 pg/mL compared with a woman with a true plasma-estradiol level of 4 pg/mL based on the data described in Example 13.72.

### Solution

We use the regression-calibration approach in Equation 13.80. From Example 13.72 we have  $\hat{\beta} = \ln(1.77) = 0.571$ . Furthermore, the 95% CI for  $\beta = [\ln(1.06), \ln(2.93)] = (0.058, 1.075)$ . The width of the 95% CI is

$$2(1.96) se(\hat{\beta}) = 3.92 se(\hat{\beta}) = 1.075 - 0.058 = 1.017$$

$$\text{and } se(\hat{\beta}) = 1.017/3.92 = 0.259$$

Furthermore, a reproducibility study was conducted among a subset of 78 of the nurses [37]. The estimated intraclass correlation coefficient for  $\ln(\text{plasma estradiol})$  was 0.68. Sixty-five of the nurses provided 3 replicates, and 13 nurses provided 2 replicates. Therefore, from step (iii) of Equation 13.80, we have the point estimate,  $\hat{\beta}^* = 0.571/0.68 = 0.840$ , with corresponding  $OR$  estimate =  $\exp(0.840) = 2.32$ . From step (iv) of Equation 13.80, we obtain  $Var(\hat{\beta}^*)$ . We have

$$A = (0.259)^2 / (0.68)^2 = 0.1455$$

To obtain  $B$ , we need to compute  $Var(r_I)$ . We have

$$Var(r_I) = 2(1 - .68)^2 [1 + (k_0 - 1)(.68)]^2 / [77k_0(k_0 - 1)]$$

To evaluate  $k_0$ , we have  $65(3) + 13(2) = 221$  replicates over the entire sample. Thus

$$\begin{aligned} k_0 &= \{221 - [65(3)^2 + 13(2)^2]\}/221\} / 77 \\ &= 218.12/77 = 2.833 \end{aligned}$$

Therefore,

$$\begin{aligned} Var(r_I) &= 2(1 - .68)^2 [1 + 1.833(.68)]^2 / [2.833(1.833)77] \\ &= 1.033/399.74 = 0.0026 \end{aligned}$$

Thus,

$$\begin{aligned} B &= [(0.571)^2 / (0.68)^4](0.0026) \\ &= 0.0039 \end{aligned}$$

It follows that

$$\begin{aligned} \text{Var}(\hat{\beta}^*) &= 0.1455 + 0.0039 = 0.1494 \\ \text{se}(\hat{\beta}^*) &= (0.1494)^{1/2} = 0.387 \end{aligned}$$

From (v), a 95% CI for  $\hat{\beta}^*$  is given by

$$\begin{aligned} 0.840 \pm 1.96 (0.387) &= 0.840 \pm 0.758 \\ &= (0.082, 1.597) \end{aligned}$$

The corresponding 95% CI for OR is  $[\exp(0.082), \exp(1.597)] = (1.09, 4.94)$ .

Thus, the measurement-error-corrected estimate of the OR relating a woman with a true estradiol level of 14 pg/mL to a woman with a true estradiol level of 4 pg/mL is 2.32 (95% CI = 1.09, 4.94). The corresponding uncorrected estimate is 1.77 (95% CI = 1.06 – 2.93). The deattenuated (corrected) point estimate is substantially larger than the uncorrected estimate, with much wider confidence limits.

Several comments are in order concerning this example. First, the design was a prospective case-control study nested within a cohort study. As discussed in Section 13.3, we cannot estimate the absolute risk of breast cancer because by design about one-third of the women were cases. However, it is possible to obtain valid estimates of relative risk for  $\ln(\text{estradiol})$ . Second, there were several other variables in the model, all of which were assumed to be measured with little or no measurement error and also to be approximately uncorrelated with the variable measured with error. Third, the point estimate and 95% CI for the relative risk differ slightly from those of [35] because of rounding error and slightly different approaches used to estimate the intraclass correlation coefficient.

In this section, we discussed methods for obtaining point and interval estimates of relative risk from logistic-regression models that are corrected for measurement error in the covariate of interest using the regression-calibration approach. In Equation 13.73, we assumed a gold-standard exposure measure was available. To implement the methods in this setting, we need both a main study of size  $n$  in which the disease and surrogate exposure are measured and a validation study of size  $n_1$  in which both the surrogate exposure and the gold-standard exposure are available on the same subjects. The validation-study sample may or may not be a subset of the main-study population. In Equation 13.80, we assumed a gold-standard exposure was not available. To implement the methodology in this setting, we need both a main-study sample, in which both the disease and the surrogate exposure are available, and a reproducibility study, in which replicate surrogate measurements are available. The reproducibility-study sample may or may not be a subset of the main-study sample. We have assumed in both Equation 13.73 and Equation 13.80 that there is only a single covariate in the main study measured with error; there may be several other covariates measured without error, but they are assumed to be approximately uncorrelated with the variable measured with error. The problem of multiple covariates measured with error is complex and beyond the scope of this text. An extension of the methods in this section, when more than one exposure variable is measured with error and/or when other covariates measured without error are correlated with the variable measured with error, is given in [38] when a gold standard is available and in [39] when a gold standard is not available. It is important to be aware that even if only a single exposure variable is measured with error, after correcting for measurement error the partial-regression coefficients of other covariates measured without error (e.g., age) may also be affected (see [39] for an example concerning

this issue). In this section, we have only focused on the effects of measurement error on the regression coefficient for the covariate measured with error.

Software to implement the measurement-error correction methods in this section is available at <http://www.hsph.harvard.edu/faculty/spiegelman/blinplus.html> when a gold standard is available and at <http://hsph.harvard.edu/faculty/spiegelman/relibpls8.html> when a gold standard is not available.

### REVIEW QUESTIONS 13J

- 1 What is the purpose of measurement-error correction?
- 2 What is the difference between a surrogate exposure and a true exposure?
- 3 What is a deattenuated estimate?
- 4 (a) What is the difference between a validation study and a reproducibility study?  
(b) Under what circumstances do we use each kind of study?  
(c) What estimated parameters from each type of study are used in correcting for measurement error?

## 13.16 Missing Data

Most epidemiologic and clinical studies have missing (or incomplete) data, for many different reasons. This presents a quandary for multivariate analyses such as multiple regression or multiple logistic regression, where to run the analysis both the dependent variable and each of the independent variables must be present. A frequent solution, which is the default in most statistical software packages, is to use the *complete-case method*, in which the analysis uses only observations with all variables present. If the amount of data missing is small, then little bias or imprecision is introduced by using this approach. However, if a nontrivial (e.g., >10%) amount of data is missing and/or if the subjects with missing data constitute a nonrandom subsample of the total study sample, then bias can potentially be introduced by using this approach.

#### Example 13.74

**Aging** The Established Populations for Epidemiologic Studies of the Elderly (EPESE) study was a collaborative study performed at four centers in the United States among people 65+ years of age in 1982–1987. Its goal was to determine the longitudinal course of aging and to identify risk factors related to the aging process. Of particular interest are risk factors that affect subsequent mortality. For this purpose, a multiple logistic-regression analysis was run among 2341 elderly participants ages 71–103 in 1988–1989 to predict mortality through 1991. The predictor variables were  $x_1$  = age (yrs);  $x_2$  = sex (coded as 1 if male and 2 if female);  $x_3$  = physical-performance score, which is a 13-item scale from 0 to 12 that indicates the number of different activities (e.g., getting up from a chair, squatting, etc., that a person can perform);  $x_4$  = self-assessed health score, which is a scale from 1 to 4 (coded as 1 if excellent, 2 if good, 3 if fair, and 4 if poor). The physical-performance scale was obtained at a home visit from 1988–1989; the self-assessed health scale was obtained either at the home visit or by telephone. Data for the physical-performance scale were missing for 550 elderly participants who were either unwilling or unable to perform the test. No data for any of the other variables were missing.

Descriptive statistics for the study population are given in Table 13.41. The results of the logistic-regression analyses for the 1791 elderly participants with complete data are given in the first column of numbers in Table 13.42.

**Table 13.41 Descriptive statistics for 2341 older residents of East Boston interviewed in 1988–1989**

	Completed the physical-performance evaluation			
	Yes		No	
Dead by 12/31/1991	No 1527	Yes 264	No 416	Yes 134
n				
Age, median (IQR) <sup>a</sup>	77 (74–81)	78 (74–84)	78 (74–83)	85 (78–90)
Male, %	32.0	45.1	30.3	47.0
Physical performance median (IQR)	8 (5–10)	6 (2–8)	—	—
Self-assessed health median (IQR)	2 (2–3)	3 (2–3)	2 (2–3)	3 (2–3)

Notes: These are the 2341 residents of East Boston who participated in the 6-year follow-up evaluation of the EPESE study (see, for example, Glynn et al. [40]). They ranged in age from 71 to 103 years at that time. Participants were asked to rate their health relative to others of their age as 1, excellent; 2, good; 3, fair; or 4, poor. Those who were able also had objective evaluations of physical performance. This was based on brief tests of balance, gait, strength, and endurance, with results summarized in an overall score ranging from 0 to 12, with higher scores indicating better function (Guralnik et al. [41])

<sup>a</sup>Interquartile range.

**Table 13.42 Comparison of effects of variables on the risk of death from alternative models; shown are logistic-regression parameters (standard errors)**

	Analytic method			
	Complete case	No physical performance	Multiple imputation	Indicator method <sup>a</sup>
n (deaths)	1791 (264)	2341 (398)	2341 (398)	2341 (398)
Age	0.033 (0.013)	0.088 (0.009)	0.057 (0.011)	0.068 (0.01)
Male	0.92 (0.15)	0.82 (0.12)	1.00 (0.14)	0.92 (0.12)
Self-assessed health	0.38 (0.095)	0.60 (0.073)	0.39 (0.087)	0.46 (0.076)
Physical performance	-0.14 (0.023)	—	-0.15 (0.024)	-0.12 (0.022)
Intercept	-4.76 (1.17)	-10.41 (0.79)	-6.60 (1.04)	-7.42 (0.92)
Indicator of missing performance				-0.47 (0.13)

<sup>a</sup>The indicator method assigns the average performance score (6.8) to those with missing values and includes an indicator variable for this group.

We see there are significant effects of age and sex on mortality, with older people and males more likely to die by 1991. In addition, participants with lower levels of physical performance and lower levels of self-assessed health (i.e., higher scores) were more likely to die by 1991 after controlling for age and sex. However, a large number of people lacked some data on physical performance. The issue is whether excluding these people affected the estimates of the regression parameters in Table 13.42.

To incorporate participants with missing data into the analysis, we use the technique of *multiple imputation*. Imputation is defined roughly as the estimation of a missing variable (or variables) as a function of other covariates that are present. To use this technique, we predict the value of physical performance as a function of other covariates in the model, including the outcome (death by 1991). We base the imputation on the subset of 1791 participants with complete data.

### Example 13.75

**Aging** Estimate the value of physical performance as a function of the other variables in Table 13.42.

### Solution

We run a multiple-regression analysis of physical performance on age, sex, self-assessed health, and death by 1991 (1 = yes, 0 = no), with results shown in Table 13.43.

If we look at the first column of Table 13.43, we see that physical-performance score decreases with age, is higher for men than women, decreases as one's self-assessed health grows worse (i.e., increase in self-assessed health scale), and is lower

**Table 13.43** Summary of linear-regression model predicting physical-performance score, and 5 draws of these regression parameters

Variable	Estimate (SE)	5 Drawn values of parameters				
		1	2	3	4	5
Age (per year)	-0.22 (0.013)	-0.23	-0.23	-0.23	-0.21	-0.22
Male gender	1.18 (0.15)	1.21	1.08	1.29	1.25	1.22
Self-assessed health	-1.38 (0.089)	-1.33	-1.43	-1.24	-1.43	-1.35
Dead by 1991	-1.30 (0.20)	-1.12	-1.47	-1.37	-1.60	-1.05
Intercept	27.2 (1.1)	27.6	27.7	27.5	26.3	27.2
Overall $R^2$ 0.30; Root MSE 2.88						

for participants who will die by 1991. These variables explain 30% of the variation in physical performance (i.e.,  $R^2 = .30$ ). The Root mean square error (MSE) (2.88) provides an estimate of the residual variation of physical-performance score after adjusting for the predictors just given. We now need to examine how to incorporate the predicted physical-performance score into the overall analysis for participants who are missing data on this variable.

Two issues arise with using the regression parameters in the first column of Table 13.43 to predict physical performance. One issue is that the underlying regression parameters are unknown; the values given in Table 13.43 are estimated parameters with associated standard errors. The second issue is that even if the regression parameters were known perfectly, there would still be residual variation around the regression predictions of individual participants as reflected by the Root MSE. For these reasons, we have used PROC MI and PROC MIANALYZE of SAS (version 9.0) to provide estimated values of physical performance for each of the 350 participants with missing physical-performance data that reflect both uncertainty in the regression-parameter estimates as well as residual variation around the predictions. In the second column of Table 13.43 are a new set of regression parameters that represent an estimate of the true regression parameters that reflect the preceding two sources of variation. They differ somewhat from the estimates in the first column, as would be expected. To complete the data set for the subjects with missing data, we use the first set of drawn parameters in Table 13.43 (second column) to obtain an initial predicted physical-performance score, then draw a random  $N(0,1)$  deviate, multiply it by Root MSE and add it to the initial prediction to obtain a final predicted physical-performance score. This process is replicated for each of the 550 subjects with missing data. The completed data set now consists of 2341 participants with complete information on all covariates. A multiple-logistic-regression analysis similar to Table 13.42 was then run based on the completed data set. The results are shown in the first column of Table 13.44. The preceding process of estimating the missing physical-performance scores and rerunning the logistic regression on the completed data set was then repeated four additional times, yielding 5 separate estimates of the linear-regression parameters (Table 13.43) and 5 separate estimates of the logistic-regression parameters (Table 13.44).

The next issue is how to combine the results from the separate imputations in Table 13.44. In general, if we have  $m$  imputations then a suggestion by Rubin [42] is to compute  $\hat{\beta} = \sum_{i=1}^m \hat{\beta}_i/m$  as an overall estimate of effect over the  $m$  imputations. The

variance of  $\hat{\beta}$  should reflect both between- and within-imputation variance. In general, if there are  $m$  imputations, then

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^m \text{Var}(\hat{\beta}_i) / m + (1 + 1/m) \sum_{i=1}^m (\hat{\beta}_i - \bar{\beta})^2 / (m-1) \equiv W + B$$

where the first component ( $W$ ) reflects within-imputation variance and the second component ( $B$ ) reflects between-imputation variance. The test statistic is then given by

$$t = \hat{\beta} / \left[ \text{Var}(\hat{\beta}) \right]^{1/2} \sim t_d \text{ under } H_0$$

where  $d = (m-1)(1+1/r)^2$  and  $r = (1+1/m)(B/W)$ .

Rubin [42] recommends  $m = 5$  imputations because additional imputations do not yield any more meaningful reductions in bias or increases in precision. The summary estimates and standard errors over the 5 imputations are provided in the last column of Table 13.44.

**Table 13.44** Summary of 5 fits of the logistic-regression model with imputed performance scores and the average estimates (shown are regression estimates, SE)

	Imputation number					
	1	2	3	4	5	Average <sup>a</sup>
Age	0.059 (0.010)	0.056 (0.010)	0.058 (0.010)	0.055 (0.010)	0.055 (0.010)	0.057 (0.011)
Male gender	0.97 (0.12)	1.01 (0.12)	1.00 (0.12)	1.03 (0.13)	1.00 (0.12)	1.00 (0.14)
Self-assessed health	0.40 (0.078)	0.39 (0.079)	0.40 (0.078)	0.37 (0.078)	0.40 (0.078)	0.39 (0.087)
Physical performance	-0.14 (0.020)	-0.15 (0.020)	-0.14 (0.020)	-0.16 (0.021)	-0.15 (0.020)	-0.15 (0.024)
Intercept	-6.86 (0.93)	-6.54 (0.94)	-6.77 (0.93)	-6.32 (0.93)	-6.49 (0.94)	-6.60 (1.04)

<sup>a</sup>The average effect is the average of the 5 estimates from the filled-in data. The standard error of the average accounts for both the average of the variances of the 5 estimates as well as the variability among the 5 estimates. Specifically,  $SE(\text{average effect}) = \sqrt{\text{average within-imputation variation} + (6/5) \text{ between-imputation variance}}$ .

The entire multiple-imputation procedure is summarized as follows.

### Equation 13.81

#### Multiple-Imputation Approach for Incorporating Missing Data into an Overall Analysis

- (1) Suppose there are  $n$  subjects with  $k$  covariates  $x_1, \dots, x_k$  and a binary outcome variable  $y$ . We assume  $y$  and  $x_1, \dots, x_{k-1}$  are present for all subjects, whereas  $x_k$  is available for  $N_{\text{obs}}$  subjects and is missing for  $N_{\text{mis}}$  subjects.
- (2) We assume for the sake of specificity that  $x_k$  is continuous.
- (3) We run a multiple-regression analysis of  $x_k$  on  $x_1, \dots, x_{k-1}$  and  $y$  of the form  $x_k = \alpha + \gamma_1 x_1 + \dots + \gamma_{k-1} x_{k-1} + \delta y + e$  based on the subjects with complete data.

- (4) For each of the  $i = 1, \dots, N_{\text{mis}}$  subjects with missing data on  $x_k$  we calculate an estimated value for  $x_k$  denoted by  $\hat{x}_{i,k,1}$  for the  $i$ th subject, which reflects error both in the estimates of the regression parameters in step 3 as well as the residual variation about the regression line (i.e.,  $e_i$ ).

- (5) We run a logistic regression of the form

$$\ln[p / (1 - p)] = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

using the observed data for  $x_k$  for the  $N_{\text{obs}}$  subjects with complete data and the estimated  $x_k$  for the  $N_{\text{mis}}$  subjects with missing data on  $x_k$  based on step 4. The resulting estimates of  $\alpha$  and  $\beta$  are denoted by  $\hat{\alpha}_1$  and  $\hat{\beta}_{1,1}, \dots, \hat{\beta}_{k,1}$  and are referred to as the *regression coefficients from the first imputed data set*.

- (6) Steps 4 and 5 are repeated for  $m - 1$  additional imputations, thus yielding a set of estimated coefficients  $\hat{\alpha}_q, \hat{\beta}_{1,q}, \dots, \hat{\beta}_{k,q}$  and associated variances  $\text{Var}(\hat{\alpha}_q), \text{Var}(\hat{\beta}_{1,q}), \dots, \text{Var}(\hat{\beta}_{k,q})$  for the  $q$ th imputed data set,  $q = 1, \dots, m$ . Note that because the regression parameters differ for each imputation in step 4, estimates of  $x_k$  also differ for each imputed data set. However, values for  $x_1, \dots, x_{k-1}$  that have no missing data remain the same.
- (7) The estimates from the  $m$  separate imputations are then combined into an overall estimate for  $\beta_j, j = 1, \dots, k$ , given by

$$\hat{\beta}_j = \sum_{q=1}^m \hat{\beta}_{j,q} / m$$

and an overall variance given by

$$\text{Var}(\hat{\beta}_j) = \sum_{q=1}^m \text{Var}(\hat{\beta}_{j,q}) / m + [(m+1)/m] \sum_{q=1}^m (\hat{\beta}_{j,q} - \hat{\beta}_j)^2 / (m-1) \equiv W + B$$

A similar procedure is used to obtain an overall estimate of  $\alpha$ .

- (8) The overall test statistic to test the hypothesis  $H_0: \beta_j = 0$  vs.  $H_1: \beta_j \neq 0$  is given by

$$t = \hat{\beta}_j / \text{se}(\hat{\beta}_j) \sim t_d \text{ under } H_0$$

where  $\text{se}(\hat{\beta}_j) = [\text{Var}(\hat{\beta}_j)]^{1/2}$ ,  $d = (m-1)(1+1/r)^2$ , and  $r = (1+1/m)(B/W)$

- (9) The  $p$ -value =  $2 \times \Pr(t_d > |t|)$ .
- (10) In the usual implementation of multiple imputation (e.g., PROC MI and MIANALYZE of SAS),  $m$  is set equal to 5.

### Example 13.76

**Aging** Implement the multiple-imputation approach for incorporating missing data for 550 subjects missing the physical-performance score in the EPESE data set described in Example 13.74.

### Solution

We refer to the third column of Table 13.42. All variables remain statistically significant, as they were using the complete-case method. Interestingly, the regression coefficients for physical performance, gender, and self-assessed health are relatively similar between the complete-case method and the multiple-imputation method. However, the effect of age is substantially larger with the multiple-imputation approach, possibly because the frailest, oldest EPESE participants were less likely to

complete the physical-performance test. Furthermore, the standard errors of all variables that had no missing data are smaller with the multiple-imputation approach than with the complete-case method because of the larger sample size. However, the standard error of the physical-performance coefficient increased with the multiple-imputation approach, reflecting uncertainty in estimation of physical-performance scores for participants with missing data on this variable.

Another possible approach to the analysis is to exclude physical performance from the multiple-logistic model (see second column of Table 13.42). However, because of the strong correlation between physical-performance score and many of the other covariates, this produces unacceptable biases in the estimates of the other regression parameters in the absence of controlling for physical-performance score.

A second possible approach is to use all participants but include an indicator variable for missing physical-performance data (= 1 if missing; = 0 if present) and to give all participants with missing physical-performance data the average value (6.8) for the complete cases (see fourth column of Table 13.42). This includes all the participants but is known to yield biased estimates (1) because the effect of physical performance is underestimated and (2) because the correlation between physical performance and other predictor variables biases other parameters as well [43].

The imputation approach described in Equation 13.81 is for a single predictor measured with error. An extension to incorporate multiple missing predictors is available [42]. PROC MI and MIANALYZE of SAS (version 9.2) offer several options for how to perform imputation in this complex setting. Finally, although the multiple-imputation methods in this section have been applied in the context of multiple logistic regression, the same approach can be used for multiple linear regression with missing covariate values.

## 13.17 Summary

In this chapter, we have examined some of the main design and analysis techniques used in epidemiologic studies. In Section 13.2, we looked at the main study designs used in epidemiologic studies, including cohort studies, case-control studies, and cross-sectional studies. In Sections 13.3 and 13.4, we then explored some common measures of effect used in these studies, including the risk difference (*RD*), the risk ratio (*RR*), the odds ratio (*OR*), and the attributable risk (*AR*). For each study design, we discussed which of these parameters are estimable and which are not. In Section 13.5, we introduced the concept of a confounder and examined standardization, which is a descriptive technique for obtaining measures of effect that are controlled for confounding variables. In Sections 13.6 and 13.7, we discussed Mantel-Haenszel-type methods, which are analytic procedures used to test hypotheses about effects of a primary exposure variable while controlling for other confounding variable(s). These techniques become cumbersome when there are many confounding variables to be controlled for. Thus, in Section 13.8 we examined multiple logistic regression, a technique similar to multiple linear regression when the outcome variable is binary. Using this technique allows one to control for many confounding variables simultaneously. In Section 13.9, we considered several extensions to logistic regression. We first considered extensions to logistic regression for matched designs. We then introduced polychotomous logistic regression (PLR), in which an outcome variable is categorical with more than two possible outcomes that are not ordered and we wish to control for other covariates. Third, we considered ordinal logistic regression, in which our outcome variable is categorical with more than two ordered categories and we wish to control for one or more covariates.

The techniques in Sections 13.1–13.9 are standard methods of design and analysis used in epidemiologic studies. In recent years, there has been much interest in extensions

of these techniques to nonstandard situations, some of which are discussed in Sections 13.10–13.16. In Section 13.10, we discussed the basic principles of meta-analysis. Meta-analysis is a popular methodology for combining results obtained from more than one study regarding a particular association of interest. In Section 13.11, we considered the emerging field of active-control (or equivalence) studies. In standard clinical trials, to demonstrate the efficacy of an agent, the active agent is usually compared with a placebo. In active-control studies, a new proposed active agent is compared with an existing active agent (which we refer to as *standard* therapy). The goal of the study is to show that the two treatments are roughly equivalent rather than that the new active treatment is superior to standard therapy. The rationale for active-control studies is that in some instances it may be unethical to randomize a subject to placebo if a prior efficacious therapy already exists (e.g., in clinical trials of drugs used to treat schizophrenia).

Another alternative design used in clinical studies is the cross-over design, as discussed in Section 13.12. Under the usual parallel design for a clinical trial with two treatments, each subject is randomized to only one of two possible treatments. Under a cross-over design, each subject receives both treatments but at different time periods. A washout period when no treatment is given is usually specified between the two active-treatment periods. The order of administration of the two treatments for an individual subject is randomized. The rationale for this design is that it usually requires fewer subjects than a parallel design, provided that the effect of treatment given in the first period does not carry over to the second active-treatment period. It is most appropriate for short-acting therapies with no carry-over effect.

In Section 13.13, we considered the statistical treatment of clustered binary data. Clustered binary data occur in clinical trials or observational studies when the unit of randomization is different from the unit of analysis. For example, in some lifestyle interventions (e.g., dietary interventions), the unit of randomization might be a school or school district, but the unit of analysis is the individual child. Modifications to ordinary techniques for analyzing  $2 \times 2$  tables (discussed in Section 10.2) were introduced to account for the correlation of response from different children in the same school or school district. In addition, we considered regression methods for clustered binary data based on GEE techniques. We also considered methods for longitudinal data analysis in Section 13.14.

In Section 13.15, we considered the emerging field of measurement-error-correction methods. These techniques provide generalizations of standard techniques, such as logistic regression, that account for the common occurrence that a noisy convenient exposure is often used in epidemiologic studies, such as a single blood-pressure measurement, when what is really desired is a more accurate “gold standard” measurement (e.g., the “true” blood pressure) conceptualized as the average of a large number of blood-pressure measurements for an individual subject. Using these techniques, we can estimate the logistic regression that would have been obtained if the gold-standard exposure had been available for all subjects instead of the surrogate exposure. We also introduced the concept of a validity study and a reproducibility study, which are ancillary studies that seek to estimate the relationship between the true and surrogate exposure, in the case when the gold standard is measurable and when it is not, respectively.

Finally, in Section 13.16, we described approaches for handling missing data in epidemiologic studies. Data commonly are missing. The default option in most statistical packages is the complete-case method, in which only subjects with complete data on all predictor variables are included. However, this approach may introduce some bias if subjects with missing data differ systematically from subjects with complete data. A more sophisticated approach based on multiple-imputation methods is introduced as a possible alternative that incorporates data from both subjects with and without missing data into the overall analysis.

## PROBLEMS

### Gynecology

In a 1985 study of the relationship between contraceptive use and infertility, 89 of 283 infertile women, compared with 640 of 3833 control (fertile) women, had used an intrauterine device (IUD) at some time in their lives [44].

**\*13.1** Use the normal-theory method to test for significant differences in contraceptive-use patterns between the two groups.

**\*13.2** Use the contingency-table method to perform the test in Problem 13.1.

**\*13.3** Compare your results in Problems 13.1 and 13.2.

**\*13.4** Compute a 95% CI for the difference in the proportion of women who have ever used IUDs between the infertile and fertile women in Problem 13.1.

**\*13.5** Compute the *OR* in favor of ever using an IUD for infertile women vs. fertile women.

**\*13.6** Provide a 95% CI for the true *OR* corresponding to your answer to Problem 13.5.

**13.7** What is the relationship between your answers to Problems 13.2 and 13.6?

### Renal Disease

**13.8** Refer to Problem 10.30. Estimate the *RR* for total mortality of the study group vs. the control group. Provide 95% confidence limits for the *RR*.

### Infectious Disease

Refer to Table 13.8.

**13.9** Perform a significance test to examine the association between OC use and bacteriuria after controlling for age.

**13.10** Estimate the *OR* in favor of bacteriuria for OC users vs. non-OC users after controlling for age.

**13.11** Provide a 95% CI for the *OR* estimate in Problem 13.10.

**13.12** Is the association between bacteriuria and OC use comparable among different age groups? Why or why not?

**13.13** Suppose you did not control for age in the preceding analyses. Calculate the crude (unadjusted for age) odds ratio in favor of bacteriuria for OC users vs. non-OC users.

**13.14** How do your answers to Problems 13.10 and 13.13 relate to each other? Explain any differences found.

### Endocrinology

A study was performed looking at the risk of fractures in three rural Iowa communities according to whether their drinking water was "higher calcium," "higher fluorides," or "control" as determined by water samples. Table 13.45 presents data comparing the rate of fractures (over 5 years) between the higher-calcium vs. the control communities for women ages 20–35 and 55–80, respectively [45].

**\*13.15** What test can be used to compare the fracture rates in these two communities while controlling for age?

**\*13.16** Implement the test in Problem 13.15, and report a *p*-value (two-sided).

**\*13.17** Estimate the *OR* relating higher calcium and fractures while controlling for age.

**\*13.18** Provide a 95% CI for the estimate obtained in Problem 13.17.

### Pulmonary Disease

Read "Influence of Passive Smoking and Parental Phlegm on Pneumonia and Bronchitis in Early Childhood," by J. R. T. Colley, W. W. Holland, and R. T. Corkhill, in *Lancet*, November 2, 1974, pages 1031–1034, and answer the following questions based on it.

**13.19** Perform a statistical test comparing the incidence rates of pneumonia and bronchitis for children in their first year of life in families in which both parents are smokers vs. families in which both parents are nonsmokers.

**13.20** Compute an *OR* to compare the incidence rates of pneumonia and bronchitis in families in which both parents are smokers vs. families in which both parents are nonsmokers.

**13.21** Compute a 95% CI corresponding to the *OR* computed in Problem 13.20.

**13.22** Compare the incidence rates of pneumonia and bronchitis for children in their first year of life in families in which one parent is a smoker vs. families in which both parents are nonsmokers.

**13.23** Answer Problem 13.20 comparing families in which one parent is a smoker with families in which both parents are nonsmokers.

**Table 13.45 Relationship of calcium content of drinking water to the rate of fractures in rural Iowa**

Ages 20–35	Number of women with fractures	Total number of women	Ages 55–80	Number of women with fractures	Total number of women
Control	3	37	Control	11	121
Higher calcium	1	33	Higher calcium	21	148

**13.24** Compute a 95% CI corresponding to the OR estimate computed in Problem 13.23.

**13.25** Is there a significant trend in the percentage of children with pneumonia and bronchitis in the first year of life according to the number of smoking parents? Report a *p*-value.

**13.26** Suppose we wish to compare the incidence rates of disease for children in their first and second years of life in families in which both parents are nonsmokers. Rates of 7.8% and 8.1% based on samples of size 372 and 358, respectively, are presented in Table II. Would it be reasonable to use a chi-square test to compare these rates?

**13.27** Perform a statistical test comparing the incidence rates of pneumonia and bronchitis for children in the first year of life when stratified by number of cigarettes per day. (Use the groupings in Table IV.) Restrict your analyses to families in which one or both parents are current smokers. Specifically, is there a consistent trend in the incidence rates as the number of cigarettes smoked increases after controlling for the respiratory symptom history of the parents?

**13.28** Does the number of siblings in the family affect the incidence rate of pneumonia and bronchitis for children in the first year of life? Use the data in Table VI to control for the confounding effects of (1) parental smoking habits and (2) parental respiratory-symptom history.

## Mental Health

Refer to Problem 10.32.

**\*13.29** Estimate the OR relating widowhood to mortality based on all the data in Table 10.28.

**\*13.30** Provide a 95% CI for the OR.

## Hypertension

A study was conducted in Wales relating blood-pressure and blood-lead levels [46]. It was reported that 4 of 455 men with blood-lead levels  $\leq 11 \text{ } \mu\text{g}/100 \text{ mL}$  had elevated SBP ( $\geq 160 \text{ mm Hg}$ ), whereas 16 of 410 men with blood-lead levels  $\geq 12 \text{ } \mu\text{g}/100 \text{ mL}$  also had elevated SBP. It was also reported that 6 of 663 women with blood-lead levels  $\leq 11 \text{ } \mu\text{g}/100 \text{ mL}$  had elevated SBP, whereas 1 of 192 women with blood-lead levels  $\geq 12 \text{ } \mu\text{g}/100 \text{ mL}$  had elevated SBP.

**13.31** What is an appropriate procedure to test the hypothesis that there is an association between blood pressure and blood lead, while controlling for sex?

**13.32** Implement the procedure in Problem 13.31, and report a *p*-value.

**13.33** Estimate the OR relating blood pressure to blood lead, and provide a 95% CI about this estimate.

## Infectious Disease

Aminoglycoside antibiotics are particularly useful clinically in the treatment of serious gram-negative bacterial infections among hospitalized patients. Despite their potential for toxicity, as well as the continued development of newer antimicrobial agents of other classes, it seems likely that the clinical use of aminoglycosides will continue to be widespread. The choice of a particular aminoglycoside antibiotic for a given patient depends on several factors, including the specific clinical situation, differences in antimicrobial spectrum and cost, and risks of side effects, particularly nephrotoxicity and auditory toxicity. Many randomized, controlled trials have been published that compare the various aminoglycoside antibiotics with respect to efficacy, nephrotoxicity, and, to a lesser extent, auditory toxicity. These individual trials have varied widely with respect to their design features and their conclusions. A major limitation to their interpretability is that the majority of the individual trials have lacked an adequate sample size to detect the small-to-moderate differences between treatment groups that are most plausible. As a result, the individual trials published to date have generally not permitted firm conclusions, especially concerning the relative potential for toxicity of aminoglycosides.

In these circumstances, one method to estimate the true effects of these agents more precisely is to conduct an overview, or *meta-analysis*, of the data from all randomized trials. In this way, a true increase in risk could emerge that otherwise would not be apparent in any single trial due to small sample size. Therefore, a quantitative overview of the results of all published randomized controlled trials that assessed the efficacy and toxicity of individual aminoglycoside antibiotics was undertaken.

Forty-five randomized clinical trials, published between June 1975 and September 1985, were identified that compared two or more of five aminoglycoside antibiotics: amikacin, gentamicin, netilmicin, sisomicin, and tobramycin. Thirty-seven of these trials could provide data suitable for comparative purposes.

The specific endpoints of interest were efficacy, nephrotoxicity, and auditory toxicity (ototoxicity). Efficacy was defined as bacterial or clinical response to treatment as reported in each individual trial. Nephrotoxicity was defined as the percentage of toxic events to the kidney reported, regardless of whether the published paper suggested some explanation other than the use of the study drug, such as use of another potentially nephrotoxic agent, or the presence of an underlying disease affecting kidney function. Auditory toxicity was defined as reported differences between pre- and post-treatment audiograms.

The data are organized into three Data Sets: EFF.DAT, NEPHRO.DAT, and OTO.DAT, all on the Companion Website. A separate record is presented for each antibiotic studied for each endpoint. The format is given in the files EFF.DOC, NEPHRO.DOC, and OTO.DOC, on the Companion Website.

Columns 1–8: Study name

10–11: Study number (number on reference list)

13: Endpoint (1 = efficacy; 2 = nephrotoxicity; 3 = ototoxicity)

15: antibiotic (1 = amikacin; 2 = gentamicin; 3 = netilmicin; 4 = sisomicin; 5 = tobramycin)

17–19: Sample size

21–23: Number cured (for efficacy) or number with side effect (for nephrotoxicity or ototoxicity)

## Renal Disease

Refer to Data Set NEPHRO.DAT on the Companion Website.

**13.34** Use methods of meta-analysis to assess whether there are differences in nephrotoxicity between each pair of antibiotics. Obtain point estimates and 95% CIs for the *OR*, and provide a two-sided *p*-value.

## Otolaryngology

Refer to Data Set OTO.DAT on the Companion Website.

**13.35** Answer the question in Problem 13.34 to assess whether there are differences in ototoxicity between each pair of antibiotics.

## Infectious Disease

Refer to Data Set EFF.DAT on the Companion Website.

**13.36** Answer the question in Problem 13.34 to assess whether there are differences in efficacy between each pair of antibiotics.

## Cardiology

A recent study compared the use of percutaneous transluminal coronary angioplasty (PTCA) with medical therapy in the treatment of single-vessel coronary-artery disease. A total of 105 patients were randomly assigned to PTCA and 107 to medical therapy. Over a period of 6 months, MI occurred in 5 patients in the PTCA group and 3 patients in the medical-therapy group.

\***13.37** Estimate the *RR* of MI for patients assigned to PTCA vs. patients assigned to medical therapy, and provide a 95% CI for this estimate.

At the 6-month clinic visit, 61 of 96 patients seen in the PTCA group and 47 of 102 patients seen in the medical-therapy group were angina free.

\***13.38** Answer Problem 13.37 for the endpoint of being angina free at 6 months.

## Sports Medicine

Refer to Problem 10.55. In this problem, we described Data Set TENNIS1.DAT (on the Companion Website), which is an observational study relating episodes of tennis elbow to other risk factors.

**13.39** Use logistic-regression methods to compare participants with 1 + episodes of tennis elbow vs. participants with 0 episodes of tennis elbow, considering multiple risk factors in the same model.

**13.40** Use linear-regression methods to predict the number of episodes of tennis elbow as a function of several risk factors in the same model.

## Hypertension

A drug company proposes to introduce a new antihypertensive agent that is aimed at elderly hypertensive participants with prior heart disease. Because this is a high-risk group, the company is hesitant to withhold antihypertensive therapy from these patients and instead proposes an equivalence study comparing the new agent (drug A) with the current antihypertensive therapy used by those participants. Hence the participants will be randomized to either maintenance of their current therapy or replacement of their current therapy with drug A. Suppose the endpoint is total cardiovascular disease (CVD) mortality, and it is assumed that under their current therapy 15% of participants will die of CVD over the next 5 years.

**13.41** Suppose drug A will be considered equivalent to the current therapy if the 5-year CVD mortality is not worse than 20%. How many participants must be enrolled in the study to ensure at least an 80% chance of demonstrating equivalence if equivalence will be based on a one-sided 95% CI approach, an equal number of subjects are randomized to drug A and current therapy, and the underlying mortality rates of the two therapies are the same?

**13.42** Suppose in the actual study that 200 participants are randomized to each group. Forty-four participants who receive drug A and 35 participants who receive current therapy die of CVD in the next 5 years. Can the treatments be considered equivalent? Why or why not?

**13.43** How much power did the study described in Problem 13.42 have of demonstrating equivalence under the assumptions in Problem 13.41?

## Cardiovascular Disease

Sudden death is an important, lethal cardiovascular endpoint. Most previous studies of risk factors for sudden death have focused on men. Looking at this issue for women is important as well. For this purpose, data were used from the Framingham Heart Study [47]. Several potential risk factors, such as age, blood pressure, and cigarette smoking, are of interest and need to be controlled for simultaneously. Therefore, a multiple logistic-regression model was fitted to these data, as shown in Table 13.46.

**13.44** Assess the statistical significance of the individual risk factors.

**13.45** What do these statistical tests mean in this instance?

**13.46** Compute the *OR* relating the additional risk of sudden death per 100-centiliter (cL) decrease in vital capacity after adjustment for the other risk factors.

**Table 13.46 Multiple logistic-regression model relating 2-year incidence of sudden death in females without prior CHD (data taken from the Framingham Heart Study) to several risk factors**

Risk factor	Regression coefficient, $\hat{\beta}_I$	$se(\hat{\beta}_I)$
Constant	-15.3	
Systolic blood pressure (mm Hg)	0.0019	0.0070
Framingham relative weight (%)	-0.0060	0.0100
Cholesterol (mg/100 mL)	0.0056	0.0029
Glucose (mg/100 mL)	0.0066	0.0038
Cigarette smoking (cigarettes/day)	0.0069	0.0199
Hematocrit (%)	0.111	0.049
Vital capacity (cL)	-0.0098	0.0036
Age (years)	0.0686	0.0225

Source: Arthur Schatzkin et al., "Sudden Death in the Framingham Heart Study: Differences in Incidence and Risk Factors by Sex and Coronary Disease Status, *American Journal of Epidemiology*, 1984 120: 888-899. Reprinted by permission of Oxford University Press.

**13.47** Provide a 95% CI for the estimate in Problem 13.46.

### Hepatic Disease

Refer to Data Set HORMONE.DAT on the Companion Website.

**13.48** Use logistic-regression methods to assess whether presence of biliary secretions during the second period (any or none) is related to the type of hormone used during the second period.

**13.49** Answer the same question as in Problem 13.48 for the presence of pancreatic secretions.

**13.50** Use logistic-regression methods to assess whether the presence of biliary secretions during the second period is related to dose of hormone used during the second period (do separate analyses for each active hormone—hormones 2–5).

**13.51** Answer the same question as in Problem 13.50 for the presence of pancreatic secretions.

### Otolaryngology

Refer to Data Set EAR.DAT (see Table 3.11, also on the Companion Website).

**13.52** Consider a subject "cured" if (1) the subject is a unilateral case and the ear clears by 14 days or (2) the subject is a bilateral case and both ears are clear by 14 days. Run a logistic regression with outcome variable = cured and independent variables (1) antibiotic, (2) age, and (3) type of case (unilateral or bilateral). Assess goodness of fit of the model you obtain.

**13.53** Use correlated binary data methods to relate clearance of an ear by 14 days to antibiotic type and age. Use the ear as the unit of analysis. (*Hint:* Use generalized estimating equation methods.)

### Sports Medicine

Refer to Data Set TENNIS2.DAT, on the Companion Website.

**13.54** Assess whether there are significant treatment effects regarding pain during maximum activity.

**13.55** Assess whether there are significant treatment effects regarding pain 12 hours after maximum activity.

**13.56** Assess whether there are significant treatment effects regarding pain on an average day.

**13.57** Assess whether there are significant carry-over effects for the endpoint in Problem 13.54.

**13.58** Assess whether there are significant carry-over effects for the endpoint in Problem 13.55.

**13.59** Assess whether there are significant carry-over effects for the endpoint in Problem 13.56.

### Hypertension

Refer to Data Set ESTROGEN.DAT on the Companion Website. The format is in Table 13.47.

Three separate two-period cross-over studies were performed, based on different groups of subjects. Study 1 compared 0.625 mg estrogen with placebo. Study 2 compared 1.25 mg estrogen with placebo. Study 3 compared 1.25 mg

**Table 13.47 Format of Data Set ESTROGEN.DAT**

Variable	Column	Comments
Subject	1–2	
Treatment	4	1 = placebo, 2 = 0.625 mg estrogen, 3 = 1.25 mg estrogen
Period	6	
Mean SBP	8–10	mm Hg
Mean DBP	12–14	mm Hg

estrogen with 0.625 mg estrogen. Subjects received treatment for 4 weeks in each active-treatment period; a 2-week washout period separated the two active-treatment periods.

**13.60** Assess whether there are any significant treatment or carry-over effects of SBP and DBP in study 1.

**13.61** Answer Problem 13.60 for study 2.

**13.62** Answer Problem 13.60 for study 3.

**13.63** Suppose we are planning a new study similar in design to study 1. How many participants do we need to study to detect an underlying 3-mm Hg treatment effect for SBP with 80% power assuming there is no carry-over effect and we perform a two-sided test with  $\alpha = .05$ ? (*Hint:* Use the sample standard deviation of the difference scores from study 1 as an estimate of the true standard deviation of the difference scores in the proposed study.)

**13.64** Answer Problem 13.63 for an underlying 2-mm Hg treatment effect for DBP.

**13.65** Answer Problem 13.63 for a new study similar in design to study 2.

**13.66** Answer Problem 13.64 for a new study similar in design to study 2.

### Otolaryngology

A longitudinal study was conducted among children in the Greater Boston Otitis Media Study [48]. Based on all doctor visits during the first year of life, children were classified as having 1+ episodes vs. 0 episodes of otitis media (OTM). A separate classification was performed for the right and left ears. Several risk factors were studied as possible predictors of OTM. One such risk factor was a sibling history of ear infection, with relevant data displayed in Table 13.48.

**13.67** Assess whether a sibling history of ear infection is associated with OTM incidence in the first year of life. (*Hint:* Use clustered binary data methods based on Equation 13.59.)

**13.68** Provide a 95% CI for the true difference in incidence rates for children with siblings between those with and without a sibling history of ear infection.

**13.69** Answer the questions in Problems 13.67 and 13.68 using generalized estimating equation methods, and compare results with the solution to Problems 13.67–13.68.

### Otolaryngology

Consider Data Set EAR.DAT (see Table 3.11).

Suppose we use the ear as the unit of analysis, where the outcome is a success if an ear clears by 14 days and a failure otherwise.

**13.70** Compare the percentage of cleared ears between the cefaclor-treated and the amoxicillin-treated groups. Report a two-tailed  $p$ -value.

**13.71** Compare the percentage of cleared ears among children 2–5 years of age vs. the percentage of cleared ears among children <2 years of age. Report a two-tailed  $p$ -value.

**13.72** Compare the percentage of cleared ears among children 6+ years of age vs. the percentage of cleared ears among children <2 years of age. Report a two-tailed  $p$ -value.

### Cancer, Nutrition

A logistic-regression analysis similar to that presented in Example 13.70 was run relating breast-cancer incidence in 1980–1984 to calorie-adjusted total fat (heretofore referred to as *total fat intake*) as reported on a 1980 FFQ. In addition, age in 5-year categories and alcohol in categories (0, 0.1–4.9, 5.0–14.9, 15+ g/day) were also controlled for. The regression coefficient for a 10-g/day increase in total fat intake was  $-0.163$  with standard error =  $0.135$ .

**13.73** Obtain a point estimate and a 95% CI for the relative risk of breast cancer comparing women whose total fat intake differs by 10 g/day.

The validation-study data discussed in Section 13.15 are available in Data Set VALID.DAT.

**Table 13.48 Association between sibling history (Hx) of ear infection and number of episodes of OTM in the first year of life**

Group 1			Group 2		
Sibling Hx ear infection = yes			Sibling Hx ear infection = no		
Right ear	Left ear	<i>n</i>	Right ear	Left ear	<i>n</i>
–	–	76	–	–	115
+	–	21	+	–	20
–	+	20	–	+	18
+	+	<u>77</u>	+	+	<u>91</u>
Total		194	Total		244

Note: + = 1+ episodes of OTM in the first year of life in a specific ear; – = 0 episodes of OTM in the first year of life in a specific ear.

**13.74** Use the data for total fat to fit the linear regression of DR total fat intake on FFQ total fat intake. Obtain the regression coefficient, standard error, and *p*-value from this regression.

**13.75** Using the results from Problems 13.73 and 13.74, obtain an estimate of the *RR* of breast cancer, comparing women who differ by 10 g/day on total fat intake on the DR, assuming age and alcohol intake have no measurement error and are not correlated with total fat intake.

**13.76** Obtain a 95% CI for the point estimate in Problem 13.75.

**13.77** Compare the measurement-error-corrected *RR* and CI in Problems 13.75 and 13.76 with the uncorrected *RR* and CI in Problem 13.73.

### Cancer, Endocrinology

In the study presented in Example 13.72, other hormones were considered in addition to plasma estradiol. Table 13.49 presents the uncorrected relative-risk estimates and 95% CIs for several other hormones [35].

**13.78** Obtain the uncorrected logistic-regression coefficients and standard errors for each of the hormones in Table 13.49.

The hormones in Table 13.49 were also included in the reproducibility study mentioned in Example 13.72 [37]. The

intraclass correlation coefficient and sample size used for each hormone are given in Table 13.50.

**13.79** Obtain the measurement-error-corrected logistic-regression coefficient and standard error for each of the hormones in Table 13.49.

**13.80** Using the results from Problem 13.79, obtain a measurement-error-corrected *OR* and 95% CI for each of the hormones.

**13.81** How do the results from Problem 13.80 compare with the results in Table 13.49?

### Environmental Health, Pediatrics

Refer to Data Set LEAD.DAT on the Companion Website. One goal of the study was to assess the effect of lead level in 1972 (variable name LD72) on neurological and psychological measures of health, while controlling for age and sex. One problem is that lead-level data in 1972 are incomplete for some children (coded as 99).

**13.82** Use the complete-case method to relate lead levels in 1972 to full-scale IQ score (variable name IQF), while controlling for age and sex.

**13.83** Repeat the analysis in Problem 13.82 using multiple-imputation methods.

**13.84** Compare your results in Problems 13.82 and 13.83.

**Table 13.49** Relative-risk estimates and 95% CIs for breast-cancer incidence from 1989 to June 1, 1994, in a nested case-control study among 11,169 postmenopausal women in the NHS not taking hormone-replacement therapy in 1989, comparing women at the median value of the fourth quartile vs. women at the median value of the first quartile of the hormone distributions

Hormone	Median value 1st quartile	Median value 4th quartile	<i>RR</i>	95% CI
Free estradiol (%)	1.33	1.82	1.69	1.03–2.80
Estrone (pg/mL)	17	45	1.91	1.15–3.16
Testosterone (ng/dL)	12	37	1.65	1.00–2.71

Note: Comparing women at the median value of the fourth quartile vs. women at the median value of the first quartile, where the quartiles are determined from the distribution of hormones among controls.

**Table 13.50** Intraclass correlation coefficients (ICCs) for selected hormones from the NHS reproducibility study, 1989

Hormone	ICC	Number of subjects with			Total number of measurements
		3 replicates	2 replicates		
Free estradiol (%)	0.80	79	0		237
Estrone (pg/mL)	0.74	72	6		228
Testosterone (ng/dL)	0.88	79	0		237

**Table 13.51 Comparison of fracture incidence between raloxifene- and placebo-treated women**

	No pre-existing fractures			Pre-existing fractures		
	New fractures	No new fractures	Total	New fractures	No new fractures	Total
Raloxifene	34	1466	1500	Raloxifene	103	597
Placebo	68	1432	1500	Placebo	170	630
Total	102	2898	3000	Total	273	1227
						1500

### Endocrinology

A study of raloxifene and incidence of fractures was conducted among women with evidence of osteoporosis. The women were initially divided into two groups: those with and those without pre-existing fractures. The women were then randomized to raloxifene or placebo and followed for 3 years to determine the incidence of new vertebral fractures, with the results shown in Table 13.51.

**13.85** Among those with no pre-existing fractures, test whether raloxifene affects the incidence of new fractures.

**13.86** Among those with no pre-existing fractures, compute the relative risk of new fractures among those randomized to raloxifene vs. placebo, along with its associated 95% CI.

**13.87** Test the association of study agent with new fractures combining both groups of those with and without preexisting fractures.

**13.88** Combining both groups, compute the standardized RR for raloxifene vs. placebo and new fractures. (*Hint:* Use the total population as the standard.)

**13.89** Is pre-existing fracture a confounder in these data?

### Hypertension

Suppose that 200 obese ( $BMI \geq 25$ ) children and 500 normal-weight ( $BMI < 25$ ) children are identified in a school-based screening for hypertension. Eighteen of the obese children and 10 of the normal-weight children are hypertensive.

**13.90** What type of study is this?

**13.91** What is the RR for hypertension? What is a 95% CI associated with this estimate?

**13.92** Suppose that 30% of children are obese by the above definition. What percent of hypertension is attributable to obesity? Provide a 95% CI associated with this estimate.

### Hypertension

An important issue is whether there are racial differences in hypertension among children. We define hypertension as being above the 95th percentile for either systolic blood pressure (SBP) or diastolic blood pressure (DBP) among children of the same age, height, and sex. Since some of the children were observed at multiple visits, a GEE model was run of hypertension on ethnic group.

There were three ethnic groups considered: Caucasian, African-American, and Hispanic. The results among boys are given in Table 13.52.

**Table 13.52 Relationship between hypertension and ethnic group among 27,009 boys in the Pediatric Task Force Data**

Variable	Regression coefficient	se
Intercept	-2.07	0.026
African-American	0.049	0.041
Hispanic	0.328	0.059

**13.93** What is the estimated OR for hypertension comparing Hispanic boys vs. Caucasian boys? (Call this  $OR_1$ .) What is a 95% CI for this estimate?

One issue is that BMI, which may vary among ethnic groups, is positively related to hypertension. Hence, a second logistic regression model was run, as shown in Table 13.53.

**Table 13.53 Relationship between hypertension, ethnic group, and BMI among 27,009 boys in the Pediatric Task Force Data**

Variable	Regression coefficient	se
Intercept	-4.277	0.090
African-American	0.009	0.042
Hispanic	0.186	0.060
BMI ( $kg/m^2$ )	0.107	0.004

**13.94** What is the estimated OR for hypertension comparing Hispanic boys vs. Caucasian boys based on the results in Table 13.53? (Call this  $OR_2$ .) Provide a 95% CI for this estimate. What is the difference in interpretation between  $OR_1$  and  $OR_2$ ?

**13.95** Suppose the average BMI of Hispanic boys is higher than that for Caucasian boys. Is BMI a confounder of the association between ethnic group and hypertension? If so, is it a positive or negative confounder?

One assumption of the model in Table 13.53 is that the association between hypertension and ethnic group would be

the same for all levels of BMI. To test this assumption a third logistic model was run as presented in Table 13.54.

**Table 13.54 Possible effect modification of the association between hypertension and ethnic group by BMI among 27,009 boys in the Pediatric Task Force Data**

Variable	Regression coefficient	se
Intercept	-2.169	0.028
African-American	0.063	0.044
Hispanic	0.231	0.066
BMI-20*	0.123	0.006
African-American × (BMI-20)	-0.035	0.009
Hispanic × (BMI-20)	-0.024	0.012

\* BMI minus 20

**13.96** Is BMI an effect modifier of the association between hypertension and ethnic group? Why or why not?

**13.97** What is the estimated OR for hypertension comparing Hispanic vs. Caucasian boys with  $BMI = 25 \text{ kg/m}^2$  (call this  $OR_3$ )?

### Cardiovascular Disease

A study was performed relating baldness pattern to MI (heart attack) among men in the Atherosclerosis Risk in Communities (ARIC) study [49]. Baldness pattern and prevalent MI were determined at the same examination during the period 1996–1998. Baldness was categorized into 5 categories (none/frontal/mild vertex/moderate vertex/severe vertex). For this example, we focus on the comparison of severe vertex baldness to no baldness. The data in Table 13.55 were reported by age group.

**Table 13.55 Association between severe vertex baldness and MI in the ARIC study**

Age group	Baldness	MI	No MI	Total
$\leq 60$ years	Severe vertex	49	280	329
	None	71	639	710
	Total	120	919	1039
$> 60$ years	Severe vertex	131	656	787
	None	144	782	926
	Total	275	1438	1713

**13.98** What type of study was this?

**13.99** What is the estimated OR for MI comparing men with severe vertex baldness vs. no baldness after controlling for age?

**13.100** Is there a significant association between MI and severe vertex baldness after controlling for age? Please report a two-tailed  $p$ -value.

**13.101** What is the OR between MI and severe vertex baldness in (i) men  $\leq 60$  and (ii) men  $> 60$ ? If these are the true ORs, is age an effect modifier of the association between baldness and MI? Why or why not?

### Renal Disease

Refer to Data Set SWISS.DAT on the Companion Website. **13.102** Use methods of longitudinal data analysis to compare the rates of change in serum creatinine over time by treatment group.

**13.103** Compare your results with those obtained using ordinary ANOVA methods based on slopes in Problem 12.49.

### Cardiovascular Disease

Refer to Table 13.7.

**13.104** Assess the crude association between MI and OC use without taking age into account (provide a two-tailed  $p$ -value).

**13.105** Assess the association between MI and OC use after controlling for age (provide a two-tailed  $p$ -value).

**13.106** Estimate the OR between OC use and MI after controlling for age, and provide a 95% CI about this estimate.

**13.107** Is there evidence of effect modification of the OC–MI relationship by age?

### Mental Health

Refer to Table 10.28.

**13.108** Estimate the OR between widowhood and mortality, and provide a 95% CI about this estimate.

### Cancer

A case-control study was performed early in the NHS to assess the possible association between oral contraceptive (OC) use and ovarian cancer [50]. Forty seven ovarian cancer cases were identified at or before baseline (1976). For each case, 10 controls matched by year of birth and with intact ovaries at the time of the index woman's diagnosis were randomly chosen from questionnaire respondents free from ovarian cancer. The data in Table 13.56 were presented.

**Table 13.56 Duration of OC use by age at diagnosis among women with ovarian cancer and controls**

		Duration OC use		
Age at diagnosis		Never	<3 years	3+ years
Under 35	Case	9	2	0
	Control	55	42	12
35–44	Case	13	2	4
	Control	127	27	30
45+	Case	12	3	2
	Control	129	18	23

**13.109** Use logistic regression methods to assess whether there is an association between ovarian cancer risk and duration of OC use while controlling for age. Provide a two-sided  $p$ -value. Assume that the average duration of use in the < 3 years group = 1.5 years and in the 3+ years group = 4 years. Also, provide an estimate of the  $OR$  relating ovarian cancer risk per year of use of OCs and a 95% CI.

**13.110** Use logistic regression methods to assess whether there is an association between ever use of OCs and ovarian cancer risk, while controlling for age. Also, provide an estimate of the  $OR$  and a 95% CI about this estimate.

### Ophthalmology

The data in Table 13.57 were presented relating body mass index (BMI) to progression of advanced age-related macular degeneration (AMD), a common eye disease in the elderly that results in significant visual loss [51].

**Table 13.57 Association between BMI and progression of AMD**

BMI	Progression	Nonprogression
<25	72	423
≥25	209	762

**13.111** What is the attributable risk ( $AR$ ) of high BMI ( $\geq 25$ ) for progression of AMD?

**13.112** Provide a 95% CI about this estimate.

### Cardiovascular Disease

The Women's Health Study randomly assigned 39,876 initially healthy women ages 45 years or older to receive either 100 mg of aspirin on alternate days or placebo and monitored them for 10 years for a major cardiovascular event [52]. Table 13.58 shows the results stratified by age at randomization.

**Table 13.58 Incidence of CVD by treatment group and age in the Women's Health Study**

Age	Treatment group	CVD=yes	CVD=no
45–54	Aspirin	163	11,847
	Placebo	161	11,854
55–64	Aspirin	183	5693
	Placebo	186	5692
≥65	Aspirin	131	1917
	Placebo	175	1874

Use logistic regression methods to characterize the relationship between aspirin assignment and the odds of CVD, by doing the following.

**13.113** Obtain the crude  $OR$  estimate, and provide a 95% CI for the crude  $OR$ .

**13.114** Test the null hypothesis of no association between aspirin assignment and CVD.

**13.115** Evaluate whether age confounds the CVD–aspirin relationship by using dummy variables for age categories; calculate the age-adjusted  $OR$  estimate and 95% CI.

**13.116** Evaluate whether age is an effect modifier of the relationship between aspirin and CVD.

### Cancer

The data file BLOOD.DAT (on the Companion Website) contains data from a case–control study assessing several plasma risk factors for breast cancer. The women were matched approximately by age at the blood draw, fasting status and, if possible, current PMH use at the time of the blood draw. There was 1 case and either 1 or 2 controls per matched set, although some of the matched sets are incomplete due to missing data. The matching variable is matchid.

Use logistic regression methods to assess the association between testosterone and breast cancer risk after controlling for age at the blood draw and current PMH use and taking the matching into account.

Perform the analysis in two ways:

**13.117** Treat testosterone as a continuous variable (suitably transformed if necessary).

**13.118** Treat testosterone as a categorical variable in quartiles, with the 1st quartile as the reference group.

**13.119** Discuss your results from Problems 13.117 and 13.118.

### Cancer

Results from a population-based case–control study of ovarian cancer were recently reported from the North Carolina Case–Control Study based on data collected from

1999–2008 [53]. Cases were women with ovarian cancer who were ages 20–74 from 48 North Carolina counties; controls were frequency matched by age and race and were recruited from the same geographic regions using random-digit dialing. Controls could not have a bilateral oophorectomy. The data in Table 13.59 were reported concerning the association between age at menarche (age when periods begin) and ovarian cancer.

**Table 13.59 Association between age at menarche and ovarian cancer**

Age at menarche	Caucasians		African-Americans	
	Cases	Controls	Cases	Controls
<12	181	157	28	53
≥12	562	708	82	136

**13.120** For both Caucasians and African-Americans, estimate the *OR* between late age at menarche ( $\geq 12$ )

and ovarian cancer and provide a 95% CI about this estimate.

Two logistic models were run with Stata Version 11 using these data.

$$\text{Logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2,$$

where  $x_1$  = age at menarche (1 represents  $\geq 12$ , 0 represents  $< 12$ ),

$x_2$  = race (1 = African-American, 0 = Caucasian).

$$\text{Logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

**13.121** Estimate the *OR* for the association between age at menarche and ovarian cancer after controlling for race, and provide a 95% CI about this estimate.

**13.122** What does the variable  $\beta_1$  mean in the 2nd logistic regression (Table 13.61)? How does it differ from the meaning of  $\beta_1$  in the 1st logistic regression (Table 13.60)?

**13.123** Assess whether the effect of age at menarche is different for Caucasian vs. African-American women. Report a *p*-value (two-tailed).

**Table 13.60 Logistic regression of ovarian cancer on age at menarche and race**

```
. logit case ageatmenarche race [fweight=freq]
Logistic regression
Number of obs      =      1907
LR chi2(2)        =      15.76
Prob > chi2       =     0.0004
Pseudo R2         =     0.0060

-----+
case |      Coef.    Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
ageatmenar-e |  -.2860765  .1113846   -2.57  0.010  -.5043864  -.0677666
      race |  -.4082626  .1304354   -3.13  0.002  -.6639112  -.152614
      _cons |   .0735805  .1010801     0.73  0.467  -.1245329  .271694
-----+
. gen agemenarche_race=ageatmenarche * race
```

**Table 13.61 Logistic regression of ovarian cancer on age at menarche, race, and age at menarche × race**

```
. logit case ageatmenarche race agemenarche_race [fweight=freq]
Logistic regression
Number of obs      =      1907
LR chi2(3)        =      18.66
Prob > chi2       =     0.0003
Pseudo R2         =     0.0071

-----+
case |      Coef.    Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
ageatmenar-e |  -.3731935  .1228254   -3.04  0.002  -.6139268  -.1324601
      race |  -.7803386  .2578292   -3.03  0.002  -1.285675  -.2750027
      agemenarche
      * race |   .5053452  .29869     1.69  0.091  -.0800764  1.090767
      _cons |   .1422512  .1090609     1.30  0.192  -.0715043  .3560067
-----+
```

## REFERENCES

- [1] Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19, 251–253.
- [2] Shapiro, S., Slone, D., Rosenberg, L., et al. (1979). Oral contraceptive use in relation to myocardial infarction. *Lancet*, 1, 743–747.
- [3] Evans, D. A., Hennekens, C. H., Miao, L., Laughlin, L. W., Chapman, W. G., Rosner, B., Taylor, J. O., & Kass, E. H. (1978). Oral contraceptives and bacteriuria in a community-based study. *New England Journal of Medicine*, 299, 536–537.
- [4] Sandler, D. P., Everson, R. B., & Wilcox, A. J. (1985). Passive smoking in adulthood and cancer risk. *American Journal of Epidemiology*, 121(1), 37–48.
- [5] Robins, J. M., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large strata limiting models. *Biometrics*, 42, 311–323.
- [6] Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., & Badr, S. (1993). The occurrence of sleep-disordered breathing among middle-aged adults. *New England Journal of Medicine*, 328, 1230–1235.
- [7] Pesch, B., Ranft, U., Jakubis, P., Nieuwenhuijsen, M. J., Hergemoller, A., Unfried, K., Jakubis, M., Miskovic, P., Keegan, T., & The EXPASCAN Study Group. (2002). Environmental arsenic exposure from a coal-burning power plant as a potential risk factor for nonmelanoma skin carcinoma: Results from a case control study in the District of Prievidza, Slovakia. *American Journal of Epidemiology*, 155(9), 798–809.
- [8] Lipnick, R. J., Buring, J. E., Hennekens, C. H., Rosner, B., Willett, W., Bain, C., Stampfer, M. J., Colditz, G. A., Peto, R., & Speizer, R. E. (1986). Oral contraceptives and breast cancer: A prospective cohort study. *JAMA*, 255, 58–61.
- [9] Munoz, A., & Rosner, B. (1984). Power and sample size estimation for a collection of 2 x 2 tables. *Biometrics*, 40, 995–1004.
- [10] McCormack, W. M., Rosner, B., McComb, D. E., Ervard, J. R., & Zinner, S. H. (1985). Infection with *Chlamydia trachomatis* in female college students. *American Journal of Epidemiology*, 121(1), 107–115.
- [11] Dawber, T. R. (1980). *The Framingham Study*. Cambridge, MA: Harvard University Press.
- [12] Rosner, B., Spiegelman, D., & Willett, W. C. (1992). Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology*, 136, 1400–1413.
- [13] Takkouche, B., Regueira-Mendez, C., Garcias-Closas, R., Figueiras, A., Gestal-Otero, J. J., & Hernan, M. A. (2002). Intake of wine, beer, and spirits and the risk of clinical common cold. *American Journal of Epidemiology*, 155(9), 853–858.
- [14] Colditz, G., Rosner, B. A., Chen, W. Y., Holmes, M. D., & Hankinson, S. E. (2004). Risk factors for breast cancer according to estrogen and progesterone receptor status. *Journal of the National Cancer Institute*, 96, 218–228.
- [15] Breslow, N., & Day, N. E. (1980). *Statistical Methods in Cancer Research: Vol. 1 – The Analysis of Case-Control Studies*. Lyon, France: IARC Scientific Publications.
- [16] Marshall, R. J., & Chisholm, E. M. (1985). Hypothesis testing in the polychotomous logistic model with an application to detecting gastrointestinal cancer. *Statistics in Medicine*, 4, 337–344.
- [17] Buring, J. E., Evans, D. A., Mayrent, S. L., Rosner, B., Colton, T., & Hennekens, C. H. (1988). Randomized trials of aminoglycoside antibiotics. *Reviews of Infectious Disease*, 10(5), 951–957.
- [18] DerSimonian, R., & Laird, N. M. (1986). Meta analysis in clinical trials. *Controlled Clinical Trials*, 7, 177–188.
- [19] Bodey, G. P., Chang, H-Y., Rodriguez, V., & Stewart, D. (1975). Feasibility of administering aminoglycoside antibiotics by continuous intravenous infusion. *Antimicrobial Agents and Chemotherapy*, 8, 328–333.
- [20] Hedges, L. V., & Olkin, I. (1985). *Statistical methods in meta analysis*. London: Academic Press.
- [21] Makuch, R., & Simon, R. (1978). Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports*, 62, 1037–1040.
- [22] Temple, R. (1996). Problems in interpreting active control equivalence trials. *Accountability in Research*, 4, 267–275.
- [23] Rothman, K. J., & Michels, K. B. (1994). The continuing unethical use of placebo controls. *New England Journal of Medicine*, 331(16), 394–398.
- [24] Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.
- [25] Grizzle, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics*, 21, 467–480.
- [26] Rowe, N. H., Brooks, S. L., Young, S. K., Spencer, J., Petrick, T. J., Buchanan, R. A., Drach, J. C., & Shipman, C. (1979). A clinical trial of topically applied 3 percent vidarabine against recurrent herpes labialis. *Oral Pathology*, 47, 142–147.
- [27] Banting, D. W., Ellen, R. P., & Fillery, E. D. (1985). A longitudinal study of root caries: Baseline and incidence data. *Journal of Dental Research*, 64, 1141–1144.
- [28] Imrey, P. B. (1986). Considerations in the statistical analyses of clinical trials in periodontics. *Journal of Clinical Periodontology*, 13, 517–528.
- [29] Fleiss, J. L., Park, M. H., & Chilton, N. W. (1987). Within-mouth correlations and reliabilities for probing depth and attachment level. *Journal of Periodontology*, 58, 460–463.
- [30] Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- [31] Statistical Analysis System, Version 9.2, Cary N.C.
- [32] Laird, N., & Ware, J. (1982). Random effects models for longitudinal data: An overview of recent results. *Biometrics*, 38, 963–974.
- [33] Rosner, B., Munoz, A., Tager, L., Speizer, F., & Weiss, S. (1985). The use of an autoregressive model for the analysis of longitudinal data in epidemiologic studies. *Statistics in Medicine*, 4, 457–467.
- [34] Rosner, B., Willett, W. C., & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and

confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8, 1051–1069.

[35] Hankinson, S. E., Willett, W. C., Manson, J. E., Colditz, G. A., Hunter, D. J., Spiegelman, D., Barbieri, R. L., & Speizer, F. E. (1998). Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women. *Journal of the National Cancer Institute*, 90, 1292–1299.

[36] Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, 54, 67–82.

[37] Hankinson, S. E., Manson, J. E., Spiegelman, D., Willett, W. C., Longcope, C., & Speizer, F. E. (1995). Reproducibility of plasma hormone levels in postmenopausal women over a 2–3 year period. *Cancer, Epidemiology, Biomarkers and Prevention*, 4, 649–654.

[38] Rosner, B., Spiegelman, D., & Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology*, 132, 734–745.

[39] Rosner, B., Spiegelman, D., & Willett, W. C. (1992). Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology*, 136, 1400–1413.

[40] Glynn, R. J., Beckett, L. A., Hebert, L. E., Morris, M. C., Scherr, P. A., & Evans, D. A. (1999). Current and remote blood pressure and cognitive decline. *JAMA*, 281, 438–445.

[41] Guralnik, J. M., Simonsick, E. M., Ferrucci, L., Glynn, R. J., Berkman, L. F., Blazer, D. G., Scherr, P. A., & Wallace, R. B. (1994). A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission. *Journal of Gerontology—Medical Science*, 49, M85–M94.

[42] Rubin, D. B. (1987). *Multiple imputation for non-response in surveys*. New York: Wiley.

[43] Vach, W., & Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values in confounding variables. *American Journal of Epidemiology*, 134(8), 895–907.

[44] Cramer, D. W., Schiff, I., Schoenbaum, S. C., Gibson, M., Belisle, J., Albrecht, B., Stillman, R. J., Berger, M. J.,

Wilson, E., Stadel, B. V., & Seibel, M. (1985). Tubal infertility and the intrauterine device. *New England Journal of Medicine*, 312(15), 941–947.

[45] Sowers, M. F. R., Clark, M. K., Jannausch, M. L., & Wallace, R. B. (1991). A prospective study of bone mineral content and fracture in communities with differential fluoride exposure. *American Journal of Epidemiology*, 133(7), 649–660.

[46] Elwood, P. C., Yarnell, J. W. G., Oldham, P. D., Catford, J. C., Nutbeam, D., Davey-Smith, G., & Toothill, C. (1988). Blood pressure and blood lead in surveys in Wales. *American Journal of Epidemiology*, 127(5), 942–945.

[47] Schatzkin, A., Cupples, L. A., Heeren, T., Morelock, S., & Kannel, W. B. (1984). Sudden death in The Framingham Heart Study: Differences in incidence and risk factors by sex and coronary disease status. *American Journal of Epidemiology*, 120(6), 888–899.

[48] Teele, D. W., Klein, J. O., & Rosner, B. (1989). Epidemiology of otitis media during the first seven years of life in children in Greater Boston: A prospective, cohort study. *Journal of Infectious Disease*, 160(1), 83–94.

[49] Shahar, E., Heiss, G., Rosamond, W. D., & Szklo, M. (2008). Baldness and myocardial infarction in men: The Atherosclerosis Risk in Communities Study. *American Journal of Epidemiology*, 167, 676–683.

[50] Willett, W. C., Bain, C., Hennekens, C. H., Rosner, B., & Speizer, F. E. (1981). Oral contraceptives and risk of ovarian cancer. *Cancer*, 48, 1684–1687.

[51] Seddon, J. M., Francis, P. J., George, S., Schultz, D. W., Rosner, B., & Klein, M. L. (2007). Association of CFH Y402H and LOC 387715 A69S with progression of age-related macular degeneration. *Journal of the American Medical Association*, 297(16), 1793–1800.

[52] Ridker, P. M., Cook, N. R., Lee, I.-M., Gordon, D., Gaziano, J. M., Manson, J. E., Hennekens, C. H., & Buring, J. E. (2005). A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *New England Journal of Medicine*, 352, 1293–1304.

[53] Moorman, P. G., Palmieri, R. T., Akushevich, L., Berchuck, A., & Schildkraut, J. M. (2009). Ovarian cancer risk factors in African-American and white women. *American Journal of Epidemiology*, 70, 598–606.

# Hypothesis Testing: Person-Time Data

## 14.1 Measure of Effect for Person-Time Data

In Chapter 10, we discussed the analysis of categorical data, where the person was the unit of analysis. In a prospective-study design, we identified groups of exposed and unexposed individuals at baseline and compared the proportion of subjects who developed disease over time between the two groups. We referred to these proportions as *incidence rates*, although a more technically appropriate term would be **cumulative incidence rates** (see Definition 3.20). Cumulative incidence (*CUMI*) rates are proportions based on the person as the unit of analysis and must range between 0 and 1. In computing cumulative incidence rates, we implicitly assume all subjects are followed for the same period of time  $T$ . This is not always the case, as the following example shows.

### Example 14.1

**Cancer** A hypothesis of much recent interest is the possible association between the use of oral contraceptives (OC) and the development of breast cancer. To address this issue, data were collected in the Nurses' Health Study (NHS) in which disease-free women were classified in 1976 according to OC status (current user/past user/never user). A questionnaire was mailed out every 2 years in which OC status was updated, and breast-cancer status was ascertained over the next 2 years. The amount of time that each woman was a current user or a never user of OCs (ignoring past use) was calculated and this *person-time* was accumulated over the entire cohort of nurses. Thus, each nurse contributed a different amount of person-time to the analysis. The data are presented in Table 14.1 for current and never users of OCs among women 45–49 years of age. How should these data be used to assess any differences in the incidence rate of breast cancer by OC-use group?

**Table 14.1** Relationship between breast-cancer incidence and OC use among 45- to 49-year-old women in the NHS

OC-use group	Number of cases	Number of person-years
Current users	9	2935
Never users	239	135,130

The first issue to consider is the appropriate unit of analysis for each group. If the woman is used as the unit of analysis, then the problem is that different women may contribute different amounts of person-time to the analysis, and the assumption

of a constant probability of an event for each woman is violated. If the person-year is used as the unit of analysis (i.e., one person followed for 1 year), then because each woman can contribute more than 1 person-year to the analysis, the important assumption of independence for the binomial distribution is violated.

To allow for varying follow-up time for each individual, we define the concept of incidence density.

**Definition 14.1**

The **incidence density** in a group is defined as the number of events in that group divided by the total person-time accumulated during the study in that group.

The denominator used in computing incidence density is the person-year. Unlike cumulative incidence, incidence density may range from 0 to  $\infty$ .

**Example 14.2**

**Cancer** Compute the estimated incidence density among current and never OC users in Table 14.1.

**Solution**

The incidence density among current users =  $9/2935 = .00307$  events per person-year = 307 events per 100,000 person-years. The incidence density among never users =  $239/135,130 = .00177$  events per person-year = 177 events per 100,000 person-years.

In following a subject, the incidence density may remain the same or may vary over time (e.g., as a subject ages over time, the incidence density generally increases). How can we relate cumulative incidence over time  $t$  to incidence density? Suppose for simplicity that incidence density remains the same over some time period  $t$ . If  $CUMI(t)$  = cumulative incidence over time  $t$  and  $\lambda$  = incidence density, then it can be shown using calculus methods that

**Equation 14.1**

$$CUMI(t) = 1 - e^{-\lambda t}$$

If the cumulative incidence is low (<.1), then we can approximate  $e^{-\lambda t}$  by  $1 - \lambda t$  and  $CUMI(t)$  by

**Equation 14.2**

$$CUMI(t) \approx 1 - (1 - \lambda t) = \lambda t$$

This relationship is summarized as follows.

**Equation 14.3**
**Relationship Between Cumulative Incidence and Incidence Density**

Suppose we follow a group of individuals with constant incidence density  $\lambda$  = number of events per person-year. The exact cumulative incidence over time period  $t$  is

$$CUMI(t) = 1 - e^{-\lambda t}$$

If the cumulative incidence is low (<.1), then the cumulative incidence can be approximated by

$$CUMI(t) = \lambda t$$

Later in this chapter we refer to incidence density by the more common term *incidence rate* ( $\lambda$ ) and distinguish it from the cumulative incidence over some time period  $t = \text{CUMI}(t)$ . The former can range from 0 to  $\infty$ , whereas the latter is a proportion and must vary between 0 and 1.

**Example 14.3**

**Cancer** Suppose the incidence density of breast cancer in 40- to 44-year-old premenopausal women is 200 events per 100,000 person-years. What is the cumulative incidence of breast cancer over 5 years among 40-year-old initially disease-free women?

**Solution**

From Equation 14.3, we have  $\lambda = 200/10^5$ ,  $t = 5$  years. Thus the exact cumulative incidence is given by

$$\begin{aligned}\text{CUMI}(5) &= 1 - e^{-(200/10^5)5} \\ &= 1 - e^{-1000/10^5} \\ &= 1 - e^{-10^{-2}} = 1 - e^{-0.01} = .00995 = 995/10^5\end{aligned}$$

The approximate cumulative incidence is given by

$$\text{CUMI}(5) \approx (200/10^5) \times 5 = .01 = 1000/10^5$$

## 14.2 One-Sample Inference for Incidence-Rate Data

### Large-Sample Test

**Example 14.4**

**Cancer, Genetics** A registry is set up during the period 1990–1994 of women with a suspected genetic marker for breast cancer but who have not yet had breast cancer. Five hundred 60- to 64-year-old women are identified and followed until December 31, 2000. Thus the length of follow-up is variable. The total length of follow-up is 4000 person-years, during which 28 new cases of breast cancer occurred. Is the incidence rate of breast cancer different in this group from that in the general population of 60- to 64-year-old women if the expected incidence rate is  $400/10^5$  person-years in this age group?

We want to test the hypothesis  $H_0: ID = ID_0$  vs.  $H_1: ID \neq ID_0$ , where  $ID$  = the unknown incidence density (rate) in the genetic-marker group and  $ID_0$  = the known incidence density (rate) in the general population. We base our test on the observed number of breast cancer cases, which we denote by  $a$ . We assume  $a$  approximately follows a Poisson distribution. Under  $H_0$ ,  $a$  has mean =  $t(ID_0)$  and variance =  $t(ID_0)$ , where  $t$  = total number of person-years. If we assume the normal approximation to the Poisson distribution is valid, then this suggests the following test procedure.

**Equation 14.4**

### One-Sample Inference for Incidence-Rate Data (Large-Sample Test)

Suppose that  $a$  events are observed over  $t$  person-years of follow-up and that  $ID$  = underlying incidence density (rate). To test the hypothesis  $H_0: ID = ID_0$  vs.  $H_1: ID \neq ID_0$ ,

- (1) Compute the test statistic

$$X^2 = \frac{(a - \mu_0)^2}{\mu_0} \sim \chi^2_1 \text{ under } H_0$$

where

$$\mu_0 = t(ID_0)$$

- (2) For a two-sided test at level  $\alpha$ ,  $H_0$  is rejected if

$$X^2 > \chi_{1,1-\alpha}^2$$

and accepted if

$$X^2 \leq \chi_{1,1-\alpha}^2$$

- (3) The exact  $p$ -value =  $Pr(\chi_1^2 > X^2)$ .

- (4) This test should only be used if  $\mu_0 = t(ID_0) \geq 10$ .

### Example 14.5

#### Solution

**Cancer, Genetics** Perform a significance test based on the data in Example 14.4.

We have that  $a = 28$ ,  $\mu_0 = (400/10^5)(4000) = 16$ . Thus the test statistic is given by

$$\begin{aligned} X^2 &= \frac{(28-16)^2}{16} \\ &= \frac{144}{16} = 9.0 \sim \chi_1^2 \text{ under } H_0 \end{aligned}$$

Because  $\chi_{1,995}^2 = 7.88$ ,  $\chi_{1,999}^2 = 10.83$ , and  $7.88 < 9.0 < 10.83$ , it follows that  $.001 < p < .005$ . Thus there is a significant excess of breast cancers in the genetic-marker group.

### Exact Test

Suppose the number of events is too small to apply the large-sample test in Equation 14.4. In this case, an exact test based on the Poisson distribution must be used. If  $\mu = t(ID)$ , then we can restate the hypotheses in the form:  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$  and apply the one-sample Poisson test as follows.

### Equation 14.5

#### One-Sample Inference for Incidence-Density (Rate)

#### Data (Small-Sample Test)

Suppose that  $a$  events are observed over  $t$  person-years of follow-up and that  $ID$  = underlying incidence density (rate). We wish to test the hypothesis  $H_0: ID = ID_0$  vs.  $H_1: ID \neq ID_0$ .

Under  $H_0$ , the observed number of events ( $a$ ) follows a Poisson distribution with parameter  $\mu_0 = t(ID_0)$ . Thus the exact two-sided  $p$ -value is given by

$$\min\left(2 \times \sum_{k=0}^a \frac{e^{-\mu_0} \mu_0^k}{k!}, 1\right) \text{ if } a < \mu_0$$

$$\min\left[2 \times \left(1 - \sum_{k=0}^{a-1} \frac{e^{-\mu_0} \mu_0^k}{k!}\right), 1\right] \text{ if } a \geq \mu_0$$

### Example 14.6

**Cancer, Genetics** Suppose 125 of the 500 women in Example 14.4 also have a family history of breast cancer in addition to having the genetic marker. Eight cases of breast cancer are observed in this subgroup over 1000 person-years. Does this

subgroup have a breast-cancer incidence significantly different from the general population?

**Solution**

The expected number of breast cancers in this subgroup =  $1000(400/10^5) = 4$ . Thus the expected number of cases is too small to use the large-sample test in Equation 14.4. Instead, we use the small-sample test in Equation 14.5. We have  $a = 8$ ,  $\mu_0 = 4$ . Because  $8 \geq 4$ , we have

$$p\text{-value} = 2 \times \left( 1 - \sum_{k=0}^{7} \frac{e^{-4} 4^k}{k!} \right)$$

From the Poisson tables (Table 2 in the Appendix), this is given by

$$\begin{aligned} p\text{-value} &= 2 \times [1 - (.0183 + .0733 + \dots + .0595)] \\ &= 2 \times (1 - .9489) \\ &= .102 \end{aligned}$$

Hence, breast-cancer incidence in this subgroup is not significantly different from that in the general population. A larger number of person-years of follow-up are needed to obtain more power in this case.

## Confidence Limits for Incidence Rates

To obtain confidence limits for  $ID$ , we obtain confidence limits for the expected number of events ( $\mu$ ) based on the Poisson distribution and then divide each confidence limit by  $t$  = number of person-years of follow-up. Specifically, we have  $\hat{\mu} = a$ ,  $Var(\hat{\mu}) = a$ . Thus, if the normal approximation to the Poisson distribution holds, then a  $100\% \times (1 - \alpha)$  confidence interval (CI) for  $\mu$  is given by  $a \pm z_{1-\alpha/2} \sqrt{a}$ . The corresponding  $100\% \times (1 - \alpha)$  CI for  $ID = (a \pm z_{1-\alpha/2} \sqrt{a}) / t$ . Otherwise, we obtain exact confidence limits for  $\mu$  from Table 8 in the Appendix and divide each confidence limit by  $t$  to obtain the corresponding CI for  $ID$ . The procedure is summarized as follows.

**Equation 14.6**
**Point and Interval Estimation for Incidence Rates**

Suppose that  $a$  events are observed over  $t$  person-years of follow-up.

- (1) A point estimate of the incidence density rate =  $\hat{ID} = a/t$ .
- (2) To obtain a two-sided  $100\% \times (1 - \alpha)$  CI for  $\mu$ ,
  - (a) If  $a \geq 10$ , then compute  $a \pm z_{1-\alpha/2} \sqrt{a} = (c_1, c_2)$ ;
  - (b) If  $a < 10$ , then obtain  $(c_1, c_2)$  from Appendix Table 8 by referring to the  $a$  row and the  $1 - \alpha$  column.
- (3) The corresponding two-sided  $100\% \times (1 - \alpha)$  CI for  $ID$  is given by  $(c_1/t, c_2/t)$ .

**Example 14.7**

**Cancer, Genetics** Obtain a point estimate and a two-sided 95% CI for  $ID$  based on the data in Example 14.4.

**Solution**

We have  $a = 28$ ,  $t = 4000$ . Hence the point estimate of  $ID = 28/4000 = .007 = 700/10^5$  person-years =  $\hat{ID}$ . Because  $a \geq 10$ , to obtain a 95% CI for  $\mu$ , we refer to 2(a) in Equation 14.6 and obtain the confidence limits for  $\mu$  given by

$$28 \pm 1.96 \sqrt{28} = 28 \pm 10.4 = (17.6, 38.4) = (c_1, c_2)$$

The corresponding 95% CI for  $ID = (17.6/4000, 38.4/4000) = (0.00441, 0.00959)$  or  $(441/10^5 \text{ person-years}, 959/10^5 \text{ person-years})$ . This interval excludes the null rate of  $400/10^5 \text{ person-years}$  given in Example 14.4.

**Example 14.8** **Cancer, Genetics** Obtain a point estimate and a two-sided 95% CI for  $ID$  based on the data in Example 14.6.

**Solution** From Example 14.6, the expected number of breast-cancer cases in this subgroup (i.e.,  $\mu_0 = 4$ ). The point estimate of  $ID = 8/1000 = .008 = 800/10^5 \text{ person-years}$ . Because  $a = 8 < 10$ , we use 2(b) of Equation 14.6 to obtain a 95% CI for  $ID$ . Referring to the  $a = 8$  row and 0.95 column of Table 8 in the Appendix, we have a 95% CI for  $\mu = (3.45, 15.76)$ . The corresponding 95% CI for  $ID = (3.45/1000, 15.76/1000) = (345/10^5 \text{ person-years}, 1576/10^5 \text{ person-years})$ . This interval includes the general population rate of  $400/10^5 \text{ person-years}$ .

In this section, we have learned about incidence density (also called *incidence rate*), which is expressed as the number of events per unit time, and have distinguished it from cumulative incidence, which is the probability of an event occurring over time  $t$ . We have considered one-sample inference for incidence rates. The inference procedures are based on modeling the number of events over time  $t$  by a Poisson distribution. We have used a large-sample test based on the normal approximation to the Poisson distribution when the expected number of events is  $\geq 10$  and a small-sample test based on the exact Poisson probabilities when the expected number of events is  $< 10$ . Finally, we also discussed methods of point and interval estimation for incidence rates. In the next section, we extend this discussion to investigate methods for comparing incidence rates from two samples.

On the flowchart at the end of this chapter (Figure 14.15, p. 803), we answer yes to (1) person-time data? and (2) one-sample problem? This leads us to the box labeled “Use one-sample test for incidence rates.”

### REVIEW QUESTIONS 14A

- 1 What is the difference between incidence density and cumulative incidence?
- 2 Suppose we observe 20 cases of ovarian cancer among 10,000 women ages 50–69, each of whom is followed for 10 years. Provide a point estimate and a 95% CI for the incidence density.
- 3 Suppose we observe 8 cases of ovarian cancer among a subset of 2000 women from the group in Review Question 14A.2, each of whom is followed for 10 years. The subgroup consists of women who are overweight (body-mass index [BMI]  $\geq 25 \text{ kg/m}^2$ ). Provide a point estimate and a 95% CI for the incidence density in this subgroup.

## 14.3 Two-Sample Inference for Incidence-Rate Data

### Hypothesis Testing—General Considerations

The question we address in this section is, How can we compare the underlying incidence rates between two different exposure groups?

The approach we will take is to use a *conditional* test. Specifically, suppose we consider the case of two exposure groups and have the general table in Table 14.2.

**Table 14.2 General observed table for comparing incidence rates between two groups**

Exposure group	Number of events	Person-time
1	$a_1$	$t_1$
2	$a_2$	$t_2$
Total	$a_1 + a_2$	$t_1 + t_2$

We wish to test the hypothesis  $H_0: ID_1 = ID_2$  vs.  $H_1: ID_1 \neq ID_2$ , where  $ID_1$  = true incidence density in group 1 = the number of events per unit of person-time in group 1, and  $ID_2$  is the comparable rate in group 2. Under the null hypothesis, a fraction  $t_1/(t_1 + t_2)$  of the total number of events ( $a_1 + a_2$ ) would be expected to occur in group 1 and a fraction  $t_2/(t_1 + t_2)$  of the total number of events to occur in group 2. Thus under  $H_0$ , conditional on the observed total number of events =  $a_1 + a_2$ , the expected number of events in each group is given by

**Equation 14.7**

$$\text{Expected number of events in group 1} = E_1 = (a_1 + a_2)t_1/(t_1 + t_2)$$

$$\text{Expected number of events in group 2} = E_2 = (a_1 + a_2)t_2/(t_1 + t_2)$$

**Example 14.9**

**Cancer** Compute the expected number of events among current and never users for the OC-breast-cancer data in Table 14.1.

**Solution**

We have  $a_1 = 9$ ,  $a_2 = 239$ ,  $t_1 = 2935$  person-years,  $t_2 = 135,130$  person-years. Therefore, under  $H_0$ , from Equation 14.7,  $2935/(2935 + 135,130) = .0213$  of the cases would be expected to occur among current OC users and  $135,130/(2935 + 135,130) = .9787$  of the cases to occur among never OC users. Thus

$$E_1 = .0213(248) = 5.27$$

$$E_2 = .9787(248) = 242.73$$

## Normal-Theory Test

To assess statistical significance, the number of events in group 1 under  $H_0$  is treated as a binomial random variable with parameters  $n = a_1 + a_2$  and  $p_0 = t_1/(t_1 + t_2)$ . Under this assumption, the hypotheses can be stated as  $H_0: p = p_0$  vs.  $H_1: p \neq p_0$ , where  $p$  = the true proportion of events that are expected to occur in group 1. We also assume that the normal approximation to the binomial distribution is valid. Using the normal approximation to the binomial distribution, the observed number of events in group 1 =  $a_1$  is normally distributed with mean =  $np_0 = (a_1 + a_2)t_1/(t_1 + t_2) = E_1$  and variance =  $np_0q_0 = (a_1 + a_2)t_1t_2/(t_1 + t_2)^2 = V_1$ .  $H_0$  is rejected if  $a_1$  is much smaller or larger than  $E_1$ . This is an application of the large-sample one-sample binomial test, given by the following.

**Equation 14.8**

### Comparison of Incidence Rates (Large-Sample Test)

To test the hypothesis  $H_0: ID_1 = ID_2$  vs.  $H_1: ID_1 \neq ID_2$ , where  $ID_1$  and  $ID_2$  are the true incidence densities in groups 1 and 2, use the following procedure:

## (1) Compute the test statistic

$$z = \frac{a_1 - E_1 - .5}{\sqrt{V_1}} \text{ if } a_1 > E_1$$

$$= \frac{a_1 - E_1 + .5}{\sqrt{V_1}} \text{ if } a_1 \leq E_1$$

where  $E_1 = (a_1 + a_2)t_1/(t_1 + t_2)$ 

$$V_1 = (a_1 + a_2)t_1t_2/(t_1 + t_2)^2$$

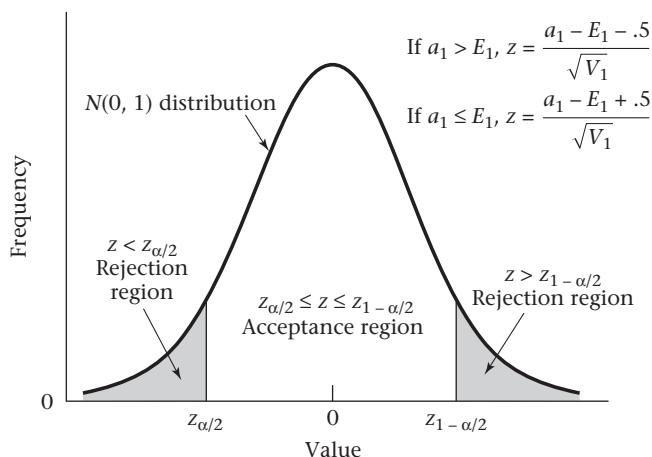
 $a_1, a_2$  = number of events in groups 1 and 2 $t_1, t_2$  = amount of person-time in groups 1 and 2Under  $H_0$ ,  $z \sim N(0, 1)$ (2) For a two-sided level  $\alpha$  test,if  $z > z_{1-\alpha/2}$  or  $z < z_{\alpha/2}$ , then reject  $H_0$ .if  $z_{\alpha/2} \leq z \leq z_{1-\alpha/2}$ , then accept  $H_0$ .(3) Use this test only if  $V_1 \geq 5$ .(4) The  $p$ -value for this test is given by

$$2 \times [1 - \Phi(z)] \quad \text{if } z \geq 0$$

$$2 \times \Phi(z) \quad \text{if } z < 0$$

The critical regions and  $p$ -value are illustrated in Figures 14.1 and 14.2, respectively.

**Figure 14.1** Acceptance and rejection regions for the two-sample test for incidence rates (normal-theory method)

**Example 14.10**

**Cancer** Assess the statistical significance of the OC-breast-cancer data in Table 14.1.

**Solution**

From Example 14.9,  $a_1 = 9$ ,  $a_2 = 239$ ,  $t_1 = 2935$ ,  $t_2 = 135,130$ ,  $E_1 = 5.27$ ,  $E_2 = 242.73$ . Furthermore,

$$V_1 = \frac{(a_1 + a_2)t_1 t_2}{(t_1 + t_2)^2}$$

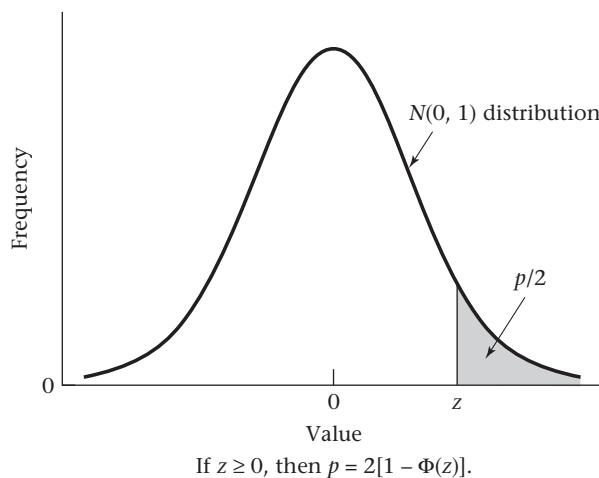
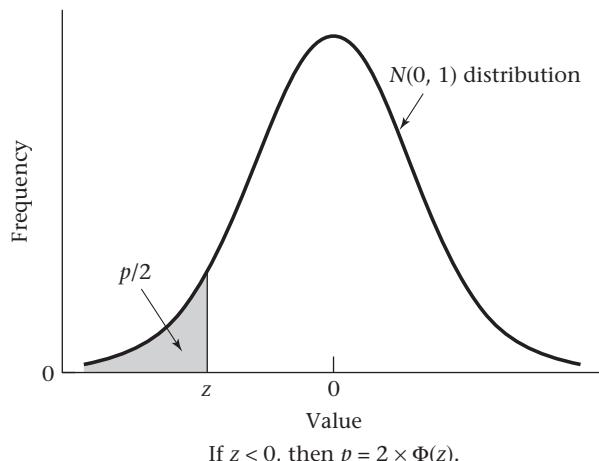
$$= \frac{(9 + 239)(2935)(135,130)}{(2935 + 135,130)^2} = \frac{9.8358 \times 10^{10}}{138,065^2} = 5.16$$

Because  $V_1 \geq 5$ , we can use the large-sample test in Equation 14.8.

Therefore, because  $a_1 > E_1$ ,  $z = \frac{9 - 5.27 - .5}{\sqrt{5.16}} = \frac{3.23}{2.27} = 1.42 \sim N(0,1)$

The  $p$ -value =  $2 \times [1 - \Phi(1.42)] = 2 \times (1 - .9223) = .155$ . Thus the results are not statistically significant and there is no significant difference in the incidence rate of breast cancer between current OC users and never OC users in this age group.

**Figure 14.2 Computation of the  $p$ -value for the two-sample test for incidence rates (normal-theory method)**



## Exact Test

Suppose the number of events is too small to apply the normal-theory test (i.e.,  $V_1 < 5$ ). In this case, an exact test based on the binomial distribution must be used. From the earlier discussion (p. 731), under  $H_0$ , the number of events in group 1 ( $a_1$ ) will follow a binomial distribution with parameters  $n = a_1 + a_2$  and  $p = p_0 = t_1/(t_1 + t_2)$ . We want to test the hypothesis  $H_0: p = p_0$  vs.  $H_1: p \neq p_0$ , where  $p$  is the underlying proportion of events that occur in group 1. This is an application of the exact one-sample binomial test.  $H_0$  will be rejected if the observed number of events  $a_1$  is much smaller or much larger than the expected number of events  $= E_1 = np_0$ . The following test procedure is used.

### Equation 14.9

#### Comparison of Incidence Rates—Exact Test

Let  $a_1, a_2$  be the observed number of events and  $t_1, t_2$  the amount of person-time in groups 1 and 2, respectively. Let  $p$  = true proportion of events in group 1. To test the hypothesis  $H_0: ID_1 = ID_2$  (or equivalently,  $p = p_0$ ) vs.  $H_1: ID_1 \neq ID_2$  (or equivalently,  $p \neq p_0$ ), where

$ID_1$  = true incidence density in group 1

$ID_2$  = true incidence density in group 2

$p_0 = t_1/(t_1 + t_2)$ ,  $q_0 = 1 - p_0$

using a two-sided test with significance level  $\alpha$ , use the following procedure:

(1) If  $a_1 < (a_1 + a_2)p_0$ ,

$$\text{then } p\text{-value} = 2 \times \sum_{k=0}^{a_1} \binom{a_1 + a_2}{k} p_0^k q_0^{a_1+a_2-k}$$

(2) If  $a_1 \geq (a_1 + a_2)p_0$ ,

$$\text{then } p\text{-value} = 2 \times \sum_{k=a_1}^{a_1+a_2} \binom{a_1 + a_2}{k} p_0^k q_0^{a_1+a_2-k}$$

(3) This test is valid in general for comparing two incidence densities but is particularly useful when  $V_1 < 5$ , in which case the normal-theory test in Equation 14.8 should not be used.

The computation of the  $p$ -value is illustrated in Figure 14.3.

### Example 14.11

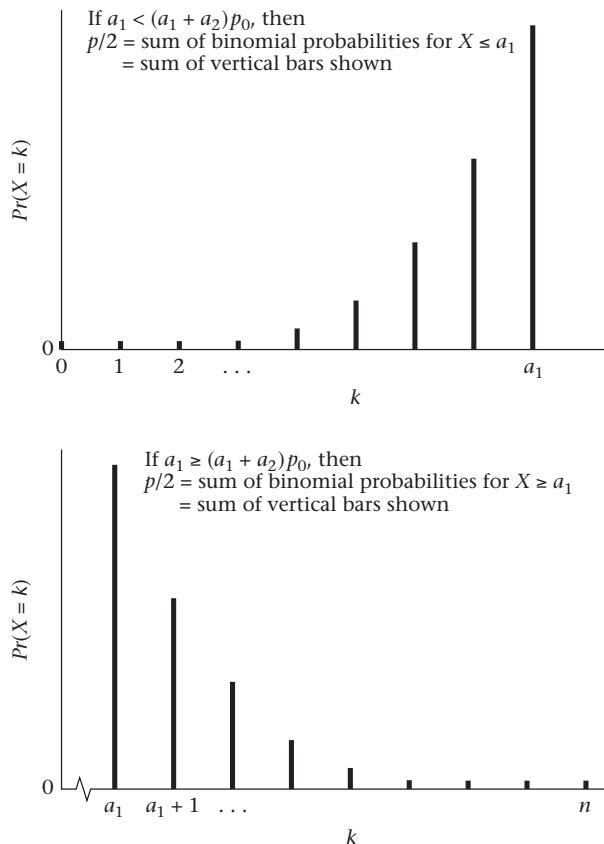
**Cancer** Suppose we have the data in Table 14.3 relating OC use and incidence of breast cancer among women ages 30–34. Assess the statistical significance of the data.

### Table 14.3

#### Relationship between breast-cancer incidence and OC use among 30- to 34-year-old women in the NHS

OC-use group	Number of cases	Number of person-years
Current users	3	8250
Never users	9	17,430

**Figure 14.3 Illustration of the  $p$ -value for the two-sample test for incidence rates, exact method (two-sided alternative)**



**Solution**

Note that  $a_1 = 3$ ,  $a_2 = 9$ ,  $t_1 = 8250$ ,  $t_2 = 17,430$ . Thus

$$V_1 = \frac{12(8250)(17,430)}{(8250 + 17,430)^2} = 2.62 < 5$$

Because  $V_1 < 5$ , the small-sample test must be used. From Equation 14.9,  $p_0 = 8250/25,680 = .321$ ,  $n = a_1 + a_2 = 12$ . Because  $a_1 = 3 < 12(.321) = 3.9$ , we have

$$p\text{-value} = 2 \times \sum_{k=0}^{12} \binom{12}{k} (.321)^k (.679)^{12-k}$$

To evaluate the  $p$ -value, let  $X$  be a random variable representing the number of events in group 1, and use the BINOMDIST function of Excel as follows:

<b>n</b>	<b>12</b>
<b>p</b>	<b>0.321262</b>

<b>k</b>	<b>Pr(X=k)</b>
0	0.009559
1	0.054296
2	0.141346
3	0.223008
<b>Pr(X≤3)</b>	<b>0.428209</b>
<b>p-value</b>	<b>0.856418</b>

Thus

$$p\text{-value} = 2 \times (.0096 + .0543 + .1413 + .2230) = 2 \times .4282 = .856$$

Therefore, there is no significant effect of current OC use on breast-cancer incidence in the 30- to 34-year-old age group.

## The Rate Ratio

In Section 13.3, we defined the risk ratio (*RR*) as a measure of effect for the comparison of two proportions. We applied this measure to compare cumulative incidence between two exposure groups in a prospective study, where the person was the unit of analysis. A similar concept can be employed to compare two incidence rates based on person-time data.

### Definition 14.2

Let  $\lambda_1, \lambda_2$  be incidence rates for an exposed and an unexposed group, respectively. The **incidence rate ratio** is defined as  $\lambda_1/\lambda_2$  and denoted by *IRR*. Usually, the *IRR* is abbreviated as just “rate ratio.”

### Example 14.12

**Cancer** Suppose the incidence rate of breast cancer is  $500/10^5$  person-years among 40- to 49-year-old premenopausal women with a family history of breast cancer (either a mother or a sister history of breast cancer) (group 1) compared with  $200/10^5$  person-years among 40- to 49-year-old premenopausal women with no family history (group 2). What is the rate ratio of group 1 vs. group 2?

### Solution

The rate ratio =  $(500/10^5)/(200/10^5) = 2.5$ .

What is the relationship between the rate ratio (*IRR*) based on incidence rates and the risk ratio (*RR*) based on cumulative incidence? Suppose each person in a cohort is followed for  $T$  years, with incidence rate  $\lambda_1$  in the exposed group and  $\lambda_2$  in the unexposed group. If the cumulative incidence is low, then the cumulative incidence will be approximately  $\lambda_1 T$  in the exposed group and  $\lambda_2 T$  in the unexposed group. Thus the *RR* will be approximately  $\lambda_1 T / (\lambda_2 T) = \lambda_1 / \lambda_2$  = rate ratio = *IRR*.

How can we estimate the rate ratio from observed data? Suppose we have the number of events and person-years shown in Table 14.2. The estimated incidence rate in the exposed group =  $a_1/t_1$  and in the unexposed group =  $a_2/t_2$ . A point estimate of the rate ratio is given by  $\widehat{IRR} = (a_1/t_1)(a_2/t_2)$ . To obtain an interval estimate, we assume approximate normality of  $\ln(\widehat{IRR})$ . It can be shown that

### Equation 14.10

$$\text{Var}[\ln(\widehat{IRR})] \approx \frac{1}{a_1} + \frac{1}{a_2}$$

Therefore, a two-sided  $100\% \times (1 - \alpha)$  CI for  $\ln(\text{IRR})$  is given by

$$(d_1, d_2) = \ln(\widehat{IRR}) \pm z_{1-\alpha/2} \sqrt{\frac{1}{a_1} + \frac{1}{a_2}}$$

If we take the antilog of  $d_1$  and  $d_2$ , we obtain a two-sided  $100\% \times (1 - \alpha)$  CI for *IRR*. This is summarized as follows.

### Equation 14.11

#### Point and Interval Estimation of the Rate Ratio

Suppose we have observed  $a_1$  events in  $t_1$  person-years for an exposed group and  $a_2$  events in  $t_2$  person-years for an unexposed group. A point estimate of the rate ratio is given by

$$\widehat{IRR} = (a_1/t_1)/(a_2/t_2)$$

A two-sided  $100\% \times (1 - \alpha)$  CI for  $IRR$  is given by  $(c_1, c_2)$  where

$$c_1 = e^{d_1}, c_2 = e^{d_2} \text{ and}$$

$$d_1 = \ln(\widehat{IRR}) - z_{1-\alpha/2} \sqrt{\frac{1}{a_1} + \frac{1}{a_2}}$$

$$d_2 = \ln(\widehat{IRR}) + z_{1-\alpha/2} \sqrt{\frac{1}{a_1} + \frac{1}{a_2}}$$

This interval should only be used if  $V_1 = (a_1 + a_2)t_1t_2/(t_1 + t_2)^2$  is  $\geq 5$ .

### Example 14.13

**Cancer** Obtain a point estimate and associated 95% CI for the rate ratio relating OC use to breast-cancer incidence based on the data in Table 14.1.

#### Solution

From Table 14.1, the estimated rate ratio is

$$\widehat{IRR} = \frac{9/2935}{239/135,130} = 1.73$$

To obtain an interval estimate, we refer to Equation 14.11. A 95% CI for  $\ln(IRR)$  is  $(d_1, d_2)$ , where

$$\begin{aligned} d_1 &= \ln(\widehat{IRR}) - 1.96 \sqrt{\frac{1}{9} + \frac{1}{239}} \\ &= 0.550 - 0.666 = -0.115 \end{aligned}$$

$$\begin{aligned} d_2 &= \ln(\widehat{IRR}) + 1.96 \sqrt{\frac{1}{9} + \frac{1}{239}} \\ &= 0.550 + 0.666 = 1.216 \end{aligned}$$

Therefore  $c_1 = e^{-0.115} = 0.89$ ,  $c_2 = e^{1.216} = 3.37$ . Thus the 95% CI for  $IRR$  is  $(0.89, 3.37)$ .

In this section, we have introduced the rate ratio, which is a measure of effect for comparing two incidence rates. It is analogous to but not the same as the risk ratio. The latter was introduced in Chapter 13 as a measure of effect for comparing two cumulative incidence rates. Inference procedures for comparing two incidence rates are based on the one-sample binomial test, where the number of units of analysis = the number of events over the two samples combined, and the probability of success  $p$  = the probability that a subject is in group 1, given that he (she) has had an event. We considered both large-sample and small-sample inference procedures based on the normal approximation to the binomial distribution and exact binomial probabilities, respectively. In the next section, we consider power and sample-size estimation procedures for comparing two incidence rates.

On the flowchart (Figure 14.15, p. 803), we answer yes to (1) person-time data? no to (2) one-sample problem? and yes to (3) incidence rates remain constant over time? and (4) two-sample problem? This leads us to the box labeled "Use two-sample test for comparison of incidence rates, if no confounding is present, or methods for stratified person-time data, if confounding is present." In this section, we assume no confounding is present. In Section 14.5, we consider two-sample inference for incidence rates when confounding is present.

### REVIEW QUESTIONS 14B

- 1 What is a rate ratio? How does it differ from a risk ratio?
- 2 Suppose we are studying a gene that has been linked to coronary heart disease. We find in a pilot study that in subjects with the *C/C* genotype 20 events have occurred over 2000 person-years, whereas in subjects with the *T/T* genotype 30 events have occurred over 2500 person-years.
  - (a) Perform a test to compare the incidence density in the *T/T*-genotype group vs. the *C/C*-genotype group, and report a two-tailed *p*-value.
  - (b) Provide an estimate for the rate ratio for coronary heart disease for people with the *T/T* genotype compared with people with the *C/C* genotype, and obtain a corresponding 95% CI.

## 14.4 Power and Sample-Size Estimation for Person-Time Data

### Estimation of Power

#### Example 14.14

**Cancer** Suppose researchers propose to enroll 10,000 postmenopausal women who have not had any previous cancer for a clinical trial, where 5000 are randomized to receive estrogen-replacement therapy (ERT) and 5000 are randomized to placebo. The endpoint is breast-cancer incidence. Participants are enrolled from January 1, 2005 to December 31, 2006 and are followed until December 31, 2010, for an average of 5 years of follow-up for each participant (range, 4 to 6 years of follow-up). The expected incidence rate in the control group is  $300/10^5$  person-years. If it is hypothesized that ERT increases the incidence rate of breast cancer by 25%, then how much power does the proposed study have?

We base our power calculations on the comparison of incidence rates as given in Equation 14.8. We want to test the hypothesis  $H_0$ : rate ratio (*IRR*) = 1 vs.  $H_1$ : *IRR*  $\neq$  1, where  $\text{IRR} = ID_1/ID_2$ .

As discussed in Equation 14.9, another way to state the hypothesis is as follows:  $H_0$ :  $p = p_0$  vs.  $H_1$ :  $p \neq p_0$ , where  $p_0 = t_1/(t_1 + t_2)$ . This is a one-sample binomial test considered in Section 7.10, where  $n$  = total number of events over both groups and  $p$  = probability that an individual person with an event comes from group 1. The issue is: What specific value of  $p$  under  $H_1$  (call it  $p_1$ ) corresponds to a rate ratio of *IRR*? To derive this, note that

$$\text{Expected number of events in group 1} = 1 - \exp(-ID_1 t_1) \approx t_1 ID_1$$

$$\text{Expected number of events in group 2} = 1 - \exp(-ID_2 t_2) \approx t_2 ID_2$$

Thus the expected proportion of events in group 1 is

#### Equation 14.12

$$p = t_1 ID_1 / (t_1 ID_1 + t_2 ID_2)$$

If we divide the numerator and denominator of Equation 14.12 by  $ID_2$ , we obtain

#### Equation 14.13

$$p = t_1 IRR / (t_1 IRR + t_2)$$

Under  $H_1$ , IRR will differ from 1, and we denote  $p$  in Equation 14.13 by  $p_1$ . Under  $H_0$ ,  $p = t_1/(t_1 + t_2)$ , which we denote by  $p_0$ . We now apply the power formula for the one-sample binomial test (Equation 7.45) and obtain

**Equation 14.14**

$$\text{Power} = \Phi \left[ \sqrt{\frac{p_0 q_0}{p_1 q_1}} \left( z_{\alpha/2} + \frac{|p_0 - p_1| \sqrt{m}}{\sqrt{p_0 q_0}} \right) \right]$$

where  $m = \text{expected number of events over both groups combined} = m_1 + m_2$

We now want to relate the expected number of events in groups 1 and 2 ( $m_1, m_2$ ) to the number of participants available in each group ( $n_1, n_2$ ). Recall from Equation 14.1 that

**Equation 14.15**

$$CUMI(t) = 1 - \exp(-IDt^*)$$

where  $t^* = \text{average number of person-years per subject}$

Applying Equation 14.15 to each group, we have

**Equation 14.16**

$$\text{Cumulative incidence in group 1} = m_1 / n_1 = 1 - \exp(-ID_1 t_1^*)$$

$$\text{Cumulative incidence in group 2} = m_2 / n_2 = 1 - \exp(-ID_2 t_2^*)$$

or

**Equation 14.17**

$$m_1 = n_1[1 - \exp(-ID_1 t_1^*)]$$

$$m_2 = n_2[1 - \exp(-ID_2 t_2^*)]$$

Substituting Equation 14.17 into Equation 14.14, we obtain the following power formula.

**Equation 14.18****Power for the Comparison of Two Incidence Rates**

Suppose we want to test the hypothesis  $H_0: ID_1 = ID_2$  vs.  $H_1: ID_2 \neq ID_1$ , where  $ID_1, ID_2$  are incidence densities in groups 1 and 2. The power of the test for the specific alternative  $ID_1/ID_2 = IRR$  with two-sided significance level =  $\alpha$  is given by

$$\text{Power} = \Phi \left[ \sqrt{\frac{p_0 q_0}{p_1 q_1}} \left( z_{\alpha/2} + \frac{|p_0 - p_1| \sqrt{m}}{\sqrt{p_0 q_0}} \right) \right]$$

where

$$p_0 = t_1/(t_1 + t_2)$$

$$p_1 = t_1 \text{IRR}/(t_1 \text{IRR} + t_2)$$

$$m = \text{expected number of events in the two groups combined} = m_1 + m_2$$

$$m_1 = n_1[1 - \exp(-ID_1 t_1^*)]$$

$$m_2 = n_2[1 - \exp(-ID_2 t_2^*)]$$

$n_1, n_2$  = number of subjects available in groups 1 and 2, respectively  
 $t_1, t_2$  = total number of person-years in groups 1 and 2, respectively  
 $t_1^*, t_2^*$  = average number of person-years per subject in groups 1 and 2,  
 respectively =  $(t_1/n_1$  and  $t_2/n_2$  in groups 1 and 2)  
 $ID_1, ID_2$  = incidence density in groups 1 and 2, respectively, under  $H_1$

**Example 14.15** **Cancer** Answer the question posed in Example 14.14.

**Solution**

We have  $n_1 = n_2 = 5000$  subjects,  $t_1^* = t_2^* = 5$ ,  $ID_2 = 300/10^5$  person-years,  $ID_1 = 1.25 \times 300/10^5 = 375/10^5$  person-years,  $IRR = 1.25$ . Thus, from Equation 14.18,

$$\begin{aligned}
 m_1 &= 5000[1 - \exp[-(375/10^5)5]] \\
 &= 5000(1 - .98142) = 92.9 \\
 m_2 &= 5000[1 - \exp[-(300/10^5)5]] \\
 &= 5000(1 - .98511) = 74.4 \\
 m &= 92.9 + 74.4 = 167.3 \\
 t_1 &= t_2 = 5000 \times 5 = 25,000 \\
 p_0 &= 1/2 \\
 p_1 &= 25,000(1.25)/[25,000(1.25) + 25,000] \\
 &= 31,250/56,250 = .556
 \end{aligned}$$

Thus

$$\begin{aligned}
 \text{Power} &= \Phi\left[\sqrt{\frac{.5(.5)}{.556(.444)}}\left(-1.96 + \frac{|.50 - .556| \sqrt{167.3}}{\sqrt{.5(.5)}}\right)\right] \\
 &= \Phi\left[1.00062\left(-1.96 + \frac{0.7186}{.50}\right)\right] \\
 &= \Phi[1.0062(-0.5228)] \\
 &= \Phi(-0.5260) = .299
 \end{aligned}$$

Therefore, the study only has about 30% power to test the hypotheses.

### Sample-Size Estimation

Clearly, the study proposed in Example 14.14 is too small to have sufficient power to test the hypotheses proposed. The issue is how large a study would be needed to have a prespecified (say, 80%) level of power. For this purpose, if we prespecify a power of  $1 - \beta$ , then we can solve for the required total number of events  $m$  from the one-sample binomial test. Specifically, from Equation 7.46 we have

**Equation 14.19**

$$m = \frac{(\sqrt{p_0 q_0} z_{1-\alpha/2} + \sqrt{p_1 q_1} z_{1-\beta})^2}{|p_0 - p_1|^2}$$

To convert from the required number of events ( $m$ ) to the required number of subjects ( $n$ ), we refer to Equation 14.17 and get

**Equation 14.20**

$$m = m_1 + m_2 = n_1[1 - \exp(-ID_1 t_1^*)] + n_2[1 - \exp(-ID_2 t_2^*)]$$

If we prespecify the ratio of sample sizes in the two groups—i.e.,  $n_2/n_1 = k$ ; then it follows from Equation 14.20 that

**Equation 14.21**

$$n_1 = \frac{m}{\left\{1 - \exp(-ID_1 t_1^*) + k \left[1 - \exp(-ID_2 t_2^*)\right]\right\}}$$

$$n_2 = kn_1$$

Combining Equations 14.19–14.21 yields the following sample-size formula.

**Equation 14.22****Sample-Size Estimation for the Comparison of Two Incidence Rates**

Suppose we want to test the hypothesis  $H_0: ID_1 = ID_2$  vs.  $H_1: ID_1 \neq ID_2$ , where  $ID_1$  and  $ID_2$  are the incidence densities in groups 1 and 2, respectively. We assume that the ratio of sample sizes in the two samples is prespecified as  $k = n_2/n_1$ . If we conduct a two-sided test with significance level  $\alpha$  and power  $1 - \beta$ , then we need a total expected number of events over both groups of  $m$ , where

$$m = \frac{\left(\sqrt{p_0 q_0} Z_{1-\alpha/2} + \sqrt{p_1 q_1} Z_{1-\beta}\right)^2}{|p_0 - p_1|^2}$$

where

$$p_0 = t_1/(t_1 + t_2)$$

$$p_1 = t_1 IRR/(t_1 IRR + t_2)$$

$$IRR = ID_1/ID_2$$

$t_1, t_2$  = total number of person-years in groups 1 and 2, respectively

$ID_1, ID_2$  = incidence densities in groups 1 and 2, respectively, under  $H_1$

The corresponding number of subjects in each group is

$$n_1 = \frac{m}{(k+1) - \exp(-ID_1 t_1^*) - k \exp(-ID_2 t_2^*)}$$

$$n_2 = kn_1$$

**Example 14.16**

**Cancer** How many participants need to be enrolled in the study proposed in Example 14.14 to have 80% power, if a two-sided test with significance level of .05 is used and an equal number of participants are enrolled in each group?

**Solution**

We have  $\alpha = .05$ ,  $1 - \beta = .80$ ,  $k = 1$ . Also, from the solution to Example 14.15, we have  $p_0 = .50$ ,  $p_1 = .556$ . Thus, from Equation 14.22, the required total number of events is

$$m = \frac{[\sqrt{.5(.5)}(1.96) + \sqrt{.556(.444)}(0.84)]^2}{(.50 - .556)^2}$$

$$= \frac{1.397^2}{.056^2}$$

$$= \frac{1.9527}{.00309} = 632.7 \text{ or } 633 \text{ events}$$

Thus we need a total of 633 events to achieve 80% power. From Example 14.15, we have  $t_1^* = t_2^* = 5$  years,  $ID_1 = 375/10^5$  person-years,  $ID_2 = 300/10^5$  person-years. Also,

because the sample size in each group is the same, we have  $k = 1$ . Therefore, from Equation 14.22, the required number of participants in each group is

$$\begin{aligned} n_1 = n_2 &= \frac{633}{2 - \exp[(-375/10^5)5] - \exp[(-300/10^5)5]} \\ &= \frac{633}{2 - .98142 - .98511} \\ &= \frac{633}{.0335} = 18,916.2 \text{ or } 18,917 \text{ participants} \end{aligned}$$

Thus we need to enroll 18,917 participants in each group or a total of 37,834 participants to have 80% power. This is about four times as large a study as the one originally contemplated in Example 14.14. This study would be expected to yield

$$\begin{aligned} m_1 &= 18,917\{1 - \exp[(-375/10^5)5]\} \\ &= 18,917(1 - .98142) = 351 \text{ events in the ERT group} \\ m_2 &= 18,917\{1 - \exp[(-300/10^5)5]\} \\ &= 18,917(1 - .98511) = 282 \text{ events in the control group} \end{aligned}$$

for a total of 633 events. This is one study design used in the Women's Health Initiative, a large multicenter set of clinical trials that have similar sample sizes, numbers of events, and time frames to those posed in Examples 14.14–14.16.

In this section, we have considered power and sample-size formulas for comparing two incidence rates. The formulas are special cases of similar formulas used for the one-sample binomial test in Equations 7.45 and 7.46, respectively. If the number of person-years of follow-up is the same for each subject, then these formulas should be approximately the same as the corresponding power and sample-size formulas for comparing two proportions, which are given in Equations 10.15 and 10.14, respectively. However, the advantage of the methods in this section is that they allow for a variable length of follow-up for individual subjects, which is more realistic in many clinical-trial situations. In the next section, we consider inference procedures for comparing incidence rates between two groups, while controlling for confounding variables.

## 14.5 Inference for Stratified Person-Time Data

### Hypothesis Testing

It is very common in analyzing person-time data to control for confounding variables before assessing the relationship between the main exposure of interest and disease. Confounding variables may include age and sex as well as other covariates related to exposure, disease, or both.

#### Example 14.17

**Cancer** An issue of continuing interest is the effect of postmenopausal hormone (PMH) replacement on cardiovascular and cancer outcomes in postmenopausal women. Data were collected from postmenopausal women in the Nurses' Health Study (NHS) to address this issue. Women were mailed an initial questionnaire in 1976 and follow-up questionnaires every 2 years thereafter. Data from 1976 to 1986, encompassing 352,871 person-years of follow-up and 707 incident cases of breast cancer, are given in Table 14.4 [1].

**Table 14.4** Current and past use of PMH replacement and risk of breast cancer among postmenopausal participants in the NHS

Age	Never users		Current users			Past users		
	No. of cases	Person-years	No. of cases	Person-years	IRR	No. of cases	Person-years	IRR
39–44	5	4722	12	10,199	1.11	4	3835	0.99
45–49	26	20,812	22	14,044	1.25	12	8921	1.08
50–54	129	71,746	51	24,948	1.14	46	26,256	0.97
55–59	159	73,413	72	21,576	1.54	82	39,785	0.95
60–64	35	15,773	23	4876	2.13	29	11,965	1.09

There were 23,607 women who were postmenopausal and did not have any type of cancer (except for nonmelanoma skin cancer) in 1976. Other women became postmenopausal during the follow-up period. Follow-up was terminated at the diagnosis of breast cancer, death, or the date of the last questionnaire return. Thus each woman had a variable duration of follow-up. Because breast-cancer incidence and possibly PMH replacement are related to age, it was important to control for age in the analysis.

We can use methods similar to the Mantel-Haenszel test used for cumulative incidence data (or generally for count data), as presented in Chapter 13.

Suppose we have  $k$  strata, where the number of events and the amount of person-time in the  $i$ th stratum are as shown in Table 14.5.

**Table 14.5** General observed table for the number of events and person-time in the  $i$ th stratum,  $i = 1, \dots, k$ 

Exposure group	Number of events	Person-time
Exposed	$a_{1i}$	$t_{1i}$
Unexposed	$a_{2i}$	$t_{2i}$
Total	$a_{1i} + a_{2i}$	$t_{1i} + t_{2i}$

Let's denote the incidence rate of disease among the exposed by  $p_{1i}$  and among the unexposed by  $p_{2i}$ . Therefore, the expected number of events among the exposed =  $p_{1i}t_{1i}$  and among the unexposed =  $p_{2i}t_{2i}$ . Let  $p_i$  = the expected proportion of the total number of events over both groups that are among the exposed. We can relate  $p_i$  to  $p_{1i}$  and  $p_{2i}$  by

$$\text{Equation 14.23} \quad p_i = \frac{p_{1i}t_{1i}}{p_{1i}t_{1i} + p_{2i}t_{2i}} = \frac{t_{1i}}{t_{1i} + t_{2i}} \quad \text{if } p_{1i} = p_{2i}, \text{ which we denote by } p_i^{(0)}$$

We assume the rate ratio relating disease to exposure is the same for each stratum and denote it by  $IRR$ . Therefore,  $IRR = p_{1i}/p_{2i}$  and is the same for each  $i = 1, \dots, k$ . If we divide the numerator and denominator of Equation 14.23 by  $p_{2i}$  and substitute  $IRR$  for  $p_{1i}/p_{2i}$ , we obtain

**Equation 14.24**

$$\begin{aligned} p_i &= \frac{(p_{1i} / p_{2i})t_{1i}}{(p_{1i} / p_{2i})t_{1i} + t_{2i}} \\ &= \frac{IRRt_{1i}}{IRRt_{1i} + t_{2i}}, \text{ which we denote by } p_i^{(1)}. \end{aligned}$$

We want to test the hypothesis  $H_0: IRR = 1$  vs.  $H_1: IRR \neq 1$  or, equivalently,  $H_0: p_i = p_i^{(0)}$  vs.  $H_1: p_i = p_i^{(1)}, i = 1, \dots, k$ . We base our test on  $A = \sum_{i=1}^k a_{1i}$  = total observed number of events for the exposed. Under  $H_0$ , we assume the total observed number of events for the  $i$ th stratum ( $a_{1i} + a_{2i}$ ) is fixed. Therefore, under  $H_0$ ,

**Equation 14.25**

$$\begin{aligned} E(a_{1i}) &= (a_{1i} + a_{2i})p_i^{(0)} = (a_{1i} + a_{2i})t_{1i}/(t_{1i} + t_{2i}) \\ Var(a_{1i}) &= (a_{1i} + a_{2i})p_i^{(0)}(1 - p_i^{(0)}) = (a_{1i} + a_{2i})t_{1i}t_{2i}/(t_{1i} + t_{2i})^2 \end{aligned}$$

and  $E(A) = \sum_{i=1}^k E(a_{1i}), Var(A) = \sum_{i=1}^k Var(a_{1i})$ . Under  $H_1$ ,  $A$  is larger than  $E(A)$  if  $IRR > 1$  and is smaller than  $E(A)$  if  $IRR < 1$ . We use the test statistic  $X^2 = [|A - E(A)| - 0.5]^2 / Var(A)$ , which follows a  $\chi^2_1$  distribution under  $H_0$  and reject  $H_0$  for large values of  $X^2$ . The test procedure is summarized as follows.

**Equation 14.26****Hypothesis Testing for Stratified Person-Time Data**

Let  $p_{1i}, p_{2i}$  = incidence rate of disease for the exposed and unexposed groups in the  $i$ th stratum, respectively. Let  $a_{1i}, t_{1i}$  = the number of events and person-years for the exposed group in the  $i$ th stratum,  $a_{2i}, t_{2i}$  = the number of events and person-years for the unexposed group in the  $i$ th stratum,  $i = 1, \dots, k$ .

We assume  $IRR = p_{1i}/p_{2i}$  is constant across all strata. To test the hypothesis  $H_0: IRR = 1$  vs.  $H_1: IRR \neq 1$  using a two-sided test with significance level =  $\alpha$ :

(1) We compute the total observed number of events among the exposed over all strata =  $A = \sum_{i=1}^k a_{1i}$ .

(2) We compute the total expected number of events under  $H_0$  among the exposed over all strata =  $E(A) = \sum_{i=1}^k E(a_{1i})$ , where  

$$E(a_{1i}) = (a_{1i} + a_{2i})t_{1i}/(t_{1i} + t_{2i}), i = 1, \dots, k$$

(3) We compute  $Var(A) = \sum_{i=1}^k Var(a_{1i})$  under  $H_0$ , where  

$$Var(a_{1i}) = (a_{1i} + a_{2i})t_{1i}t_{2i}/(t_{1i} + t_{2i})^2, i = 1, \dots, k$$

(4) We compute the test statistic

$$X^2 = \frac{(|A - E(A)| - 0.5)^2}{Var(A)}$$

which follows a chi-square distribution with 1 degree of freedom ( $df$ ) under  $H_0$ .

- (5) If  $X^2 > \chi^2_{1,1-\alpha}$ , then we reject  $H_0$ .  
 If  $X^2 \leq \chi^2_{1,1-\alpha}$ , then we accept  $H_0$ .
- (6) The  $p$ -value =  $Pr(\chi^2_1 > X^2)$ .
- (7) This test should only be used if  $Var(A) \geq 5$ .

**Example 14.18**

**Cancer** Test the hypothesis that there is a significant association between breast-cancer incidence and current use of PMH replacement based on the data in Table 14.4.

**Solution**

We compare current users of PMH replacement (the exposed group) with never users of PMH replacement (the unexposed group) using the method in Equation 14.26. For 39- to 44-year-old women, we have

$$a_{11} = 12 \\ E(a_{11}) = \frac{(12+5)10,199}{10,199 + 4722} = 17 \times .684 = 11.62 \\ Var(a_{11}) = 17 \times .684 \times .316 = 3.677$$

Similar computations are performed for each of the other four age groups, whereby

$$A = 12 + 22 + 51 + 72 + 23 = 180 \\ E(A) = 11.62 + 19.34 + 46.44 + 52.47 + 13.70 = 143.57 \\ Var(A) = 3.677 + 11.548 + 34.459 + 40.552 + 10.462 = 100.698 \\ X^2 = \frac{(|180 - 143.57| - 0.5)^2}{100.698} = \frac{35.93^2}{100.698} = 12.82 \sim \chi^2_1 \text{ under } H_0$$

Because  $X^2 > 10.83 = \chi^2_{1,.999}$ , it follows that  $p < .001$ . Therefore, there is a highly significant association between breast-cancer incidence and current PMH replacement therapy.

**Estimation of the Rate Ratio**

We use a similar approach to that considered in estimating a single rate ratio in Section 14.3. We obtain estimates of the  $\ln(\text{rate ratio})$  in each stratum and then compute a weighted average of the stratum-specific estimates to obtain an overall estimate of the  $\ln(\text{rate ratio})$ . Specifically, let

**Equation 14.27**

$$\widehat{IRR}_i = (a_{1i}/t_{1i})/(a_{2i}/t_{2i})$$

be the estimate of the rate ratio in the  $i$ th stratum. From Equation 14.10, we have

**Equation 14.28**

$$Var[\ln(\widehat{IRR}_i)] \doteq \frac{1}{a_{1i}} + \frac{1}{a_{2i}}$$

To obtain an overall estimate of  $\ln(IRR)$ , we now compute a weighted average of  $\ln(\widehat{IRR}_i)$  where the weights are the inverse of the variance of  $\ln(\widehat{IRR}_i)$  and then take the antilog of the weighted average.

**Equation 14.29**

$$\ln(\widehat{IRR}) = \frac{\sum_{i=1}^k w_i \ln(\widehat{IRR}_i)}{\sum_{i=1}^k w_i}$$

where  $w_i = 1/\text{Var}[\ln(\widehat{IRR}_i)]$ . We then take the antilog of  $\ln(\widehat{IRR})$  to obtain an estimate of  $IRR$ .

To obtain confidence limits for the rate ratio, we use Equation 14.29 to obtain the variance of  $\ln(\widehat{IRR})$  as follows.

**Equation 14.30**

$$\begin{aligned}\text{Var}[\ln(\widehat{IRR})] &= \frac{1}{\left(\sum_{i=1}^k w_i\right)^2} \text{Var}\left[\sum_{i=1}^k w_i \ln(\widehat{IRR}_i)\right] \\ &= \frac{1}{\left(\sum_{i=1}^k w_i\right)^2} \sum_{i=1}^k w_i^2 \text{Var}[\ln(\widehat{IRR}_i)] \\ &= \frac{1}{\left(\sum_{i=1}^k w_i\right)^2} \sum_{i=1}^k w_i^2 (1/w_i) \\ &= \frac{1}{\left(\sum_{i=1}^k w_i\right)^2} \sum_{i=1}^k w_i = \frac{1}{\sum_{i=1}^k w_i}\end{aligned}$$

Thus a two-sided  $100\% \times (1 - \alpha)$  CI for  $\ln(IRR)$  is given by

$$\ln(\widehat{IRR}) \pm z_{1-\alpha/2} \times \sqrt{1/\sum_{i=1}^k w_i}$$

We then take the antilog of each of the confidence limits for  $\ln(IRR)$  to obtain confidence limits for  $IRR$ . This procedure is summarized as follows.

**Equation 14.31**

#### Point and Interval Estimation of the Rate Ratio (Stratified Data)

Let  $a_{1i}, t_{1i}$  = number of events and person-years for the exposed in the  $i$ th stratum

$a_{2i}, t_{2i}$  = number of events and person-years for the unexposed in the  $i$ th stratum

A point estimate of the rate ratio ( $RR$ ) is given by  $\widehat{IRR} = e^c$ , where

$$c = \frac{\sum_{i=1}^k w_i \ln(\widehat{IRR}_i)}{\sum_{i=1}^k w_i}$$

$$\widehat{IRR}_i = \frac{a_{1i} / t_{1i}}{a_{2i} / t_{2i}}$$

$$w_i = \left( \frac{1}{a_{1i}} + \frac{1}{a_{2i}} \right)^{-1}$$

A two-sided  $100\% \times (1 - \alpha)$  CI for  $IRR$  is given by  $(e^{c_1}, e^{c_2})$ , where

$$c_1 = \ln(\widehat{IRR}) - z_{1-\alpha/2} \sqrt{1 / \sum_{i=1}^k w_i}$$

$$c_2 = \ln(\widehat{IRR}) + z_{1-\alpha/2} \sqrt{1 / \sum_{i=1}^k w_i}$$

This interval should only be used if  $\text{Var}(A)$  as given in Equation 14.26 is  $\geq 5$ .

### Example 14.19

**Cancer** Obtain a point estimate and associated 95% confidence limits for the rate ratio for breast-cancer incidence rate for current vs. never users of estrogen-replacement therapy, based on the data in Table 14.4.

### Solution

The computations are summarized in Table 14.6. For example, for the age group 39–44,

$$\widehat{IRR}_1 = \frac{12/10,199}{5/4722} = 1.11$$

$$\ln(\widehat{IRR}_1) = 0.105$$

$$\text{Var}[\ln(\widehat{IRR}_1)] = \frac{1}{12} + \frac{1}{5} = 0.283$$

$$w_i = 1/0.283 = 3.53$$

**Table 14.6** Breast cancer vs. estrogen-replacement therapy, estimation of the rate ratio after stratification by age

Age group	$\widehat{IRR}_i$	$\ln(\widehat{IRR})$	$\text{Var}[\ln(\widehat{IRR})]$	$w_i$	$w_i \ln(\widehat{IRR})$
39–44	1.11	0.105	0.283	3.53	0.372
45–49	1.25	0.226	0.084	11.92	2.697
50–54	1.14	0.128	0.027	36.55	4.691
55–59	1.54	0.432	0.020	49.56	21.423
60–64	2.13	0.754	0.072	13.88	10.467
Total				115.43	39.650

Note:

$$\ln(\widehat{IRR}) = \frac{39.650}{115.43} = 0.343, \widehat{IRR} = \exp(0.343) = 1.41$$

$$\text{Var}[\ln(\widehat{IRR})] = 1/115.43 = 0.0087, \text{se}[\ln(\widehat{IRR})] = \sqrt{0.0087} = 0.0931$$

$$95\% \text{ CI for } \ln(\widehat{IRR}) = 0.343 \pm 1.96(0.0931) = 0.343 \pm 0.182 = (0.161, 0.526)$$

$$95\% \text{ CI for } IRR = [\exp(0.161), \exp(0.526)] = (1.17, 1.69)$$

Similar computations are performed for each of the other age strata. Thus the overall estimate of the rate ratio = 1.41 with 95% confidence limits = (1.17, 1.69). This indicates the incidence of breast cancer is estimated to be about 40% higher for current

users of estrogen-replacement therapy than for never users even after controlling for age. Note that the crude  $IRR = (180/75,643)/(354/186,466) = 1.25 < 1.41$ , which implies that age is a **negative confounder**. Age is a negative confounder because the percentage of postmenopausal women using estrogen-replacement therapy decreases with increasing age, whereas the incidence rate of breast cancer increases with increasing age.

### Testing the Assumption of Homogeneity of the Rate Ratio across Strata

An important assumption made in the estimation methods in Equation 14.31 is that the underlying rate ratio is the same in all strata. If this assumption is not true, then it makes little sense to estimate a common rate ratio. If the rate ratios in different strata are all in the same direction relative to the null hypothesis (i.e., all rate ratios  $> 1$  or all rate ratios  $< 1$ ), then the hypothesis-testing procedure in Equation 14.26 is still valid with only a slight loss of power. However, if the rate ratios are in different directions in different strata, or are null in some strata, then the power of the hypothesis-testing procedure is greatly diminished.

To test this assumption, we use similar methods to those used for testing the assumption of homogeneity of the odds ratio in different strata for count data given in Chapter 13. Specifically, we want to test the hypothesis  $H_0: IRR_1 = \dots = IRR_k$  vs.  $H_1$ : at least two of the  $IRR_i$ 's are different. We base our hypothesis test on the test statistic

$$X_{\text{het}}^2 = \sum_{i=1}^k w_i [\ln(\widehat{IRR}_i) - \ln(\widehat{IRR})]^2 \sim \chi_{k-1}^2 \text{ under } H_0$$

and will reject  $H_0$  for large values of  $X_{\text{het}}^2$ . The test procedure is summarized as follows.

#### Equation 14.32

##### Chi-Square Test for Homogeneity of Rate Ratios across Strata

Suppose we have incidence-rate data and wish to control for the confounding effect of another variable(s) that comprises  $k$  strata. To test the hypothesis  $H_0: IRR_1 = \dots = IRR_k$  vs.  $H_1$ : at least two of the  $IRR_i$ 's differ, with significance level  $\alpha$ , we use the following procedure:

- (1) We compute the test statistic

$$X_{\text{het}}^2 = \sum_{i=1}^k w_i [\ln(\widehat{IRR}_i) - \ln(\widehat{IRR})]^2 \sim \chi_{k-1}^2 \text{ under } H_0$$

where  $\widehat{IRR}_i$  = estimated rate ratio in the  $i$ th stratum

$\widehat{IRR}$  = estimate of the overall rate ratio as given in Equation 14.31

$w_i = 1 / \text{Var}[\ln(\widehat{IRR}_i)]$  as defined in Equation 14.31

- (2) If  $X_{\text{het}}^2 > \chi_{k-1,1-\alpha}^2$ , we reject  $H_0$ .

If  $X_{\text{het}}^2 \leq \chi_{k-1,1-\alpha}^2$ , we accept  $H_0$ .

- (3) The  $p$ -value is given by  $p = \Pr(\chi_{k-1}^2 > X_{\text{het}}^2)$ .

- (4) An alternative computational form for the test statistic in step 1 is

$$\sum_{i=1}^k w_i [\ln(\widehat{IRR}_i)]^2 - \left( \sum_{i=1}^k w_i \right) [\ln(\widehat{IRR})]^2$$

**Example 14.20**

**Cancer** Test for the assumption of the homogeneity of the rate ratio over the five age strata based on the data in Table 14.4.

**Solution**

We have that

$$\begin{aligned} X_{het}^2 &= \sum_i [\ln(\widehat{IRR}_i) - \ln(\widehat{IRR})]^2 = \sum_i w_i [\ln(\widehat{IRR}_i)]^2 - \left( \sum_i w_i \right) [\ln(\widehat{IRR})]^2 \sim \chi_{k-1}^2 \\ &= 18.405 - 115.43(0.343)^2 = 18.405 - 13.619 = 4.786 \sim \chi_4^2 \text{ under } H_0 \\ p\text{-value} &= Pr(\chi_4^2 > 4.786) = .31 \end{aligned}$$

Thus there is no significant heterogeneity. However, it appears from Table 14.6 that the rate ratios are increasing with age. Thus the test procedure in Equation 14.32 may not be sensitive to variation in the rate ratios in a specific direction with respect to the confounding variable(s). One possible explanation is that the average duration of use generally increases with increasing age. Thus the apparent increase in risk with increasing age may actually represent an increase in risk with increasing duration of use. To properly account for the effects of both age and duration of use, it is more appropriate to use Cox regression analyses, which are discussed later in this chapter.

In this section, we have considered a method for comparing incidence rates between two groups while controlling for a single categorical exposure variable. This method can also be used if there is more than one covariate to be controlled for, but it would be tedious to do so with many covariates. Instead, Poisson regression analysis can be used to accomplish this. This is a generalization of logistic regression for incidence-rate data but is beyond the scope of this text.

On the flowchart (Figure 14.15, p. 803), we answer yes to (1) person-time data? no to (2) one-sample problem? and yes to (3) incidence rates remain constant over time? and (4) two-sample problem? which leads us to the box labeled "Use two-sample test for comparison of incidence rates, if no confounding is present, or methods for stratified person-time data, if confounding is present." In this section, we have considered techniques for performing two-sample inference for incidence rates when confounding is present.

### REVIEW QUESTIONS 14C

Suppose we want to study the effect of cigarette smoking on the incidence rate of primary open-angle glaucoma. The following data were obtained from the Nurses' Health Study (NHS) over the time period 1980–1996 and the Health Professionals' Follow-Up Study (a study of male health professionals) over the time period 1986–1996 [2].

**Table 14.7** Association between open-angle glaucoma and cigarette smoking in the Nurses' Health Study and the Health Professionals' Follow-Up Study

	Never smokers		Current smokers	
	Number of cases	Person-years	Number of cases	Person-years
Women	150	360,348	38	147,476
Men	71	117,804	10	17,706

- 1 If we assume the age distributions of current smokers and never smokers are comparable within each study, then test the hypothesis that current cigarette smoking is associated with the incidence of open-angle glaucoma. Report a two-tailed *p*-value.

- 2 Estimate the rate ratio relating current cigarette smoking to the incidence of open-angle glaucoma, and provide a 95% CI.
- 3 Test for the homogeneity of the rate ratio between men and women.

In the actual analysis, we would have to control for age and possibly other factors. The technique for doing so is discussed later in this chapter in our work on the Cox proportional-hazards model.

## 14.6 Power and Sample-Size Estimation for Stratified Person-Time Data

### Sample-Size Estimation

In Section 14.4, we studied how to obtain power and sample-size estimates for comparing two incidence rates. In this section, we extend this discussion to allow for power and sample-size estimation for comparison of incidence rates while controlling for confounding variables.

#### Example 14.21

**Cancer** Suppose we want to study whether the positive association between breast-cancer incidence and postmenopausal hormone (PMH) use is also present in another study population. We assume that the age-specific incidence rates of breast cancer for unexposed women (i.e., never users of PMH replacement) and the age distribution and percentage of women using PMH replacement within specific age groups (as reflected by the percentage of total person-years realized by specific age-exposure groups) are the same as in the NHS (Table 14.4). We also assume the true rate ratio within each age group = 1.5. How many participants do we need to enroll if the average participant is followed for 5 years and we want 80% power using a two-sided test with  $\alpha = .05$ ?

The sample-size estimate depends on

- (1) The age-specific incidence rates of disease in the unexposed group
- (2) The distribution of total person-years within specific age-exposure groups
- (3) The true rate ratio under the alternative hypothesis
- (4) The type I and type II errors

The sample-size estimate is given as follows.

#### Equation 14.33

### Sample-Size Estimation for Incidence-Rate Data

Suppose we have  $s$  strata. Let  $p_i$  = the probability that a case in the  $i$ th stratum comes from the exposed group. We wish to test the hypothesis

$$H_0: p_i = t_{1i}/(t_{1i} + t_{2i}) = p_i^{(0)} \text{ vs. } H_1: p_i = IRR t_{1i}/(IRR t_{1i} + t_{2i}) = p_i^{(1)}, i = 1, \dots, s$$

where

$t_{1i}$  = number of person-years of follow-up among exposed participants in the  $i$ th stratum

$t_{2i}$  = number of person-years of follow-up among unexposed participants in the  $i$ th stratum

The equivalent hypotheses are  $H_0: IRR = 1$  vs.  $H_1: IRR \neq 1$ , where  $IRR = ID_{1i}/ID_{2i}$  = ratio of incidence densities of exposed compared with unexposed participants

in the  $i$ th stratum.  $IRR$  is assumed to be the same for all strata. To test these hypotheses using a two-sided test with significance level  $\alpha$  and power of  $1 - \beta$  vs. a true rate ratio of  $IRR$  under  $H_1$  requires a total expected number of cases over both groups =  $m$ , where

$$m = \frac{(z_{1-\alpha/2} \sqrt{C} + z_{1-\beta} \sqrt{D})^2}{(A - B)^2}$$

where

$$\begin{aligned} A &= \sum_{i=1}^s \lambda_i p_i^{(0)} = \sum_{i=1}^s A_i \\ B &= \sum_{i=1}^s \lambda_i p_i^{(1)} = \sum_{i=1}^s B_i \\ C &= \sum_{i=1}^s \lambda_i p_i^{(0)} [1 - p_i^{(0)}] = \sum_{i=1}^s C_i \\ D &= \sum_{i=1}^s \lambda_i p_i^{(1)} [1 - p_i^{(1)}] = \sum_{i=1}^s D_i \end{aligned}$$

$\lambda_i = G_i/G$  = proportion of cases in the  $i$ th stratum

where

$$G_i = \frac{\theta_i (k_i p_{2i} + p_{1i})}{k_i + 1}, i = 1, \dots, s, G = \sum_{i=1}^s G_i$$

$p_{2i}$  = rate of disease among the unexposed participants in stratum  $i$   
 $= 1 - \exp(-ID_{2i}T_{2i})$ ,  $i = 1, \dots, s$

$p_{1i}$  = rate of disease among exposed participants in stratum  $i$   
 $= 1 - \exp[-IRR ID_{2i}T_{1i}]$ ,  $i = 1, \dots, s$

$T_{1i}$  = average length of follow-up per exposed participant in stratum  $i$

$T_{2i}$  = average length of follow-up per unexposed participant in stratum  $i$

$ID_{2i}$  = incidence density among unexposed participants in the  $i$ th stratum,  $i = 1, \dots, s$

$k_i = t_{2i}/t_{1i}$ ,  $i = 1, \dots, s$

$\theta_i = n_i/n$  = overall proportion of participants in the  $i$ th stratum,  $i = 1, \dots, s$

$$p_i^{(0)} = \frac{t_{1i}}{t_{1i} + t_{2i}}$$

= proportion of cases in the  $i$ th stratum that are exposed under  $H_0$ ,  
 $i = 1, \dots, s$

$$p_i^{(1)} = \frac{t_{1i}IRR}{t_{1i}IRR + t_{2i}}$$

= proportion of cases in the  $i$ th stratum that are exposed under  $H_1$ ,  
 $i = 1, \dots, s$

The required total sample size ( $n$ ) is

$$n = \frac{m}{\sum_{i=1}^s \theta_i (k_i p_{2i} + p_{1i}) / (k_i + 1)} = \frac{m}{\sum_{i=1}^s G_i} = \frac{m}{G}$$

**Example 14.22**

**Cancer** Compute the required sample size for the study proposed in Example 14.21 using a two-sided test with  $\alpha = .05$ , power = 80%,  $IRR = 1.5$ , if there are 5 years of follow-up for each participant, the incidence rate of disease among the unexposed is the same as in Table 14.4 for never users, and the age-exposure distribution of person-years is the same as in Table 14.4.

**Solution**

We need to compute  $ID_{2,i}$ ,  $p_{2,i}$ ,  $p_{1,i}$ ,  $k_i$ ,  $\theta_i$ ,  $p_i^{(0)}$ , and  $p_i^{(1)}$  for each of the five age strata in Table 14.4. For example, in the first age group (age group = 39–44 years old) we have

$$\begin{aligned} ID_{2,1} &= 105.9/10^5 \text{ person-years} \\ p_{2,1} &= 1 - \exp[-5(105.9/10^5)] = .00528 \\ p_{1,1} &= 1 - \exp[-5(1.5)(105.9/10^5)] = .00791 \\ k_1 &= 4722/10,199 = 0.463 \\ p_1^{(0)} &= 10,199/(10,199 + 4722) \\ &= 10,199/14,921 = .684 \\ p_1^{(1)} &= 10,199(1.5)/[10,199(1.5) + 4722] \\ &= 15,299/20,021 = .764 \end{aligned}$$

The number of participants in each age group is not given in Table 14.4. However, if we assume the average length of follow-up is the same for each age group, then

$$\theta_i \equiv t_i / \sum_{i=1}^s t_i$$

where  $t_i = t_{1i} + t_{2i}$ . Thus

$$\theta_1 = 14,921/262,109 = .0569$$

The computations for each age group are summarized in Table 14.8.

Therefore, the required expected total number of events is

$$\begin{aligned} m &= \frac{(z_{.975}\sqrt{.1894} + z_{.80}\sqrt{.2197})^2}{(.2734 - .3566)^2} \\ &= \frac{[1.96(.4352) + 0.84(.4687)]^2}{.0832^2} \\ &= \frac{1.2468^2}{.0832^2} = 224.5 \text{ or } 225 \text{ events} \end{aligned}$$

The corresponding total number of participants is

$$n = \frac{225}{1.04 \times 10^{-2}} = 21,560.4 \text{ or } 21,561 \text{ participants}$$

This constitutes approximately 107,805 person-years among current and never users combined. From Table 14.4, we see that 90,762 person-years (25.7%) are realized by past users of PMH replacement out of a total of 352,871 person-years. Thus, we need to accrue  $107,805/(1 - .257) = 145,135$  person-years or enroll 29,027 postmenopausal women and follow them for 5 years to achieve 80% power in the comparison of current users with never users, if the underlying  $IRR = 1.5$  between these groups.

**Table 14.8** Computations needed for sample-size estimate in Example 14.22

<i>i</i>	Age group	<i>ID</i> <sub>2<i>i</i></sub> <sup>a</sup>	<i>p</i> <sub>2<i>i</i></sub>	<i>p</i> <sub>1<i>i</i></sub>	<i>t</i> <sub>1<i>i</i></sub>	<i>t</i> <sub>2<i>i</i></sub>	<i>k</i> <sub><i>i</i></sub>	<i>θ</i> <sub><i>i</i></sub>	<i>p</i> <sub><i>i</i></sub> <sup>(0)</sup>	<i>p</i> <sub><i>i</i></sub> <sup>(1)</sup>	<i>G</i> <sub><i>i</i></sub>	<i>λ</i> <sub><i>i</i></sub>
1	39–44	105.9	.00528	.00791	10,199	4722	0.463	0.0569	.684	.764	$4.03 \times 10^{-4}$	.039
2	45–49	124.9	.00623	.00933	14,044	20,812	1.482	0.1330	.403	.503	$9.94 \times 10^{-4}$	.095
3	50–54	179.8	.00895	.01339	24,948	71,746	2.876	0.3689	.258	.343	$3.72 \times 10^{-3}$	.357
4	55–59	216.6	.01077	.01611	21,576	73,413	3.403	0.3624	.227	.306	$4.34 \times 10^{-3}$	.416
5	60–64	221.9	.01103	.01650	4876	15,773	3.235	0.0788	.236	.317	$9.71 \times 10^{-4}$	.093
Total												$1.04 \times 10^{-2}$
<i>i</i>	Age group	<i>A</i> <sub><i>i</i></sub>	<i>B</i> <sub><i>i</i></sub>	<i>C</i> <sub><i>i</i></sub>	<i>D</i> <sub><i>i</i></sub>							
1	39–44	.0264	.0295	.0084	.0070							
2	45–49	.0384	.0479	.0229	.0238							
3	50–54	.0921	.1223	.0683	.0804							
4	55–59	.0945	.1273	.0731	.0884							
5	60–64	.0220	.0295	.0168	.0201							
Total		.2734	.3566	.1894	.2197							

<sup>a</sup>Per 10<sup>6</sup> person-years.

## Estimation of Power

In some instances, the size of the study population and the average duration of follow-up are fixed by design and we want to assess the power that can be obtained for a given rate ratio. In this instance, we can solve for the power as a function of the total number of person-years  $\left( T = \sum_{j=1}^s \sum_{i=1}^n t_{ji} \right)$ , stratum-specific incidence rates among the unexposed (*ID*<sub>2*i*</sub>), distribution of person-years by stratum and exposure status (*t*<sub>1*i*</sub>, *t*<sub>2*i*</sub>), projected rate ratio (*IRR*), and type I error ( $\alpha$ ). The power formula follows.

### Equation 14.34

#### Estimation of Power for Stratified Incidence-Rate Data

Suppose we want to compare the incidence rate of disease between exposed and unexposed participants and want to control for one (or more) covariates that can, as a group, be represented by a single categorical variable with *k* categories. Using the same notation as in Equation 14.33, if we wish to test the hypothesis  $H_0$ : *IRR* (rate ratio) = 1 vs.  $H_1$ : *IRR*  $\neq$  1, using a two-sided test with significance level  $\alpha$ , then the power vs. the specific rate ratio = *IRR* under the alternative hypothesis is given by

$$\text{Power} = \Phi \left[ \frac{\sqrt{m} |B - A| - z_{1-\alpha/2} \sqrt{C}}{\sqrt{D}} \right]$$

where *m* = the total expected number of events given by

$$m = \left( \sum_{i=1}^s G_i \right) n$$

*n* = total number of exposed and unexposed individuals over all strata, and *A*, *B*, *C*, *D*, and *G*<sub>*i*</sub> are defined in Equation 14.33.

**Example 14.23**

**Cancer** Suppose we enroll 25,000 postmenopausal women and expect that 75% of the person-time is attributable to current or never PMH use with an average follow-up of 5 years per woman. If the same assumptions are made as in Example 14.22, then how much power will the study have if the true rate ratio = 1.5?

**Solution**

Because the exposure-stratum-specific incidence rates ( $p_{1i}$ ,  $p_{2i}$ ) and person-year distribution are the same as in Example 14.22, we can use the same values for  $p_{1i}$ ,  $p_{2i}$ ,  $p_i^{(0)}$ ,  $p_i^{(1)}$ , and  $\lambda_i$ . Thus, from Table 14.8,  $A = .2734$ ,  $B = .3566$ ,  $C = .1894$ , and  $D = .2197$ .

To compute  $m$ , we note from Table 14.8 that  $\sum_{i=1}^5 G_i = 1.04 \times 10^{-2}$ . Also, the number of women who are current or never users =  $25,000(.75) = 18,750$ . Thus

$$m = 1.04 \times 10^{-2} (18,750) = 195.7 \text{ or } 196 \text{ events}$$

To compute power, we refer to Equation 14.34 and obtain

$$\begin{aligned}\text{Power} &= \Phi\left[\frac{\sqrt{196}(.3566 - .2734) - Z_{.975}\sqrt{.1894}}{\sqrt{.2197}}\right] \\ &= \Phi\left[\frac{1.1651 - 1.96(.4352)}{.4687}\right] \\ &= \Phi\left(\frac{0.3120}{0.4687}\right) \\ &= \Phi(0.666) = .747\end{aligned}$$

Thus the study would have about 75% power. Note that, from Example 14.22, to achieve 80% power we needed to accrue 145,135 person-years among all postmenopausal women or 107,805 person-years among current or never PMH users to yield an expected 225 events. If we actually accrue 125,000 person-years among all postmenopausal women, as in this example, then this will result in 93,750 person-years among current or never PMH users, which yields an expected 196 events and, as a result, obtains about 75% power.

Another approach to power estimation is to ignore the effect of age and base the power computation on the comparison of overall incidence rates between current and never PMH users. However, from Example 14.19 we see that the age-adjusted  $IRR = 1.41$ , whereas the crude  $IRR = 1.25$ , between breast-cancer incidence and current PMH use. In this example, age is a negative confounder because it is positively related to breast-cancer incidence and is negatively related to PMH use. Thus, the power based on crude rates is lower than the appropriate power based on rates stratified by age. In the case of a positive confounder, the power based on crude rates is higher than the power based on rates stratified by age. In general, if confounding is present, then it is important to base power calculations on Equation 14.34, which takes confounding into account, rather than Equation 14.18, which does not.

**REVIEW QUESTIONS 14D**

- Consider the study described in Review Questions 14C. Suppose we are planning a new study among both men and women, assessing the association between cigarette smoking and the incidence of open-angle glaucoma. Let's classify participants as current smokers or never smokers at baseline and follow each participant for 5 years. We want to conduct a two-sided test with  $\alpha = .05$  and power = 80% to detect an  $IRR$  of 0.80 comparing current smokers with never smokers. If the sex distribution and the sex-specific incidence rates of open-angle glaucoma among never

smokers are the same as in Review Questions 14C, then how many participants do we need to enroll? (*Hint:* Recall that in Review Questions 14C, each woman was followed for an average of 16 years and each man was followed for an average of 10 years.)

- 2 Suppose we have a total cohort (men and women combined) of 100,000 people and each is followed for 5 years. How much power will the study have to detect an *IRR* of 0.8 under the same assumptions as in Review Question 14D.1? (*Hint:* Assume the sex by smoking distribution is the same as in Table 14.7.)

## 14.7 Testing for Trend: Incidence-Rate Data

In Sections 14.1–14.6, we were concerned with the comparison of incidence rates between an exposed and an unexposed group, possibly after controlling for other relevant covariates. In some instances, there are more than two exposure categories and researchers want to assess whether incidence rates are increasing or decreasing in a consistent manner as the level of exposure increases.

**Example 14.24**

**Cancer** The data in Table 14.9 display the relationship between breast-cancer incidence and parity (the number of children) by age, based on NHS data from 1976 to 1990. We see that within a given parity group, breast-cancer incidence rises sharply with age. Thus it is important to control for age in the analysis. Also, within a given age group, breast-cancer incidence is somewhat higher for women with 1 birth than for nulliparous women (women with 0 births). However, for parous women (women with at least 1 child), breast-cancer incidence seems to decline with increasing parity. How should we assess the relationship between breast-cancer incidence and parity?

It is reasonable to study parous women as a group and to model  $\ln(\text{breast-cancer incidence})$  for the  $i$ th age group and the  $j$ th parity group as a linear function of parity.

**Table 14.9** Relationship of breast-cancer incidence to parity after controlling for age, NHS, 1976–1990

Age group	Parity			
	0	1	2	3+
30–39	13/15,265 (85)	18/20,098 (90)	72/87,436 (82)	60/86,452 (69)
40–49	44/30,922 (142)	73/31,953 (228)	245/140,285 (175)	416/262,068 (159)
50–59	102/35,206 (290)	94/31,636 (297)	271/103,399 (262)	608/262,162 (232)
60–69	32/11,594 (276)	50/10,264 (487)	86/29,502 (292)	176/64,448 (273)

<sup>a</sup>Per 100,000 person-years.

**Equation 14.35**

$$\ln(p_{ij}) = \alpha_i + \beta(j - 1)$$

where  $\alpha_i$  represents  $\ln(\text{incidence})$  for women in the  $i$ th age group with 1 child and  $\beta$  represents the increase in  $\ln(\text{incidence})$  for each additional child. Notice that  $\beta$  is assumed to be the same for each age group ( $i$ ). In general, if we have  $k$  exposure groups we might assign a score  $S_j$  for the  $j$ th exposure group, which might represent average exposure within that group, and consider a model of the form

**Equation 14.36**

$$\ln(p_{ij}) = \alpha_i + \beta S_j$$

We want to test the hypothesis  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ . The models in Equations 14.35 and 14.36 represent a more efficient use of the data than comparing individual pairs of groups because we can use all the data to test for an overall trend. Comparing pairs of groups might yield contradictory results and would often have less power than the overall test for trend. We use a “weighted regression approach” where incidence rates based on a larger number of cases are given more weight. The procedure is summarized as follows.

**Equation 14.37****Test for Trend: Incidence-Rate Data**

Suppose we have an exposure variable  $E$  with  $k$  levels of exposure, where the  $j$ th exposure group is characterized by a score  $S_j$ , which may represent the average level of exposure within that group, if available. If no obvious scoring method is available, then integer scores  $1, \dots, k$  may be used instead. If  $p_{ij}$  = true incidence rate for the  $i$ th stratum and  $j$ th level of exposure,  $i = 1, \dots, s; j = 1, \dots, k$ ;  $\hat{p}_{ij}$  = the corresponding observed incidence rate and we assume that

$$\ln(p_{ij}) = \alpha_i + \beta S_j$$

then, to test the hypothesis  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ , using a two-sided test with significance level  $\alpha$ :

- (1) We compute a point estimate of  $\beta$  given by  $\hat{\beta} = L_{xy}/L_{xx}$ , where

$$L_{xy} = \sum_{i=1}^s \sum_{j=1}^k w_{ij} S_j \ln(\hat{p}_{ij}) - \left( \sum_{i=1}^s \sum_{j=1}^k w_{ij} S_j \right) \left[ \sum_{i=1}^s \sum_{j=1}^k w_{ij} \ln(\hat{p}_{ij}) \right] / \sum_{i=1}^s \sum_{j=1}^k w_{ij}$$

$$L_{xx} = \sum_{i=1}^s \sum_{j=1}^k w_{ij} S_j^2 - \left( \sum_{i=1}^s \sum_{j=1}^k w_{ij} S_j \right)^2 / \sum_{i=1}^s \sum_{j=1}^k w_{ij}$$

$w_{ij} = a_{ij}$  = number of cases in the  $i$ th stratum and  $j$ th level of exposure

- (2) The standard error of  $\hat{\beta}$  is given by

$$se(\hat{\beta}) = 1/\sqrt{L_{xx}}$$

- (3) We compute the test statistic

$$z = \hat{\beta}/se(\hat{\beta}) \sim N(0, 1) \text{ under } H_0$$

- (4) If  $z > z_{1-\alpha/2}$  or  $z < -z_{1-\alpha/2}$ , then we reject  $H_0$ ;

if  $-z_{1-\alpha/2} \leq z \leq z_{1-\alpha/2}$ , then we accept  $H_0$ .

- (5) The two-sided  $p$ -value =  $2\Phi(z)$  if  $z < 0$

$$= 2[1 - \Phi(z)] \text{ if } z \geq 0$$

(6) A two-sided  $100\% \times (1 - \alpha)$  CI for  $\beta$  is given by

$$\hat{\beta} \pm z_{1-\alpha/2} se(\hat{\beta})$$

### Example 14.25

**Cancer** Assess whether there is a significant trend between breast-cancer incidence and parity for parous women based on the data in Table 14.9.

#### Solution

We have four age strata (30–39, 40–49, 50–59, 60–69) ( $s = 4$ ) and three exposure (parity) groups ( $k = 3$ ) to which we assign scores of 1, 2, 3, respectively. The  $\ln(\text{incidence rate})$ , score, and weight are given for each of the 12 cells in Table 14.10. We then proceed as in Equation 14.37, as follows:

$$\begin{aligned}\sum_{i=1}^4 \sum_{j=1}^3 w_{ij} &= 2169 \\ \sum_{i=1}^4 \sum_{j=1}^3 w_{ij} \ln(\hat{p}_{ij}) &= -13,408.7 \\ \sum_{i=1}^4 \sum_{j=1}^3 w_{ij} S_j &= 5363 \\ \sum_{i=1}^4 \sum_{j=1}^3 w_{ij} S_j^2 &= 14,271 \\ \sum_{i=1}^4 \sum_{j=1}^3 w_{ij} S_j \ln(\hat{p}_{ij}) &= -33,279.2\end{aligned}$$

**Table 14.10**  $\ln(\text{incidence rate})$  of breast cancer and weight used in weighted regression analysis

Age group	$i$	Parity ( $j$ )	$\ln(\hat{p}_{ij})$	$w_{ij}$
30–39	1	1	-7.018	18
30–39	1	2	-7.102	72
30–39	1	3	-7.273	60
40–49	2	1	-6.082	73
40–49	2	2	-6.350	245
40–49	2	3	-6.446	416
50–59	3	1	-5.819	94
50–59	3	2	-5.944	271
50–59	3	3	-6.067	608
60–69	4	1	-5.324	50
60–69	4	2	-5.838	86
60–69	4	3	-5.903	176

$$L_{xx} = 14,271 - 5363^2 / 2169 = 1010.6$$

$$L_{xy} = -33,279.2 - (-13,408.7)(5363) / 2169 = -125.2$$

$$\hat{\beta} = -125.2 / 1010.6 = -0.124$$

$$se(\hat{\beta}) = \sqrt{1/1010.6} = 0.031$$

$$z = \hat{\beta} / se(\hat{\beta}) = -0.124 / 0.031 = -3.94 \sim N(0, 1)$$

$$p\text{-value} = 2 \times \Phi(-3.94) < 0.001$$

Thus there is a significant inverse association between  $\ln(\text{breast-cancer incidence})$  and parity among parous women. Breast-cancer incidence declines by  $(1 - e^{-0.124}) \times 100\% = 11.7\%$  for each additional birth up to 3 births within a given age group.

In this section, we have considered the problem of relating the incidence density to a categorical exposure variable  $E$ , where  $E$  has more than two categories and the categories of  $E$  correspond to an ordinal scale with an associated score variable  $S_j$  for the  $j$ th category. The procedure is similar to the chi-square test for trend given in Chapter 10, except that here we are modeling trends in incidence rates based on person-time data, whereas in Chapter 10 we were modeling trends in proportions based on count data (which as a special case might correspond to cumulative incidence), where the person is the unit of analysis.

On the flowchart (Figure 14.15, p. 803), we answer yes to (1) person-time data? no to (2) one-sample problem? yes to (3) incidence rates remain constant over time? and no to (4) two-sample problem? which leads to (5) interested in test of trend over more than two exposure groups. This path leads us to the box labeled “Use test of trend for incidence rates.”

## 14.8 Introduction to Survival Analysis

Sections 14.1–14.7 discussed methods for comparing incidence rates between two groups, where the period of follow-up may differ for the two groups considered. One assumption made in performing these analyses is that incidence rates remain *constant* over time. In many instances this assumption is not warranted and one wants to compare the number of disease events between two groups where the incidence of disease varies over time.

### Example 14.26

**Health Promotion** Consider Data Set SMOKE.DAT on the Companion Website. In this data set, 234 smokers who expressed a willingness to quit smoking were followed for 1 year to estimate the cumulative incidence of recidivism; that is, the proportion of smokers who quit for a time but who started smoking again. One hypothesis is that older smokers are less likely than younger smokers to be successful quitters (and more likely to be recidivists). How can this hypothesis be tested?

The data in Table 14.11 were obtained after subdividing the study population by age ( $>40/\leq 40$ ).

**Table 14.11 Number of days quit smoking by age**

Age	Number of days quit smoking					Total
	$\leq 90$	91–180	181–270	271–364	365	
$>40$	92	4	4	1	19	120
$\leq 40$	88	7	3	2	14	114
Total	180	11	7	3	33	234
Percentage	76.9	4.7	3.0	1.3	14.1	

We can compute the estimated incidence rate of disease (recidivism) within each 90-day period for the combined study population. Let's assume participants who started smoking within a given period did so at the midpoint of the respective period. Thus the number of person-days within days 1–90 = 180(45) + 54(90) = 12,960, and the incidence rate of recidivism = 180/12,960 = 0.014 events per person-day. For days 91–180, there were 11 recidivists and 43 successful quitters. Hence, the number of person-days = 11(45) + 43(90) = 4365 person-days and the incidence rate = 11/4365 = 0.0025 events per person-day. Similarly, the incidence rate over days 181–270 = 7/[7(45) + 36(90)] = 7/3555 = 0.0020 events per person-day. Finally, the incidence rate over days 271–365 = 3/[3(47) + 33(95)] = 3/3276 = 0.00092 events per person-day. Thus, the incidence rate of recidivism is much higher in the first 90 days and declines throughout the 365-day period. Incidence rates that vary substantially over time are more commonly called **hazard rates**.

In Example 14.26, we have assumed, for simplicity, that the hazard remains constant during each 90-day period. One nice way of comparing incidence rates between two groups is to plot their hazard functions.

#### Example 14.27

**Health Promotion** Plot the hazard function for subjects age >40 and age ≤40, respectively.

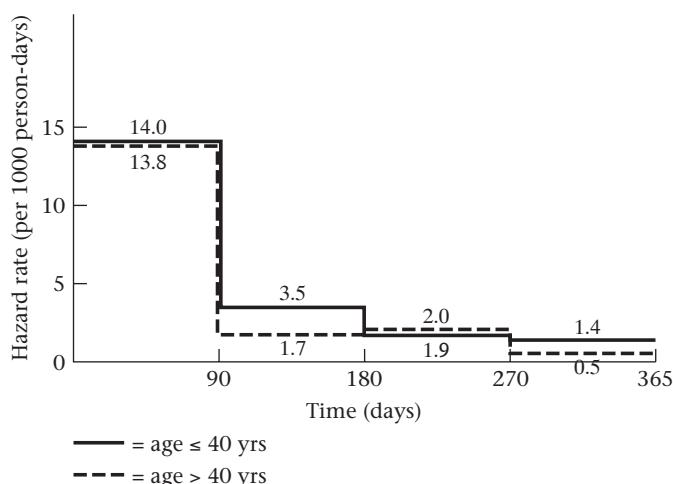
#### Solution

The hazard functions are plotted in Figure 14.4. There is actually a slight tendency for younger smokers ( $\leq 40$ ) to be more likely to start smoking (i.e., become recidivists) than older smokers ( $>40$ ), particularly after the first 90 days.

Hazard functions are used extensively in biostatistical work to assess mortality risk.

Another way of comparing disease incidence between two groups is by their cumulative incidence. If the incidence rate of disease over time is constant, as was assumed in Sections 14.1–14.7, then the cumulative incidence over time  $t$  is given exactly by  $1 - e^{-\lambda t}$  and approximately by  $\lambda t$  if the cumulative incidence is low (see Equation 14.3). We could also compute the probability of not developing disease = 1 – cumulative incidence =  $e^{-\lambda t} \approx 1 - \lambda t$ . The probability of not developing disease is

**Figure 14.4** Hazard rates (per 1000 person-days) by age



commonly called the **survival probability**. We can plot the survival probability as a function of time. This function is the survival function.

**Definition 14.3** The **survival function**  $S(t)$  gives the probability of survival up to time  $t$  for each  $t \geq 0$ .

The hazard at time  $t$ , denoted by  $h(t)$ , can be expressed in terms of the survival function  $S(t)$  as follows:

**Definition 14.4** The **hazard function**  $h(t)$  is the *instantaneous* probability of having an event at time  $t$  (per unit time), i.e., the instantaneous incidence rate, given that one has survived (i.e., has not had an event) up to time  $t$ . In particular,

$$h(t) = \left[ \frac{S(t) - S(t + \Delta t)}{\Delta t} \right] / S(t) \text{ as } \Delta t \text{ approaches 0}$$

**Example 14.28** **Demography** Based on U.S. life table data in 1986 there were 100,000 men at age 0 of whom 80,908 survived to age 60, 79,539 survived to age 61, 34,789 survived to age 80, and 31,739 survived to age 81. Compute the approximate mortality hazard at ages 60 and 80, respectively, for U.S. men in 1986.

**Solution** There were 80,908 men who survived to age 60, and 79,539 men who survived to age 61. Therefore, the hazard at age 60 is approximately,

$$h(60) = \frac{80,908 - 79,539}{80,908} = .017$$

Similarly, because 34,789 men survived to age 80, and 31,739 men survived to age 81, the hazard at age 80 is approximately given by

$$h(80) = \frac{34,789 - 31,739}{34,789} = .088$$

Thus, in words, the probability of dying in the next year is 1.7% given that one has survived to age 60, and 8.8% given that one has survived to age 80. The percentages 1.7% and 8.8% represent the approximate hazard at ages 60 and 80, respectively. To improve the approximation, shorter time intervals than 1 year would need to be considered.

#### REVIEW QUESTION 14E

1 What is a hazard function?

### 14.9 Estimation of Survival Curves: The Kaplan-Meier Estimator

To estimate the survival probability when the incidence rate varies over time, we could use a more complex parametric survival model than the exponential model given in Equation 14.3. (See [3] for a good description of other parametric survival models.) For this purpose, we will discuss the Weibull model later in this chapter.

However, a more common approach is to use a nonparametric method referred to as the **product-limit** or **Kaplan-Meier estimator**.

Suppose individuals in the study population are assessed at times  $t_1, \dots, t_k$  where the times do not have to be equally spaced. If we want to compute the probability of surviving up to time  $t$ , we can write this probability in the form.

**Equation 14.38**

$$\begin{aligned} S(t_i) &= \text{Prob}(\text{surviving to time } t_i) = \text{Prob}(\text{surviving to time } t_1) \\ &\quad \times \text{Prob}(\text{surviving to time } t_2 | \text{survived to time } t_1) \\ &\quad \vdots \\ &\quad \times \text{Prob}(\text{surviving to time } t_j | \text{survived to time } t_{j-1}) \\ &\quad \vdots \\ &\quad \times \text{Prob}(\text{surviving to time } t_i | \text{survived to time } t_{i-1}) \end{aligned}$$

**Example 14.29**

**Health Promotion** Estimate the survival curve for people age  $>40$  and age  $\leq 40$  for the participants depicted in Table 14.11.

**Solution**

We have for persons age  $>40$ ,

$$S(90) = 1 - \frac{92}{120} = .233$$

$$S(180) = S(90) \times \left(1 - \frac{4}{28}\right) = .200$$

$$S(270) = S(180) \times \left(1 - \frac{4}{24}\right) = .167$$

$$S(365) = S(270) \times \left(1 - \frac{1}{20}\right) = .158$$

For people age  $\leq 40$ , we have

$$S(90) = 1 - \frac{88}{114} = .228$$

$$S(180) = S(90) \times \left(1 - \frac{7}{26}\right) = .167$$

$$S(270) = S(180) \times \left(1 - \frac{3}{19}\right) = .140$$

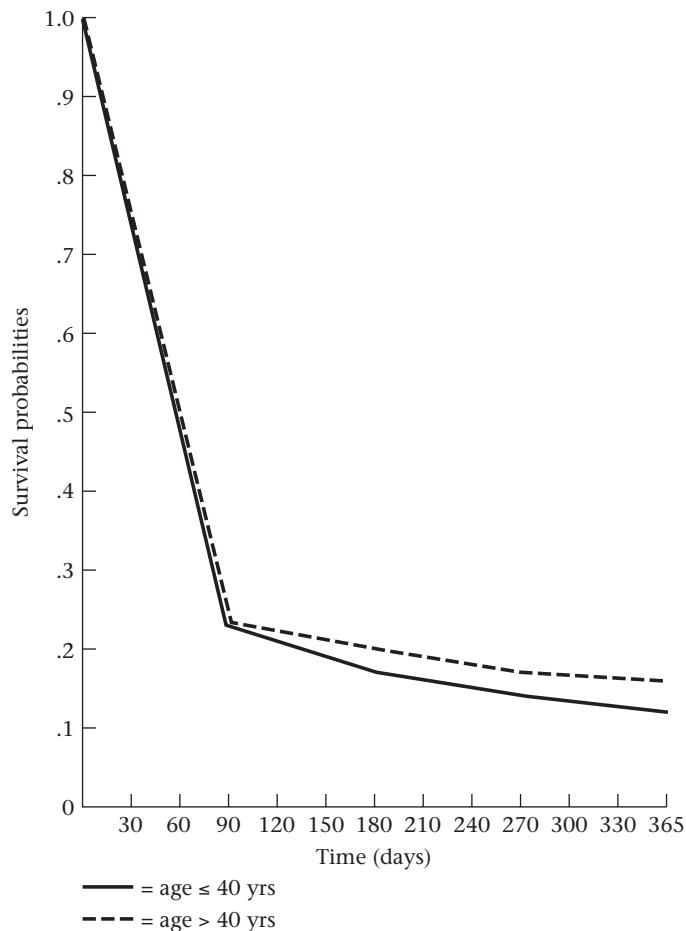
$$S(365) = S(270) \times \left(1 - \frac{2}{16}\right) = .123$$

These survival curves are plotted in Figure 14.5.

Participants age  $>40$  have a slightly higher estimated survival probability (i.e., probability of remaining a quitter) after the first 90 days.

## The Treatment of Censored Data

In Example 14.26, all members of the study population were followed until they started smoking again or for 1 year, whichever occurred first. In other instances, some participants are not followed for the maximum period of follow-up but have not yet had an event (have not yet failed).

**Figure 14.5** Survival probabilities by age**Example 14.30**

**Ophthalmology** A clinical trial was conducted to test the efficacy of different vitamin supplements in preventing visual loss in patients with retinitis pigmentosa (RP) [4]. Visual loss was measured by loss of retinal function as characterized by a 50% decline in the electroretinogram (ERG) 30 Hz amplitude, a measure of the electrical activity in the retina. In normal people, the normal range for ERG 30 Hz amplitude is  $>50 \mu\text{V}$  (microvolts). In patients with RP, ERG 30 Hz amplitude is usually  $<10 \mu\text{V}$  and is often  $<1 \mu\text{V}$ . Approximately 50% of patients with ERG 30 Hz amplitudes near  $0.05 \mu\text{V}$  are legally blind compared with  $<10\%$  of patients whose ERG 30 Hz amplitudes are near  $1.3 \mu\text{V}$  (the average ERG amplitude for patients in this clinical trial). Patients in the study were randomized to one of four treatment groups:

Group 1 received 15,000 IU of vitamin A and 3 IU (a trace amount) of vitamin E.

Group 2 received 75 IU (a trace amount) of vitamin A and 3 IU of vitamin E.

Group 3 received 15,000 IU of vitamin A and 400 IU of vitamin E.

Group 4 received 75 IU of vitamin A and 400 IU of vitamin E.

Let's call these four groups the A group, trace group, AE group, and E group, respectively. We want to compare the proportion of patients who fail (i.e., lose 50%

of initial ERG 30 Hz amplitude) in different treatment groups. Patients were enrolled in 1984–1987, and follow-up was terminated in September 1991. Because follow-up was terminated at the same point in *chronological time*, the period of follow-up differed for each patient. Patients who entered early in the study were followed for 6 years, whereas patients who enrolled later in the study were followed for 4 years. In addition, some patients dropped out of the study before September 1991 and had not failed. Dropouts were due to death, other diseases, side effects possibly due to the study medications, or unwillingness to comply (take study medications). How can we estimate the hazard and survival functions in each treatment group in the presence of variable follow-up for each patient?

#### **Definition 14.5**

We refer to patients who do not reach a disease endpoint during their period of follow-up as **censored observations**. A participant has been censored at time  $t$  if the participant has been followed up to time  $t$  and has not failed.

We assume censoring is noninformative; that is, patients who are censored have the same underlying survival curve after their censoring time as patients who are not censored.

#### **Definition 14.6**

**Right censored data** are data in which a subject is known to have survived for at least  $t$  weeks but the failure time after this point is unknown. For example, in a cancer clinical trial in which the endpoint is recurrence, a subject may have remained in remission (i.e., survived) for 13 weeks during the study and did not experience a recurrence (i.e., fail) as of the end of the study. The survival time for this subject is denoted by 13+ weeks. This is distinct from another subject who remained in remission for 13 weeks and then had a recurrence at this time (i.e., failed). The survival time for this subject is 13 weeks.

Thus each subject has two important data values that are used in the analysis of survival data,  $(T_i, C_i)$ , where  $T_i$  = survival time during the study and  $C_i$  = censoring indicator = 1 if the subject failed during the study and = 0 if the subject was censored. Thus, the first subject is denoted by (13, 0) or 13+, while the second subject is denoted by (13, 1) or just 13.

### Other Types of Censoring

Other types of censoring are also possible.

#### **Definition 14.7**

**Left Censoring** In a study of age of legal blindness among subjects with RP, there may be a subset of subjects who are already legally blind at the start of the study; the precise age of legal blindness is unknown. This is known as **left censoring**.

#### **Definition 14.8**

**Interval Censoring** In a study of age of developing breast cancer among postmenopausal women, it may be known that a woman is breast cancer free at age 50 (first questionnaire) and has developed breast cancer by age 52 (next questionnaire), but the precise age at diagnosis is unknown. These data would be **interval censored**.

In this text, we will focus on **right censoring**, which is the most common.

To estimate the survival function in the presence of censoring, suppose  $S_{i-1}$  patients have survived through time  $t_{i-1}$  and are not censored at time  $t_{i-1}$ . Among these patients,  $S_i$  patients survive,  $d_i$  patients fail, and  $l_i$  patients are censored at time  $t_i$ . Thus  $S_{i-1} = S_i + d_i + l_i$ . We can estimate the probability of surviving to time  $t_i$  given that a patient has survived up to time  $t_{i-1}$  by  $(1 - d_i/S_{i-1}) = [1 - d_i/(S_i + d_i + l_i)]$ . The  $l_i$  patients who are censored at time  $t_i$  do not contribute to the estimation of the survival function at time  $> t_i$ . However, these patients do contribute to the estimation of the survival function at time  $\leq t_i$ . We can summarize this procedure as follows.

**Equation 14.39****Kaplan-Meier (Product-Limit) Estimator of Survival (Censored Data)**

Suppose that  $S_{i-1}$  subjects have survived up to time  $t_{i-1}$  and are not censored at time  $t_{i-1}$ , of whom  $S_i$  survive,  $d_i$  fail, and  $l_i$  are censored at time  $t_i$ ,  $i = 1, \dots, k$ . The **Kaplan-Meier estimator** of the survival probability at time  $t_i$  is

$$\hat{S}(t_i) = \left(1 - \frac{d_1}{S_0}\right) \times \left(1 - \frac{d_2}{S_1}\right) \times \dots \times \left(1 - \frac{d_i}{S_{i-1}}\right), \quad i = 1, \dots, k$$

**Example 14.31**

**Ophthalmology** Estimate the survival probability at each of years 1–6 for participants receiving 15,000 IU of vitamin A (i.e., groups A and AE combined) and participants receiving 75 IU of vitamin A (i.e., groups E and trace combined), respectively, based on the data set mentioned in Example 14.30.

**Solution**

The calculations are given in Table 14.12. For example, for the participants receiving 15,000 IU of vitamin A the survival probability at year 1 = .9826. The probability of surviving to year 2 given that one survives to year 1 = 159/165 = .9636. Thus the survival probability at year 2 = .9826 × .9636 = .9468, etc. The survival probabilities for the participants receiving 15,000 IU of vitamin A tend to be higher than for the participants receiving 75 IU of vitamin A, particularly at year 6.

In the above example, we assume that subjects who are censored at time  $t$  are followed up to time  $t$  and have not failed as of time  $t$ . There are other methods for calculating survival probabilities, such as assuming that censored participants are only followed for half of the time interval during which they were measured. These methods are beyond the scope of this text.

**Interval Estimation of Survival Probabilities**

In Equation 14.39, we derived a point estimate of the survival probability at specific time points. We can derive an interval estimate as well. To obtain an interval estimate for  $S(t)$ , we consider  $Var\{\ln[\hat{S}(t)]\}$ , which is given by

**Equation 14.40**

$$Var\{\ln[\hat{S}(t_i)]\} = \sum_{j=1}^i \frac{d_j}{S_{j-1}(S_{j-1} - d_j)}$$

We then can obtain an approximate two-sided 100%  $\times (1 - \alpha)$  CI for  $\ln[S(t_i)]$  given by

$$\ln[\hat{S}(t_i)] \pm z_{1-\alpha/2} \times \sqrt{Var\{\ln[\hat{S}(t_i)]\}} = (c_1, c_2)$$

**Table 14.12** Survival probabilities for participants receiving 15,000 IU of vitamin A and 75 IU of vitamin A, respectively

Time	Fail	Censored	Survive	Total	Prob(survive to time $t_i$   survived up to time $t_{i-1}$ )	$\hat{S}(t_i)$	$\hat{h}(t_i)$
<b>15,000 IU of vitamin A daily</b>							
1 yr = $t_1$	3	4	165	172	.9826	.9826	0.0174
2 yr = $t_2$	6	0	159	165	.9636	.9468	0.0364
3 yr = $t_3$	15	1	143	159	.9057	.8575	0.0943
4 yr = $t_4$	21	26	96	143	.8531	.7316	0.1469
5 yr = $t_5$	15	35	46	96	.8438	.6173	0.1563
6 yr = $t_6$	5	41	0	46	.8913	.5502	0.1087
<b>75 IU of vitamin A daily</b>							
1 yr = $t_1$	8	0	174	182	.9560	.9560	0.0440
2 yr = $t_2$	13	3	158	174	.9253	.8846	0.0747
3 yr = $t_3$	21	2	135	158	.8671	.7670	0.1329
4 yr = $t_4$	21	28	86	135	.8444	.6477	0.1556
5 yr = $t_5$	13	31	42	86	.8488	.5498	0.1512
6 yr = $t_6$	13	29	0	42	.6905	.3796	0.3095

Note: A person fails if his or her ERG 30 Hz amplitude declines by at least 50% from baseline to any follow-up visit, regardless of any subsequent ERG values obtained after the visit where the failure occurs.

The corresponding two-sided  $100\% \times (1 - \alpha)$  CI for  $S(t_i)$  is given by  $(e^{c_1}, e^{c_2})$ . This procedure, known as Greenwood's formula, is summarized as follows.

#### Equation 14.41

#### Interval Estimation of Survival Probabilities

Suppose that  $S_{i-1}$  subjects have survived up to time  $t_{i-1}$  and are not censored at time  $t_{i-1}$ , of whom  $S_i$  survive,  $d_i$  fail, and  $l_i$  are censored at time  $t_i$ ,  $i = 1, \dots, k$ . A two-sided  $100\% \times (1 - \alpha)$  CI for the survival probability at time  $t_i$  (i.e.,  $S(t_i)$ ) is given by  $(e^{c_1}, e^{c_2})$ , where

$$c_1 = \ln[\hat{S}(t_i)] - z_{1-\alpha/2} se\{\ln[\hat{S}(t_i)]\}$$

$$c_2 = \ln[\hat{S}(t_i)] + z_{1-\alpha/2} se\{\ln[\hat{S}(t_i)]\}$$

$\hat{S}(t_i)$  is obtained from the Kaplan-Meier estimator in Equation 14.39, and

$$se\{\ln[\hat{S}(t_i)]\} = \sqrt{\sum_{j=1}^i \frac{d_j}{S_{j-1}(S_{j-1} - d_j)}}, \quad i = 1, \dots, k$$

#### Example 14.32

**Ophthalmology** Obtain a 95% CI for the survival probability at 6 years for the patients assigned to the 15,000 IU/day vitamin A group based on the data in Table 14.12.

#### Solution

From Table 14.12, we have  $\hat{S}(6) = .5502$ . Thus

$$\ln[\hat{S}(6)] = \ln(.5502) = -0.5975$$

$$Var\{\ln[\hat{S}(6)]\} = \frac{3}{172(169)} + \frac{6}{165(159)} + \frac{15}{159(144)} + \frac{21}{143(122)} + \frac{15}{96(81)} + \frac{5}{46(41)} \\ = 6.771 \times 10^{-3}$$

Thus a 95% CI for  $\ln[S(6)]$  is given by

$$-0.5975 \pm 1.96 \sqrt{6.771 \times 10^{-3}} = -0.5975 \pm 0.1613 = (-0.7588, -0.4362)$$

The corresponding 95% CI for  $S(6)$  is  $(e^{-0.7588}, e^{-0.4362}) = (.4682, .6465)$ .

## Estimation of the Hazard Function: The Product-Limit Method

In Example 14.27, we considered estimation of the hazard function in the context of the smoking-cessation data in Table 14.11. In this example, we estimated the hazard for each group for each approximately 90-day period. We assumed recidivists would resume smoking randomly throughout the 90-day period. Thus to compute the hazard in the first 90 days for participants with age >40, we assume at day 45 half the recidivists ( $92/2 = 46$ ) have resumed smoking. Thus there remain  $120 - 46 = 74$  participants who are still quitters. Among these participants  $92/90 = 1.022$  participants resume smoking by day 46. Thus the estimated hazard rate at day 45 =  $1.022/74 = .0138$  events per person-day. In a similar manner, the hazard for each 90-day period was approximated by the hazard at the period midpoint. This approach to hazard estimation is often called the *actuarial method*.

In epidemiology, another approach is often used. A key assumption of the actuarial method is that events occur randomly throughout a defined follow-up period. Another approach is to assume an event occurs at the precise time it is either observed (e.g., if an abnormality such as a heart murmur is observed at a physical examination) or reported (e.g., if the patient reports a specific symptom, such as dizzy spells or breathlessness). This approach is called the **product-limit method**.

### Example 14.33

**Ophthalmology** Estimate the hazard at each year of follow-up for each vitamin A dose group based on the data in Table 14.12 using the product-limit method.

### Solution

For participants taking 15,000 IU of vitamin A daily, the estimated hazard at year 1 = number of participants with events at year 1/number of participants available for examination at year 1 =  $3/172 = .0174$ . At year 2, 165 participants were examined, of whom 6 failed. Thus the estimated hazard at year 2 =  $6/165 = .0364$ , . . . , etc. In general, the estimated hazard at year  $t_i = h(t_i) = d_i/S_{i-1} = 1 - \text{Prob}(\text{survive to time } t_i | \text{survive to time } t_{i-1})$ . The estimated hazard function for each group by year is given in the last column of Table 14.12.

In the rest of this chapter, we use the product-limit approach for hazard estimation. In this section, we introduced the basic concepts of survival analysis. The primary outcome measures in this type of analysis are the *survival function*, which provides the probability of surviving to time  $t$ , and the *hazard function*, which provides the instantaneous rate of disease per unit time given that the person has survived to time  $t$ . A unique aspect of survival data is that usually not all participants are followed for the same length of time. Thus we introduced the concept of a *censored observation*, which is a participant who has not failed by time  $t$  but is not followed any longer, so the actual time of failure for censored observations is unknown. Finally, we learned about the *Kaplan-Meier product-limit method*, which is a technique for estimating the survival and hazard functions in the presence of censored data.

On the flowchart (Figure 14.15, p. 803), we answer yes to (1) person-time data? and no to both (2) one-sample problem? and (3) incidence rates remain constant over time? This path leads us to the box labeled “Use survival-analysis methods.”

In the next section, we continue our discussion of survival analysis and examine analytic techniques for comparing survival curves from two independent samples.

## REVIEW QUESTIONS 14F

- 1 What is the Kaplan-Meier estimator of the survival function?
- 2 What do we mean by a “censored observation”?
- 3 The following study was performed among women with breast cancer who had been treated with tamoxifen for at least 2 years. The women in the study were randomized to either exemestane or tamoxifen and were followed for an additional 3 years to assess whether a change in treatment after being on tamoxifen would provide a better (or worse) clinical outcome [5]. The women were followed for a maximum of 3 years or until they had an event (prior to 3 years). An event was defined as a recurrence of the initial breast cancer, a new breast cancer in the opposite breast, or death. The results are shown in Table 14.13.

**Table 14.13** Time course of events by treatment group in a breast-cancer trial among women who have been treated with tamoxifen for at least 2 years

Exemestane group				
Year	Number of events	Number censored	Number survived	Total
1	52	420	1696	2168
2	60	879	757	1696
3	44	713	0	757
Tamoxifen group				
Year	Number of events	Number censored	Number survived	Total
1	78	413	1682	2173
2	90	862	730	1682
3	76	654	0	730

- (a) Obtain the Kaplan-Meier survival curve for each group, and plot the curves on the same graph.
- (b) Does one group seem to be doing better than the other? Explain.
- (c) Provide a 95% CI for the survival probability at 3 years for each group.

## 14.10 The Log-Rank Test

In this section, we consider how to compare the two survival curves in Figure 14.5 for the smoking-cessation data. We could compare survival at specific time points. However, we usually gain power if we consider the entire survival curve. We could also compare the mean survival time between groups. However, survival time distributions are often highly skewed and it isn't clear how to treat censored observations in computing mean survival time. Suppose we want to compare the survival experience of two groups, the exposed and unexposed groups. Let  $h_1(t)$  = hazard at time  $t$  for participants in the exposed group, and  $h_2(t)$  = hazard at time  $t$  for participants in the unexposed group. We assume the hazard ratio is a constant  $\exp(\beta)$ .

$$h_1(t)/h_2(t) = \exp(\beta)$$

We want to test the hypothesis  $H_0: \beta = 0$  vs.  $H_1: \beta \neq 0$ . If  $\beta = 0$ , then the survival curves of the two groups are the same. If  $\beta > 0$ , then exposed participants are consistently at greater risk for disease than unexposed participants, or equivalently, the survival probability of the exposed group is less than that of the unexposed group at each time  $t$ . If  $\beta < 0$ , then the exposed participants are at lower risk than the unexposed participants and their survival probabilities are greater than those of the unexposed participants. This is a similar hypothesis-testing situation to that in Equation 14.8, where we were interested in comparing two incidence rates. The difference is that in Equation 14.8 we assumed the incidence rate for a group was constant over time, whereas in Equation 14.42 we let the hazard rate for each group vary over time but maintain a constant hazard ratio at each time  $t$ .

Consider the data in Table 14.11. These data could be analyzed in terms of cumulative incidence over 1 year; that is, the percentage of older vs. younger ex-smokers who were successful quitters for 1 year could be compared. However, if incidence changes greatly over the year this is not as powerful as the log-rank test described later in Equation 14.43. Using this procedure, *when* an event occurs (in this case the event is recidivism) rather than simply *whether* it occurs is taken into account.

To implement this procedure, the total period of follow-up is subdivided into shorter time periods over which incidence is relatively constant. In Example 14.26, the ideal situation would be to subdivide the 1-year interval into 365 daily time intervals. However, to illustrate the method time has been subdivided into 3-month intervals. For each time interval, the number of people who have been successful quitters up to the beginning of the interval are identified. These people are at risk for recidivism during this time interval. This group is then categorized according to whether they remained successful quitters or became recidivists during the time interval. For each time interval, the data are displayed as a  $2 \times 2$  contingency table relating age to incidence of recidivism over the time interval.

### Example 14.34

**Health Promotion** Display the smoking-cessation data in Table 14.11 in the form of incidence rates by age for each of the four time intervals, 0–90 days, 91–180 days, 181–270 days, and 271–365 days.

### Solution

For the first time interval, 0–90 days, 120 older smokers were successful quitters at time 0, of whom 92 became recidivists during the 0- to 90-day period; similarly, of 114 younger smokers, 88 became recidivists during this time period. These data are shown in a  $2 \times 2$  contingency table in Table 14.14. For the second time period, 91–180 days, 28 older smokers were successful quitters at day 90, of whom 4 became recidivists during the period from day 91 to day 180; similarly, 26 younger smokers were successful quitters at 90 days, of whom 7 became recidivists from day 91 to day 180. Thus the second  $2 \times 2$  contingency table would look like Table 14.15. Similarly,  $2 \times 2$  contingency tables for the time periods 181–270 days and 271–365 days can be developed, as in Tables 14.16 and 14.17, respectively.

**Table 14.14** Incidence rates by age for the 0- to 90-day period

Age	Outcome		Total
	Recidivist	Successful quitter	
>40	92	28	120
≤40	88	26	114
Total	180	54	234

**Table 14.15** Incidence rates by age for the 91- to 180-day period

Age	Outcome		Total
	Recidivist	Successful quitter	
>40	4	24	28
≤40	7	19	26
Total	11	43	54

**Table 14.16** Incidence rates by age for the 181- to 270-day period

Age	Outcome		Total
	Recidivist	Successful quitter	
>40	4	20	24
≤40	3	16	19
Total	7	36	43

**Table 14.17** Incidence rates by age for the 271- to 365-day period

Age	Outcome		Total
	Recidivist	Successful quitter	
>40	1	19	20
≤40	2	14	16
Total	3	33	36

If age has no association with recidivism, then the incidence rate for recidivism for older and younger smokers within each of the four time intervals should be the same. Conversely, if it is harder for older smokers than younger smokers to remain quitters, then the incidence rate of recidivism should be consistently higher for older smokers within each of the four time intervals considered. Note that incidence is allowed to vary over different time intervals under either hypothesis. To accumulate evidence over the entire period of follow-up, the Mantel-Haenszel procedure in Equation 13.16, based on the  $2 \times 2$  tables in Tables 14.14 to 14.17, is used. This procedure is called the *log-rank test* and is summarized as follows.

**Equation 14.43****The Log-Rank Test**

To compare incidence rates for an event between two exposure groups, where incidence varies over the period of follow-up ( $T$ ), use the following procedure:

- (1) Subdivide  $T$  into  $k$  smaller time intervals, over which incidence is homogeneous.
- (2) Compute a  $2 \times 2$  contingency table corresponding to each time interval relating incidence over the time interval to exposure status (+/−). Consider censored subjects at a particular time as having a slightly longer follow-up time than subjects who fail at a given time. The  $i$ th table is displayed in Table 14.18,

where  $n_{i1}$  = the number of exposed people who have not yet had the event at the beginning of the  $i$ th time interval and were not censored at the beginning of the interval

$n_{i2}$  = the number of unexposed people who have not yet had the event at the beginning of the  $i$ th time interval and were not censored at the beginning of the interval

$a_i$  = the number of exposed people who had an event during the  $i$ th time interval

$b_i$  = the number of exposed people who did not have an event during the  $i$ th time interval

and  $c_i$ ,  $d_i$  are defined similarly for unexposed people.

- (3) Perform the Mantel-Haenszel test over the collection of  $2 \times 2$  tables defined in step 2. Specifically, compute the test statistic

$$X_{LR}^2 = \frac{(|O - E| - .5)^2}{Var_{LR}}$$

where

$$\begin{aligned} O &= \sum_{i=1}^k a_i \\ E &= \sum_{i=1}^k E_i = \sum_{i=1}^k \frac{(a_i + b_i)(a_i + c_i)}{n_i} \\ Var_{LR} &= \sum_{i=1}^k V_i = \sum_{i=1}^k \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)} \end{aligned}$$

which under  $H_0$  follows a chi-square distribution with one  $df$ .

- (4) For a two-sided test with significance level  $\alpha$ ,

if  $X_{LR}^2 > \chi_{1,1-\alpha}^2$ , then reject  $H_0$ .

If  $X_{LR}^2 \leq \chi_{1,1-\alpha}^2$ , then accept  $H_0$ .

- (5) The exact  $p$ -value for this test is given by

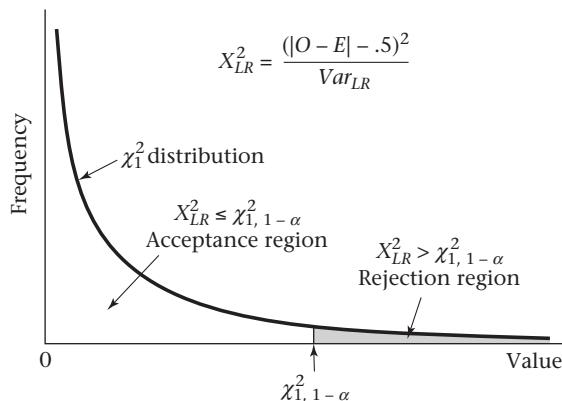
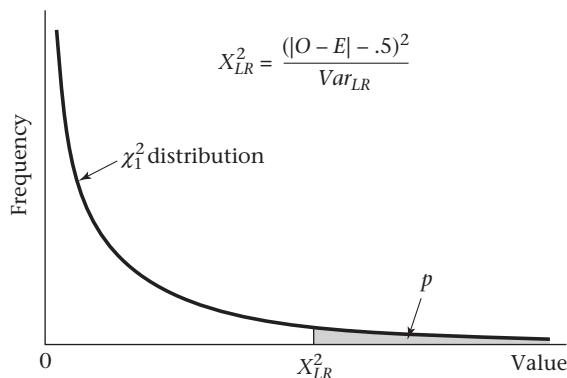
$$p\text{-value} = Pr(\chi_1^2 > X_{LR}^2)$$

- (6) This test should be used only if  $Var_{LR} \geq 5$ .

The acceptance and rejection regions for the log-rank test are shown in Figure 14.6. Computation of the exact  $p$ -value is given in Figure 14.7.

**Table 14.18 Relationship of disease incidence to exposure status over the  $i$ th time interval**

		Event		Total
Exposure	+	+	-	
		$a_i$	$b_i$	$n_{i1}$
-		$c_i$	$d_i$	$n_{i2}$
		$a_i + c_i$	$b_i + d_i$	$n_i$

**Figure 14.6** Acceptance and rejection regions for the log-rank test**Figure 14.7** Computation of the *p*-value for the log-rank test**Example 14.35**

**Health Promotion** Evaluate the statistical significance of the possible association between age and incidence of recidivism based on the smoking cessation data in Table 14.11.

**Solution**

Refer to the four  $2 \times 2$  tables (Tables 14.14–14.17) developed in Example 14.34. We have

$$O = 92 + 4 + 4 + 1 = 101$$

$$E = \frac{120 \times 180}{234} + \frac{28 \times 11}{54} + \frac{24 \times 7}{43} + \frac{20 \times 3}{36}$$

$$= 92.308 + 5.704 + 3.907 + 1.667 = 103.585$$

$$Var_{LR} = \frac{120 \times 114 \times 180 \times 54}{234^2 \times 233} + \frac{28 \times 26 \times 11 \times 43}{54^2 \times 53}$$

$$+ \frac{24 \times 19 \times 7 \times 36}{43^2 \times 42} + \frac{20 \times 16 \times 3 \times 33}{36^2 \times 35}$$

$$= 10.422 + 2.228 + 1.480 + 0.698 = 14.829$$

Because  $\text{Var}_{LR} \geq 5$ , the log-rank test can be used. The test statistic is given by

$$X_{LR}^2 = \frac{(|101 - 103.585| - .5)^2}{14.829} = \frac{2.085^2}{14.829} = 0.29 \sim \chi_1^2 \text{ under } H_0$$

Because  $\chi_{1,95}^2 = 3.84 > 0.29$ , it follows that the  $p$ -value  $> .05$ , and there is no significant difference in recidivism rates between younger and older smokers.

The data set in Table 14.11 did not have any censored data; that is, all participants were followed until either 1 year had elapsed or they resumed smoking, whichever occurred first. However, the log-rank test can also be used if censored data are present. Here, for the year  $i$  table,  $S_i$  = number of participants who survived to time  $t_i$  and  $I_i$  = number of participants who were censored at time  $t_i$  and did not fail are combined into one group, because they all survived from time  $t_{i-1}$  to time  $t_i$ .

### Example 14.36

**Ophthalmology** Compare the survival curves for participants receiving 15,000 IU of vitamin A vs. participants receiving 75 IU of vitamin A, given the data in Table 14.12.

#### Solution

We have 6 contingency tables corresponding to years 1, 2, 3, 4, 5, and 6:

#### Year 1

Vitamin A dose	Fail	Survive	Total
15,000 IU	3	169	172
75 IU	8	174	182
	11	343	354

#### Year 2

Vitamin A dose	Fail	Survive	Total
15,000 IU	6	159	165
75 IU	13	161	174
	19	320	339

#### Year 3

Vitamin A dose	Fail	Survive	Total
15,000 IU	15	144	159
75 IU	21	137	158
	36	281	317

#### Year 4

Vitamin A dose	Fail	Survive	Total
15,000 IU	21	122	143
75 IU	21	114	135
	42	236	278

**Year 5**

Vitamin A dose	Fail	Survive	Total
15,000 IU	15	81	96
75 IU	13	73	86
	28	154	182

**Year 6**

Vitamin A dose	Fail	Survive	Total
15,000 IU	5	41	46
75 IU	13	29	42
	18	70	88

We use PROC LIFETEST of SAS to perform the log-rank test. Given Equation 14.43, we have  $O = 65$ ,  $E = 78.432$ , and  $Var_{LR} = 33.656$ . The chi-square statistic using PROC LIFETEST is 5.36 with 1  $df$  and yields a  $p$ -value of .021. Note that PROC LIFETEST does not use a continuity correction. Thus the test statistic is

$$X_{LR, \text{uncorrected}}^2 = \frac{(O - E)^2}{Var_{LR}}$$

as opposed to

$$X_{LR}^2 = \frac{(|O - E| - .5)^2}{Var_{LR}}$$

given in Equation 14.43. In this example,  $X_{LR}^2 = 4.97$ , with  $p$ -value = .026. The Wilcoxon and likelihood ratio (*LR*) procedures are other approaches for comparing survival curves provided by PROC LIFETEST. These approaches are not discussed in this text because the log-rank test is more widely used. PROC LIFETEST can also provide survival probabilities by treatment group similar to Table 14.12.

Therefore, there is a significant difference between the survival curves of the two groups. Because  $O = \text{observed number of events in the } 15,000 \text{ IU group} = 65 < E = \text{expected number of events in the } 15,000 \text{ IU group} = 78.432$ , it follows that the 15,000 IU group had a better survival experience than the 75 IU group. Stated another way, there were significantly fewer failures in the 15,000 IU group than in the 75 IU group.

In this section, we have presented the log-rank test, which is a procedure for comparing survival curves from two independent samples. It is similar to the Mantel-Haenszel test and can be used to compare survival curves with and without censored data. It allows one to compare the entire survival curve, which provides more power than focusing on survival at specific points in time.

On the flowchart (Figure 14.15, p. 803), we answer yes to (1) person-time data? and no to (2) one-sample problem? and (3) incidence rates remain constant over time? This path leads to the box labeled “Use survival-analysis methods.” We then answer yes to (4) interested in comparison of survival curves of two groups with limited control of covariates? This leads to the box labeled “Use log-rank test.”

**REVIEW QUESTIONS 14G**

- 1 (a)** What is the log-rank test?
- (b)** How does it differ from comparing survival curves between two groups at specific points in time?
- 2** Refer to the data in Review Question 14F.3.
  - (a)** Use the log-rank test to compare the survival experience of the exemestane group vs. the tamoxifen group. Report a two-tailed  $p$ -value.
  - (b)** What is your overall interpretation of the results?

**14.11 The Proportional-Hazards Model**

The log-rank test is a very powerful method for analyzing data when the time to an event is important rather than simply whether or not the event occurs. The test can be used if variable periods of follow-up are available for each individual and/or if some data are censored. It can also be extended to allow one to look at the relationship between survival and a single primary exposure variable, while controlling for the effects of one or more other covariate(s). This can be accomplished by stratifying the data according to the levels of the other covariates; computing the observed number, expected number, and variance of the number of failures in each stratum; summing the respective values over all strata; and using the same test statistic as in Equation 14.43. However, if there are many strata and/or if there are several risk factors of interest, a more convenient approach is to use a method of regression analysis for survival data.

Many different models can be used to relate survival to a collection of other risk factors. One of the most frequently used models, first proposed by D. R. Cox [6], is called a *proportional-hazards model*.

**Equation 14.44****Proportional-Hazards Model**

Under a **proportional-hazards model**, the hazard  $h(t)$  is modeled as

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_k x_k)$$

where  $x_1, \dots, x_k$  are a collection of independent variables, and  $h_0(t)$  is the baseline hazard at time  $t$ , representing the hazard for a person with the value 0 for all the independent variables. The hypothesis  $H_0: \beta_j = 0$  vs.  $H_1: \beta_j \neq 0$  can be tested as follows:

- (1) Compute the test statistic  $z = \hat{\beta}_j / se(\hat{\beta}_j)$ .
- (2) To conduct a two-sided level  $\alpha$  significance test,
 

if  $z < z_{\alpha/2}$  or  $z > z_{1-\alpha/2}$ , then reject  $H_0$ ;  
 if  $z_{\alpha/2} \leq z \leq z_{1-\alpha/2}$ , then accept  $H_0$ .
- (3) The exact  $p$ -value is given by

$$\begin{aligned} 2 \times [1 - \Phi(z)] &\quad \text{if } z \geq 0 \\ 2 \times \Phi(z) &\quad \text{if } z < 0 \end{aligned}$$

By dividing both sides of Equation 14.44 by  $h_0(t)$  and taking logarithms, a proportional-hazards model can be written in the form

$$\ln \left[ \frac{h(t)}{h_0(t)} \right] = \beta_1 x_1 + \cdots + \beta_k x_k$$

This representation lets us interpret the coefficients of a proportional-hazards model in a similar manner to that of a multiple logistic-regression model. In particular, if  $x_j$  is a dichotomous independent variable, then the following principle applies.

#### Equation 14.45

##### Estimation of Hazard Ratio for Proportional-Hazards Models for Dichotomous Independent Variables

Suppose we have a dichotomous independent variable ( $x_j$ ) that is coded as 1 if present and 0 if absent. For the proportional-hazards model in Equation 14.44 the quantity  $\exp(\beta_j)$  represents the ratio of hazards for two people, one with the risk factor present and the other with the risk factor absent, given that both people have the same values for all other covariates. The hazard ratio or relative hazard can be interpreted as the *instantaneous* relative risk of an event per unit time for a person with the risk factor present compared with a person with the risk factor absent, given that both individuals have survived to time  $t$  and are the same on all other covariates.

A two-sided  $100\% \times (1 - \alpha)$  CI for  $\beta_j$  is given by  $(e^{c_1}, e^{c_2})$ , where

$$c_1 = \hat{\beta}_j - z_{1-\alpha/2} se(\hat{\beta}_j)$$

$$c_2 = \hat{\beta}_j + z_{1-\alpha/2} se(\hat{\beta}_j)$$

Similarly, if  $x_j$  is a continuous independent variable, then the following interpretation of the regression coefficient  $\beta_j$  is used.

#### Equation 14.46

##### Estimation of Hazard Ratio for Proportional-Hazards Models for Continuous Independent Variables

Suppose there is a continuous independent variable ( $x_j$ ). Consider two people who differ by the quantity  $\Delta$  on the  $j$ th independent variable and are the same for all other independent variables. The quantity  $\exp(\beta_j \Delta)$  represents the ratio of hazards between the two individuals. The hazard ratio can also be interpreted as the *instantaneous* relative risk of an event per unit time for an individual with risk-factor level  $x_j + \Delta$  compared with someone with risk-factor level  $x_j$ , given that both people have survived to time  $t$  and are the same for all other covariates.

A two-sided  $100\% \times (1 - \alpha)$  CI for  $\beta_j \Delta$  is given by  $(e^{c_1}, e^{c_2})$  where

$$c_1 = \Delta[\hat{\beta}_j - z_{1-\alpha/2} se(\hat{\beta}_j)]$$

$$c_2 = \Delta[\hat{\beta}_j + z_{1-\alpha/2} se(\hat{\beta}_j)]$$

Note that the hazard for a subject ( $h(t)$ ) can vary over time, but the ratio of hazards between 2 subjects, one of whom has covariate values  $(x_1, \dots, x_k)$  and the other of

whom has covariate values of 0 for all covariates is given by  $\exp(\sum_{j=1}^k \beta_j x_j)$ , which is

the same for all  $t$ . The Cox proportional-hazards model can also be thought of as an extension of multiple logistic regression where the time when an event occurs is taken into account, rather than simply whether an event occurs.

**Example 14.37**

**Health Promotion** Fit a proportional-hazards model to the smoking-cessation data in Example 14.26 using the risk factors sex and adjusted  $\log_{10}$ (CO concentration), which is an index of inhalation of smoking prior to quitting. Assess the statistical significance of the results, and interpret the regression coefficients.

**Solution**

The SAS PHREG (Proportional Hazards Regression model) procedure has been used to fit the Cox model to the smoking-cessation data. For ease of interpretation, sex was recoded as (1 = male, 0 = female) from the original coding of (1 = male, 2 = female). For this example, the actual time of starting smoking was used as the “survival time” based on the raw data in Data Set SMOKE.DAT, on the Companion Website, rather than on the grouped data given in Table 14.11. The results are given in Table 14.19.

**Table 14.19**

**Proportional-hazards model fitted to the smoking-cessation data in SMOKE.DAT**

Risk factor	Regression coefficient ( $\hat{\beta}_j$ )	Standard error $se(\hat{\beta}_j)$	$z$ [ $\hat{\beta}_j/se(\hat{\beta}_j)$ ]
$\log_{10}$ CO (adjusted) <sup>a</sup>	0.833	0.350	2.380
Sex (1 = M, 0 = F)	-0.117	0.135	-0.867

<sup>a</sup>This variable represents CO values adjusted for minutes elapsed since last cigarette smoked prior to quitting.

To assess the significance of each regression coefficient, compute the test statistic given in Equation 14.44 as follows:

$$z(\log_{10}\text{CO}) = 0.833/0.350 = 2.380$$

$$p(\log_{10}\text{CO}) = 2 \times [1 - \Phi(2.380)] = 2 \times (1 - .9913) = .017$$

$$z(\text{sex}) = -0.117/0.135 = -0.867$$

$$p(\text{sex}) = 2 \times \Phi(-0.867) = 2 \times [1 - \Phi(0.867)] = 2 \times (1 - .8069) = .386$$

Thus there is a significant effect of CO concentration on the hazard or risk of recidivism (i.e., propensity to start smoking again), with the higher the CO concentration, the higher the hazard (risk of recidivism). Based on these data, there is no significant effect of sex on the risk of recidivism.

The effect of CO can be quantified in terms of relative risk. Specifically, if two people of the same sex who differ by one unit on adjusted  $\log_{10}$ CO are considered (i.e., who differ by 10-fold in CO concentration), then the instantaneous relative risk of recidivism for a person with adjusted  $\log_{10}\text{CO} = x_j + 1$  (person A) compared with a person with adjusted  $\log_{10}\text{CO} = x_j$  (person B) is given by

$$RR = \exp(0.833) = 2.30$$

Thus, given that person A and person B have not started smoking up to time  $t$ , person A is 2.3 times as likely to start smoking over a short period of time as person B.

The Cox proportional-hazards model can also be used with censored data.

**Example 14.38**

**Ophthalmology** Use the Cox proportional-hazards model to compare the survival curves for subjects receiving a high dose (15,000 IU) vs. a low dose (75 IU) of vitamin A, based on the data in Table 14.12.

**Solution**

We have used the SAS program PROC PHREG to compare the survival curves. In this case, there is only a single binary covariate  $x$  defined by

$$x = \begin{cases} 1 & \text{if high dose A} \\ 0 & \text{if low dose A} \end{cases}$$

The output from the program is given in Table 14.20. We see that subjects on 15,000 IU of vitamin A (denoted by high\_a) have a significantly lower hazard than subjects on 75 IU of vitamin A ( $p = .031$ , denoted by  $Pr > \text{ChiSq}$ ). The hazard ratio is estimated by  $e^{\hat{\beta}} = e^{-0.35173} = 0.703$ . Thus the failure rate at any point in time is approximately 30% lower for patients on 15,000 IU of vitamin A than for patients on 75 IU of vitamin A. We can obtain 95% confidence limits for the hazard ratio by  $(e^{c_1}, e^{c_2})$ , where

$$c_1 = \hat{\beta} - 1.96\text{se}(\hat{\beta}) = -0.352 - 1.96(0.163) = -0.672$$

$$c_2 = \hat{\beta} + 1.96\text{se}(\hat{\beta}) = -0.352 + 1.96(0.163) = -0.032$$

Thus the 95% CI =  $(e^{-0.672}, e^{-0.032}) = (0.51, 0.97)$ . The estimated survival curve(s) by year (lenfl30) are given at the bottom of Table 14.20 separately for the high-dose A group (high\_a = 1) (rows 1–7) and the low-dose A group (high\_a = 0) (rows 8–14), and are plotted in Figure 14.8. It is estimated that by year 6, 47% of subjects in the high-dose group (1 – .53) and 60% of subjects in the low-dose group (1 – .40) will have failed; that is, their ERG amplitude will decline by at least 50%.

If there are no ties—that is, if all subjects have a unique failure time—then the Cox proportional-hazards model with a single binary covariate and the log-rank test provide identical results. There are several different methods for handling ties with the Cox proportional-hazards model; in general, in the presence of ties, the Cox proportional-hazards model and the log-rank test do not yield identical  $p$ -values, particularly in data sets with many tied observations, as in Table 14.20. Similarly, if there are many ties then the survival curve estimated using the proportional-hazards model will not be exactly the same as that obtained from the Kaplan-Meier product-limit method.

**Table 14.20** Cox proportional-hazards model run on the RP data set in Table 14.12

The PHREG Procedure			
Model Information			
<b>Data Set:</b>			WORK.TIMES2
<b>Dependent Variable:</b>			lenfl30
<b>Censoring Variable:</b>			fail30
<b>Censoring Value(s):</b>			0
<b>Ties Handling:</b>			BRESLOW
<b>Number of Observations Read</b>			354
<b>Number of Observations Used</b>			354
Summary of the Number of Event and Censored Values			
Percent			
Total	Event	Censored	Censored
354	154	200	56.50

(continued)

**Table 14.20** Cox proportional-hazards model run on the RP data set in Table 14.12 (Continued)

Convergence Status						
Convergence criterion (GCONV=1E-8) satisfied						
Model Fit Statistics						
Criterion	Without Covariates		With Covariates			
-2 LOG L	1690.482		1685.777			
AIC	1690.482		1687.777			
SBC	1690.482		1690.814			
Testing Global Null Hypothesis: BETA=0						
Test	Chi-Square		DF	Pr > ChiSq		
Likelihood Ratio	4.7053		1	0.0301		
Score	4.6915		1	0.0303		
Wald	4.6436		1	0.0312		
Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
high_a	1	-0.35173	0.16322	4.6436	0.0312	0.703
The SAS System						
obs	high_a	lenfl30		s		
1	1	0		1.00000		
2	1	1		0.97436		
3	1	2		0.92916		
4	1	3		0.84090		
5	1	4		0.73395		
6	1	5		0.63853		
7	1	6		0.52667		
8	0	0		1.00000		
9	0	1		0.96375		
10	0	2		0.90082		
11	0	3		0.78167		
12	0	4		0.64423		
13	0	5		0.52851		
14	0	6		0.40194		

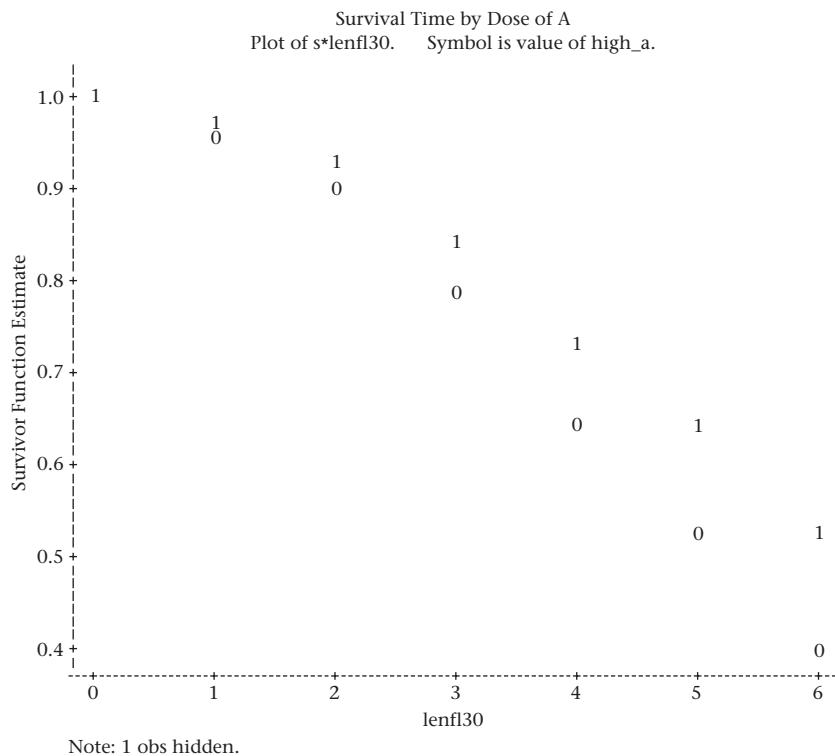
**Example 14.39**

**Ophthalmology** The Cox proportional-hazards model can also be used to control for the effects of other covariates as well as for the other treatment (denoted by high\_e) which is defined by

$$\text{high\_e} = \begin{cases} 1 & \text{if patient received 400 IU of vitamin E daily} \\ 0 & \text{if patient received 3 IU of vitamin E daily} \end{cases}$$

The other covariates considered were

**Figure 14.8** Survival curve for patients receiving 15,000 IU of vitamin A (high\_A = 1) and 75 IU of vitamin A (high\_A = 0)



agebas = age at the baseline visit – 30 (in years)

$$\text{sex} = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

er30oucn =  $\ln(\text{ERG } 30 \text{ Hz amplitude at baseline}) - 0.215$

blretlcn = baseline serum retinol – 50.0 (vitamin A) ( $\mu\text{g/dL}$ )

blvitecn = baseline serum alpha-tocopherol – 0.92 (vitamin E) ( $\text{mg/dL}$ )

dretinncn = dietary intake of retinol at baseline – 3624 (vitamin A) (IU)

dvtmnecn = dietary intake of alpha-tocopherol at baseline – 11.89 (vitamin E) (IU)

The mean value was subtracted from each covariate so as to minimize the time to convergence of the iterative algorithm used to fit Cox regression methods. The results are given in Table 14.21.

We see there are significant effects of both treatments, but in opposite directions. The estimated hazard ratio for subjects given 15,000 IU of vitamin A vs. subjects given 75 IU of vitamin A was 0.70 ( $p = .032$ ), whereas the estimated hazard ratio for subjects given 400 IU of vitamin E vs. subjects given 3 IU of vitamin E was 1.45 ( $p = .024$ ). Thus vitamin A has a significant protective effect and vitamin E has a significant harmful effect even after controlling for other baseline risk factors. Subjects given high-dose vitamin A were about a third less likely to fail than subjects given low-dose vitamin A, whereas subjects given high-dose vitamin E were about 50%

**Table 14.21 Effects of treatments administered in RP clinical trial, while controlling for the effects of other baseline covariates using the SAS PHREG procedure**

The PHREG Procedure											
Model Information											
Data Set:		WORK.TIMES2									
Dependent Variable:		lenfl30									
Censoring Variable:		fail130									
Censoring Value(s):		0									
Ties Handling:		BRESLOW									
Number of Observations Read		354									
Number of Observations Used		354									
Summary of the Number of Event and Censored Values											
		Percent									
Total	Event	Censored		Censored							
354	153	200		56.50							
Convergence Status											
Convergence criterion (GCONV=1E-8) satisfied											
Model Fit Statistics											
		Without Covariates		With Covariates							
Criterion											
-2 LOG L		1690.482		1672.718							
AIC		1690.482		1690.718							
SBC		1690.482		1718.051							
Testing Global Null Hypothesis: BETA=0											
Test		Chi-Square	DF	Pr > Chi-Sq							
Likelihood Ratio		17.7640	9	0.0380							
Score		17.1355	9	0.0466							
Wald		16.9629	9	0.0493							
Analysis of Maximum Likelihood Estimates											
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Sq	Hazard Ratio					
agebas	1	-0.00560	0.01108	0.2552	0.6135	0.994					
sex	1	-0.03554	0.17413	0.0417	0.8363	0.965					
er30oucn	1	-0.09245	0.07333	1.5895	0.2074	0.912					
blretlcn	1	-0.01599	0.00953	2.8178	0.0932	0.984					
blvitecn	1	0.44933	0.43161	1.0838	0.2978	1.567					
dretinccn	1	-0.0000443	0.0000479	0.8562	0.3548	1.000					
dvtmneecn	1	-0.01103	0.01404	0.6180	0.4318	0.989					
high_a	1	-0.35307	0.16473	4.5939	0.0321	0.703					
high_e	1	0.37258	0.16534	5.0781	0.0242	1.451					

more likely to fail than subjects given low-dose vitamin E. None of the other baseline covariates were statistically significant, although baseline serum retinol was the closest to being significant ( $p = .09$ ) with subjects with a high serum retinol being less likely to fail.

## Testing the Assumptions of the Cox Proportional-Hazards Model

In this section, we have been introduced to the Cox proportional-hazards model. This technique is analogous to multiple logistic regression and lets us estimate the hazard ratio of a primary exposure variable while controlling for the effects of other covariates. An important assumption of this method is that the hazard ratio of the primary exposure (and also of any other covariates in the model) remains constant over time. This assumption can be tested by introducing a cross-product term of a specific variable of interest ( $x$ ) by time ( $t$ ) into the model and testing for statistical significance, and is considered below.

### Example 14.40

**Cardiovascular Disease** The Physicians' Health Study was a randomized trial among 22,071 male U.S. physicians that simultaneously tested the efficacy of 325 mg of aspirin (ASA, or acetylsalicylic acid) taken every other day to prevent coronary heart disease and 50 mg of beta-carotene taken every day to prevent cancer using a  $2 \times 2$  factorial design, where subjects were randomized to four groups (1 = ASA active and beta-carotene placebo; 2 = ASA placebo and beta-carotene active; 3 = ASA active and beta-carotene active; 4 = ASA placebo and beta-carotene placebo). In addition, risk-factor information was collected at baseline to be used for subsequent risk prediction. The following variables were used for risk prediction of myocardial infarction (MI), a type of coronary heart disease: age, body-mass index (BMI), current cigarette smoking, history of high blood pressure, history of high cholesterol, and history of diabetes. These variables were defined based on the baseline (1982) questionnaire.

A total of 18,662 participants in the study with complete covariate information were followed from 1982 through 2001, although the randomized aspirin component of the study ended in 1990 and the randomized beta-carotene component ended in 1995. To be eligible for the study, participants could not have had a history of MI at baseline. However, 1215 participants developed an MI over the 20-year follow-up period. The issue is whether the baseline covariates have similar predictive power over the entire 20-year follow-up period.

### Solution

The fitted model is given in the first column of Table 14.22. We see there are significant effects of BMI, history of high blood pressure (hypertension), history of cholesterol, history of diabetes, and current cigarette smoking ( $p < .001$  for all covariates).

The model assumes the hazard ratio for the baseline covariates remains the same over the 20-year follow-up period. Because some variables may change over time, it would be prudent to test this assumption. For this purpose, a second model was fit, including terms of the form  $BMI \times [\ln(\text{time}) - 2.8]$ , ..., current cigarette smoking  $\times [\ln(\text{time}) - 2.8]$ , in addition to the variables in the first column of Table 14.22, where time = follow-up time (yrs) from 1982 to time  $t$ . The Cox proportional-hazards model allows for "time-dependent covariates," which means the values of the variables are allowed to change over time. Thus a variable such as  $BMI \times [\ln(\text{time}) - 2.8]$  is an example of a time-dependent covariate because its value changes over the follow-up period. The results from fitting this model are given in the second column of Table 14.22.

**Table 14.22 Hazard ratio of MI among 18,662 participants in the Physicians' Health Study; 1,215 developed MI over a median follow-up of 20.1 years**

Variable	Model without time interactions			Model with time interactions		
	Hazard ratio	(95% CI)	p-value	Hazard ratio	(95% CI)	p-value
Age, per year	1.05	(1.04–1.06)	<.001	1.05 <sup>a</sup>	(1.04–1.05)	<.001
BMI, per 1 kg/m <sup>2</sup>	1.06	(1.04–1.08)	<.001	1.06 <sup>a</sup>	(1.03–1.09)	<.001
Hypertension	1.55	(1.37–1.75)	<.001	1.43 <sup>a</sup>	(1.22–1.68)	<.001
High cholesterol	1.39	(1.20–1.61)	<.001	1.19 <sup>a</sup>	(0.98–1.46)	.079
Diabetes mellitus	2.00	(1.57–2.55)	<.001	1.51 <sup>a</sup>	(1.05–2.15)	.024
Current smoking	1.72	(1.47–2.01)	<.001	1.74 <sup>a</sup>	(1.42–2.14)	<.001
Age × ln(time) <sup>b</sup>	—			0.996 <sup>c</sup>	(0.99–1.00)	.16
BMI × ln(time) <sup>b</sup>	—			0.999 <sup>c</sup>	(0.98–1.02)	.89
Hptn <sup>d</sup> × ln(time) <sup>b</sup>	—			0.92 <sup>c</sup>	(0.82–1.04)	.17
Chol <sup>d</sup> × ln(time) <sup>b</sup>	—			0.85 <sup>c</sup>	(0.74–0.97)	.015
DM <sup>d</sup> × ln(time) <sup>b</sup>	—			0.80 <sup>c</sup>	(0.66–0.97)	.022
Smoking × ln(time) <sup>b</sup>	—			1.02 <sup>c</sup>	(0.87–1.19)	.84

<sup>a</sup>Relative risk for this variable at the geometric mean follow-up time (16.4 years).<sup>b</sup>In these interactions, ln(time) is centered at its mean value of 2.8 ln(years).<sup>c</sup>Change in effects of this variable per unit increase in ln(time).<sup>d</sup>Hptn = hypertension; Chol = high cholesterol; DM = diabetes mellitus.

We see there are several violations of the proportional-hazards assumption. Specifically, the terms high cholesterol × [ln(time) – 2.8] ( $p = .015$ ) and history of diabetes × [ln(time) – 2.8] ( $p = .022$ ) are statistically significant. The estimated relative risks for each of these variables are less than 1, which implies that the effects of high cholesterol and diabetes are much stronger at the beginning of the study and weaker toward the end of the study, possibly because the risk-factor status of the men changed over time or because the effects of these variables are weaker for older men.

If a violation of the proportional-hazards assumption is found, then the prudent approach is to present separate analyses for different periods of time (e.g., first 10 years, last 10 years), and/or to consider updating the risk factors, if possible, so that a man's risk-factor status could be allowed to change over time. For this purpose, a Cox proportional-hazards model using baseline covariates was fit separately for follow-up time during the first 10 years of the study (1982–1991) and during the last 10 years of the study (1992–2001). The results are shown in Table 14.23. We see that the effects of high cholesterol and diabetes are much stronger in the first 10 years than

**Table 14.23 Hazard ratio of MI in the Physicians' Health Study according to baseline risk factors stratified by time; 629 MIs in the first 10 years of follow-up and 586 MIs after 10 years**

Variable	629 events in the first 10 years			586 events after 10 years		
	Hazard ratio	(95% CI)	p-value	Hazard ratio	(95% CI)	p-value
Age, per year	1.05	(1.05–1.06)	<.001	1.04	(1.04–1.05)	<.001
BMI, per 1 kg/m <sup>2</sup>	1.06	(1.03–1.08)	<.001	1.06	(1.04–1.09)	<.001
Hypertension	1.69	(1.43–2.00)	<.001	1.39	(1.16–1.66)	<.001
High cholesterol	1.56	(1.29–1.90)	<.001	1.21	(0.97–1.52)	.093
Diabetes mellitus	2.17	(1.62–2.93)	<.001	1.67	(1.09–2.56)	.020
Current smoking	1.69	(1.36–2.09)	<.001	1.75	(1.39–2.20)	<.001

in the last 10 years of the study. Hypertension showed a similar trend in Table 14.23, which is consistent with the nonsignificant inverse *RR* of hypertension  $\times [\ln(\text{time}) - 2.8]$  in Table 14.22. Age, BMI, and current smoking seem to behave similarly in the two time periods.

On the flowchart (Figure 14.15, p. 803), we answer yes to (1) person-time data? and no to each of (2) one-sample problem? and (3) incidence rates remain constant over time? which leads to the box labeled “Use survival-analysis methods.” We then answer no to (4) interested in comparison of survival curves of two groups with limited control of covariates? which leads to (5) interested in effects of several risk factors on survival. We then answer no to (6) willing to assume survival curve comes from a Weibull distribution. This leads us to the box labeled “Use Cox proportional-hazards model.”

In the next section, we consider methods of power and sample-size estimation for proportional-hazards models.

### REVIEW QUESTIONS 14H

- 1** What is the difference between the Cox proportional-hazards model and a multiple-logistic-regression model? When do we use each?
- 2** When do we use the Cox proportional-hazards model, and when do we use the log-rank test?
- 3** **(a)** What does the term *proportional hazards* mean?  
**(b)** How can we check whether the proportional-hazards assumption is correct?
- 4** Suppose we are studying the effect of current smoking on the incidence of lung cancer. We fit a Cox proportional-hazards model with age, sex, and current smoking as covariates. Suppose the regression coefficient for current smoking = 2.5 with *se* = 1.0. What does the regression coefficient of 2.5 mean? (*Hint:* Interpret the results in terms of an estimated-hazard ratio for current smoking and an associated 95% CI.)

## 14.12 Power and Sample-Size Estimation under the Proportional-Hazards Model

### Estimation of Power

#### Example 14.41

**Ophthalmology** Suppose the investigators consider repeating the study described in Example 14.30 to be sure the protective effect of vitamin A was not a random occurrence. The study design of the new study would have only two vitamin A treatment groups, 15,000 IU per day and 75 IU per day. The investigators feel they can recruit 200 patients in each group who were not involved in the previous study. As in the previous study, the participants would be enrolled over a 2-year period and followed for a maximum of 6 years. How much power would the study have to detect an *RR* of 0.7, where the endpoint is a 50% decline in ERG 30 Hz amplitude comparing the 15,000 IU per day group with the 75 IU per day group?

Several methods have been proposed for estimation of power and sample size for clinical trials based on survival curves that satisfy the proportional-hazards assumption. We present the method of Freedman [7] because it is relatively easy to implement and has fared relatively well in comparative simulation studies [8]. The method is as follows.

**Equation 14.47****Estimation of Power for the Comparison of Survival Curves Between Two Groups under the Cox Proportional-Hazards Model**

Suppose we want to compare the survival curves between an experimental group ( $E$ ) and a control group ( $C$ ) in a clinical trial with  $n_1$  participants in the  $E$  group and  $n_2$  participants in the  $C$  group, with a maximum follow-up of  $t$  years. We wish to test the hypothesis  $H_0$ :  $IRR = 1$  vs.  $H_1$ :  $IRR \neq 1$ , where  $IRR$  = underlying hazard ratio for the  $E$  group vs. the  $C$  group. We postulate a hazard ratio of  $IRR$  under  $H_1$  and will conduct a two-sided test with significance level  $\alpha$ . If the ratio of participants in group 1 compared with group 2 =  $n_1/n_2 = k$ , then the power of the test is

$$\text{Power} = \Phi\left(\frac{\sqrt{km}|IRR - 1|}{kIRR + 1} - Z_{1-\alpha/2}\right)$$

where

$m$  = expected total number of events over both groups

$$= n_1 p_E + n_2 p_C$$

$n_1, n_2$  = number of participants in groups 1 and 2 (i.e., the  $E$  and  $C$  groups)

$p_C$  = probability of failure in group  $C$  over the maximum time period of the study ( $t$  years)

$p_E$  = probability of failure in group  $E$  over the maximum time period of the study ( $t$  years)

To calculate  $p_C$  and  $p_E$ , we let

- (1)  $\lambda_i = Pr(\text{failure at time } i \text{ among participants in the } C \text{ group, given that a participant has survived to time } i - 1 \text{ and is not censored at time } i - 1) = \text{approximate hazard at time } i \text{ in the } C \text{ group, } i = 1, \dots, t$
- (2)  $IRR\lambda_i = Pr(\text{failure at time } i \text{ among participants in the } E \text{ group, given that a participant has survived to time } i - 1 \text{ and is not censored at time } i - 1) = \text{approximate hazard at time } i \text{ in the } E \text{ group, } i = 1, \dots, t$
- (3)  $\delta_i = Pr(\text{a participant is censored at time } i \text{ given that he was followed up to time } i \text{ and has not failed}), i = 0, \dots, t$ , which is assumed the same in each group

It follows from (1), (2), and (3) that

$$p_C = \sum_{i=1}^t \lambda_i A_i C_i = \sum_{i=1}^t D_i$$

$$p_E = \sum_{i=1}^t (IRR\lambda_i) B_i C_i = \sum_{i=1}^t E_i$$

where

$$A_i = \prod_{j=1}^{i-1} (1 - \lambda_j)$$

$$B_i = \prod_{j=0}^{i-1} (1 - IRR\lambda_j)$$

$$C_i = \prod_{k=0}^{i-1} (1 - \delta_k)$$

Note that the power formula in Equation 14.47 depends on the total number of events over both groups ( $m$ ), as well as the rate ratio ( $IRR$ ). Hence, if analysis of the data is based on the Cox proportional-hazards model

$$\lambda(t) = \lambda_0(t) \exp(\beta x)$$

where  $x = 1$  if a subject is in group  $E$  and  $= 0$  if a subject is in group  $C$ , then the power depends on  $m$ , which is a function of both the baseline hazard function  $\lambda_0(t)$  (in this case, the hazard function for the control group), as well as the hazard function for the experimental group  $= \lambda_0(t) \exp(\beta) = \lambda_0(t) IRR$ . The power formula also assumes that the central-limit theorem is valid and hence is appropriate for large samples.

**Example 14.42**

**Ophthalmology** Compute the power for the study proposed in Example 14.41.

**Solution**

We have  $IRR = 0.7$ ,  $\alpha = .05$ ,  $z_{1-\alpha/2} = z_{.975} = 1.96$ ,  $k = 1$ ,  $n_1 = n_2 = 200$ ,  $t = 6$ . To compute  $p_C$  and  $p_E$ , we must obtain  $\lambda_i$ ,  $IRR\lambda_i$ , and  $\delta_i$ . We use the data in Table 14.12.

In this example, the 75 IU per day group is group  $C$  and the 15,000 IU per day group is group  $E$ . Also, no participants are censored at year 0 (i.e., all participants were followed for at least 1 year). We have  $\lambda_1 = 8/182 = 0.0440$ ,  $\lambda_2 = 13/174 = 0.0747$ ,  $\dots$ ,  $\lambda_6 = 13/42 = 0.3095$ . Also,  $\delta_0 = 0$ ,  $\delta_1 = 0$ ,  $\delta_2 = 3/161 = .0186$ ,  $\dots$ ,  $\delta_5 = 31/73 = .4247$ ,  $\delta_6 = 29/29 = 1.0$ . The computations are shown in Table 14.24.

**Table 14.24**

**Calculation of  $p_C$  and  $p_E$  for Example 14.42**

$i$	$\lambda_i$	$IRR\lambda_i$	$\delta_i$	$A_i$	$B_i$	$C_i$	$D_i$	$E_i$
0	0.0	0.0	0.0	—	—	—	—	—
1	0.0440	0.0308	0.0	1.0	1.0	1.0	0.0440	0.0308
2	0.0747	0.0523	0.0186	0.9560	0.9692	1.0	0.0714	0.0507
3	0.1329	0.0930	0.0146	0.8846	0.9185	0.9814	0.1154	0.0839
4	0.1556	0.1089	0.2456	0.7670	0.8331	0.9670	0.1154	0.0877
5	0.1512	0.1058	0.4247	0.6477	0.7424	0.7295	0.0714	0.0573
6	0.3095	0.2167	1.0	0.5498	0.6638	0.4197	0.0714	0.0604
Total							0.4890	0.3707

Thus  $p_C = .4890$ ,  $p_E = .3707$ , and  $m = 200(.4890 + .3707) = 171.9$ . Finally,

$$\begin{aligned}\text{Power} &= \Phi\left(\frac{\sqrt{171.9}|0.7 - 1|}{0.7 + 1} - 1.96\right) \\ &= \Phi\left[\frac{13.11(0.3)}{1.7} - 1.96\right] \\ &= \Phi(2.314 - 1.96) \\ &= \Phi(0.354) = .638\end{aligned}$$

Thus the study would have about 64% power.

## Estimation of Sample Size

Similarly, we can ask the following question: How many participants are needed in each group to achieve a specified power of  $1 - \beta$ ? The sample size can be obtained by solving for  $m$  and as a result  $n_1$ ,  $n_2$  based on the power formula in Equation 14.47. The result is as follows.

### Equation 14.48

#### Sample-Size Estimation for the Comparison of Survival Curves Between Two Groups under the Cox Proportional-Hazards Model

Suppose we wish to compare the survival curves between an experimental group (group  $E$ ) and a control group (group  $C$ ) in a clinical trial in which the ratio of participants in group  $E$  ( $n_1$ ) to group  $C$  ( $n_2$ ) is given by  $k$  and the maximum length of follow-up =  $t$ . We postulate a hazard ratio of  $IRR$  for group  $E$  compared with group  $C$  and wish to conduct a two-sided test with significance level  $\alpha$ . The number of participants needed in each group to achieve a power of  $1 - \beta$  is

$$n_1 = \frac{mk}{kp_E + p_C}, \quad n_2 = \frac{m}{kp_E + p_C}$$

where

$$m = \frac{1}{k} \left( \frac{kIRR+1}{IRR-1} \right)^2 (z_{1-\alpha/2} + z_{1-\beta})^2$$

and  $p_E$ ,  $p_C$  are the probabilities of failure over time  $t$  in groups  $E$  and  $C$ , respectively, given in Equation 14.47.

### Example 14.43

**Ophthalmology** Estimate the required number of participants needed in each group to achieve 80% power for the study proposed in Example 14.41.

### Solution

From Example 14.42 we have  $p_E = .3707$ ,  $p_C = .4890$ , and  $k = 1$ . Also, from Equation 14.48 we have

$$\begin{aligned} m &= \left( \frac{1.7}{0.3} \right)^2 (z_{.975} + z_{.80})^2 \\ &= 32.11(1.96 + 0.84)^2 \\ &= 32.11(7.84) = 251.8 \text{ events over both groups combined} \end{aligned}$$

Thus

$$\begin{aligned} n_1 = n_2 &= \frac{251.8}{.3707 + .4890} \\ &= \frac{251.8}{.8597} = 293 \text{ participants per group} \end{aligned}$$

Therefore, we need to recruit 293 participants in each group, or 586 participants in total, to achieve 80% power.

It may seem counterintuitive that we were able to achieve statistical significance based on the original study of 354 participants in total over both groups, and yet the sample-size requirement for the new study is about 50% larger. The reason is

that the results from the original study were only borderline significant ( $p = .03$ ). If the  $p$ -value was exactly .05, and we used the same effect size in the proposed new study, then we would only achieve 50% power. Our power was slightly larger (64%) because the  $p$ -value was somewhat smaller than .05. To achieve 80% power, we need a larger sample size to allow for random fluctuations around the true effect size in finite-sample clinical trials.

The methods of power and sample-size estimation given in this section assume a proportional-hazards model relating the survival curves between the  $E$  and  $C$  groups. If the proportional-hazards assumption is not satisfied, then more complicated methods of power and sample-size estimation are needed. Other approaches for sample-size and power estimation are given in [8] and [9].

**Example 14.44**

**Cancer** Apply the method of sample-size estimation in Equation 14.48 to the study proposed in Example 14.14. Compare the results with those obtained in Example 14.16 using Equation 14.22.

**Solution**

We have  $IRR = 1.25$ ,  $\alpha = .05$ ,  $\beta = .20$ , and  $k = 1$ . Thus

$$\begin{aligned} m &= \frac{(1.25+1)^2}{(1.25-1)^2} (1.96 + 0.84)^2 \\ &= 635.04 \end{aligned}$$

This is very similar to the required total number of events in Example 14.16 (633). To find the corresponding sample-size estimate in each group, we use the Kaplan-Meier estimator to estimate  $p_C$  and  $p_E$  as follows:

$$\begin{aligned} p_C &= 1 - \left[ 1 - (300 \times 10^{-5}) \right]^5 = 1 - .98509 = .01491 \\ p_E &= 1 - \left[ 1 - (375 \times 10^{-5}) \right]^5 = 1 - .98139 = .01861 \end{aligned}$$

Thus

$$n_1 = n_2 = \frac{635.04}{.01861 + .01491} = 18,945 \text{ participants per group}$$

or 37,890 participants in total, to have 80% power.

This is also very similar to the total sample size in Example 14.16 (37,834 participants). Thus, although somewhat different approaches were used to derive these sample-size formulas, the results in this example are very similar, which gives confidence in the validity of each approach.

## 14.13 Parametric Survival Analysis

**Example 14.45**

**Cancer** A clinical trial to evaluate the efficacy of maintenance chemotherapy for leukemia patients was conducted [3].

After reaching remission through chemotherapy treatment, subjects were randomized to 2 groups:

- (a) a maintenance chemotherapy group
- (b) control group

The primary goal of the study was to compare the survival experience of the 2 treatment groups with survival defined as maintenance of remission. The preliminary data as of 10/74 are given in Table 14.25. How should the data be analyzed?

**Table 14.25** Length of complete remission (weeks) for leukemia treatment data<sup>a</sup>

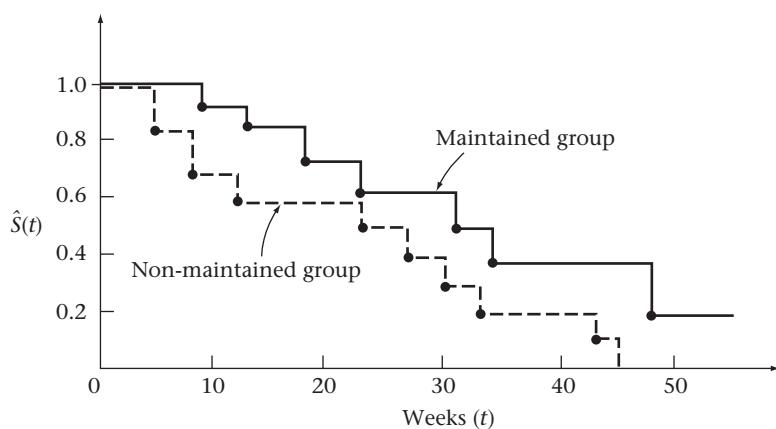
Maintained group ( $n = 11$ )	Non-maintained group ( $n = 12$ )
9	5
13	5
13+	8
18	8
23	12
28+	16+
31	23
34	27
45+	30
48	33
161+	43
	45

<sup>a</sup>13+ weeks indicates a censored observation. The patient was in remission for 13 weeks and was not followed any further.

In this example, survival is defined in terms of maintenance of remission.

The estimated survival curves for each group based on the Kaplan-Meier estimator are plotted in Figure 14.9.

**Figure 14.9** Survival curves by treatment group in the leukemia treatment trial



The maintained group appears to have a better survival profile since the probability of survival is higher for the maintained group than the control group at each point in time. Because the sample sizes are small, we will consider both parametric and nonparametric methods of analysis to maximize power.

## Weibull Survival Model

**Definition 14.9** The Weibull survival function is defined by

$$S(t) = e^{-(\lambda t)^\gamma}$$

where  $\lambda$  and  $\gamma$  are  $> 0$ .

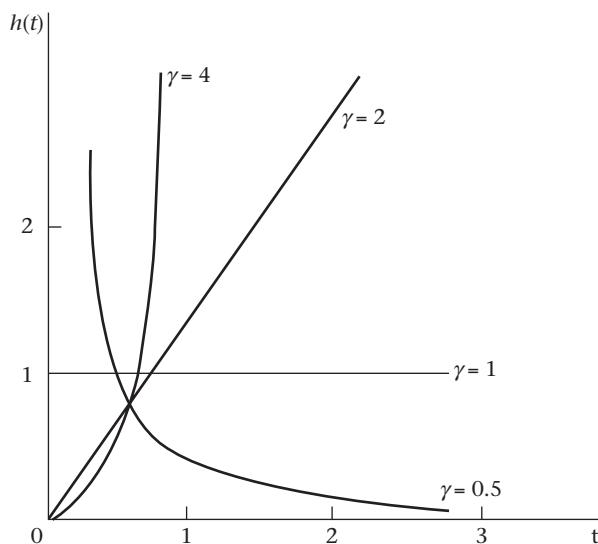
The parameter  $\gamma$  is referred to as the *shape* parameter and  $1/\lambda$  is referred to as the *scale* parameter. The corresponding hazard function is given by

**Equation 14.49**

$$h(t) = \lambda \gamma (\lambda t)^{\gamma-1}$$

A plot of the hazard function for different values of  $\gamma$  holding  $\lambda$  fixed (at 1) is given in Figure 14.10 [10].

**Figure 14.10** Hazard functions of the Weibull distribution with  $\lambda = 1$



The hazard increases over time if  $\gamma > 1$ , decreases over time if  $\gamma < 1$ , and remains constant if  $\gamma = 1$ . If  $\gamma = 1$ , then the Weibull model is also called an exponential survival model.

## Estimation of the Parameters of the Weibull Model

The parameters of the Weibull model are estimated either by least squares or by maximum likelihood methods based on an iterative algorithm.

The estimation methods take account of censoring and can be implemented with either right censored, left censored, or interval censored data. We will only discuss the right censored case here.

We have used MINITAB to fit the Weibull model to the survival data for each treatment group in the leukemia recurrence example (Example 14.45). The results are shown in Table 14.26 together with the fitted percentiles and are displayed in Figure 14.11.

**Table 14.26** Estimation of the survival curve by treatment group for the data in Example 14.45 using the MINITAB Weibull distribution program

Results for: leukemia.logrank.mtw

Distribution Analysis Survival Time by Group

Variable: Survival Time

Group = 1 (maintained group)

Censoring Information Count

Uncensored value 7

Right censored value 4

Censoring value: Censored = 0

Estimation Method: Least Squares (failure time(X) on rank (Y))

Distribution: Weibull

#### Parameter Estimates

Parameter	Estimate	Standard Error	95.0% Normal CI	
			Lower	Upper
Shape	1.78230	0.236879	1.37357	2.31266
Scale	38.8114	8.33681	25.4753	59.1289

Log-Likelihood = -43.100

Goodness of Fit

Anderson-Darling (adjusted) = 13.551

Correlation Coefficient = 0.993

#### Characteristics of Distribution

	Estimate	Standard Error	95.0% Normal CI	
			Lower	Upper
Mean (MTTF)	34.5316	7.23079	22.9075	52.0541
Standard Deviation	20.0298	2.75775	15.2926	26.2344
Median	31.5974	7.49082	19.8544	50.2859
First Quartile (Q1)	19.2917	5.67183	10.8421	34.3263
Third Quartile (Q3)	46.6177	9.13894	31.7454	68.4574
Interquartile Range (IQR)	27.3260	3.96947	20.5555	36.3266

#### Table of Percentiles

Percent	Percentile	Standard Error	95.0% Normal CI	
			Lower	Upper
1	2.93790	1.55662	1.04001	8.29925
2	4.34681	2.08455	1.69813	11.1268
3	5.47287	2.46377	2.26476	13.2254
4	6.45015	2.76925	2.78050	14.9629
5	7.33178	3.02909	3.26242	16.4770

(continued)

**Table 14.26** Estimation of the survival curve by treatment group for the data in Example 14.45 using the MINITAB Weibull distribution program (*Continued*)

Percent	Percentile	Standard Error	95.0% Normal CI	
			Lower	Upper
6	8.14541	3.25736	3.71977	17.8365
7	8.90766	3.46228	4.15828	19.0815
8	9.62945	3.64908	4.58183	20.2378
9	10.3184	3.82135	4.99318	21.3231
10	10.9802	3.98169	5.39441	22.3500
20	16.7289	5.20412	9.09226	30.7797
30	21.7647	6.08838	12.5788	37.6586
40	26.6244	6.82551	16.1088	44.0044
50	31.5974	7.49082	19.8544	50.2859
60	36.9537	8.13020	24.0095	56.8762
70	43.0716	8.78640	28.8769	64.2439
80	50.6896	9.52430	35.0738	73.2580
90	61.9711	10.5187	44.4334	86.4308
91	63.5464	10.6522	45.7515	88.2626
92	65.2721	10.7975	47.1975	90.2686
93	67.1864	10.9580	48.8036	92.4935
94	69.3445	11.1383	50.6163	95.0023
95	71.8311	11.3457	52.7067	97.8946
96	74.7857	11.5923	55.1920	101.335
97	78.4653	11.9012	58.2870	105.629
98	83.4334	12.3238	62.4613	111.447
99	91.4300	13.0249	69.1558	120.878

## Distribution Analysis: Survival Time by Group

Variable: Survival Time

Group = 2 (control group)

Censoring Information Count

Uncensored value 11

Right censored value 1

Censoring value: Censored = 0

Estimation Method: Least Squares (failure time(X) on rank (Y))

Distribution: Weibull

**Parameter Estimate**

Parameter	Estimate	Standard Error	95.0% Normal CI	
			Lower	Upper
Shape	1.41762	0.394237	0.821947	2.44499
Scale	25.0227	5.50354	16.2600	38.5078

Log-Likelihood = -44.236

Goodness of Fit

Anderson-Darling (adjusted) = 1.695

Correlation Coefficient = 0.957

(continued)

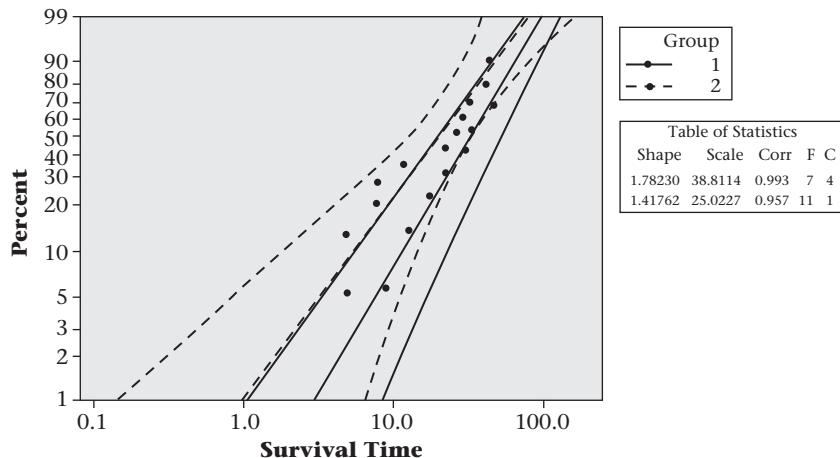
**Table 14.26** Estimation of the survival curve by treatment group for the data in Example 14.45 using the MINITAB Weibull distribution program (Continued)**Characteristics of Distribution**

	Estimate	Standard Error	95.0% Normal CI	
			Lower	Upper
Mean (MTTF)	22.7625	4.90629	14.9194	34.7288
Standard Deviation	16.2840	5.45149	8.44887	31.3851
Median	19.3220	4.72679	11.9624	31.2093
First Quartile (Q1)	10.3908	3.73968	5.13212	21.0377
Third Quartile (Q3)	31.5065	6.82059	20.6125	48.1582
Interquartile Range (IQR)	21.1157	5.82343	12.2986	36.2540

**Table of Percentiles**

Percent	Percentile	Standard Error	95.0% Normal CI	
			Lower	Upper
1	0.975124	0.946333	0.145543	6.53325
2	1.59573	1.33639	0.309098	8.23802
3	2.13177	1.62000	0.480711	9.45356
4	2.62090	1.84782	0.658149	10.4370
5	3.07895	2.03989	0.840348	11.2810
6	3.51450	2.20657	1.02669	12.0306
7	3.93286	2.35408	1.21676	12.7119
8	4.33761	2.48648	1.41030	13.3411
9	4.73133	2.60660	1.60708	13.9293
10	5.11594	2.71650	1.80695	14.4845
20	8.68603	3.48247	3.95869	19.0586
30	12.0922	3.95544	6.36898	22.9583
40	15.5793	4.33114	9.03460	26.8650
50	19.3220	4.72679	11.9624	31.2093
60	23.5262	5.26964	15.1667	36.4933
70	28.5234	6.15135	18.6910	43.5283
80	35.0045	7.74129	22.6923	53.9968
90	45.0654	11.1125	27.7939	73.0698
91	46.5104	11.6726	28.4398	76.0631
92	48.1039	12.3095	29.1316	79.4323
93	49.8842	13.0437	29.8808	83.2789
94	51.9070	13.9055	30.7040	87.7521
95	54.2578	14.9419	31.6266	93.0834
96	57.0784	16.2316	32.6898	99.6624
97	60.6313	17.9228	33.9687	108.222
98	65.4965	20.3489	35.6255	120.414
99	73.4839	24.5783	38.1495	141.545

**Figure 14.11** Estimated survival curve by treatment group with 95% confidence limits based on the data in Example 14.45 using the MINITAB Weibull Distribution Program



### Estimation of Percentiles of the Weibull Survival Function

Note that Table 14.26 gives percentiles for the cumulative distribution function (c.d.f.) for each group. Thus, the  $100\% \times (1 - p)$ th percentile of the c.d.f. ( $F(t)$ ) corresponds to a survival probability of  $p$ . Therefore, to solve for the survival time ( $t_p$ ) associated with a survival probability of  $p$ , we set

$$1 - p = S(t_p) = e^{-(\lambda t_p)^\gamma}$$

then take logs of both sides of the equation and multiply by  $-1$ , yielding:

$$-\ln(1 - p) = (\lambda t_p)^\gamma$$

We now raise each side of the equation to the  $1/\gamma$  power and divide by  $\lambda$  to obtain:

**Equation 14.50**

$$\begin{aligned} t_p &= [-\ln(1 - p)]^{1/\gamma} / \lambda = (1/\lambda)[-\ln(1 - p)]^{1/\gamma} \\ &= \text{scale parameter} \times [-\ln(1 - p)]^{1/\text{shape parameter}} \end{aligned}$$

#### Example 14.46

**Cancer** Estimate the time corresponding to the 90th percentile for group 1 (maintained group) in the leukemia recurrence data in Example 14.45.

#### Solution

From Table 14.26 we see that the scale parameter = 38.8114 and the shape parameter = 1.7823 for the maintained group (group 1). Hence, using equation 14.50 we obtain:

$$\begin{aligned} t_{.90} &= 38.8114 [-\ln(0.1)]^{1/1.7823} \\ &= 38.8114 (2.30)^{0.56} = 61.97 \text{ weeks} \end{aligned}$$

Thus, 90% of group 1 subjects have failed by 61.97 weeks, with only 10% surviving beyond this time, or  $\hat{S}(61.97) = 0.10$ .

Note that the percentiles are higher for group 1 (maintained group) than for group 2 (control group) (e.g., 90th percentile for group 2 = 45.1 weeks).

This is also clearly shown in Figure 14.11, where the percentiles for group 1 are consistently to the right of those from group 2.

A reasonable question is whether the survival curve for the maintained group is significantly different from the control group when both groups are modeled using a Weibull distribution. We discuss this in Section 14.14.

### Assessing Goodness of Fit of the Weibull Model

From Definition 14.9, we have that  $S(t) = e^{-(t/\alpha)^\gamma}$  where  $\alpha = 1/\lambda$ .

If we take logs of both sides of the equation and multiply by  $-1$  we obtain:

$$-\ln[S(t)] = (t/\alpha)^\gamma$$

If we take logs a second time we obtain:

**Equation 14.51**

$$\ln\{-\ln[S(t)]\} = \gamma \ln t - \gamma \ln \alpha$$

The transformation on the left-hand side is referred to as a complementary log-log transformation.

Thus if the Weibull model holds, there should be a linear relationship between  $\ln\{-\ln[S(t)]\}$  and  $\ln t$  with slope equal to the shape parameter  $\gamma$ .

Therefore, if a plot of  $\ln\{-\ln[S(t)]\}$  vs.  $\ln t$  is approximately linear, the Weibull model should provide a good fit.

**Example 14.47**

**Cancer** Assess the goodness of fit of the Weibull model for the leukemia recurrence data in Example 14.45.

**Solution**

We use MINITAB, clicking on Statistics/Reliability/Survival/Distribution analysis (Right Censoring)/Nonparametric Distribution Analysis, entering *c2* for variables (survival time), *c3* for censoring variable, and By variable *c4*. Under storage we click times for probabilities and survival probabilities so that *t* and *S(t)* will be stored in different columns for each group. We then compute  $\log\{-\log[S(t)]\}$  and  $\log t$  in separate columns for each group using the CALC/Calculator command. The results are shown in Table 14.27.

**Table 14.27** Summary statistics used to assess goodness of fit for the leukemia recurrence data in Example 14.45

**Data Display**

Row	<i>t</i> _1	<i>S(t)</i> _1	$\log\{-\log(S(t))\}_1$	$\log(t)_1$
1	9	0.909091	-2.35062	2.19722
2	13	0.818182	-1.60609	2.56495
3	18	0.715909	-1.09601	2.89037
4	23	0.613636	-0.71672	3.13549
5	31	0.490909	-0.34039	3.43399
6	34	0.368182	-0.00082	3.52636
7	48	0.184091	0.52610	3.87120

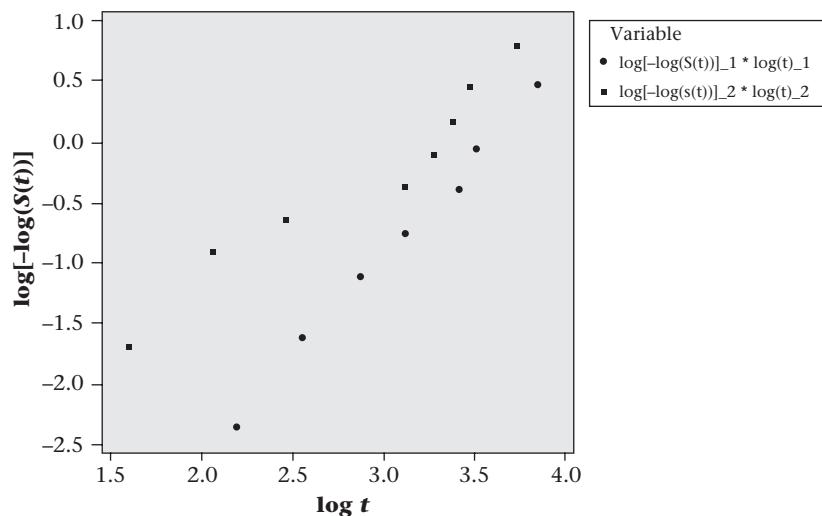
(continued)

**Table 14.27** Summary statistics used to assess goodness of fit for the leukemia recurrence data in Example 14.45 (Continued)

Data Display				
Row	t_2	S(t)_2	log[-log(S(t))_2]	log(t)_2
1	5	0.833333	-1.70198	1.60944
2	8	0.666667	-0.90272	2.07944
3	12	0.583333	-0.61805	2.48491
4	23	0.486111	-0.32668	3.13549
5	27	0.388889	-0.05714	3.29584
6	30	0.291667	0.20876	3.40120
7	33	0.194444	0.49324	3.49651
8	43	0.097222	0.84619	3.76120
9	45	0.000000	*	3.80666

We now plot  $\log[-\log(S(t))]$  vs.  $\log t$  separately for each group as shown in Figure 14.12.

**Figure 14.12** Scatterplot of  $\log[-\log(S(t))]$  vs.  $\log t$  by treatment group for the leukemia recurrence data in Example 14.45



The plots look approximately linear, especially for group 1 (maintained group). Also,  $\log[-\log(S(t))]$  is consistently lower for group 1 vs. group 2 for all  $t$ , indicating a higher survival probability for group 1 vs. group 2.

Thus, in summary, the Weibull model appears to fit these data reasonably well.

## 14.14 Parametric Regression Models for Survival Data

The Weibull model is of the form

$$\text{Equation 14.52} \quad Pr(T > t) = e^{-(\lambda t)^\gamma} \equiv e^{-(t/\alpha)^\gamma}$$

where

$\alpha = 1/\lambda$  = Weibull scale parameter

$\gamma$  = Weibull shape parameter

It can be shown based on this model that the probability density function (pdf) of  $Y = \ln(T)$  is given by

**Equation 14.53**

$$f(y) = \frac{1}{\sigma} \exp\left[\frac{y - \ln \alpha}{\sigma} - \exp\left(\frac{y - \ln \alpha}{\sigma}\right)\right], -\infty < y < \infty$$

where  $\sigma = 1/\gamma = 1/\text{shape parameter}$  and  $\alpha = \text{scale parameter}$ .

To incorporate covariates we let

**Equation 14.54**

$$\alpha = e^{\beta \underline{x}} \quad \text{or} \quad \ln \alpha = \beta \underline{x} \equiv \sum_{k=1}^K \beta_k x_k$$

Thus two subjects with different covariates vectors  $\underline{x}_A$  and  $\underline{x}_B$  have the same shape parameter ( $\gamma$ ), but different scale parameters given by  $e^{\beta \underline{x}_A}$  and  $e^{\beta \underline{x}_B}$ , respectively. Based on Equations 14.53 and 14.54, if we write  $Z = (Y - \ln \alpha)/\sigma$ , we can express  $Y$  in the form

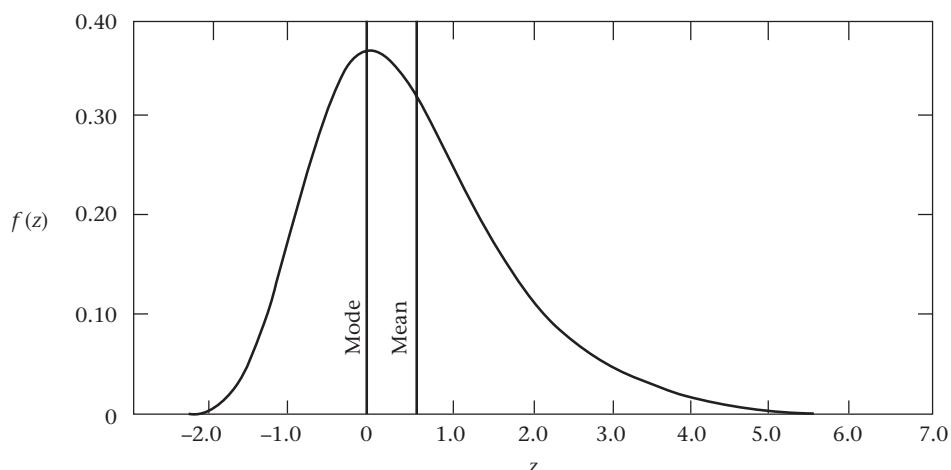
**Equation 14.55**

$$Y = \beta \underline{x} + \sigma Z$$

$\sigma Z$  can be interpreted similarly to the error term in linear regression. However, because the distribution of  $Y$  (i.e.,  $\ln_e$ (survival time)) is skewed to the right,  $Z$  is not normally distributed but instead follows a *standard extreme value distribution* with  $\text{pdf} = \exp(z - e^z)$ .

Under very general conditions, the standard extreme value distribution is used to obtain a limiting distribution for  $\max(X_1, \dots, X_n)$  in a sample of size  $n$  as  $n$  gets large. Its pdf is illustrated in Figure 14.13 [11].

**Figure 14.13** Probability density function of a standard extreme value distribution given by  $f(z) = \exp(-z - e^z)$



Thus

$$E(Y) = \beta \bar{x} + E(Y | \bar{x} = 0) \equiv \beta \bar{x} + E(Y_0)$$

For a particular covariate  $x_j$ ,

If  $\beta_j < 0$ , then the expected value of  $y$  decreases as  $x_j$  increases;

If  $\beta_j > 0$ , then the expected value of  $y$  increases as  $x_j$  increases.

Thus this type of model is sometimes called an *accelerating failure time model* since the time to failure is accelerated if  $\beta < 0$  and decelerated if  $\beta > 0$ .

It can also be shown that the Weibull model is a *proportional hazards model*.

To see this we note from Equation 14.49 that the hazard function of a Weibull distribution can be written in the form

**Equation 14.56**

$$h(t) = \lambda^\gamma \gamma t^{\gamma-1} = \left(\frac{1}{\alpha}\right)^\gamma \gamma t^{\gamma-1}$$

Substituting  $\alpha = e^{\beta \bar{x}}$  from Equation 14.54, we obtain

$$h(t) = \frac{1}{e^{\gamma \beta \bar{x}}} \gamma t^{\gamma-1}$$

Thus if we have two individuals with covariate vectors  $\bar{x}_1$  and  $\bar{x}_2$ , respectively, the hazard ratio comparing subject 2 with subject 1 is

**Equation 14.57**

$$h(t | \bar{x}_2) / h(t | \bar{x}_1) = \left( \frac{1}{e^{\gamma \beta \bar{x}_2}} \right) / \left( \frac{1}{e^{\gamma \beta \bar{x}_1}} \right) = e^{\beta(\bar{x}_1 - \bar{x}_2)\gamma}$$

which is independent of  $t$ . Hence, the hazard ratio comparing two subjects who are 1 unit apart on a variable  $x_k$  and are the same for all other covariates is given by

**Equation 14.58**

$$h(t | x_k + 1) / h(t | x_k) = e^{-\beta k \gamma}$$

Although the Weibull model is a proportional-hazards model it is different from the Cox proportional-hazards model because the hazard function is specified explicitly in terms of the parameters  $\lambda$  and  $\gamma$  (i.e., parametrically) rather than nonparametrically in the Cox model.

### Example 14.48

**Cancer** Use a Weibull parametric regression survival model to compare the survival curves of the 2 treatment groups using the leukemia recurrence data in Example 14.45.

### Solution

We have run a Weibull parametric regression model using MINITAB, where  $Y = \ln(\text{survival time})$  and

$$Y = \beta x + \sigma Z$$

where  $x = 1$  if a subject is in group 2 (control)

$= 0$  is a subject is in group 1 (maintained)

and  $Z$  is a standard extreme value distribution.

The data used for this analysis are given in Table 14.28.

To perform the analysis, we click on Statistics/Reliability/Survival/Accelerated Life Testing/Weibull. The results are shown in the Table 14.29.

**Table 14.28** Data display for leukemia recurrence data

Row	Survival		
	time	Censor	Group
1	9	1	1
2	13	1	1
3	13	0	1
4	18	1	1
5	23	1	1
6	28	0	1
7	31	1	1
8	34	1	1
9	45	0	1
10	48	1	1
11	161	0	1
12	5	1	2
13	5	1	2
14	8	1	2
15	8	1	2
16	12	1	2
17	16	0	2
18	23	1	2
19	27	1	2
20	30	1	2
21	33	1	2
22	43	1	2
23	45	1	2

**Table 14.29** MINITAB Weibull parametric survival model using the leukemia recurrence data in Example 14.45

Accelerated Life Testing: Survival Time versus Group_2						
Response Variable: Survival Time						
Censoring Information						Count
Uncensored value						18
Right censored value						5
Censoring value: Censored = 0						
Estimation Method: Maximum Likelihood						
Distribution: Weibull						
Relationship with accelerating variable(s): Linear						
Regression Table						
Predictor	Coef	Error	Z	P	95.0% Normal CI	
Intercept	4.10906	0.299890	13.70	0.000	3.52128	4.69683
Group_2	-0.929342	0.382502	-2.43	0.015	-1.67903	-0.179652
Shape	1.26430	0.225328			0.891546	1.79289
Log-Likelihood = -80.522						

We see that there is a significant effect of group ( $\beta = -0.93 \pm 0.38, p = .015$ ) with the ln(survival time) being shorter (i.e., accelerated failure) for group 2 (the control group) relative to group 1 (the maintained group). Also, the hazard ratio comparing subjects in group 2 vs. group 1 =  $e^{+0.929(1.264)} = 3.2$  or for group 1 (maintained group) vs. group 2 (control group) =  $e^{-0.929(1.264)} = 0.31$ . Thus, at any time  $t$ , subjects in the maintained group are about one-third as likely to experience a recurrence than subjects in the control group.

Note that this parametric regression analysis is somewhat different than the analysis in Table 14.26, where we fit separate Weibull survival functions in each group and as a result allowed both the shape and scale parameters to vary between the 2 groups. The approach in Table 14.26 makes it possible to compare survival probabilities at specific times but makes it difficult to compare the entire survival curve between the 2 groups. Using the regression approach, the shape parameter is forced to be the same for each group (or in general for different values of  $X$ ), while only the scale parameter is different. Hence, if we have 2 individuals who vary by 1 unit on a particular covariate (say  $x_j$ ) with all other variables held constant, then this implies that the scale parameter is different for these individuals, the hazard ratio comparing these two individuals is different from 1, and the respective survival curves are different from each other.

### Estimation of Percentiles for the Weibull Survival Distribution

In addition, we can estimate percentiles of the survival distribution for each group. For the  $p$ th percentile, we compute

$$\text{Equation 14.59} \quad t_p = [-\ln(1 - p)]^{1/\text{shape parameter}} \times \text{scale parameter}$$

where scale parameter original scale =  $\exp(\text{scale parameter log scale})$ .

#### Example 14.49

**Cancer** Estimate the 10th, 50th, and 90th percentiles for the survival time for each group in the leukemia recurrence data in Example 14.45.

#### Solution

We have the following shape and scale parameters in each group.

	Shape parameter	Scale parameter
Group 1	1.2643	$e^{4.10906} = 60.8892$
Group 2	1.2643	$e^{4.10906 - 0.9293} = 24.0399$

Hence, using Equation 14.59 we obtain:

*Group 1 (maintained group)*

$$t_{0.10} = [-\ln(0.9)]^{1/1.2643} 60.8892 = 10.3 \text{ weeks}$$

$$t_{0.50} = [-\ln(0.5)]^{1/1.2643} 60.8892 = 45.6 \text{ weeks}$$

$$t_{0.90} = [-\ln(0.1)]^{1/1.2643} 60.8892 = 117.8 \text{ weeks}$$

*Group 2 (control group)*

$$t_{0.10} = [-\ln(0.9)]^{1/1.2643} \quad 24.0399 = 4.1 \text{ weeks}$$

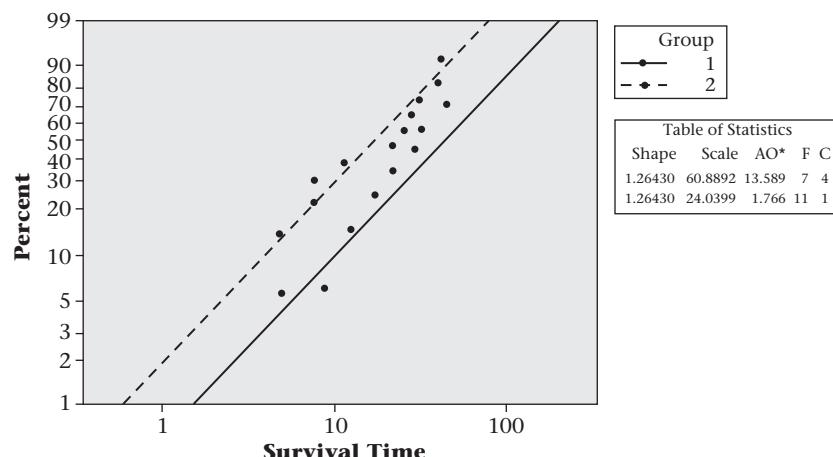
$$t_{0.50} = [-\ln(0.5)]^{1/1.2643} \quad 24.0399 = 18.0 \text{ weeks}$$

$$t_{0.90} = [-\ln(0.1)]^{1/1.2643} \quad 24.0399 = 46.5 \text{ weeks}$$

In general, the survival function is  $S(t) = e^{-(t/\text{scale parameter})^{\text{shape parameter}}}$ . For group 1, the survival function is  $S(t) = e^{-(t/60.8892)^{1.2643}}$ . For group 2, the survival function is  $S(t) = e^{-(t/24.0399)^{1.2643}}$ .

The complete set of percentiles is given in Figure 14.14.

**Figure 14.14** Plot of survival time by group using the MINITAB Weibull survival regression program



In general, the Weibull scale parameter is given by  $e^{\beta x}$ . Hence, to interpret  $\beta_k$ , if we compare 2 people who differ by 1 unit on  $x_k$  (subject 1 =  $x_k + 1$ ; subject 2 =  $x_k$ ) and are the same for all other covariates, then the estimated  $p$ th percentile for the first subject/estimated  $p$ th percentile for the second subject =  $e^{\beta_k}$ . For the leukemia data,  $\beta = -0.92$  for group 2 vs. group 1 or, equivalently,  $\beta = 0.92$  for group 1 vs. group 2. Hence, the ratio of the estimated  $p$ th percentile of the survival distribution for a person in the maintained group (group 1) vs. the  $p$ th percentile for a person in the control group (group 2) =  $e^{0.92} = 2.5$ , which is the same for all  $p$ .

We have also analyzed the leukemia recurrence data using a Cox proportional-hazards model:

$$h(t) = h_0(t) \exp(\beta x)$$

where  $x = 1$  if group 1  
 $= 0$  if group 2

The results using the SAS PROC PHREG program are given in Table 14.30. We see that the estimated hazard ratio = 2.47 for group 2 vs. group 1, but the results are not statistically significant ( $p = .078$ ).

It is interesting that comparisons of the survival curves for the 2 groups were significant using the parametric Weibull model ( $p = .015$ ), whereas they were not

**Table 14.30 Analysis of the leukemia recurrence data using the Cox proportional-hazards model**

The PHREG Procedure									
Model Information									
Data Set					WORK.LEUKEMIA				
Dependent Variable					survival				
Censoring Variable					censored				
Censoring Value(s)					0				
Ties Handling					BRESLOW				
Number of Observations Read					23				
Number of Observations Used					23				
Summary of the Number of Event and Censored Values									
Total	Event	Censored	Percent						
23	18	5	21.74						
Convergence Status									
Convergence criterion (GCONV=1E-8) satisfied.									
Model Fit Statistics									
Criterion		Without Covariates	With Covariates						
-2 LOG L		85.796	82.500						
AIC		85.796	84.500						
SBC		85.796	85.391						
Testing Global Null Hypothesis: BETA=0									
Test	Chi-Square		DF	Pr > ChiSq					
Likelihood Ratio	3.2960		1	0.0694					
Score	3.3226		1	0.0683					
Wald	3.1159		1	0.0775					
Analysis of Maximum Likelihood Estimates									
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq				
group	1	0.90422	0.51225	3.1159	0.0775				
					2.470				

significant using the nonparametric Cox regression model ( $p = .078$ ). The hazard ratio for group 2 (control) vs. group 1 (maintained) =  $\exp(0.929) = 2.53$  with the Weibull model and 2.47 with the Cox model which are similar. In general, nonparametric methods such as the Cox proportional-hazards model have the advantage of making fewer assumptions but may be slightly less efficient than their parametric counterparts, especially if the latter are supported by acceptable goodness of fit, as was the case for the leukemia data.

The methods of survival analysis (at least for the Cox model) can be extended to handle:

- (a) time-dependent covariates (i.e., covariates that change over time)
- (b) competing risks (i.e., where there are multiple types of failure that “compete” with each other). For example, if the outcome of interest is time to initial breast cancer, a competing risk would be death from another cause (e.g., heart attack).

However, this is beyond the scope of this text (see Miller [3]).

## 14.15 Summary

In this chapter, we discussed how to analyze data where the unit of analysis is person-time. The incidence rate was defined as the number of events per unit of person-time and was compared with cumulative incidence, which is the proportion of people who develop disease over a specified period of time. The incidence rate is in units of events per unit time and has no upper bound, whereas cumulative incidence is a proportion bounded between 0 and 1. We discussed procedures for comparing a single estimated incidence rate with a known incidence rate and also for comparing two incidence rates both for crude data as well as for data stratified by other potential covariates. We then discussed methods of power and sample-size estimation for study designs based on comparisons of two incidence rates, both with and without adjusting for confounding variables. Finally, these methods were extended to exposure variables with more than two levels of exposure and a test of trend was introduced to analyze data of this form.

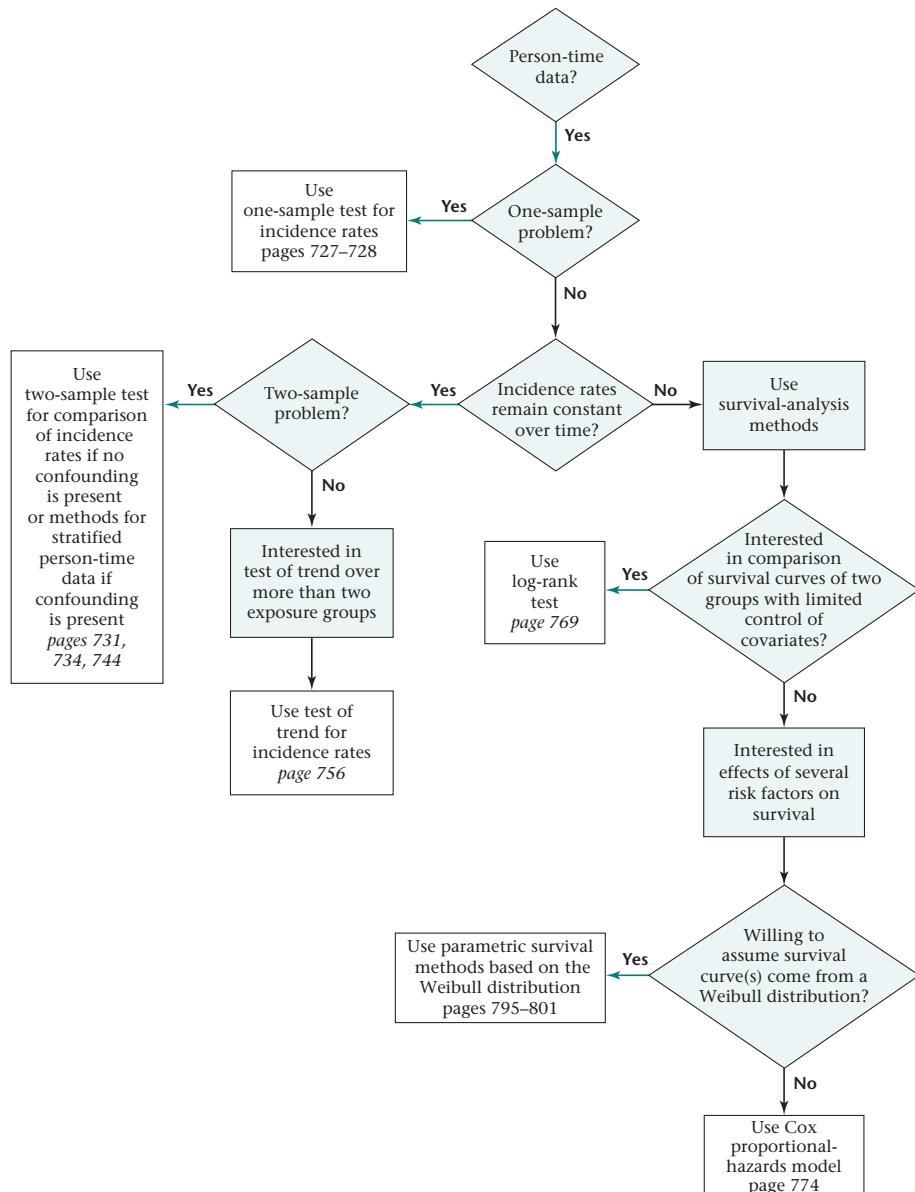
If incidence rates change greatly over time, then methods of survival analysis are appropriate. We introduced the concept of a hazard function, which is a function characterizing how incidence changes over time. We also introduced the concept of a survival curve, which is a function giving the cumulative probability of *not* having an event (i.e., surviving) as a function of time. The Kaplan-Meier estimator was introduced as a nonparametric method for estimating a survival curve. The log-rank test was then presented to let us statistically compare two survival curves (e.g., for an exposed vs. an unexposed group). If we want to study the effects of several risk factors on survival, then the Cox proportional-hazards model can be used. This method is analogous to multiple logistic regression where the time when an event occurs is considered rather than simply whether or not an event has occurred. We also discussed methods for assessing whether the assumption of proportional hazards is appropriate. Furthermore, we considered methods of power and sample-size estimation for studies where the primary method of analysis is the Cox proportional-hazards model. Finally, we considered methods of parametric survival analysis based on the Weibull distribution to estimate survival curves and to test hypotheses concerning the effect of covariates on survival curves. The methods in this chapter are outlined in the flowchart in Figure 14.15.

## PROBLEMS

### Cancer

The data relating oral-contraceptive (OC) use and the incidence of breast cancer in the age group 40–44 in the NHS are given in Table 14.31.

- \*14.1 Compare the incidence density of breast cancer in current users vs. never users, and report a *p*-value.
- \*14.2 Compare the incidence density of breast cancer in past users vs. never users, and report a *p*-value.

**Figure 14.15** Flowchart for appropriate methods of statistical inference—person-time data**Table 14.31** Relationship between breast-cancer incidence and OC use among 40- to 44-year-old women in the NHS

OC-use group	Number of cases	Number of person-years
Current users	13	4,761
Past users	164	121,091
Never users	113	98,091

\***14.3** Estimate the rate ratio comparing current users vs. never users, and provide a 95% CI about this estimate.

\***14.4** Estimate the rate ratio comparing past users vs. never users, and provide a 95% CI about this estimate.

**14.5** How much power did the study have for detecting an IRR for breast cancer of 1.5, comparing current OC users vs. never OC users among 40- to 44-year-old women if

(a) the true incidence rate of breast cancer among never users and the amount of person-time for current and never users are the same as in Table 14.31, (b) the expected

number of events for never OC users is the same as the observed number of events in Table 14.31, and (c) the average follow-up time per subject is the same for both current and never OC users?

**14.6** What is the expected number of events that need to be realized in each group to achieve 80% power to detect an  $IRR$  for breast cancer of 1.5 for current OC users vs. never OC users under the same assumptions as in Problem 14.5?

### Health Promotion

Refer to Data Set SMOKE.DAT on the Companion Website.

**14.7** Divide participants according to median  $\log_{10}$ CO (adjusted), and estimate survival curves for each subgroup.

**14.8** Compare survival curves of the two groups, using hypothesis-testing methods, and report a  $p$ -value.

A Cox proportional-hazards model was fit to these data to assess the relationship between age, sex, number of cigarettes smoked, and  $\log_{10}$ CO concentration, when considered simultaneously, on the ability to remain abstinent from smoking. The results are given in Table 14.32.

**14.9** Assess the significance of each of the variables.

**14.10** Estimate the effects of each variable in terms of hazard ratios, and provide 95% confidence limits corresponding to each point estimate.

**14.11** Compare the crude and adjusted analyses of the relationship of  $\log_{10}$ CO to recidivism in Problems 14.8 and 14.9.

### Bioavailability

Refer to Data Set BETACAR.DAT on the Companion Website.

**Table 14.32 Proportional-hazards model relating the hazard of recidivism to age, sex, number of cigarettes smoked prior to quitting, and  $\log_{10}$ CO concentration**

Risk factor	Regression coefficient ( $\hat{\beta}$ )	Standard error $se(\hat{\beta})$
Age	0.0023	0.0058
Sex(1 = M/0 = F)	-0.127	0.143
Number of cigarettes smoked	-0.0038	0.0050
$\log_{10}$ CO (adjusted) <sup>a</sup>	0.912	0.366

<sup>a</sup>This variable represents CO values adjusted for minutes elapsed since last cigarette smoked prior to quitting.

**14.12** Suppose we regard a preparation as being bioavailable for a subject at the first week when level of plasma carotene increases by 50% from the baseline level (based on an average of the first and second baseline determinations). Use survival-analysis methods to estimate the proportion of subjects for whom the preparation is not bioavailable at different points in time.

**14.13** Assess whether there are significant differences among the survival curves obtained in Problem 14.12. (*Hint:* Use a dummy-variable approach with proportional-hazards models.)

**14.14** Answer the same question as in Problem 14.12 if the criterion for bioavailability is a 100% increase in plasma-carotene level from baseline.

**14.15** Answer the same question posed in Problem 14.13 if the criterion for bioavailability is a 100% increase in plasma-carotene level from baseline.

### Ophthalmology

In Table 14.33, we present data from the RP clinical trial described in Example 14.30 concerning effect of high-dose vitamin E (400 IU/day) vs. low-dose vitamin E (3 IU/day) on survival (where failure is loss of at least 50% of initial ERG 30 Hz amplitude).

**14.16** Estimate the hazard function by year for each group.

**14.17** Estimate the survival probability by year for each group.

**14.18** Obtain a 95% CI for the survival probability at year 6 for each group.

**14.19** Compare overall survival curves of the two groups, and obtain a  $p$ -value.

**14.20** Suppose a new study is planned, with 200 patients randomly assigned to each of a 400 IU per day vitamin E group and a 3 IU per day vitamin E group. If the survival experience in the 3 IU per day group is assumed the same as in Table 14.33, the relative hazard for the 400 IU/day group vs. the 3 IU/day group = 1.5, and the censoring experience of both groups are assumed the same as for the 3 IU per day group in Table 14.33, then how much power would a new study have if the maximum duration of follow-up is 4 years (rather than 6 years as in the original study) and a two-sided test is used with  $\alpha = .05$ ?

**14.21** How many subjects need to be enrolled in each group (assume equal sample size in each group) to achieve 80% power if a two-sided test is used with  $\alpha = .05$  and the same assumptions are made as in Problem 14.20?

### Infectious Disease

Suppose the rate of allergic reactions in a certain population is constant over time.

**\*14.22** A person is selected randomly from the population and followed for 1.5 years. If the true rate of allergic reactions is 5 reactions per 100 person-years, what is the probability that the subject will have at least one allergic reaction during the follow-up period (i.e., cumulative incidence)?

**\*14.23** Two hundred subjects are selected randomly from the population and followed for various lengths of time. The average length of follow-up is 1.5 years. Suppose that at the end of the study, the estimated rate is 4 per 100 person-years. How many events must have been observed in order to yield the estimated rate of 4 per 100 person-years?

**\*14.24** Provide a 95% CI for the underlying rate, based on the observed data in Problem 14.23. Express the answer in units of number of events per 100 person-years.

**Table 14.33 Number of patients who failed, were censored, or survived by year in the 400 IU vitamin E group and 3 IU vitamin E group, respectively, RP clinical trial**

	Fail	Censored	Survive	Total
<b>400 IU of vitamin E daily</b>				
0–1 yr	7	3	170	180
1–2 yr	9	2	159	170
2–3 yr	22	2	135	159
3–4 yr	24	27	84	135
4–5 yr	13	32	39	84
5–6 yr	11	28	0	39
<b>3 IU of vitamin E daily</b>				
0–1 yr	4	1	169	174
1–2 yr	10	3	156	169
2–3 yr	14	1	141	156
3–4 yr	16	27	98	141
4–5 yr	15	34	49	98
5–6 yr	7	42	0	49

Note: A person fails if his or her ERG amplitude declines by at least 50% from baseline to any follow-up visit.

## Cancer

The data in Table 14.34 provide the relationship between breast-cancer incidence rate and menopausal status by age, based on Nurses' Health Study (NHS) data from 1976 to 1990.

**Table 14.34 Relationship between breast-cancer incidence rate and menopausal status after controlling for age, NHS, 1976–1990**

Age	Premenopausal	Postmenopausal
	Cases/person-years (incidence rate) <sup>a</sup>	Cases/person-years (incidence rate)
35–39	124/131,704 (94)	15/14,795 (101)
40–44	264/179,132 (147)	47/43,583 (108)
45–49	304/151,548 (201)	163/90,965 (179)
50–54	159/61,215 (260)	401/184,597 (217)
55–59	25/6133 (408)	490/180,458 (272)

<sup>a</sup>Per 100,000 person-years.

**14.25** Assess whether there is a difference between the incidence rate of breast cancer for premenopausal vs. postmenopausal women, while controlling for age. Report a *p*-value.

**14.26** Estimate the rate ratio for postmenopausal vs. premenopausal women after controlling for age. Provide a 95% CI for the rate ratio.

## Renal Disease

Refer to Data Set SWISS.DAT on the Companion Website.

**14.27** Suppose a serum-creatinine level of  $\geq 1.5$  mg/dL is considered a sign of possible kidney toxicity. Use survival-analysis methods to assess whether there are differences in the incidence of kidney toxicity between the high-N-acetyl-p-aminophenol (NAPAP) group and the control group. In this analysis, exclude subjects who were  $\geq 1.5$  mg/dL at baseline.

**14.28** Answer Problem 14.27 comparing the low-NAPAP group with the control group.

One issue is that the groups in Problem 14.27 may not be exactly balanced by age and/or initial level of serum creatinine.

**14.29** Answer Problem 14.27 while controlling for possible age and initial-level differences between groups. Consider both parametric and nonparametric survival analysis methods.

**14.30** Answer Problem 14.28 while controlling for possible age and initial-level differences between groups.

**14.31** Assess the validity of the proportional-hazards assumption in Problems 14.29 and 14.30.

## Infectious Disease

Suppose the incidence rate of influenza (flu) during the winter of 1998–1999 (i.e., from December 21, 1998 to March 20, 1999) was 50 events per 1000 person-months among students in high schools in a particular city.

**14.32** A group of high-risk high school students was identified in the winter of 1998–1999 who had 3+ previous episodes of influenza before December 21, 1998. There were 20 students in this group, each followed for 90 days, of whom 8 developed influenza. Test the hypothesis that the high-risk students had a higher incidence rate of influenza than the average high school student during the winter of 1998–1999. Report a one-tailed *p*-value.

**14.33** Provide a two-sided 95% CI for incidence rate of flu among high-risk students during winter 1998–1999.

Among 1200 students in one high school in the city, 200 developed a new case of influenza over the 90 days from December 21, 1999 to March 20, 2000.

**14.34** What is the estimated incidence rate of flu in the 1999–2000 winter season per 1000 person-months?

**14.35** Provide a 95% CI for the rate estimated in Problem 14.34.

**14.36** Test the hypothesis that the rate of flu has changed from the 1998–1999 to 1999–2000 winter season. Report a two-tailed *p*-value.

## Orthopedics

A study was performed among patients with piriformis syndrome, a pelvic condition that involves malfunction of

the piriformis muscle (a deep buttock muscle), which often causes lumbar and buttock pain with sciatica (pain radiating down the leg). A randomized double-blind clinical trial was performed whereby patients were injected with one of three substances.

Group 1 received an injection of a combination of triamcinolone and lidocaine (the TL group).

Group 2 received a placebo injection.

Group 3 received an injection of Botox.

The randomization schedule was set up such that, approximately, for every six patients, three were assigned to group 1, one to group 2, and two to group 3. All injections were directly into the piriformis muscle. Patients were asked to come back at 2 weeks post-injection (0.5 month), 1 month post-injection, and monthly thereafter up to 17 months, although there were many missed visits. At each visit the patients rated their percentage of improvement of pain vs. baseline on a visual analog scale, with a maximum of 100% improvement (indicated by 100). Negative numbers indicate percentage of worsening. There were a total of 69 subjects and 70 legs (one patient, ID 23, had the condition in both legs). A priori it was of interest to compare the degree of efficacy between each pair of groups. Three additional covariates may influence the outcome: age (yrs), gender (1 = male, 0 = female), and the affected side (L = left, R = right). The data are in BOTOX.DAT, with a description in BOTOX.DOC, on the Companion Website.

**14.37** If the visual analog scale is treated as a continuous variable, assess whether there are any between-group differences in efficacy without considering the covariates. Try to do at least one analysis that uses the entire data set rather than focusing on specific time points.

**14.38** Repeat the analysis in Problem 14.37, but account for covariate differences between groups.

**14.39** Another way to score the visual analog scale is as a categorical variable where  $\geq 50\%$  improvement is considered a success and  $< 50\%$  improvement, remaining the same, or worsening is considered a failure. Answer the question posed in Problem 14.37 using the success/failure scoring. Note that a patient may be a success at one visit but a failure at succeeding visits. (*Hint:* Either logistic-regression methods or survival-analysis methods may be applicable here.)

**14.40** Repeat the analyses in Problem 14.39, but account for covariate differences between groups.

To reduce variability, the investigators also considered a criterion of at least 50% improvement on two successive visits as a definition of success.

**14.41** Answer Problem 14.39 under this definition of success.

**14.42** Answer Problem 14.40 under this definition of success.

## Cancer

Suppose we wish to study the association between aspirin intake and the incidence of colon cancer. We find that 10% of women take 7 aspirin tablets per week (ASA group), while 50% of women never take aspirin (control group). The ASA group is followed for 50,000 person-years, during which 34 new colon cancers occurred over a 20-year period. The control group is followed for 250,000 person-years, during which 251 new colon cancers developed over a 20-year period.

**14.43** What are the estimated incidence rates in the ASA and control groups?

**14.44** Is there a significant difference between these incidence rates? Report a *p*-value (two-tailed).

**14.45** What is the estimated rate ratio for colon cancer between the ASA and placebo groups?

**14.46** Provide a 95% CI for the rate ratio in Problem 14.45.

**14.47** Suppose we look at the subset of women with a family history of colon cancer. Aspirin might be more beneficial in this high-risk subgroup. We have a total of 5000 person-years among ASA women and 2 events. We have a total of 20,000 person-years among control women and 20 events. Is there a significant difference in the incidence rates of colon cancer between these 2 groups? Provide a *p*-value (two-tailed).

## Health Promotion

A recent article by Kenfield et al. [12] studied the relationship between various aspects of smoking and mortality among 104,519 women in the Nurses' Health Study (NHS) from 1980–2004. One issue is whether there is a mortality benefit from quitting smoking vs. continuing to smoke and, if so, how long it takes for the mortality experience of former smokers to approximate that of never smokers. The data in Table 14.35 were presented comparing former smokers with current smokers.

**Table 14.35 Relationship of time since quitting to total mortality**

	Number of deaths	Number of person-years of follow-up
Current smokers	3,602	420,761
Former smokers		
Quit < 5 yrs	889	124,095
Quit 5–9 yrs	669	113,056
Quit 10–14 yrs	590	111,701
Quit 15–19 yrs	541	117,914
Quit 20+ yrs	1,707	336,177

**14.48** What is the estimated mortality rate and 95% CI per 1000 person-years among current smokers?

**14.49** What test can be performed to compare mortality incidence between former smokers who quit <5 years ago vs. current smokers?

**14.50** Implement the test in Problem 14.49, and report a *p*-value (two-tailed).

**14.51** What is the estimated rate ratio for total mortality between former smokers who quit 20+ years ago and current smokers? Provide a 95% CI for this estimate.

**14.52** The age-adjusted rate ratio between the groups in Problem 14.51 was 0.34. Is this different from the estimated rate ratio in Problem 14.51? If so, why? (*Hint:* Assume that former smokers who quit for 20+ years are older than current smokers and mortality increases with age.)

**14.56** Implement the test in Problem 14.55, and report a *p*-value (two-tailed).

### Pulmonary Disease

A study was performed among 169,871 Chinese men and women in 1991 ages 40 years and older [13]. Baseline data were collected in 1991, and a follow-up exam was conducted in 1999–2000. One component of the follow-up exam was a mortality follow-up for subjects who died between 1991 and 1999, where the date and cause of death were determined from Chinese vital statistics data. Of particular interest were risk factors for death from chronic-obstructive pulmonary disease (COPD). A Cox proportional-hazards regression

**Table 14.36 Relationship between PMH use and breast cancer incidence among 1200 women in the NHS**

Year	Current PMH users			Never PMH users		
	Number of women			Number of women		
	In risk set	Failed <sup>a</sup>	Censored <sup>b</sup>	In risk set	Failed	Censored
1990	200	0	1	1000	0	12
1992	199	3	2	988	3	10
1994	194	2	2	975	9	22
1996	190	4	1	944	7	23
1998	185	2	50	914	5	193
2000	133	2	131	716	9	107

<sup>a</sup>Failed means developed breast cancer.

<sup>b</sup>Assume that at any given year that the failures occur just prior to the censored observations in that year.

### Cancer

A study was performed to compare breast cancer incidence between postmenopausal women who used PMH vs. women who did not. A group of 200 women who were current PMH users and 1000 women who were never PMH users in 1990 in the NHS were identified. All women were postmenopausal and free of cancer as of 1990. The 1200 women were ascertained for incident breast cancer by mail questionnaire every 2 years up to the year 2000. However, not all women had complete follow-up. For simplicity, we will assume that women can only fail every 2 years, i.e., in 1992, 1994, . . . , 2000. The results are given in Table 14.36.

**14.53** What does a censored observation in 1992 mean in the context of these data?

**14.54** Estimate the 10-year incidence of breast cancer in each group. (*Hint:* Use the product limit method.)

**14.55** What test can be used to compare the incidence of breast cancer between the 2 groups, taking into account the time when breast cancer develops and the length of follow-up of each subject?

model was used to relate risk factors in 1991 to time of death from COPD between 1991 and 1999–2000.

**14.57** What is a hazard rate in the context of this study?

**14.58** Write down the Cox proportional-hazards model. What does the term proportional hazards mean?

The results in Table 14.37 were obtained from the study.

**14.59** What is the hazard ratio for COPD mortality among men for smokers of ≥20 pack-years vs. never smokers? Provide a 95% CI.

Most risk factors seem of comparable magnitude for men and women. However, one exception is cigarette smoking.

**14.60** Test the hypothesis that the hazard ratio for smoking ≥20 pack-years vs. never smoking is significantly different (at the 5% level) for men vs. women. (*Hint:* Use a *z* statistic approach, considering the men and women as 2 independent samples.)

### Cancer

**14.61** Suppose we wish to conduct a new study of the association between ASA and colon cancer. We will

**Table 14.37 Risk factors for COPD death among 169,871 study participants, China, 1991–2000**

Baseline risk factor	Men		Women	
	$\beta$	se	$\beta$	se
Age, 10 years	1.030	0.031	0.997	0.034
Alcohol consumption (at least 12 times in past year)	-0.174	0.064	0.365	0.153
Cigarette smoking				
1–19 pack-years	-0.062	0.091	0.300	0.109
$\geq 20$ pack-years	0.166	0.067	0.571	0.100
Hypertension	0.049	0.063	0.058	0.069
No high school education	0.863	0.110	0.904	0.202
Physical inactivity	0.451	0.082	0.300	0.073
Underweight (BMI <18.5)	0.978	0.065	0.956	0.073
Living in Northern China	0.329	0.062	0.548	0.073
Living in rural China	0.761	0.071	0.582	0.074

enroll 50,000 women in the study, half of whom will be assigned to ASA and half to placebo. Each woman is followed for 5 years. The expected incidence rate of colon cancer in the ASA group =  $70/10^5$  person-years and in the placebo group =  $100/10^5$  person-years. If we conduct a two-sided test with  $\alpha = 0.05$ , how much power will the study have?

**14.62** Suppose we want to enroll  $n$  subjects per group in the previously proposed study and follow each woman for 5 years. How many subjects do we need to achieve 90% power?

## Cancer

The data set in file BREAST.DAT consists of 1200 women in the NHS. The women were ascertained in 1990 and were postmenopausal and free of any cancer as of 1990. The 1200 women were selected in such a way that 200 of the women were current postmenopausal hormone (PMH) users in 1990 and 1000 of the women had never used PMH as of 1990. The objective of the analysis was to relate breast cancer incidence from 1990 to 2000 to PMH use as of 1990. Fifty-three of the women developed breast cancer between 1990 and 2000. PMH use is characterized both by current use/never use in 1990 as well as by duration of use as of 1990. Some current users in 1990 may have duration of use of 0 as of 1990 if they just started use in 1990 or if they used other types of PMH as of 1990 (other than estrogen or estrogen plus progesterone). There are two duration variables according to type of PMH use (duration of estrogen use in months as of 1990 and duration of estrogen plus progesterone use in months as of 1990). Each woman has a date of return of the 1990 questionnaire and a follow-up date = date of diagnosis of breast cancer if a case, or date of the last questionnaire filled out up to 2000 if a control. On the data file we provide the length of follow-up = follow-up

date - date of return of the 1990 questionnaire (variable 18). Thus, the first subject (ID 10013) had a date of return of  $1087 = \text{July } 1990 = (12 \times 90 + 7)$  and a follow-up date of  $1206 = \text{June } 2000 = (12 \times 100 + 6)$ . In addition, the file contains the values of other breast cancer risk factors as of 1990. A description of the data set is given in BREAST.DOC.

Format for Breast Cancer—postmenopausal hormone file

```

1 ID
2 case 1=case, 0=control
3 age
4 age at menarche
5 age at menopause
6 age at first birth 98=nullip
7 parity
8 Benign Breast disease (bbd) 1=yes/0=no
9 family history of breast cancer 1=yes/0=no
10 BMI (kg/m**2)
11 Height (inches)
12 Alcohol use (grams/day)
13 PMH status 2=never user/3=current user
14 Duration of Estrogen use (months)
15 Duration of Estrogen + progesterone use (months)
16 Current Smoker 1=yes/0=no
17 Past smoker 1=yes/0=no
18 follow-up time (months)

```

**14.63** Compare breast cancer incidence between the two exposure groups, where group 1 is the current PMH subjects and group 2 is the never PMH subjects. (*Hint:* Both the number of events and when the events occurred should be considered in the analysis.)

**14.64** Is there a difference in incidence according to duration of use (Variables 14 and 15)?

**14.65** Compare the current PMH users vs. the never users in 1990 on other possible confounding variables.

**14.66** Perform an adjusted analysis comparing breast cancer incidence between the current PMH users vs. the never PMH users, adjusting for confounders that you found in Problem 14.65.

**14.67** Are there any interaction effects between PMH exposure and other risk factors?

### Ophthalmology

Retinitis pigmentosa (RP) is a hereditary ocular disease in which patches of pigment appear on the retina; the condition can result in substantial losses of vision and, in some cases, complete blindness. There are various modes of inheritance, including a dominant form, a recessive form, and a sex-linked form. An important discovery over the past 10 years was a set of genes that account for many of the RP cases. Specifically, mutations in the rhodopsin gene (RHO) account for many of the dominantly inherited cases; mutations in the RPGR gene (RPGR) account for many of the sex-linked cases. An important issue is whether the rate of progression is different between the RHO patients and the RPGR patients. On the data file FIELD.DAT are visual field data from approximately 100 patients in each group. Visual field is a measure of area of vision. It is measured in degrees<sup>2</sup>. Longitudinal data with varying follow-up times are provided for each patient separately for the right eye (labeled OD) and the left eye (labeled OS). Follow-up time varies from a minimum of 3 years to a maximum of about 25–30 years. (*Hint:* For simplicity, use the geometric mean field over 2 eyes as the summary measure of field for a subject at a particular visit.)

**14.68** Assess whether the baseline level of visual field differs between RHO and RPGR patients.

**14.69** Assess whether the rate of decline differs between RHO and RPGR patients.

### FIELD.DOC

Column	Variable
1–6	ID
8	group (1 = RHO/2 = RPGR)
11–14	age at visit (XX.X in years)
16	gender (1 = m/2 = f) Gender is coded as missing, but is actually male for all members of the RPGR group.
18–27	date of visit (month/day/year)
29–34	time from 1st visit in years
36–43	total field area right eye (OD) in degrees <sup>2</sup>
45–52	total field area left eye (OS) in degrees <sup>2</sup>

One possible complexity is that the age distribution of the 2 groups may not be balanced. Also, the RHO group contains both males and females while the RPGR group consists exclusively of males. Gender is also missing for some subjects in the RHO group. Gender is coded as missing for subjects in the RPGR group because they are all male.

**14.70** Answer the question posed in Problem 14.69, while accounting for possible age and gender differences between groups.

An important endpoint for RP patients is legal blindness. For visual field, legal blindness is usually defined as <20° diameter of equivalent circular field area in the better eye. The equivalent circular field area for a 20° diameter is  $\pi R^2 = \pi(10)^2 = 314$  degrees squared. For example, ID 156 reached this endpoint at the fifth visit (approximately at 5.06 years of follow-up) in the right eye and at both eyes at the sixth visit (approximately 6 years of follow-up).

**14.71** Assess whether the time to legal blindness is the same or different between the RHO group and the RPGR group. For simplicity, assume that legal blindness in a particular eye is an absorbing state (i.e., once they become legally blind in an eye, they remain legally blind). Also, do not include eyes that are legally blind at baseline in the analysis, since they have already reached the endpoint. The format of the data set is given in FIELD.DOC.

### Infectious Disease

A study was recently performed concerning the incidence of H1N1 influenza in Australia and New Zealand [14].

**14.72** It was found that among 626 Australian H1N1 patients admitted to an intensive care unit (ICU), 61 were aboriginal. If 2.5% of Australians are aboriginal, test the hypothesis that the percentage of aboriginal H1N1 cases differs from the percentage of aborigines in the general population. Provide a two-sided *p*-value.

The data in Table 14.38 were presented on H1N1 patients admitted to the ICU (no. per million inhabitants) by week and region.

**Table 14.38** H1N1 patients admitted to the ICU in 2009 (no. per million inhabitants) by week and region

Week	Region		
	Victoria	New South Wales	Queensland
06/29	4.2	1.8	0.2
07/06	2.8	4.2	0.8
07/13	4.5	6.5	3.7
07/20	1.7	4.7	6.1
07/27	0.9	4.7	6.8
08/03	1.3	2.0	6.1
08/10	1.7	1.2	5.2
2001			
Population	5,314,000	6,984,000	4,294,000

**14.73** For each region,

- estimate the incidence rate per million ( $10^6$ ) inhabitants (per week) over the 7-week period, and
- the number of cases over 7 weeks from 06/29/09 to 08/16/09 (rounded to the nearest integer).

**14.74** Using the incidence rates calculated in Problem 14.73, test whether there are significant differences in incidence rates by region over the 7-week period. Provide a two-sided  $p$ -value. (Note: Assume that the underlying incidence rate in a given region is the same over the 7 weeks.)

**14.75.** There were 722 patients admitted to the ICU for H1N1 in Australia and New Zealand from June 1 to August 31, 2009 (winter season). The total population of Australia and New Zealand is approximately  $25,000,000 = 25$  million. Suppose that the underlying incidence rate is the same for the United States as for Australia and New Zealand. If there are 250 million people who live in the United States, then what is the estimated number of H1N1 cases admitted to the ICU in the United States during the winter season (12/21/09–3/20/10), and what is a 95% CI for the number of US cases?

## REFERENCES

- [1] Colditz, G. A., Stampfer, M. J., Willett, W. C., Hennekens, C. H., Rosner, B., & Speizer, R. E. (1990). Prospective study of estrogen replacement therapy and risk of breast cancer in post-menopausal women. *Journal of the American Medical Association*, 264, 2648–2653.
- [2] Kang, J. H. (2001). Cigarette smoking, antioxidants and fats and primary open-angle glaucoma. Unpublished doctoral thesis, Harvard School of Public Health.
- [3] Miller, R. G. (1981). *Survival analysis*. New York: Wiley.
- [4] Berson, E. L., Rosner, B., Sandberg, M. A., Hayes, K. C., Nicholson, B. W., Weigel-DiFranco, C., & Willett, W. C. (1993). A randomized trial of vitamin A and vitamin E supplementation for retinitis pigmentosa. *Archives of Ophthalmology*, 111, 761–772.
- [5] Coombes, R. C., Hall, E., Gibson, L. J., Paridaens, R., Jassem, J., Delozier, T., Jones, S. E., Alvarez, I., Bertelli, G., Ortmann, O., Coates, A. S., Bajetta, E., Dodwell, D., Coleman, R. E., Fallowfield, L. J., Mickiewicz, E., Andersen, J., Lonning, P. E., Cocconi, G., Stewart, A., Stuart, N., Snowdon, C. F., Carpentieri, M., Massimini, G., Bliss, J. M., for the Intergroup Exemestane Study. (2004). A randomized trial of exemestane after two to three years of tamoxifen therapy in postmenopausal women with primary breast cancer. *New England Journal of Medicine*, 350, 1081–1092.
- [6] Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 34, 187–220.
- [7] Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using the log-rank test. *Statistics in Medicine*, 1, 121–129.
- [8] Lakatos, E., & Lan, K. K. G. (1992). A comparison of sample size methods for the log-rank statistic. *Statistics in Medicine*, 11, 179–191.
- [9] Lachin, J. M., & Foulkes, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, 42, 507–519.
- [10] Lee, E. T. (1986). *Statistical methods for survival data analysis*. Belmont, CA: Wadsworth.
- [11] David, H. A. (1980). *Order statistics*. New York: Wiley.
- [12] Kenfield, S. A., Stampfer, M. J., Rosner, B. A., & Colditz, G. A. (2008). Smoking and smoking cessation in relation to mortality in women. *Journal of the American Medical Association*, 299(17), 2037–2047.
- [13] Reilly, K. H., Gu, D., Duan, X., Wu, X., Chen, C.-S., Huang, J., Kelly, T. N., Chen, J., Liu, X., Yu, L., Bazzano, L. A., & He, J. (2008). Risk factors for chronic obstructive pulmonary disease mortality in Chinese adults. *American Journal of Epidemiology*, 167, 998–1004.
- [14] ANZIC Influenza Investigators. (2009). Critical care services and 2009 H1N1 influenza in Australia and New Zealand. *New England Journal of Medicine*, 361, 1925–1934.

# APPENDIX

## Tables

**Table 1** Exact binomial probabilities  $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$

<i>n</i>	<i>k</i>	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0283	.0488	.0768	.1128	.1563
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0313
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
8	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188

(continued on next page)

**Table 1** Exact binomial probabilities  $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$  (continued)

<i>n</i>	<i>k</i>	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
9	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0313
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
	9	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
10	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
	10	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
11	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
	11	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
12	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
	12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
13	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095
	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571

(continued on next page)

**Table 1** Exact binomial probabilities  $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$  (continued)

<i>n</i>	<i>k</i>	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	6	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095
	7	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095
	8	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571
	9	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873
	10	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009
	2	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056
	3	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222
	4	.0037	.0349	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611
	5	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222
	6	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833
	7	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095
	8	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833
	9	.0000	.0000	.0000	.0003	.0018	.0066	.0183	.0408	.0762	.1222
	10	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0093	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417
	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
16	0	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001	.0000
	1	.3706	.3294	.2097	.1126	.0535	.0228	.0087	.0030	.0009	.0002
	2	.1463	.2745	.2775	.2111	.1336	.0732	.0353	.0150	.0056	.0018
	3	.0359	.1423	.2285	.2463	.2079	.1465	.0888	.0468	.0215	.0085
	4	.0061	.0514	.1311	.2001	.2252	.2040	.1553	.1014	.0572	.0278
	5	.0008	.0137	.0555	.1201	.1802	.2099	.2008	.1623	.1123	.0667
	6	.0001	.0028	.0180	.0550	.1101	.1649	.1982	.1983	.1684	.1222
	7	.0000	.0004	.0045	.0197	.0524	.1010	.1524	.1889	.1969	.1746
	8	.0000	.0001	.0009	.0055	.0197	.0487	.0923	.1417	.1812	.1964
	9	.0000	.0000	.0001	.0012	.0058	.0185	.0442	.0840	.1318	.1746
	10	.0000	.0000	.0000	.0002	.0014	.0056	.0167	.0392	.0755	.1222
	11	.0000	.0000	.0000	.0000	.0002	.0013	.0049	.0142	.0337	.0667
	12	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0040	.0115	.0278
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0029	.0085
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

(continued on next page)

**Table 1** Exact binomial probabilities  $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$  (continued)

<i>n</i>	<i>k</i>	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
17	0	.4181	.1668	.0631	.0225	.0075	.0023	.0007	.0002	.0000	.0000
	1	.3741	.3150	.1893	.0957	.0426	.0169	.0060	.0019	.0005	.0001
	2	.1575	.2800	.2673	.1914	.1136	.0581	.0260	.0102	.0035	.0010
	3	.0415	.1556	.2359	.2393	.1893	.1245	.0701	.0341	.0144	.0052
	4	.0076	.0605	.1457	.2093	.2209	.1868	.1320	.0796	.0411	.0182
	5	.0010	.0175	.0668	.1361	.1914	.2081	.1849	.1379	.0875	.0472
	6	.0001	.0039	.0236	.0680	.1276	.1784	.1991	.1839	.1432	.0944
	7	.0000	.0007	.0065	.0267	.0668	.1201	.1685	.1927	.1841	.1484
	8	.0000	.0001	.0014	.0084	.0279	.0644	.1134	.1606	.1883	.1855
	9	.0000	.0000	.0003	.0021	.0093	.0276	.0611	.1070	.1540	.1855
	10	.0000	.0000	.0000	.0004	.0025	.0095	.0263	.0571	.1008	.1484
	11	.0000	.0000	.0000	.0001	.0005	.0026	.0090	.0242	.0525	.0944
	12	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0081	.0215	.0472
	13	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0021	.0068	.0182
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
18	0	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000	.0000
	1	.3763	.3002	.1704	.0811	.0338	.0126	.0042	.0012	.0003	.0001
	2	.1683	.2835	.2556	.1723	.0958	.0458	.0190	.0069	.0022	.0006
	3	.0473	.1680	.2406	.2297	.1704	.1046	.0547	.0246	.0095	.0031
	4	.0093	.0700	.1592	.2153	.2130	.1681	.1104	.0614	.0291	.0117
	5	.0014	.0218	.0787	.1507	.1988	.2017	.1664	.1146	.0666	.0327
	6	.0002	.0052	.0301	.0816	.1436	.1873	.1941	.1655	.1181	.0708
	7	.0000	.0010	.0091	.0350	.0820	.1376	.1792	.1892	.1657	.1214
	8	.0000	.0002	.0022	.0120	.0376	.0811	.1327	.1734	.1864	.1669
	9	.0000	.0000	.0004	.0033	.0139	.0386	.0794	.1284	.1694	.1855
	10	.0000	.0000	.0001	.0008	.0042	.0149	.0385	.0771	.1248	.1669
	11	.0000	.0000	.0000	.0001	.0010	.0046	.0151	.0374	.0742	.1214
	12	.0000	.0000	.0000	.0000	.0002	.0012	.0047	.0145	.0354	.0708
	13	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0045	.0134	.0327
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0039	.0117
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0009	.0031
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
19	0	.3774	.1351	.0456	.0144	.0042	.0011	.0003	.0001	.0000	.0000
	1	.3774	.2852	.1529	.0685	.0268	.0093	.0029	.0008	.0002	.0000
	2	.1787	.2852	.2428	.1540	.0803	.0358	.0138	.0046	.0013	.0003
	3	.0533	.1796	.2428	.2182	.1517	.0869	.0422	.0175	.0062	.0018
	4	.0112	.0798	.1714	.2182	.2023	.1491	.0909	.0467	.0203	.0074
	5	.0018	.0266	.0907	.1636	.2023	.1916	.1468	.0933	.0497	.0222
	6	.0002	.0069	.0374	.0955	.1574	.1916	.1844	.1451	.0949	.0518
	7	.0000	.0014	.0122	.0443	.0974	.1525	.1844	.1797	.1443	.0961
	8	.0000	.0002	.0032	.0166	.0487	.0981	.1489	.1797	.1771	.1442
	9	.0000	.0000	.0007	.0051	.0198	.0514	.0980	.1464	.1771	.1762
	10	.0000	.0000	.0001	.0013	.0066	.0220	.0528	.0976	.1449	.1762
	11	.0000	.0000	.0000	.0003	.0018	.0077	.0233	.0532	.0970	.1442
	12	.0000	.0000	.0000	.0000	.0004	.0022	.0083	.0237	.0529	.0961
	13	.0000	.0000	.0000	.0000	.0001	.0005	.0024	.0085	.0233	.0518
	14	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0082	.0222
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0022	.0074
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003

(continued on next page)

**Table 1** Exact binomial probabilities  $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$  (continued)

<i>n</i>	<i>k</i>	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000
	2	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002
	3	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011
	4	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046
	5	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148
	6	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370
	7	.0000	.0020	.0160	.0546	.1124	.1643	.1844	.1659	.1221	.0739
	8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201
	9	.0000	.0001	.0011	.0074	.0271	.0654	.1158	.1597	.1771	.1602
	10	.0000	.0000	.0002	.0020	.0099	.0308	.0686	.1171	.1593	.1762
	11	.0000	.0000	.0000	.0005	.0030	.0120	.0336	.0710	.1185	.1602
	12	.0000	.0000	.0000	.0001	.0008	.0039	.0136	.0355	.0727	.1201
	13	.0000	.0000	.0000	.0000	.0002	.0010	.0045	.0146	.0366	.0739
	14	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0049	.0150	.0370
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0049	.0148
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0046
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

**Table 2** Exact Poisson probabilities  $Pr(X = k) = \frac{e^{-\mu}\mu^k}{k!}$ 

<i>k</i>	$\mu$									
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
0	.6065	.3679	.2231	.1353	.0821	.0498	.0302	.0183	.0111	.0067
1	.3033	.3679	.3347	.2707	.2052	.1494	.1057	.0733	.0500	.0337
2	.0758	.1839	.2510	.2707	.2565	.2240	.1850	.1465	.1125	.0842
3	.0126	.0613	.1255	.1804	.2138	.2240	.2158	.1954	.1687	.1404
4	.0016	.0153	.0471	.0902	.1336	.1680	.1888	.1954	.1898	.1755
5	.0002	.0031	.0141	.0361	.0668	.1008	.1322	.1563	.1708	.1755
6	.0000	.0005	.0035	.0120	.0278	.0504	.0771	.1042	.1281	.1462
7	.0000	.0001	.0008	.0034	.0099	.0216	.0385	.0595	.0824	.1044
8	.0000	.0000	.0001	.0009	.0031	.0081	.0169	.0298	.0463	.0653
9	.0000	.0000	.0000	.0002	.0009	.0027	.0066	.0132	.0232	.0363
10	.0000	.0000	.0000	.0000	.0002	.0008	.0023	.0053	.0104	.0181
11	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0019	.0043	.0082
12	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0034
13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0013
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005
15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
<i>k</i>	$\mu$									
	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
0	.0041	.0025	.0015	.0009	.0006	.0003	.0002	.0001	.0001	.0000
1	.0225	.0149	.0098	.0064	.0041	.0027	.0017	.0011	.0007	.0005
2	.0618	.0446	.0318	.0223	.0156	.0107	.0074	.0050	.0034	.0023
3	.1133	.0892	.0688	.0521	.0389	.0286	.0208	.0150	.0107	.0076

(continued on next page)

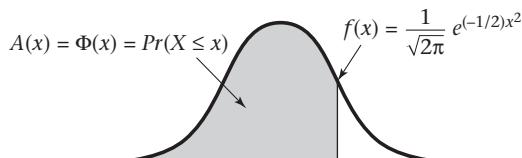
**Table 2** Exact Poisson probabilities  $\Pr(X = k) = \frac{e^{-\mu}\mu^k}{k!}$  (continued)

k	$\mu$									
	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
4	.1558	.1339	.1118	.0912	.0729	.0573	.0443	.0337	.0254	.0189
5	.1714	.1606	.1454	.1277	.1094	.0916	.0752	.0607	.0483	.0378
6	.1571	.1606	.1575	.1490	.1367	.1221	.1066	.0911	.0764	.0631
7	.1234	.1377	.1462	.1490	.1465	.1396	.1294	.1171	.1037	.0901
8	.0849	.1033	.1188	.1304	.1373	.1396	.1375	.1318	.1232	.1126
9	.0519	.0688	.0858	.1014	.1144	.1241	.1299	.1318	.1300	.1251
10	.0285	.0413	.0558	.0710	.0858	.0993	.1104	.1186	.1235	.1251
11	.0143	.0225	.0330	.0452	.0585	.0722	.0853	.0970	.1067	.1137
12	.0065	.0113	.0179	.0263	.0366	.0481	.0604	.0728	.0844	.0948
13	.0028	.0052	.0089	.0142	.0211	.0296	.0395	.0504	.0617	.0729
14	.0011	.0022	.0041	.0071	.0113	.0169	.0240	.0324	.0419	.0521
15	.0004	.0009	.0018	.0033	.0057	.0090	.0136	.0194	.0265	.0347
16	.0001	.0003	.0007	.0014	.0026	.0045	.0072	.0109	.0157	.0217
17	.0000	.0001	.0003	.0006	.0012	.0021	.0036	.0058	.0088	.0128
18	.0000	.0000	.0001	.0002	.0005	.0009	.0017	.0029	.0046	.0071
19	.0000	.0000	.0000	.0001	.0002	.0004	.0008	.0014	.0023	.0037
20	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0006	.0011	.0019
21	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0003	.0005	.0009
22	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0004
23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
k	$\mu$									
	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0003	.0002	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000
2	.0015	.0010	.0007	.0004	.0003	.0002	.0001	.0001	.0001	.0000
3	.0053	.0037	.0026	.0018	.0012	.0008	.0006	.0004	.0003	.0002
4	.0139	.0102	.0074	.0053	.0038	.0027	.0019	.0013	.0009	.0006
5	.0293	.0224	.0170	.0127	.0095	.0070	.0051	.0037	.0027	.0019
6	.0513	.0411	.0325	.0255	.0197	.0152	.0115	.0087	.0065	.0048
7	.0769	.0646	.0535	.0437	.0353	.0281	.0222	.0174	.0135	.0104
8	.1009	.0888	.0769	.0655	.0551	.0457	.0375	.0304	.0244	.0194
9	.1177	.1085	.0982	.0874	.0765	.0661	.0563	.0473	.0394	.0324
10	.1236	.1194	.1129	.1048	.0956	.0859	.0760	.0663	.0571	.0486
11	.1180	.1194	.1181	.1144	.1087	.1015	.0932	.0844	.0753	.0663
12	.1032	.1094	.1131	.1144	.1132	.1099	.1049	.0984	.0910	.0829
13	.0834	.0926	.1001	.1056	.1089	.1099	.1089	.1060	.1014	.0956
14	.0625	.0728	.0822	.0905	.0972	.1021	.1050	.1060	.1051	.1024
15	.0438	.0534	.0630	.0724	.0810	.0885	.0945	.0989	.1016	.1024
16	.0287	.0367	.0453	.0543	.0633	.0719	.0798	.0866	.0920	.0960
17	.0177	.0237	.0306	.0383	.0465	.0550	.0633	.0713	.0785	.0847
18	.0104	.0145	.0196	.0255	.0323	.0397	.0475	.0554	.0632	.0706
19	.0057	.0084	.0119	.0161	.0213	.0272	.0337	.0409	.0483	.0557
20	.0030	.0046	.0068	.0097	.0133	.0177	.0228	.0286	.0350	.0418
21	.0015	.0024	.0037	.0055	.0079	.0109	.0146	.0191	.0242	.0299
22	.0007	.0012	.0020	.0030	.0045	.0065	.0090	.0121	.0159	.0204
23	.0003	.0006	.0010	.0016	.0024	.0037	.0053	.0074	.0100	.0133
24	.0001	.0003	.0005	.0008	.0013	.0020	.0030	.0043	.0061	.0083
25	.0001	.0001	.0002	.0004	.0006	.0010	.0016	.0024	.0035	.0050
26	.0000	.0000	.0001	.0002	.0003	.0005	.0008	.0013	.0020	.0029
27	.0000	.0000	.0000	.0001	.0001	.0002	.0004	.0007	.0011	.0016
28	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0005	.0009

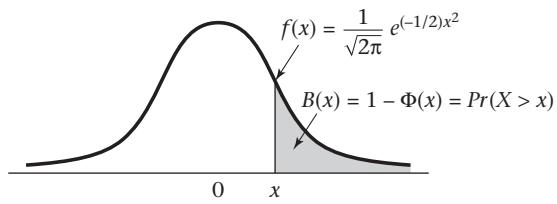
(continued on next page)

**Table 2** Exact Poisson probabilities  $Pr(X = k) = \frac{e^{-\mu}\mu^k}{k!}$  (continued)

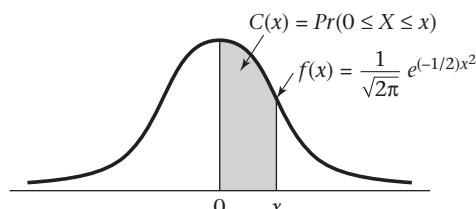
k	$\mu$										
	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0	
29	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0004	
30	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	
31	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	
32	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	
33	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	
k	$\mu$										
	15.5	16.0	16.5	17.0	17.5	18.0	18.5	19.0	19.5	20.0	
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
3	.0001	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
4	.0004	.0003	.0002	.0001	.0001	.0001	.0000	.0000	.0000	.0000	.0000
5	.0014	.0010	.0007	.0005	.0003	.0002	.0002	.0001	.0001	.0001	.0001
6	.0036	.0026	.0019	.0014	.0010	.0007	.0005	.0004	.0003	.0002	
7	.0079	.0060	.0045	.0034	.0025	.0019	.0014	.0010	.0007	.0005	
8	.0153	.0120	.0093	.0072	.0055	.0042	.0031	.0024	.0018	.0013	
9	.0264	.0213	.0171	.0135	.0107	.0083	.0065	.0050	.0038	.0029	
10	.0409	.0341	.0281	.0230	.0186	.0150	.0120	.0095	.0074	.0058	
11	.0577	.0496	.0422	.0355	.0297	.0245	.0201	.0164	.0132	.0106	
12	.0745	.0661	.0580	.0504	.0432	.0368	.0310	.0259	.0214	.0176	
13	.0888	.0814	.0736	.0658	.0582	.0509	.0441	.0378	.0322	.0271	
14	.0983	.0930	.0868	.0800	.0728	.0655	.0583	.0514	.0448	.0387	
15	.1016	.0992	.0955	.0906	.0849	.0786	.0719	.0650	.0582	.0516	
16	.0984	.0992	.0985	.0963	.0929	.0884	.0831	.0772	.0710	.0646	
17	.0897	.0934	.0956	.0963	.0956	.0936	.0904	.0863	.0814	.0760	
18	.0773	.0830	.0876	.0909	.0929	.0936	.0930	.0911	.0882	.0844	
19	.0630	.0699	.0761	.0814	.0856	.0887	.0905	.0911	.0905	.0888	
20	.0489	.0559	.0628	.0692	.0749	.0798	.0837	.0866	.0883	.0888	
21	.0361	.0426	.0493	.0560	.0624	.0684	.0738	.0783	.0820	.0846	
22	.0254	.0310	.0370	.0433	.0496	.0560	.0620	.0676	.0727	.0769	
23	.0171	.0216	.0265	.0320	.0378	.0438	.0499	.0559	.0616	.0669	
24	.0111	.0144	.0182	.0226	.0275	.0328	.0385	.0442	.0500	.0557	
25	.0069	.0092	.0120	.0154	.0193	.0237	.0285	.0336	.0390	.0446	
26	.0041	.0057	.0076	.0101	.0130	.0164	.0202	.0246	.0293	.0343	
27	.0023	.0034	.0047	.0063	.0084	.0109	.0139	.0173	.0211	.0254	
28	.0013	.0019	.0028	.0038	.0053	.0070	.0092	.0117	.0147	.0181	
29	.0007	.0011	.0016	.0023	.0032	.0044	.0058	.0077	.0099	.0125	
30	.0004	.0006	.0009	.0013	.0019	.0026	.0036	.0049	.0064	.0083	
31	.0002	.0003	.0005	.0007	.0010	.0015	.0022	.0030	.0040	.0054	
32	.0001	.0001	.0002	.0004	.0006	.0009	.0012	.0018	.0025	.0034	
33	.0000	.0001	.0001	.0002	.0003	.0005	.0007	.0010	.0015	.0020	
34	.0000	.0000	.0001	.0001	.0002	.0002	.0004	.0006	.0008	.0012	
35	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0005	.0007	
36	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0002	.0003	.0004	
37	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	
40	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	

**Table 3** The normal distribution

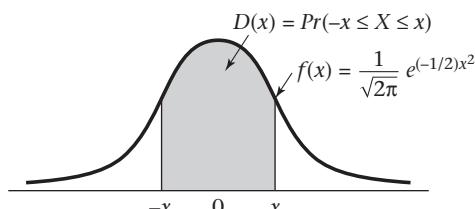
(a)



(b)



(c)



(d)

$x$	$A^a$	$B^b$	$C^c$	$D^d$	$x$	$A$	$B$	$C$	$D$
0.0	.5000	.5000	.0	.0	0.32	.6255	.3745	.1255	.2510
0.01	.5040	.4960	.0040	.0080	0.33	.6293	.3707	.1293	.2586
0.02	.5080	.4920	.0080	.0160	0.34	.6331	.3669	.1331	.2661
0.03	.5120	.4880	.0120	.0239	0.35	.6368	.3632	.1368	.2737
0.04	.5160	.4840	.0160	.0319	0.36	.6406	.3594	.1406	.2812
0.05	.5199	.4801	.0199	.0399	0.37	.6443	.3557	.1443	.2886
0.06	.5239	.4761	.0239	.0478	0.38	.6480	.3520	.1480	.2961
0.07	.5279	.4721	.0279	.0558	0.39	.6517	.3483	.1517	.3035
0.08	.5319	.4681	.0319	.0638	0.40	.6554	.3446	.1554	.3108
0.09	.5359	.4641	.0359	.0717	0.41	.6591	.3409	.1591	.3182
0.10	.5398	.4602	.0398	.0797	0.42	.6628	.3372	.1628	.3255
0.11	.5438	.4562	.0438	.0876	0.43	.6664	.3336	.1664	.3328
0.12	.5478	.4522	.0478	.0955	0.44	.6700	.3300	.1700	.3401
0.13	.5517	.4483	.0517	.1034	0.45	.6736	.3264	.1736	.3473
0.14	.5557	.4443	.0557	.1113	0.46	.6772	.3228	.1772	.3545
0.15	.5596	.4404	.0596	.1192	0.47	.6808	.3192	.1808	.3616
0.16	.5636	.4364	.0636	.1271	0.48	.6844	.3156	.1844	.3688
0.17	.5675	.4325	.0675	.1350	0.49	.6879	.3121	.1879	.3759
0.18	.5714	.4286	.0714	.1428	0.50	.6915	.3085	.1915	.3829
0.19	.5753	.4247	.0753	.1507	0.51	.6950	.3050	.1950	.3899
0.20	.5793	.4207	.0793	.1585	0.52	.6985	.3015	.1985	.3969
0.21	.5832	.4168	.0832	.1663	0.53	.7019	.2981	.2019	.4039
0.22	.5871	.4129	.0871	.1741	0.54	.7054	.2946	.2054	.4108
0.23	.5910	.4090	.0910	.1819	0.55	.7088	.2912	.2088	.4177
0.24	.5948	.4052	.0948	.1897	0.56	.7123	.2877	.2123	.4245
0.25	.5987	.4013	.0987	.1974	0.57	.7157	.2843	.2157	.4313
0.26	.6026	.3974	.1026	.2051	0.58	.7190	.2810	.2190	.4381
0.27	.6064	.3936	.1064	.2128	0.59	.7224	.2776	.2224	.4448
0.28	.6103	.3897	.1103	.2205	0.60	.7257	.2743	.2257	.4515
0.29	.6141	.3859	.1141	.2282	0.61	.7291	.2709	.2291	.4581
0.30	.6179	.3821	.1179	.2358	0.62	.7324	.2676	.2324	.4647
0.31	.6217	.3783	.1217	.2434	0.63	.7357	.2643	.2357	.4713

(continued on next page)

**Table 3** The normal distribution (*continued*)

<i>x</i>	<i>A</i> <sup>a</sup>	<i>B</i> <sup>b</sup>	<i>C</i> <sup>c</sup>	<i>D</i> <sup>d</sup>	<i>x</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
0.64	.7389	.2611	.2389	.4778	1.23	.8907	.1093	.3907	.7813
0.65	.7422	.2578	.2422	.4843	1.24	.8925	.1075	.3925	.7850
0.66	.7454	.2546	.2454	.4907	1.25	.8944	.1056	.3944	.7887
0.67	.7486	.2514	.2486	.4971	1.26	.8962	.1038	.3962	.7923
0.68	.7517	.2483	.2517	.5035	1.27	.8980	.1020	.3980	.7959
0.69	.7549	.2451	.2549	.5098	1.28	.8997	.1003	.3997	.7995
0.70	.7580	.2420	.2580	.5161	1.29	.9015	.0985	.4015	.8029
0.71	.7611	.2389	.2611	.5223	1.30	.9032	.0968	.4032	.8064
0.72	.7642	.2358	.2642	.5285	1.31	.9049	.0951	.4049	.8098
0.73	.7673	.2327	.2673	.5346	1.32	.9066	.0934	.4066	.8132
0.74	.7703	.2297	.2703	.5407	1.33	.9082	.0918	.4082	.8165
0.75	.7734	.2266	.2734	.5467	1.34	.9099	.0901	.4099	.8198
0.76	.7764	.2236	.2764	.5527	1.35	.9115	.0885	.4115	.8230
0.77	.7793	.2207	.2793	.5587	1.36	.9131	.0869	.4131	.8262
0.78	.7823	.2177	.2823	.5646	1.37	.9147	.0853	.4147	.8293
0.79	.7852	.2148	.2852	.5705	1.38	.9162	.0838	.4162	.8324
0.80	.7881	.2119	.2881	.5763	1.39	.9177	.0823	.4177	.8355
0.81	.7910	.2090	.2910	.5821	1.40	.9192	.0808	.4192	.8385
0.82	.7939	.2061	.2939	.5878	1.41	.9207	.0793	.4207	.8415
0.83	.7967	.2033	.2967	.5935	1.42	.9222	.0778	.4222	.8444
0.84	.7995	.2005	.2995	.5991	1.43	.9236	.0764	.4236	.8473
0.85	.8023	.1977	.3023	.6047	1.44	.9251	.0749	.4251	.8501
0.86	.8051	.1949	.3051	.6102	1.45	.9265	.0735	.4265	.8529
0.87	.8078	.1922	.3078	.6157	1.46	.9279	.0721	.4279	.8557
0.88	.8106	.1894	.3106	.6211	1.47	.9292	.0708	.4292	.8584
0.89	.8133	.1867	.3133	.6265	1.48	.9306	.0694	.4306	.8611
0.90	.8159	.1841	.3159	.6319	1.49	.9319	.0681	.4319	.8638
0.91	.8186	.1814	.3186	.6372	1.50	.9332	.0668	.4332	.8664
0.92	.8212	.1788	.3212	.6424	1.51	.9345	.0655	.4345	.8690
0.93	.8238	.1762	.3238	.6476	1.52	.9357	.0643	.4357	.8715
0.94	.8264	.1736	.3264	.6528	1.53	.9370	.0630	.4370	.8740
0.95	.8289	.1711	.3289	.6579	1.54	.9382	.0618	.4382	.8764
0.96	.8315	.1685	.3315	.6629	1.55	.9394	.0606	.4394	.8789
0.97	.8340	.1660	.3340	.6680	1.56	.9406	.0594	.4406	.8812
0.98	.8365	.1635	.3365	.6729	1.57	.9418	.0582	.4418	.8836
0.99	.8389	.1611	.3389	.6778	1.58	.9429	.0571	.4429	.8859
1.00	.8413	.1587	.3413	.6827	1.59	.9441	.0559	.4441	.8882
1.01	.8438	.1562	.3438	.6875	1.60	.9452	.0548	.4452	.8904
1.02	.8461	.1539	.3461	.6923	1.61	.9463	.0537	.4463	.8926
1.03	.8485	.1515	.3485	.6970	1.62	.9474	.0526	.4474	.8948
1.04	.8508	.1492	.3508	.7017	1.63	.9484	.0516	.4484	.8969
1.05	.8531	.1469	.3531	.7063	1.64	.9495	.0505	.4495	.8990
1.06	.8554	.1446	.3554	.7109	1.65	.9505	.0495	.4505	.9011
1.07	.8577	.1423	.3577	.7154	1.66	.9515	.0485	.4515	.9031
1.08	.8599	.1401	.3599	.7199	1.67	.9525	.0475	.4525	.9051
1.09	.8621	.1379	.3621	.7243	1.68	.9535	.0465	.4535	.9070
1.10	.8643	.1357	.3643	.7287	1.69	.9545	.0455	.4545	.9090
1.11	.8665	.1335	.3665	.7330	1.70	.9554	.0446	.4554	.9109
1.12	.8686	.1314	.3686	.7373	1.71	.9564	.0436	.4564	.9127
1.13	.8708	.1292	.3708	.7415	1.72	.9573	.0427	.4573	.9146
1.14	.8729	.1271	.3729	.7457	1.73	.9582	.0418	.4582	.9164
1.15	.8749	.1251	.3749	.7499	1.74	.9591	.0409	.4591	.9181
1.16	.8770	.1230	.3770	.7540	1.75	.9599	.0401	.4599	.9199
1.17	.8790	.1210	.3790	.7580	1.76	.9608	.0392	.4608	.9216
1.18	.8810	.1190	.3810	.7620	1.77	.9616	.0384	.4616	.9233
1.19	.8830	.1170	.3830	.7660	1.78	.9625	.0375	.4625	.9249
1.20	.8849	.1151	.3849	.7699	1.79	.9633	.0367	.4633	.9265
1.21	.8869	.1131	.3869	.7737	1.80	.9641	.0359	.4641	.9281
1.22	.8888	.1112	.3888	.7775	1.81	.9649	.0351	.4649	.9297

(continued on next page)

**Table 3** The normal distribution (*continued*)

<i>x</i>	<i>A</i> <sup>a</sup>	<i>B</i> <sup>b</sup>	<i>C</i> <sup>c</sup>	<i>D</i> <sup>d</sup>	<i>x</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
1.82	.9656	.0344	.4656	.9312	2.39	.9916	.0084	.4916	.9832
1.83	.9664	.0336	.4664	.9327	2.40	.9918	.0082	.4918	.9836
1.84	.9671	.0329	.4671	.9342	2.41	.9920	.0080	.4920	.9840
1.85	.9678	.0322	.4678	.9357	2.42	.9922	.0078	.4922	.9845
1.86	.9686	.0314	.4686	.9371	2.43	.9925	.0075	.4925	.9849
1.87	.9693	.0307	.4693	.9385	2.44	.9927	.0073	.4927	.9853
1.88	.9699	.0301	.4699	.9399	2.45	.9929	.0071	.4929	.9857
1.89	.9706	.0294	.4706	.9412	2.46	.9931	.0069	.4931	.9861
1.90	.9713	.0287	.4713	.9426	2.47	.9932	.0068	.4932	.9865
1.91	.9719	.0281	.4719	.9439	2.48	.9934	.0066	.4934	.9869
1.92	.9726	.0274	.4726	.9451	2.49	.9936	.0064	.4936	.9872
1.93	.9732	.0268	.4732	.9464	2.50	.9938	.0062	.4938	.9876
1.94	.9738	.0262	.4738	.9476	2.51	.9940	.0060	.4940	.9879
1.95	.9744	.0256	.4744	.9488	2.52	.9941	.0059	.4941	.9883
1.96	.9750	.0250	.4750	.9500	2.53	.9943	.0057	.4943	.9886
1.97	.9756	.0244	.4756	.9512	2.54	.9945	.0055	.4945	.9889
1.98	.9761	.0239	.4761	.9523	2.55	.9946	.0054	.4946	.9892
1.99	.9767	.0233	.4767	.9534	2.56	.9948	.0052	.4948	.9895
2.00	.9772	.0228	.4772	.9545	2.57	.9949	.0051	.4949	.9898
2.01	.9778	.0222	.4778	.9556	2.58	.9951	.0049	.4951	.9901
2.02	.9783	.0217	.4783	.9566	2.59	.9952	.0048	.4952	.9904
2.03	.9788	.0212	.4788	.9576	2.60	.9953	.0047	.4953	.9907
2.04	.9793	.0207	.4793	.9586	2.61	.9955	.0045	.4955	.9909
2.05	.9798	.0202	.4798	.9596	2.62	.9956	.0044	.4956	.9912
2.06	.9803	.0197	.4803	.9606	2.63	.9957	.0043	.4957	.9915
2.07	.9808	.0192	.4808	.9615	2.64	.9959	.0041	.4959	.9917
2.08	.9812	.0188	.4812	.9625	2.65	.9960	.0040	.4960	.9920
2.09	.9817	.0183	.4817	.9634	2.66	.9961	.0039	.4961	.9922
2.10	.9821	.0179	.4821	.9643	2.67	.9962	.0038	.4962	.9924
2.11	.9826	.0174	.4826	.9651	2.68	.9963	.0037	.4963	.9926
2.12	.9830	.0170	.4830	.9660	2.69	.9964	.0036	.4964	.9929
2.13	.9834	.0166	.4834	.9668	2.70	.9965	.0035	.4965	.9931
2.14	.9838	.0162	.4838	.9676	2.71	.9966	.0034	.4966	.9933
2.15	.9842	.0158	.4842	.9684	2.72	.9967	.0033	.4967	.9935
2.16	.9846	.0154	.4846	.9692	2.73	.9968	.0032	.4968	.9937
2.17	.9850	.0150	.4850	.9700	2.74	.9969	.0031	.4969	.9939
2.18	.9854	.0146	.4854	.9707	2.75	.9970	.0030	.4970	.9940
2.19	.9857	.0143	.4857	.9715	2.76	.9971	.0029	.4971	.9942
2.20	.9861	.0139	.4861	.9722	2.77	.9972	.0028	.4972	.9944
2.21	.9864	.0136	.4864	.9729	2.78	.9973	.0027	.4973	.9946
2.22	.9868	.0132	.4868	.9736	2.79	.9974	.0026	.4974	.9947
2.23	.9871	.0129	.4871	.9743	2.80	.9974	.0026	.4974	.9949
2.24	.9875	.0125	.4875	.9749	2.81	.9975	.0025	.4975	.9950
2.25	.9878	.0122	.4878	.9756	2.82	.9976	.0024	.4976	.9952
2.26	.9881	.0119	.4881	.9762	2.83	.9977	.0023	.4977	.9953
2.27	.9884	.0116	.4884	.9768	2.84	.9977	.0023	.4977	.9955
2.28	.9887	.0113	.4887	.9774	2.85	.9978	.0022	.4978	.9956
2.29	.9890	.0110	.4890	.9780	2.86	.9979	.0021	.4979	.9958
2.30	.9893	.0107	.4893	.9786	2.87	.9979	.0021	.4979	.9959
2.31	.9896	.0104	.4896	.9791	2.88	.9980	.0020	.4980	.9960
2.32	.9898	.0102	.4898	.9797	2.89	.9981	.0019	.4981	.9961
2.33	.9901	.0099	.4901	.9802	2.90	.9981	.0019	.4981	.9963
2.34	.9904	.0096	.4904	.9807	2.91	.9982	.0018	.4982	.9964
2.35	.9906	.0094	.4906	.9812	2.92	.9982	.0018	.4982	.9965
2.36	.9909	.0091	.4909	.9817	2.93	.9983	.0017	.4983	.9966
2.37	.9911	.0089	.4911	.9822	2.94	.9984	.0016	.4984	.9967
2.38	.9913	.0087	.4913	.9827	2.95	.9984	.0016	.4984	.9968

(continued on next page)

**Table 3** The normal distribution (continued)

<i>x</i>	<i>A</i> <sup>a</sup>	<i>B</i> <sup>b</sup>	<i>C</i> <sup>c</sup>	<i>D</i> <sup>d</sup>	<i>x</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
2.96	.9985	.0015	.4985	.9969	3.49	.9998	.0002	.4998	.9995
2.97	.9985	.0015	.4985	.9970	3.50	.9998	.0002	.4998	.9995
2.98	.9986	.0014	.4986	.9971	3.51	.9998	.0002	.4998	.9996
2.99	.9986	.0014	.4986	.9972	3.52	.9998	.0002	.4998	.9996
3.00	.9987	.0013	.4987	.9973	3.53	.9998	.0002	.4998	.9996
3.01	.9987	.0013	.4987	.9974	3.54	.9998	.0002	.4998	.9996
3.02	.9987	.0013	.4987	.9975	3.55	.9998	.0002	.4998	.9996
3.03	.9988	.0012	.4988	.9976	3.56	.9998	.0002	.4998	.9996
3.04	.9988	.0012	.4988	.9976	3.57	.9998	.0002	.4998	.9996
3.05	.9989	.0011	.4989	.9977	3.58	.9998	.0002	.4998	.9997
3.06	.9989	.0011	.4989	.9978	3.59	.9998	.0002	.4998	.9997
3.07	.9989	.0011	.4989	.9979	3.60	.9998	.0002	.4998	.9997
3.08	.9990	.0010	.4990	.9979	3.61	.9998	.0002	.4998	.9997
3.09	.9990	.0010	.4990	.9980	3.62	.9999	.0001	.4999	.9997
3.10	.9990	.0010	.4990	.9981	3.63	.9999	.0001	.4999	.9997
3.11	.9991	.0009	.4991	.9981	3.64	.9999	.0001	.4999	.9997
3.12	.9991	.0009	.4991	.9982	3.65	.9999	.0001	.4999	.9997
3.13	.9991	.0009	.4991	.9983	3.66	.9999	.0001	.4999	.9997
3.14	.9992	.0008	.4992	.9983	3.67	.9999	.0001	.4999	.9998
3.15	.9992	.0008	.4992	.9984	3.68	.9999	.0001	.4999	.9998
3.16	.9992	.0008	.4992	.9984	3.69	.9999	.0001	.4999	.9998
3.17	.9992	.0008	.4992	.9985	3.70	.9999	.0001	.4999	.9998
3.18	.9993	.0007	.4993	.9985	3.71	.9999	.0001	.4999	.9998
3.19	.9993	.0007	.4993	.9986	3.72	.9999	.0001	.4999	.9998
3.20	.9993	.0007	.4993	.9986	3.73	.9999	.0001	.4999	.9998
3.21	.9993	.0007	.4993	.9987	3.74	.9999	.0001	.4999	.9998
3.22	.9994	.0006	.4994	.9987	3.75	.9999	.0001	.4999	.9998
3.23	.9994	.0006	.4994	.9988	3.76	.9999	.0001	.4999	.9998
3.24	.9994	.0006	.4994	.9988	3.77	.9999	.0001	.4999	.9998
3.25	.9994	.0006	.4994	.9988	3.78	.9999	.0001	.4999	.9998
3.26	.9994	.0006	.4994	.9989	3.79	.9999	.0001	.4999	.9998
3.27	.9995	.0005	.4995	.9989	3.80	.9999	.0001	.4999	.9999
3.28	.9995	.0005	.4995	.9990	3.81	.9999	.0001	.4999	.9999
3.29	.9995	.0005	.4995	.9990	3.82	.9999	.0001	.4999	.9999
3.30	.9995	.0005	.4995	.9990	3.83	.9999	.0001	.4999	.9999
3.31	.9995	.0005	.4995	.9991	3.84	.9999	.0001	.4999	.9999
3.32	.9995	.0005	.4995	.9991	3.85	.9999	.0001	.4999	.9999
3.33	.9996	.0004	.4996	.9991	3.86	.9999	.0001	.4999	.9999
3.34	.9996	.0004	.4996	.9992	3.87	.9999	.0001	.4999	.9999
3.35	.9996	.0004	.4996	.9992	3.88	.9999	.0001	.4999	.9999
3.36	.9996	.0004	.4996	.9992	3.89	.9999	.0001	.4999	.9999
3.37	.9996	.0004	.4996	.9992	3.90	1.0000	.0000	.5000	.9999
3.38	.9996	.0004	.4996	.9993	3.91	1.0000	.0000	.5000	.9999
3.39	.9997	.0003	.4997	.9993	3.92	1.0000	.0000	.5000	.9999
3.40	.9997	.0003	.4997	.9993	3.93	1.0000	.0000	.5000	.9999
3.42	.9997	.0003	.4997	.9994	3.94	1.0000	.0000	.5000	.9999
3.43	.9997	.0003	.4997	.9994	3.95	1.0000	.0000	.5000	.9999
3.45	.9997	.0003	.4997	.9994	3.96	1.0000	.0000	.5000	.9999
3.46	.9997	.0003	.4997	.9995	3.97	1.0000	.0000	.5000	.9999
3.47	.9997	.0003	.4997	.9995	3.98	1.0000	.0000	.5000	.9999
3.48	.9997	.0003	.4997	.9995	3.99	1.0000	.0000	.5000	.9999

<sup>a</sup> $A(x) = \Phi(x) = P(X \leq x)$ , where  $X$  is a standard normal distribution.<sup>b</sup> $B(x) = 1 - \Phi(x) = P(X > x)$ , where  $X$  is a standard normal distribution.<sup>c</sup> $C(x) = P(0 \leq X \leq x)$ , where  $X$  is a standard normal distribution.<sup>d</sup> $D(x) = P(-x \leq X \leq x)$ , where  $X$  is a standard normal distribution.

**Table 4 Table of 1000 random digits**

01	32924	22324	18125	09077	26	96772	16443	39877	04653
02	54632	90374	94143	49295	27	52167	21038	14338	01395
03	88720	43035	97081	83373	28	69644	37198	00028	98195
04	21727	11904	41513	31653	29	71011	62004	81712	87536
05	80985	70799	57975	69282	30	31217	75877	85366	55500
06	40412	58826	94868	52632	31	64990	98735	02999	35521
07	43918	56807	75218	46077	32	48417	23569	59307	46550
08	26513	47480	77410	47741	33	07900	65059	48592	44087
09	18164	35784	44255	30124	34	74526	32601	24482	16981
10	39446	01375	75264	51173	35	51056	04402	58353	37332
11	16638	04680	98617	90298	36	39005	93458	63143	21817
12	16872	94749	44012	48884	37	67883	76343	78155	67733
13	65419	87092	78596	91512	38	06014	60999	87226	36071
14	05207	36702	56804	10498	39	93147	88766	04148	42471
15	78807	79243	13729	81222	40	01099	95731	47622	13294
16	69341	79028	64253	80447	41	89252	01201	58138	13809
17	41871	17566	61200	15994	42	41766	57239	50251	64675
18	25758	04625	43226	32986	43	92736	77800	81996	45646
19	06604	94486	40174	10742	44	45118	36600	68977	68831
20	82259	56512	48945	18183	45	73457	01579	00378	70197
21	07895	37090	50627	71320	46	49465	85251	42914	17277
22	59836	71148	42320	67816	47	15745	37285	23768	39302
23	57133	76610	89104	30481	48	28760	81331	78265	60690
24	76964	57126	87174	61025	49	82193	32787	70451	91141
25	27694	17145	32439	68245	50	89664	50242	12382	39379

**Table 5** Percentage points of the *t* distribution ( $t_{d,u}$ )<sup>a</sup>

Degrees of freedom, <i>d</i>	<i>u</i>								
	.75	.80	.85	.90	.95	.975	.99	.995	.9995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$\infty$	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

<sup>a</sup>The *u*th percentile of a *t* distribution with *d* degrees of freedom.

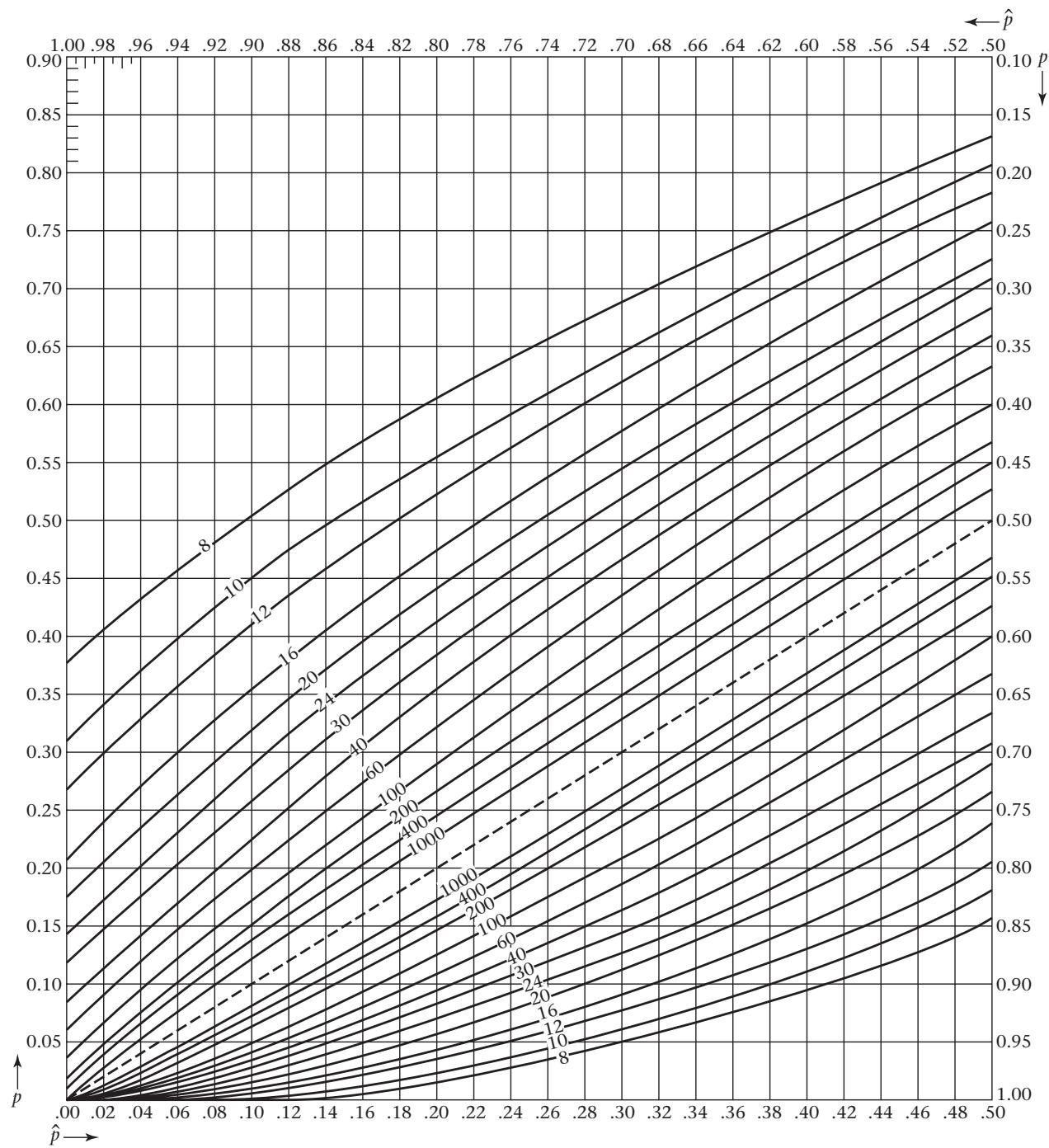
Source: Table 5 is taken from Table III of Fisher and Yates: "Statistical Tables for Biological, Agricultural and Medical Research," published by Longman Group Ltd., London (previously published by Oliver and Boyd Ltd., Edinburgh). Reprinted by permission of Pearson Education Ltd.

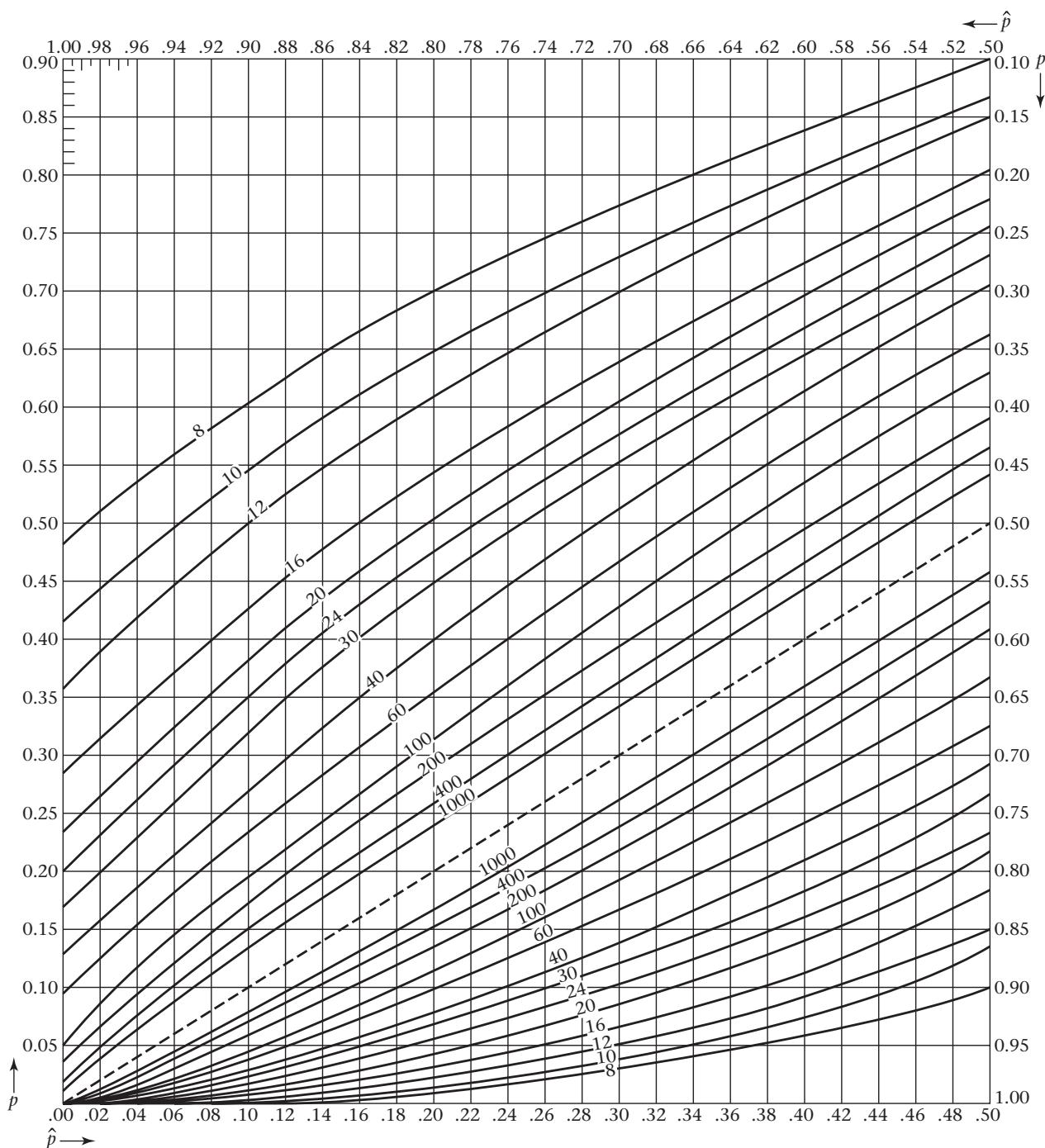
**Table 6 Percentage points of the chi-square distribution ( $\chi^2_{d,u}$ )<sup>a</sup>**

d	u													
	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995	.999
1	0.0 <sup>b</sup> 393 <sup>b</sup>	0.0 <sup>b</sup> 157 <sup>c</sup>	0.0 <sup>b</sup> 982 <sup>d</sup>	0.00393	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.0100	0.0201	0.0506	0.103	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.81
3	0.0717	0.115	0.216	0.352	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

<sup>a</sup> $\chi^2_{d,u}$  = *u*th percentile of a  $\chi^2$  distribution with *d* degrees of freedom.<sup>b</sup> = 0.0000393<sup>c</sup> = 0.000157<sup>d</sup> = 0.000982Source: Reproduced in part with permission of the Biometrika Trustees, from Table 3 of *Biometrika Tables for Statisticians*, Volume 2, edited by E. S. Pearson and H. O. Hartley, published for the Biometrika Trustees, Cambridge University Press, Cambridge, England, 1972.

**Table 7a** Exact two-sided 100%  $\times (1 - \alpha)$  confidence limits for binomial proportions ( $\alpha = .05$ )



**Table 7b** Exact two-sided 100%  $\times (1 - \alpha)$  confidence limits for binomial proportions ( $\alpha = .01$ )

Source: Tables 7a and 7b have been reproduced with permission of Biometrika Trustees, from Table 41 of *Biometrika Tables for Statisticians*, 3rd edition, Volume 1, edited by E. S. Pearson and H. O. Hartley. Published for the Biometrika Trustees, Cambridge, 1966.

**Table 8** Confidence limits for the expectation of a Poisson variable ( $\mu$ )

(1 - $\alpha$ )	Confidence level ( $1 - \alpha$ )											
	0.998		0.99		0.98		0.95		0.90		(1 - 2 $\alpha$ )	
	x	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	x
<b>0</b>	0.00000	6.91	0.00000	5.30	0.0000	4.61	0.0000	3.69	0.0000	3.00	0.00000	<b>0</b>
<b>1</b>	.00100	9.23	.00501	7.43	.0101	6.64	.0253	5.57	.0513	4.74	.00100	<b>1</b>
<b>2</b>	.0454	11.23	.103	9.27	.149	8.41	.242	7.22	.355	6.30	.0454	<b>2</b>
<b>3</b>	.191	13.06	.338	10.98	.436	10.05	.619	8.77	.818	7.75	.191	<b>3</b>
<b>4</b>	.429	14.79	.672	12.59	.823	11.60	1.09	10.24	1.37	9.15	.429	<b>4</b>
<b>5</b>	0.739	16.45	1.08	14.15	1.28	13.11	1.62	11.67	1.97	10.51	0.739	<b>5</b>
<b>6</b>	1.11	18.06	1.54	15.66	1.79	14.57	2.20	13.06	2.61	11.84	1.11	<b>6</b>
<b>7</b>	1.52	19.63	2.04	17.13	2.33	16.00	2.81	14.42	3.29	13.15	1.52	<b>7</b>
<b>8</b>	1.97	21.16	2.57	18.58	2.91	17.40	3.45	15.76	3.98	14.43	1.97	<b>8</b>
<b>9</b>	2.45	22.66	3.13	20.00	3.51	18.78	4.12	17.08	4.70	15.71	2.45	<b>9</b>
<b>10</b>	2.96	24.13	3.72	21.40	4.13	20.14	4.80	18.39	5.43	16.96	2.96	<b>10</b>
<b>11</b>	3.49	25.59	4.32	22.78	4.77	21.49	5.49	19.68	6.17	18.21	3.49	<b>11</b>
<b>12</b>	4.04	27.03	4.94	24.14	5.43	22.82	6.20	20.96	6.92	19.44	4.04	<b>12</b>
<b>13</b>	4.61	28.45	5.58	25.50	6.10	24.14	6.92	22.23	7.69	20.67	4.61	<b>13</b>
<b>14</b>	5.20	29.85	6.23	26.84	6.78	25.45	7.65	23.49	8.46	21.89	5.20	<b>14</b>
<b>15</b>	5.79	31.24	6.89	28.16	7.48	26.74	8.40	24.74	9.25	23.10	5.79	<b>15</b>
<b>16</b>	6.41	32.62	7.57	29.48	8.18	28.03	9.15	25.98	10.04	24.30	6.41	<b>16</b>
<b>17</b>	7.03	33.99	8.25	30.79	8.89	29.31	9.90	27.22	10.83	25.50	7.03	<b>17</b>
<b>18</b>	7.66	35.35	8.94	32.09	9.62	30.58	10.67	28.45	11.63	26.69	7.66	<b>18</b>
<b>19</b>	8.31	36.70	9.64	33.38	10.35	31.85	11.44	29.67	12.44	27.88	8.31	<b>19</b>
<b>20</b>	8.96	38.04	10.35	34.67	11.08	33.10	12.22	30.89	13.25	29.06	8.96	<b>20</b>
<b>21</b>	9.62	39.38	11.07	35.95	11.82	34.36	13.00	32.10	14.07	30.24	9.62	<b>21</b>
<b>22</b>	10.29	40.70	11.79	37.22	12.57	35.60	13.79	33.31	14.89	31.42	10.29	<b>22</b>
<b>23</b>	10.96	42.02	12.52	38.48	13.33	36.84	14.58	34.51	15.72	32.59	10.96	<b>23</b>
<b>24</b>	11.65	43.33	13.25	39.74	14.09	38.08	15.38	35.71	16.55	33.75	11.65	<b>24</b>
<b>25</b>	12.34	44.64	14.00	41.00	14.85	39.31	16.18	36.90	17.38	34.92	12.34	<b>25</b>
<b>26</b>	13.03	45.94	14.74	42.25	15.62	40.53	16.98	38.10	18.22	36.08	13.03	<b>26</b>
<b>27</b>	13.73	47.23	15.49	43.50	16.40	41.76	17.79	39.28	19.06	37.23	13.73	<b>27</b>
<b>28</b>	14.44	48.52	16.24	44.74	17.17	42.98	18.61	40.47	19.90	38.39	14.44	<b>28</b>
<b>29</b>	15.15	49.80	17.00	45.98	17.96	44.19	19.42	41.65	20.75	39.54	15.15	<b>29</b>
<b>30</b>	15.87	51.08	17.77	47.21	18.74	45.40	20.24	42.83	21.59	40.69	15.87	<b>30</b>
<b>35</b>	19.52	57.42	21.64	53.32	22.72	51.41	24.38	48.68	25.87	46.40	19.52	<b>35</b>
<b>40</b>	23.26	63.66	25.59	59.36	26.77	57.35	28.58	54.47	30.20	52.07	23.26	<b>40</b>
<b>45</b>	27.08	69.83	29.60	65.34	30.88	63.23	32.82	60.21	34.56	57.69	27.08	<b>45</b>
<b>50</b>	30.96	75.94	33.66	71.27	35.03	69.07	37.11	65.92	38.96	63.29	30.96	<b>50</b>

Note: If  $X$  is the random variable denoting the observed number of events and  $\mu_1, \mu_2$  are the lower and upper confidence limits for its expectation,  $\mu$ , then  $P(\mu_1 \leq \mu \leq \mu_2) = 1 - \alpha$ .

Source: *Biometrika Tables for Statisticians*, 3rd edition, Volume 1, edited by E. S. Pearson and H. O. Hartley. Published for the Biometrika Trustees, Cambridge University Press, Cambridge, England, 1966.

**Table 9 Percentage points of the *F* distribution ( $F_{d_1, d_2, p}$ )**

		df for numerator, $d_1$										
		df for numerator, $d_1$										
$d_2$	$p$	1	2	3	4	5	6	7	8	12	24	$\infty$
1	.90	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	60.71	62.00	63.33
	.95	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	243.9	249.1	254.3
	.975	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	976.7	997.2	1018.
	.99	4052.	5000.	5403.	5625.	5764.	5859.	5928.	5981.	6106.	6235.	6366.
	.995	16211.	20000.	21615.	22500.	23056.	23437.	23715.	23925.	24426.	24940.	25464.
	.999	405280.	500000.	540380.	562500.	576400.	585940.	592870.	598140.	610670.	623500.	636620.
2	.90	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.41	9.45	9.49
	.95	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.41	19.45	19.50
	.975	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.42	39.46	39.50
	.99	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.42	99.46	99.50
	.995	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.5	199.5
	.999	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.5	999.5
3	.90	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.22	5.18	5.13
	.95	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.74	8.64	8.53
	.975	17.44	16.04	15.44	15.10	14.88	14.74	14.62	14.54	14.34	14.12	13.90
	.99	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.05	26.60	26.13
	.995	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13	43.39	42.62	41.83
	.999	167.00	148.5	141.1	137.1	134.6	132.8	131.6	130.6	128.3	125.9	123.5
4	.90	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.90	3.83	3.76
	.95	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.91	5.77	5.63
	.975	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.75	8.51	8.26
	.99	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.37	13.93	13.46
	.995	31.33	26.28	24.26	23.16	22.46	21.98	21.62	21.35	20.70	20.03	19.32
	.999	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	47.41	45.77	44.05
5	.90	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.27	3.19	3.10
	.95	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.68	4.53	4.36
	.975	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.52	6.28	6.02
	.99	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	9.89	9.47	9.02
	.995	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.38	12.78	12.14
	.999	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	26.42	25.13	23.79
6	.90	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.90	2.82	2.72
	.95	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.00	3.84	3.67
	.975	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.37	5.12	4.85
	.99	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.72	7.31	6.88
	.995	18.64	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.03	9.47	8.88
	.999	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	17.99	16.90	15.75
7	.90	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.67	2.58	2.47
	.95	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.57	3.41	3.23
	.975	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.67	4.42	4.14
	.99	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.47	6.07	5.65
	.995	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.18	7.65	7.08
	.999	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	13.71	12.73	11.70
8	.90	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.50	2.40	2.29
	.95	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.28	3.12	2.93
	.975	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.20	3.95	3.67
	.99	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.67	5.28	4.86
	.995	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.01	6.50	5.95
	.999	25.42	18.49	15.83	14.39	13.49	12.86	12.40	12.04	11.19	10.30	9.33
9	.90	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.38	2.28	2.16
	.95	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.07	2.90	2.71
	.975	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.87	3.61	3.33
	.99	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.11	4.73	4.31
	.995	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.23	5.73	5.19
	.999	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	9.57	8.72	7.81
10	.90	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.28	2.18	2.06
	.95	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.91	2.74	2.54
	.975	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.62	3.37	3.08
	.99	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.71	4.33	3.91
	.995	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.66	5.17	4.64
	.999	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.45	7.64	6.76
12	.90	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.15	2.04	1.90
	.95	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.69	2.51	2.30
	.975	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.28	3.02	2.72

(continued on next page)

**Table 9 Percentage points of the  $F$  distribution ( $F_{d_1, d_2, p}$ ) (continued)**

df for denominator, $d_2$		df for numerator, $d_1$										
		1	2	3	4	5	6	7	8	12	24	
14	.99	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.16	3.78	3.36
	.995	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	4.91	4.43	3.90
	.999	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.00	6.25	5.42
	.90	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.05	1.94	1.80
	.95	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.53	2.35	2.13
	.975	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.05	2.79	2.49
16	.99	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.80	3.43	3.00
	.995	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.43	3.96	3.44
	.999	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.13	5.41	4.60
	.90	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	1.99	1.87	1.72
	.95	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.42	2.24	2.01
	.975	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	2.89	2.63	2.32
18	.99	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.55	3.18	2.75
	.995	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.10	3.64	3.11
	.999	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.55	4.85	4.06
	.90	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	1.93	1.81	1.66
	.95	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.34	2.15	1.92
	.975	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.77	2.50	2.19
20	.99	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.37	3.00	2.57
	.995	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	3.86	3.40	2.87
	.999	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.13	4.45	3.67
	.90	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.89	1.77	1.61
	.95	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.28	2.08	1.84
	.975	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.68	2.41	2.09
30	.99	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.23	2.86	2.42
	.995	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.68	3.22	2.69
	.999	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	4.82	4.15	3.38
	.90	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.77	1.64	1.46
	.95	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.09	1.89	1.62
	.975	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.41	2.14	1.79
40	.99	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	2.84	2.47	2.01
	.995	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.18	2.73	2.18
	.999	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.00	3.36	2.59
	.90	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.71	1.57	1.38
	.95	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.00	1.79	1.51
	.975	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.29	2.01	1.64
60	.99	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.66	2.29	1.80
	.995	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	2.95	2.50	1.93
	.999	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	3.64	3.01	2.23
	.90	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.66	1.51	1.29
	.95	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.92	1.70	1.39
	.975	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.17	1.88	1.48
120	.99	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.50	2.12	1.60
	.995	8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	2.74	2.29	1.69
	.999	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.32	2.69	1.89
	.90	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.60	1.45	1.19
	.95	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.83	1.61	1.25
	.975	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.05	1.76	1.31
$\infty$	.99	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.34	1.95	1.38
	.995	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.54	2.09	1.43
	.999	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.02	2.40	1.54
	.90	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.55	1.38	1.00
	.95	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.75	1.52	1.00
	.975	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	1.94	1.64	1.00
$\infty$	.99	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.18	1.79	1.00
	.995	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.36	1.90	1.00
	.999	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	2.74	2.13	1.00

Note:  $F_{d_1, d_2, p}$  =  $p$ th percentile of an  $F$  distribution with  $d_1$  and  $d_2$  degrees of freedom.

Source: This table has been reproduced in part with the permission of the Biometrika Trustees, from *Biometrika Tables for Statisticians*, Volume 2, edited by E. S. Pearson and H. O. Hartley, published for the Biometrika Trustees, Cambridge University Press, Cambridge, England, 1972.

**Table 10 Critical values for the ESD (Extreme Studentized Deviate) outlier statistic ( $ESD_{n,1-\alpha}$ ,  $\alpha = .05, .01$ )**

n	1 - $\alpha$		n	1 - $\alpha$	
	.95	.99		.95	.99
5	1.72	1.76	25	2.82	3.14
6	1.89	1.97	26	2.84	3.16
7	2.02	2.14	27	2.86	3.18
8	2.13	2.28	28	2.88	3.20
9	2.21	2.39	29	2.89	3.22
10	2.29	2.48	30	2.91	3.24
11	2.36	2.56	35	2.98	3.32
12	2.41	2.64	40	3.04	3.38
13	2.46	2.70	45	3.09	3.44
14	2.51	2.75	50	3.13	3.48
15	2.55	2.81	60	3.20	3.56
16	2.59	2.85	70	3.26	3.62
17	2.62	2.90	80	3.31	3.67
18	2.65	2.93	90	3.35	3.72
19	2.68	2.97	100	3.38	3.75
20	2.71	3.00	150	3.52	3.89
21	2.73	3.03	200	3.61	3.98
22	2.76	3.06	300	3.72	4.09
23	2.78	3.08	400	3.80	4.17
24	2.80	3.11	500	3.86	4.23

Note: For values of  $n$  not found in the table, the percentiles can be evaluated using the formula  $ESD_{n,1-\alpha} =$

$$\frac{t_{n-2,p}(n-1)}{\sqrt{n(n-2+t_{n-2,p}^2)}} \text{ where } p = 1 - [\alpha/(2n)].$$

**Table 11 Two-tailed critical values for the Wilcoxon signed-rank test**

n <sup>a</sup>	.10		.05		.02		.01	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
1	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—
5	0	15	—	—	—	—	—	—
6	2	19	0	21	—	—	—	—
7	3	25	2	26	0	28	—	—
8	5	31	3	33	1	35	0	36
9	8	37	5	40	3	42	1	44
10	10	45	8	47	5	50	3	52
11	13	53	10	56	7	59	5	61
12	17	61	13	65	9	69	7	71
13	21	70	17	74	12	79	9	82
14	25	80	21	84	15	90	12	93
15	30	90	25	95	19	101	15	105

<sup>a</sup>n = number of untied pairs.

Source: Figures from "Documenta Geigy Scientific Tables," 6th edition. Reprinted with the kind permission of CIBA-GEIGY Limited, Basel, Switzerland.

**Table 12** Two-tailed critical values for the Wilcoxon rank-sum test

$\alpha = .10$ $n_1^a$									$\alpha = .05$ $n_1$									
$n_2^b$	4	5	6	7	8	9	4	5	6	7	8	9	4	5	6	7	8	9
	$T_l^c$	$T_r^d$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$
4	11–25	17–33	24–42	32–52	41–63	51–75	10–26	16–34	23–43	31–53	40–64	49–77						
5	12–28	19–36	26–46	34–57	44–68	54–81	11–29	17–38	24–48	33–58	42–70	52–83						
6	13–31	20–40	28–50	36–62	46–74	57–87	12–32	18–42	26–52	34–64	44–76	55–89						
7	14–34	21–44	29–55	39–66	49–79	60–93	13–35	20–45	27–57	36–69	46–82	57–96						
8	15–37	23–47	31–59	41–71	51–85	63–99	14–38	21–49	29–61	38–74	49–87	60–102						
9	16–40	24–51	33–63	43–76	54–90	66–105	14–42	22–53	31–65	40–79	51–93	62–109						
10	17–43	26–54	35–67	45–81	56–96	69–111	15–45	23–57	32–70	42–84	53–99	65–115						
11	18–46	27–58	37–71	47–86	59–101	72–117	16–48	24–61	34–74	44–89	55–105	68–121						
12	19–49	28–62	38–76	49–91	62–106	75–123	17–51	26–64	35–79	46–94	58–110	71–127						
13	20–52	30–65	40–80	52–95	64–112	78–129	18–54	27–68	37–83	48–99	60–116	73–134						
14	21–55	31–69	42–84	54–100	67–117	81–135	19–57	28–72	38–88	50–104	62–122	76–140						
15	22–58	33–72	44–88	56–105	69–123	84–141	20–60	29–76	40–92	52–109	65–127	79–146						
16	24–60	34–76	46–92	58–110	72–128	87–147	21–63	30–80	42–96	54–114	67–133	82–152						
17	25–63	35–80	47–97	61–114	75–133	90–153	21–67	32–83	43–101	56–119	70–138	84–159						
18	26–66	37–83	49–101	63–119	77–139	93–159	22–70	33–87	45–105	58–124	72–144	87–165						
19	27–69	38–87	51–105	65–124	80–144	96–165	23–73	34–91	46–110	60–129	74–150	90–171						
20	28–72	40–90	53–109	67–129	83–149	99–171	24–76	35–95	48–114	62–134	77–155	93–177						
21	29–75	41–94	55–113	69–134	85–155	102–177	25–79	37–98	50–118	64–139	79–161	95–184						
22	30–78	43–97	57–117	72–138	88–160	105–183	26–82	38–102	51–123	66–144	81–167	98–190						
23	31–81	44–101	58–122	74–143	90–166	108–189	27–85	39–106	53–127	68–149	84–172	101–196						
24	32–84	45–105	60–126	76–148	93–171	111–195	27–89	40–110	54–132	70–154	86–178	104–202						
25	33–87	47–108	62–130	78–153	96–176	114–201	28–92	42–113	56–136	72–159	89–183	107–208						
26	34–90	48–112	64–134	81–157	98–182	117–207	29–95	43–117	58–140	74–164	91–189	109–215						
27	35–93	50–115	66–138	83–162	101–187	120–213	30–98	44–121	59–145	76–169	93–195	112–221						
28	36–96	51–119	67–143	85–167	103–193	123–219	31–101	45–125	61–149	78–174	96–200	115–227						
29	37–99	53–122	69–147	87–172	106–198	126–225	32–104	47–128	63–153	80–179	98–206	118–233						
30	38–102	54–126	71–151	89–177	109–203	129–231	33–107	48–132	64–158	82–184	101–211	121–239						
31	39–105	55–130	73–155	92–181	111–209	132–237	34–110	49–136	66–162	84–189	103–217	123–246						
32	40–108	57–133	75–159	94–186	114–214	135–243	34–114	50–140	67–167	86–194	106–222	126–252						
33	41–111	58–137	77–163	96–191	117–219	138–249	35–117	52–143	69–171	88–199	108–228	129–258						
34	42–114	60–140	78–168	98–196	119–225	141–255	36–120	53–147	71–175	90–204	110–234	132–264						
35	43–117	61–144	80–172	100–201	122–230	144–261	37–123	54–151	72–180	92–209	113–239	135–270						
36	44–120	62–148	82–176	102–206	124–236	148–266	38–126	55–155	74–184	94–214	115–245	137–277						
37	45–123	64–151	84–180	105–210	127–241	151–272	39–129	57–158	76–188	96–219	117–251	140–283						
38	46–126	65–155	85–185	107–215	130–246	154–278	40–132	58–162	77–193	98–224	120–256	143–289						
39	47–129	67–158	87–189	109–220	132–252	157–284	41–135	59–166	79–197	100–229	122–262	146–295						
40	48–132	68–162	89–193	111–225	135–257	160–290	41–139	60–170	80–202	102–234	125–267	149–301						
41	49–135	69–166	91–197	114–229	138–262	163–296	42–142	61–174	82–206	104–239	127–273	151–308						
42	50–138	71–169	93–201	116–234	140–268	166–302	43–145	63–177	84–210	106–244	129–279	154–314						
43	51–141	72–173	95–205	118–239	143–273	169–308	44–148	64–181	85–215	108–249	132–284	157–320						
44	52–144	74–176	96–210	120–244	146–278	172–314	45–151	65–185	87–219	110–254	134–290	160–326						
45	53–147	75–180	98–214	123–248	148–284	175–320	46–154	66–189	88–224	112–259	137–295	163–332						
46	55–149	77–183	100–218	125–253	151–289	178–326	47–157	68–192	90–228	114–264	139–301	165–339						
47	56–152	78–187	102–222	127–258	154–294	181–332	48–160	69–196	92–232	116–269	141–307	168–345						
48	57–155	79–191	104–226	129–263	156–300	184–338	48–164	70–200	93–237	118–274	144–312	171–351						
49	58–158	81–194	106–230	132–267	159–305	187–344	49–167	71–204	95–241	120–279	146–318	174–357						
50	59–161	82–198	107–235	134–272	162–310	190–350	50–170	73–207	97–245	122–284	149–323	177–363						

<sup>a</sup> $n_1$  = minimum of the two sample sizes.<sup>b</sup> $n_2$  = maximum of the two sample sizes.<sup>c</sup> $T_l$  = lower critical value for the rank sum in the first sample.<sup>d</sup> $T_r$  = upper critical value for the rank sum in the first sample.

**Table 12** Two-tailed critical values for the Wilcoxon rank-sum test (*continued*)

$n_2^b$	$\alpha = .02$ $n_1^a$							$\alpha = .01$ $n_1$						
	4	5	6	7	8	9	4	5	6	7	8	9	4	5
	$T_l^c$	$T_r^d$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$	$T_l$	$T_r$
4	—	—	15-35	22-44	29-55	38-66	48-78	—	—	—	21-45	28-56	37-67	46-80
5	10-30	16-39	23-49	31-60	40-72	50-85	—	—	15-40	22-50	29-62	38-74	48-87	
6	11-33	17-43	24-54	32-66	42-78	52-92	10-34	16-44	23-55	31-67	40-80	50-94		
7	11-37	18-47	25-59	34-71	43-85	54-99	10-38	16-49	24-60	32-73	42-86	52-101		
8	12-40	19-51	27-63	35-77	45-91	56-106	11-41	17-53	25-65	34-78	43-93	54-108		
9	13-43	20-55	28-68	37-82	47-97	59-112	11-45	18-57	26-70	35-84	45-99	56-115		
10	13-47	21-59	29-73	39-87	49-103	61-119	12-48	19-61	27-75	37-89	47-105	58-122		
11	14-50	22-63	30-78	40-93	51-109	63-126	12-52	20-65	28-80	38-95	49-111	61-128		
12	15-53	23-67	32-82	42-98	53-115	66-132	13-55	21-69	30-84	40-100	51-117	63-135		
13	15-57	24-71	33-87	44-103	56-120	68-139	13-59	22-73	31-89	41-106	53-123	65-142		
14	16-60	25-75	34-92	45-109	58-126	71-145	14-62	22-78	32-94	43-111	54-130	67-149		
15	17-63	26-79	36-96	47-114	60-132	73-152	15-65	23-82	33-99	44-117	56-136	69-156		
16	17-67	27-83	37-101	49-119	62-138	76-158	15-69	24-86	34-104	46-122	58-142	72-162		
17	18-70	28-87	39-105	51-124	64-144	78-165	16-72	25-90	36-108	47-128	60-148	74-169		
18	19-73	29-91	40-110	52-130	66-150	81-171	16-76	26-94	37-113	49-133	62-154	76-176		
19	19-77	30-95	41-115	54-135	68-156	83-178	17-79	27-98	38-118	50-139	64-160	78-183		
20	20-80	31-99	43-119	56-140	70-162	85-185	18-82	28-102	39-123	52-144	66-166	81-189		
21	21-83	32-103	44-124	58-145	72-168	88-191	18-86	29-106	40-128	53-150	68-172	83-196		
22	21-87	33-107	45-129	59-151	74-174	90-198	19-89	29-111	42-132	55-155	70-178	85-203		
23	22-90	34-111	47-133	61-156	76-180	93-204	19-93	30-115	43-137	57-160	71-185	88-209		
24	23-93	35-115	48-138	63-161	78-186	95-211	20-96	31-119	44-142	58-166	73-191	90-216		
25	23-97	36-119	50-142	64-167	81-191	98-217	20-100	32-123	45-147	60-171	75-197	92-223		
26	24-100	37-123	51-147	66-172	83-197	100-224	21-103	33-127	46-152	61-177	77-203	94-230		
27	25-103	38-127	52-152	68-177	85-203	103-230	22-106	34-131	48-156	63-182	79-209	97-236		
28	26-106	39-131	54-156	70-182	87-209	105-237	22-110	35-135	49-161	64-188	81-215	99-243		
29	26-110	40-135	55-161	71-188	89-215	108-243	23-113	36-139	50-166	66-193	83-221	101-250		
30	27-113	41-139	56-166	73-193	91-221	110-250	23-117	37-143	51-171	68-198	85-227	103-257		
31	28-116	42-143	58-170	75-198	93-227	112-257	24-120	37-148	53-175	68-204	87-233	106-263		
32	28-120	43-147	59-175	77-203	95-233	115-263	24-124	38-152	54-180	71-209	89-239	108-270		
33	29-123	44-151	61-179	78-209	97-239	117-270	25-127	39-156	55-185	72-215	90-246	110-277		
34	30-126	45-155	62-184	79-215	99-245	120-276	26-130	40-160	56-190	73-221	92-252	112-284		
35	30-130	46-159	63-189	81-220	101-251	122-283	26-134	41-164	57-195	75-226	94-258	114-291		
36	31-133	47-163	65-193	83-225	103-257	125-289	27-137	42-168	58-200	76-232	96-264	117-297		
37	32-136	48-167	66-198	84-231	105-263	127-296	28-140	43-172	60-204	78-237	98-270	119-304		
38	32-140	49-171	67-203	86-236	107-269	129-303	28-144	44-176	61-209	79-243	100-276	121-311		
39	33-143	50-175	69-207	88-241	109-275	132-309	29-147	45-180	62-214	81-248	102-282	123-318		
40	34-146	51-179	70-212	90-246	111-281	134-316	29-151	46-184	63-219	82-254	103-289	126-324		
41	34-150	52-183	72-216	91-252	113-287	137-322	30-154	46-189	65-223	84-259	105-295	128-331		
42	35-153	53-187	73-221	93-257	116-292	139-329	31-157	47-193	66-228	85-265	107-301	130-338		
43	35-157	54-191	74-226	95-262	118-298	142-335	31-161	48-197	67-233	87-270	109-307	133-344		
44	36-160	55-195	76-230	97-267	120-304	144-342	32-164	49-201	68-238	88-276	111-313	135-351		
45	37-163	56-199	77-235	98-273	122-310	147-348	32-168	50-205	69-243	90-281	113-319	137-358		
46	37-167	57-203	78-240	100-278	124-316	149-355	33-171	51-209	71-247	91-287	115-325	139-365		
47	38-170	58-207	80-244	102-283	126-322	152-361	34-174	52-213	72-252	93-292	117-331	142-371		
48	39-173	59-211	81-249	103-289	128-328	154-368	34-178	53-217	73-257	95-297	118-338	144-378		
49	39-177	60-215	82-254	105-294	130-334	157-374	35-181	54-221	74-262	96-303	120-344	146-385		
50	40-180	61-219	84-258	107-299	132-340	159-381	36-184	55-225	76-266	98-308	122-350	148-392		

Source: The data of this table are from *Documenta Geigy Scientific Tables*, 6th edition. Reprinted with the kind permission of CIBA-GEIGY Limited, Basel, Switzerland.

**Table 13** Fisher's z transformation

<i>r</i>	<i>z</i>								
.00	.000								
.01	.010	.21	.213	.41	.436	.61	.709	.81	1.127
.02	.020	.22	.224	.42	.448	.62	.725	.82	1.157
.03	.030	.23	.234	.43	.460	.63	.741	.83	1.188
.04	.040	.24	.245	.44	.472	.64	.758	.84	1.221
.05	.050	.25	.255	.45	.485	.65	.775	.85	1.256
.06	.060	.26	.266	.46	.497	.66	.793	.86	1.293
.07	.070	.27	.277	.47	.510	.67	.811	.87	1.333
.08	.080	.28	.288	.48	.523	.68	.829	.88	1.376
.09	.090	.29	.299	.49	.536	.69	.848	.89	1.422
.10	.100	.30	.310	.50	.549	.70	.867	.90	1.472
.11	.110	.31	.321	.51	.563	.71	.887	.91	1.528
.12	.121	.32	.332	.52	.576	.72	.908	.92	1.589
.13	.131	.33	.343	.53	.590	.73	.929	.93	1.658
.14	.141	.34	.354	.54	.604	.74	.950	.94	1.738
.15	.151	.35	.365	.55	.618	.75	.973	.95	1.832
.16	.161	.36	.377	.56	.633	.76	.996	.96	1.946
.17	.172	.37	.388	.57	.648	.77	1.020	.97	2.092
.18	.182	.38	.400	.58	.662	.78	1.045	.98	2.298
.19	.192	.39	.412	.59	.678	.79	1.071	.99	2.647
.20	.203	.40	.424	.60	.693	.80	1.099		

**Table 14** Two-tailed upper critical values for the Spearman rank-correlation coefficient ( $r_s$ )

$n$	$\alpha$			
	.10	.05	.02	.01
1	—	—	—	—
2	—	—	—	—
3	—	—	—	—
4	1.0	—	—	—
5	.900	1.0	1.0	—
6	.829	.886	.943	1.0
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.683	.783	.833

Source: The data for this table have been adapted with permission from E. G. Olds (1938), "Distributions of Sums of Squares of Rank Differences for Small Numbers of Individuals," *Annals of Mathematical Statistics*, 9, 133–148.

**Table 15 Critical values for the Kruskal-Wallis test statistic ( $H$ ) for selected sample sizes for  $k = 3$** 

$n_1$	$n_2$	$n_3$	$\alpha$			
			.10	.05	.02	.01
1	1	2	—	—	—	—
1	1	3	—	—	—	—
1	1	4	—	—	—	—
1	1	5	—	—	—	—
1	2	2	—	—	—	—
1	2	3	4.286	—	—	—
1	2	4	4.500	—	—	—
1	2	5	4.200	5.000	—	—
1	3	3	4.571	5.143	—	—
1	3	4	4.056	5.389	—	—
1	3	5	4.018	4.960	6.400	—
1	4	4	4.167	4.967	6.667	—
1	4	5	3.987	4.986	6.431	6.954
1	5	5	4.109	5.127	6.146	7.309
2	2	2	4.571	—	—	—
2	2	3	4.500	4.714	—	—
2	2	4	4.500	5.333	6.000	—
2	2	5	4.373	5.160	6.000	6.533
2	3	3	4.694	5.361	6.250	—
2	3	4	4.511	5.444	6.144	6.444
2	3	5	4.651	5.251	6.294	6.909
2	4	4	4.554	5.454	6.600	7.036
2	4	5	4.541	5.273	6.541	7.204
2	5	5	4.623	5.338	6.469	7.392
3	3	3	5.067	5.689	6.489	7.200
3	3	4	4.709	5.791	6.564	7.000
3	3	5	4.533	5.648	6.533	7.079
3	4	4	4.546	5.598	6.712	7.212
3	4	5	4.549	5.656	6.703	7.477
3	5	5	4.571	5.706	6.866	7.622
4	4	4	4.654	5.692	6.962	7.654
4	4	5	4.668	5.657	6.976	7.760
4	5	5	4.523	5.666	7.000	7.903
5	5	5	4.580	5.780	7.220	8.000

Source: The data for this table have been adapted from Table F of *A Nonparametric Introduction to Statistics* by C.H. Kraft and C. Van Eeden, Macmillan, New York, 1968.

**Table 16 Critical values for the studentized range statistic  $q^*$ ,  $\alpha = .05$** 

$v$	$k:$	2	3	4	5	6	7	8	9	10
1		17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2		6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99
3		4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462
4		3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826
5		3.635	4.602	5.218	5.673	6.033	5.330	6.582	6.802	6.995
6		3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493
7		3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158
8		3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918
9		3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739
10		3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599
11		3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
12		3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
13		3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
14		3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
15		3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16		2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
17		2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18		2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
19		2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038
20		2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
24		2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
30		2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
40		2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
60		2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
120		2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
$\infty$		2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474
	$k:$	11	12	13	14	15	16	17	18	19
1		5.059	51.96	53.20	54.33	55.36	56.32	57.22	58.04	58.83
2		14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57
3		9.717	9.946	10.15	10.35	10.53	10.69	10.84	10.98	11.11
4		8.027	8.208	8.373	8.525	8664	8.794	8.914	9.028	9.134
5		7.168	7.324	7.466	7.596	7.717	7.828	7.932	8.030	8.122
6		6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.508
7		6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097
8		6.054	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.802
9		5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579
10		5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405
11		5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265
12		5.511	5.615	5.710	5.798	5.878	5.953	6.023	6.089	6.151
13		5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055
14		5.364	5.463	5.554	5.637	5.714	5.786	5.852	5.915	5.974
15		5.306	5.404	5.493	5.574	5.649	5.720	5.785	5.846	5.904
16		5.256	5.352	5.439	5.520	5.593	5.662	5.727	5.786	5.843
17		5.212	5.307	5.392	5.471	5.544	5.612	5.675	5.734	5.790
18		5.174	5.267	5.352	5.429	5.501	5.568	5.630	5.688	5.242
19		5.140	5.231	5.315	5.391	5.462	5.528	5.589	5.647	5.701
20		5.108	5.199	5.282	5.357	5.427	5.493	5.553	5.610	5.663
24		5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545
30		4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.425
40		4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313
50		4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199
120		4.641	4.714	4.781	4.842	4.893	4.950	4.998	5.044	5.086
$\infty$		4.522	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974

<sup>\*</sup> $q_{k,v,05}$  = upper 5th percentile of a  $q_{k,v}$  distribution

# Answers to Selected Problems

## CHAPTER 2

**2.4–2.7** The median, mode, geometric mean, and range are each multiplied by  $c$ . **2.12**  $\bar{x} = 19.54$  mg/dL **2.13**  $s = 16.81$  mg/dL **2.15** Median = 19 mg/dL

## CHAPTER 3

**3.1** At least one parent has influenza. **3.2** Both parents have influenza. **3.3** No **3.4** At least one child has influenza. **3.5** The 1st child has influenza. **3.6**  $C = A_1 \cup A_2$  **3.7**  $D = B \cup C$  **3.8** The mother does not have influenza. **3.9** The father does not have influenza. **3.10**  $\bar{A}_1 \cap \bar{A}_2$  **3.11**  $\bar{B} \cap \bar{C}$  **3.28** .0167 **3.29** 180% **3.48** .069 **3.50** .541 **3.51** .20 **3.52** .5. This is a conditional probability. The probability in Problem 3.51 is a joint unconditional probability. **3.53** .20 **3.54** No, because  $Pr(M|F) = .6 \neq Pr(M|\bar{F}) = .2$  **3.55** .084 **3.56** .655 **3.57** .690 **3.58** .486 **3.59** .373 **3.60** No. **3.61** No. **3.64** .05 **3.65** .326 **3.66** .652 **3.67** .967 **3.68** .479 **3.69** .893 **3.70** .975 **3.71** .630. It is lower than the predictive value negative based on self-report (i.e., .893). **3.79** .95 **3.80** .99 **3.81** .913 **3.82** The new test has a 13.6% lower cost.

## CHAPTER 4

**4.1**  $Pr(0) = .72$ ,  $Pr(1) = .26$ ,  $Pr(2) = .02$  **4.2** .30 **4.3** .25 **4.4**  $F(x) = 0$  if  $x < 0$ ;  $F(x) = .72$  if  $0 \leq x < 1$ ;  $F(x) = .98$  if  $1 \leq x < 2$ ;  $F(x) = 1.0$  if  $x \geq 2$ . **4.8** 362,880 **4.11** .1042 **4.12** .2148 **4.13**  $E(X) = Var(X) = 4.0$  **4.23**  $Pr(X \geq 6) = .010$  **4.24**  $Pr(X \geq 4) = .242$  **4.26** .620 **4.27** .202 **4.28** .385 **4.29** .471 **4.30** .144 **4.31** 1.24 **4.32**  $Pr(X \geq 4) = .241 > .05$ . Thus, there is no excess risk of malformations. **4.33**  $Pr(X \geq 8) = .0006 < .05$ . Thus there is an excess risk of malformations. **4.37** .23 **4.38** .882 **4.39**  $Pr(X = 0) = .91$ ,  $Pr(X = 1) = .08$ ,  $Pr(X = 2) = .01$  **4.40** 0.100 **4.41** 0.110 **4.42** 6 months, .25; 1 year, .52 **4.43** .435 **4.44** .104 **4.45** 10.4 **4.52** Based on the Poisson distribution,  $Pr(X \geq 27) = .049 < .05$ . Thus there is a significant excess. **4.53** .0263 **4.54** If  $Y$  = number of cases of cleft palate, then based on the Poisson distribution,  $Pr(Y \geq 12) = .0532 > .05$ . This is a borderline result, because this probability is close to .05.

## CHAPTER 5

**5.1** .6915 **5.2** .3085 **5.3** .7745 **5.4** .0228 **5.5** .0441 **5.12** .079 **5.13** .0004 **5.14** .352 **5.15** .268 **5.16** .380 **5.17** .023 **5.18** .067 **5.19** .0058 **5.20** .435 **5.25** .018 **5.26** .123 **5.27** .0005 **5.28** ≥43 **5.29** ≥69 **5.30** ≥72 **5.36** .851 **5.37** Sensitivity. **5.38** .941 **5.39** Specificity. **5.40**  $\Delta = 0.2375$  mg/dL, compliance = 88% in each group **5.47** 63.5% **5.48** 32.3% **5.49** No. The distributions are very skewed.

## CHAPTER 6

**6.5** 0.079 for normal men, 0.071 for men with chronic airflow limitation. **6.15** .44 **6.16** .099 **6.17** (.25, .63)

	Point estimate	95% CI
<b>6.18</b>		
<i>E. coli</i>	25.53	(24.16, 26.90)
<i>S. aureus</i>	26.79	(24.88, 28.70)
<i>P. aeruginosa</i>	19.93	(18.60, 21.27)
<b>6.19</b>	Point estimate	95% CI
<i>E. coli</i>	25.06	(23.73, 26.38)
<i>S. aureus</i>	25.44	(24.60, 26.29)
<i>P. aeruginosa</i>	17.89	(17.09, 18.69)
<b>6.20</b>	Point estimate	95% CI
<i>E. coli</i>	1.78	(1.21, 3.42)
<i>S. aureus</i>	2.49	(1.68, 4.77)
<i>P. aeruginosa</i>	1.74	(1.17, 3.32)
<b>6.21</b>	Point estimate	95% CI
<i>E. coli</i>	1.73	(1.17, 3.31)
<i>S. aureus</i>	1.10	(0.75, 2.12)
<i>P. aeruginosa</i>	1.04	(0.70, 1.99)

**6.27**  $6/46 = .130$  **6.28** (.033, .228) **6.29** Because 10% is within the 95% CI, the two drugs are equally effective. **6.30** (6.17, 7.83) **6.31** (2.11, 9.71) **6.32**  $n \doteq 251$  **6.36** 0.544 **6.37** (0.26, 1.81) **6.38** .958 **6.39** .999 **6.52** .615 **6.53** .918 **6.54** 0.5 lb, observed proportion = .615. 1 lb, observed proportion = .935. The observed and

expected proportions are in good agreement. **6.55** Yes. **6.75** 95% CI = (2.20, 13.06). Because this interval does not include 1.8, there are an excess number of cases of bladder cancer among tire workers. **6.76** 95% CI = (1.09, 10.24). Because this interval includes 2.5, there is not an excess number of cases of stomach cancer among tire workers.

## CHAPTER 7

- 7.1**  $z = 1.732$ , accept  $H_0$  at the 5% level. **7.2**  $p = .083$
- 7.4**  $t = 1.155 - t_{11}$ ,  $p = .27$  **7.5** (0.82, 1.58) **7.6** The 95% CI contains 1.0, which is consistent with our decision to accept  $H_0$  at the 5% level. **7.15**  $z = 1.142$ , accept  $H_0$  at the 5% level. **7.16**  $p = .25$  **7.17** Accept  $H_0$  at the 5% level. **7.18**  $p = .71$  **7.22**  $z = 7.72$ ,  $p < .001$  **7.26** 31 **7.27** .770
- 7.33**  $H_0$ :  $\mu = \mu_0$  vs.  $H_1$ :  $\mu \neq \mu_0$ ;  $\sigma^2$  unknown.  $\mu$  = true mean daily iron intake for 9- to 11-year-old boys below the poverty level,  $\mu_0$  = true mean daily iron intake for 9- to 11-year-old boys in the general population. **7.34**  $t = -2.917 - t_{50}$ , reject  $H_0$  at the 5% level. **7.35**  $.001 < p < .01$  (exact  $p$ -value = .005) **7.36**  $H_0$ :  $\sigma^2 = \sigma_0^2$  vs.  $H_1$ :  $\sigma^2 \neq \sigma_0^2$ .  $\sigma^2$  = true variance in the low-income population,  $\sigma_0^2$  = true variance in the general population. **7.37**  $X^2 = 36.49 - \chi_{50}^2$ , accept  $H_0$  at the 5% level. **7.38**  $.1 < p < .2$  (exact  $p$ -value = .15) **7.39** (15.80, 34.86). The interval contains  $\sigma_0^2 = 5.56^2 = 30.91$ , so the underlying variances of the low-income and the general population are not significantly different.
- 7.47** One-sample binomial test, exact method. **7.48**  $p = .28$
- 7.49** One-sample binomial test, large-sample method.  
 $z = 3.24$ ,  $p = .0012$  **7.50** (.058, .142)

## CHAPTER 8

- 8.12** 135 girls in each group or a total of 270 overall.
- 8.13** 106 girls in each group or a total of 212 overall.
- 8.14** 96 girls in the below-poverty group, 192 girls in the above-poverty group. **8.15** Power = .401 **8.16** Power = .525 **8.17** Power = .300 **8.18** Power = .417 **8.19** Use the paired  $t$  test.  $t = -3.37 - t_{9}$ ,  $.001 < p < .01$  (exact  $p$ -value = .008) **8.20** Use the paired  $t$  test.  $t = -1.83 - t_{29}$ ,  $.05 < p < .10$  (exact  $p$ -value = .078)

8.21 Group	95% CI
Methazolamide and topical glaucoma medications	(-2.67, -0.53)
Topical drugs only	(-1.48, 0.08)

- 8.22** Use the two-sample  $t$  test with equal variances.  $t = -1.25 - t_{38}$ ,  $p > .05$  (exact  $p$ -value = .22). **8.25**  $H_0$ :  $\mu_d = 0$  vs.  $H_1$ :  $\mu_d \neq 0$ .  $\mu_d$  = mean difference in one-hour concentration (drug A – drug B) for a specific person. **8.26** Use a paired  $t$  test to test these hypotheses. **8.27**  $t = 3.67 - t_{9}$ ,  $.001 < p < .01$  (exact  $p$ -value = .005) **8.28** 3.60 mg % **8.29** (1.38, 5.82) mg % **8.41**  $H_0$ :  $\mu_1 = \mu_2$  vs.  $H_1$ :  $\mu_1 \neq \mu_2$ , where  $\mu_1$  = true mean FEV of children both of whose parents smoke,  $\mu_2$  = true mean FEV of children neither of whose parents smoke. **8.42** First, perform  $F$  test for equality of two variances,  $F = 3.06 - F_{22, 19}$ ,  $p < .05$ . Therefore, use the two-sample  $t$  test with unequal variances. **8.43**  $t = -1.17 - t_{35}$ ,

accept  $H_0$  at the 5% level. **8.44** (-0.55, 0.15) **8.45** 212 children in each group. **8.46** 176 children in each group. **8.47** .363 **8.48** .486 **8.54** The paired  $t$  test. **8.55** Raw scale,  $t = -3.49 - t_9$ ,  $.001 < p < .01$  (exact  $p$ -value = .007), In scale,  $t = -3.74 - t_9$ ,  $.001 < p < .01$  (exact  $p$ -value = .005). The In scale is preferable because the change in the raw scale seems to be related to the initial level. **8.56** Urinary protein has declined by 56.7% over 8 weeks. **8.57** 95% CI for 8-week decline = (28.2%, 73.9%) **8.63** Two-sample  $t$  test with equal variances **8.64**  $H_0$ :  $\mu_1 = \mu_2$  vs.  $H_1$ :  $\mu_1 \neq \mu_2$ ;  $\mu_1$  = mean cholesterol level for men;  $\mu_2$  = mean cholesterol level for women;  $t = -1.92 - t_{99}$ ,  $.05 < p < .10$  (exact  $p$ -value = .058) **8.65**  $H_0$ :  $\mu_1 = \mu_2$  vs.  $H_1$ :  $\mu_1 > \mu_2$ ;  $t = -1.92 - t_{99}$ ,  $.95 < p < .975$  (exact  $p$ -value = .97) **8.66** No. The twin pairs are not independent observations. **8.70**  $F$  test for the equality of two variances,  $F = 1.15 - F_{35, 29}$ ,  $p > .05$ . Therefore, use the two-sample  $t$  test with equal variances. **8.71**  $t = 1.25 - t_{64}$ ,  $.2 < p < .3$  (exact  $p$ -value = .22) **8.72** RA, .32; OA, .43 **8.73** 133 subjects in each of the RA and OA groups. **8.74** Paired  $t$  test. **8.75**  $t = 2.27 - t_{99}$ ,  $.02 < p < .05$  (exact  $p$ -value = .025) **8.76**  $F$  test for the equality of two variances,  $F = 1.99 - F_{98, 99}$ ,  $p < .05$ . Use the two-sample  $t$  test with unequal variances. **8.77**  $t = -4.20 - t_{176}$ ,  $p < .001$

## CHAPTER 9

- 9.1** Use the sign test. The critical values are  $c_1 = 6.3$  and  $c_2 = 16.7$ . Because  $c_1 \leq C \leq c_2$ , where  $C$  = number of patients who improved = 15, we accept  $H_0$  at the 5% level. **9.9** The distribution of length of stay is very skewed and far from being normal, which makes the  $t$  test not very useful here. **9.10** Use the Wilcoxon rank-sum test (large-sample test).  $R_1 = 83.5$ ,  $T = 3.10 - N(0, 1)$ ,  $p = .002$  **9.15**  $H_0$ :  $F_1 = F_2$  vs.  $H_1$ :  $F_1 \neq F_2$ , where  $F_1$  = distribution of duration of effusion for breast-fed babies,  $F_2$  = distribution of duration of effusion for bottle-fed babies. **9.16** The distribution of duration of effusion is very skewed and far from being normal. **9.17** Wilcoxon signed-rank test (large-sample test). **9.18**  $R_1 = 215$ ,  $T = 2.33 - N(0, 1)$ ,  $p = .020$ . Breast-fed babies have a shorter duration of effusion than bottle-fed babies. **9.24** The Wilcoxon signed-rank test (large-sample test). **9.25**  $R_1 = 33.5$ ,  $T = 1.76 - N(0, 1)$ ,  $p = .078$ . The mean SBP is slightly but not significantly higher with the standard cuff. **9.26** The Wilcoxon signed-rank test (large-sample test). **9.27**  $R_1 = 32.0$ ,  $T = 1.86 - N(0, 1)$ ,  $p = .062$ . Variability with the standard cuff is slightly, but not significantly, lower than with the random zero.

## CHAPTER 10

- 10.8** McNemar's test for correlated proportions. **10.9**  $X^2 = 4.76 - \chi_1^2$ ,  $.025 < p < .05$  **10.10** 87 **10.11** 13 **10.12** McNemar's test for correlated proportions, exact test;  $p = .267$  **10.13** Use chi-square test for  $2 \times 2$  tables.  $X^2 = 32.17 - \chi_1^2$ ,  $p < .001$  **10.15** Use chi-square test for  $R \times C$  tables.  $X^2 = 117.02 - \chi_2^2$ ,  $p < .001$ . There is a significant association between ethnic origin and genetic type. **10.18** Two-sample test. **10.19** Two-sided test. **10.20** Chi-square

test for  $2 \times 2$  tables. **10.21**  $X^2 = 3.48 \sim \chi_1^2$ ,  $.05 < p < .10$  **10.26** McNemar's test for correlated proportions. **10.27**  $X^2 = 4.65 \sim \chi_1^2$ ,  $.025 < p < .05$  **10.32** McNemar's test for correlated proportions (large-sample test). **10.33**  $X^2 = 6.48 \sim \chi_1^2$ ,  $.01 < p < .025$  **10.34** McNemar's test for correlated proportions (exact method). **10.35**  $p = .387$  **10.36** .9997 **10.42** .304 **10.43** .213 **10.44** 284 subjects in each group. **10.45** Cholesterol-lowering drug patients, .218; placebo pill patients, .295 **10.46** 390 subjects in each group. **10.48** 12, .273; 13, .333; 14, .303; 15, .091; 16, 0; 17, 0 **10.49** 15.13 years  $\pm$  15 years, 2 months **10.50** We use the age groups,  $\leq 12.9$ , 13.0–13.9, 14.0–14.9,  $\geq 15.0$ , and perform the chi-square goodness-of-fit test.  $X^2 = 1.27 \sim \chi_1^2$ ,  $.25 < p < .50$ . The goodness of fit of the normal model is adequate.

## CHAPTER 11

**11.1**  $y = 1894.8 + 112.1x$  **11.2**  $F = 180,750/490,818 = 0.37 \sim F_{1,7}$ ,  $p > .05$  **11.3** .05 **11.4**  $R^2$  = % variance of lymphocyte count that is explained by % reticulocytes ( $\approx 5\%$ ). **11.5** 490,818 **11.6**  $t = 0.61 \sim t_7$ ,  $p > .05$  **11.7**  $se(b) = 184.7$ ,  $se(a) = 348.5$  **11.9** Two-sample z test to compare two correlation coefficients. **11.10**  $\lambda = 3.40 \sim N(0, 1)$ , reject  $H_0$  at the 5% level. **11.11**  $p < .001$  **11.12** Use the two-sample z test to compare two correlation coefficients.  $\lambda = 3.61$ ,  $p < .001$ . The correlation coefficients are significantly different. **11.13**  $y = 1472.0 - 0.737x$ , where  $y$  = infant mortality rate,  $x$  = year. **11.14** Use F test for simple linear regression.  $F = 182.04/0.329 = 553.9 \sim F_{1,10}$ ,  $p < .001$  **11.15** 5.7 deaths per 1000 livebirths. **11.16** 0.79 deaths per 1000 livebirths. **11.17** No. If the linear relationship persisted, the expected mortality rate would eventually be projected as negative, which is impossible. **11.37** The one-sample t test for correlation **11.38**  $t = 8.75 \sim t_{901}$ ,  $p < .001$  **11.39** The two-sample z test for correlation **11.40**  $\lambda = 2.563 \sim N(0, 1)$ ,  $p = .010$  **11.41** White boys (.219, .339); black boys (.051, .227)

## CHAPTER 12

**12.1**  $F = 1643.08/160.65 = 10.23 \sim F_{2,22}$ ,  $p < .05$ . The means of the three groups are significantly different. **12.2**  $p < .001$

12.3	Groups	Test statistic	p-value
	STD, LAC	$t = 3.18 \sim t_{23}$	$.001 < p < .01$
	STD, VEG	$t = 4.28 \sim t_{23}$	$p < .001$
	LAC, VEG	$t = 1.53 \sim t_{23}$	NS

**12.4**  $t = -4.09 \sim t_{23}$ ,  $p < .001$ . The contrast is an estimate of the difference in mean protein intake between the general vegetarian population and the general nonvegetarian population. **12.6**  $F = 251.77/50.46 = 4.99 \sim F_{2,19}$ ,  $p < .05$

12.7	Groups	Test statistic	p-value
	A, B	$t = 2.67 \sim t_{19}$	$.01 < p < .02$
	A, C	$t = 2.94 \sim t_{19}$	$.001 < p < .01$
	B, C	$t = 0.82 \sim t_{19}$	NS

**12.8** A,B  $p < .05$ . A,C  $p < .05$ . B,C NS.

**12.18** Between-day variance =  $\hat{\sigma}_A^2 = 1.19$ , within-day variance =  $\hat{\sigma}^2 = 14.50$  **12.19**  $F = 16.89/14.50 = 1.16 \sim F_{9,10}$ ,  $p > .05$ . There is no significant between-day variance.

## CHAPTER 13

**13.1**  $Z = 6.19$ ,  $p < .001$  **13.2**  $X^2 = 38.34 \sim \chi_1^2$ ,  $p < .001$

**13.3** The conclusions are the same. Also,  $z^2 = X^2_{\text{corrected}} = 38.34$ . **13.4** (.090, .201) **13.5** 2.29 **13.6** (1.76, 2.98)

**13.15** The Mantel-Haenszel test **13.16**  $X^2_{\text{MH}} = 0.51 \sim \chi_1^2$ ,  $p > .05$  **13.17** 1.38 **13.18** (0.68, 2.82) **13.29** 1.40

**13.30** (1.09, 1.80) **13.37** RR = 1.70, 95% CI = (0.42, 6.93)

**13.38** RR = 1.38, 95% CI = (1.06, 1.79)

## CHAPTER 14

**14.1** Incidence density = 273.1 cases per  $10^5$  person-years for current users, 115.2 cases per  $10^5$  person-years for never users,  $z = 6.67/2.359 = 2.827 \sim N(0, 1)$ ,  $p = .005$ . There is a significant excess of breast cancer among current OC users vs. never OC users. **14.2** Incidence density = 135.4 cases per  $10^5$  person-years for past users, 115.2 cases per  $10^5$  person-years for never users.  $z = 10.47/8.276 = 1.265 \sim N(0, 1)$ ,  $p = .21$ . There is no significant excess (or deficit) of breast cancer among past OC users vs. never OC users.

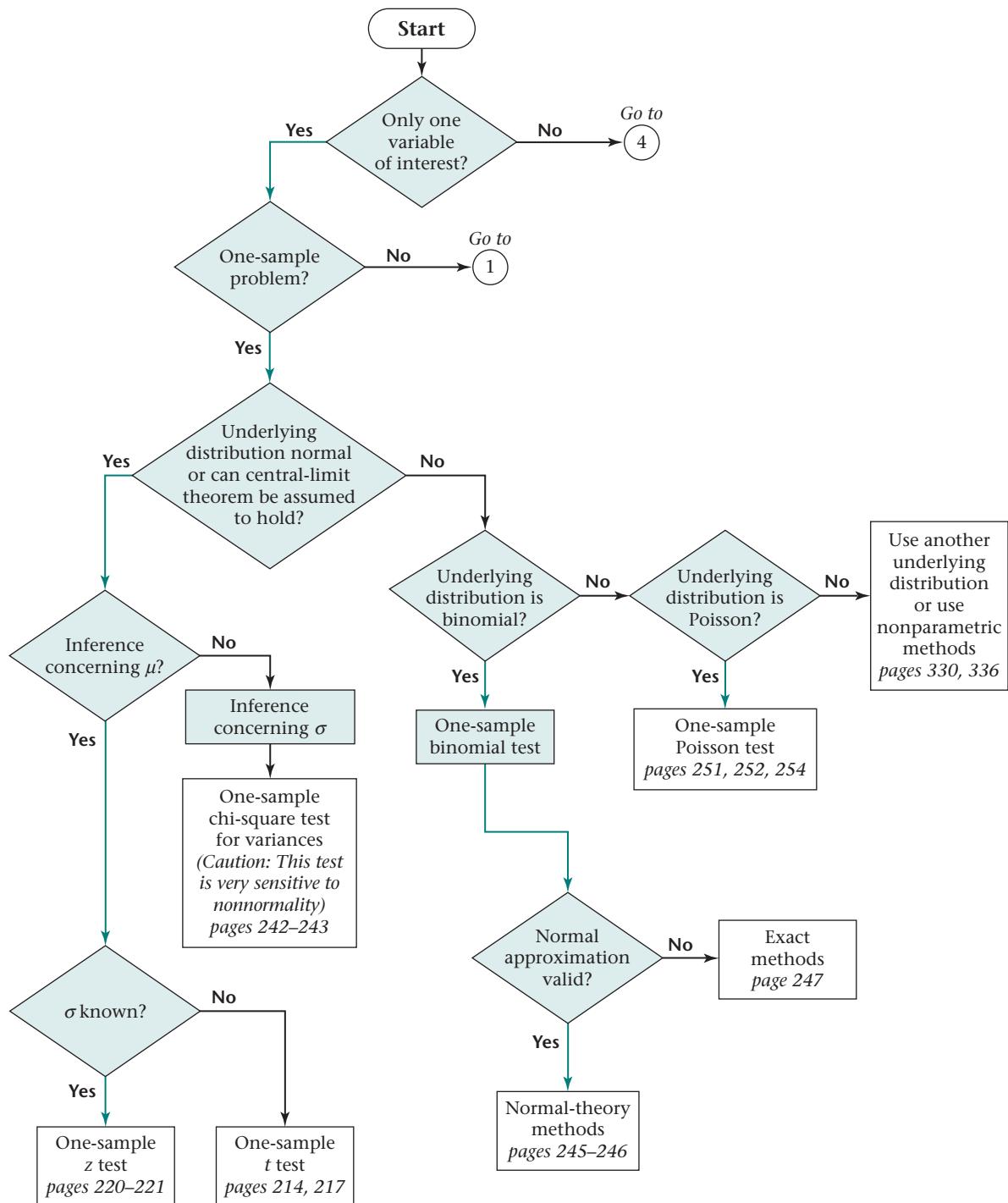
**14.3**  $\hat{RR} = 2.37$ , 95% CI = (1.34, 4.21) **14.4**  $\hat{RR} = 1.18$ , 95% CI = (0.93, 1.49) **14.22** .072 **14.23** 12

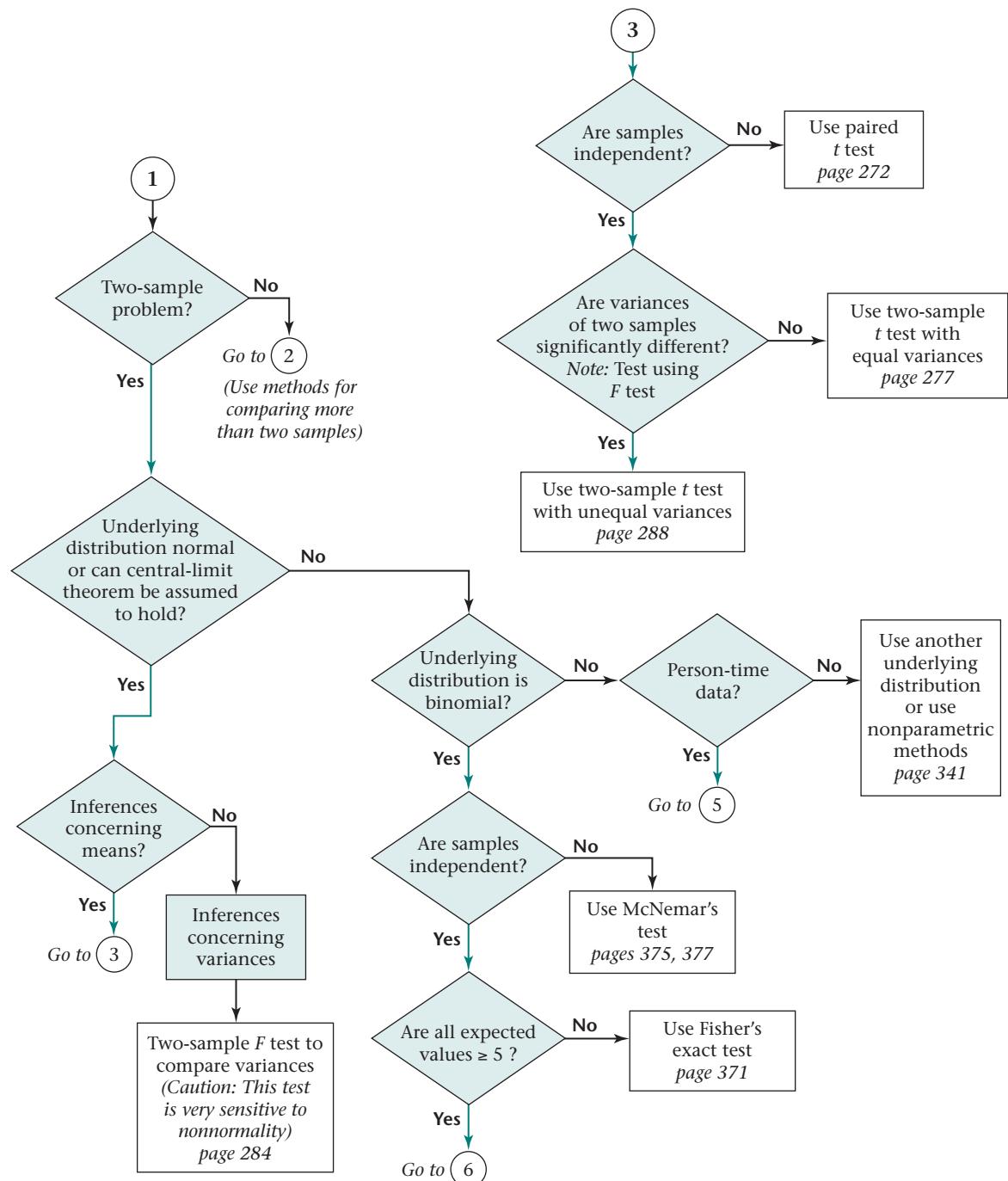
**14.24** (2.1 events per 100 person-years, 7.0 events per 100 person-years)

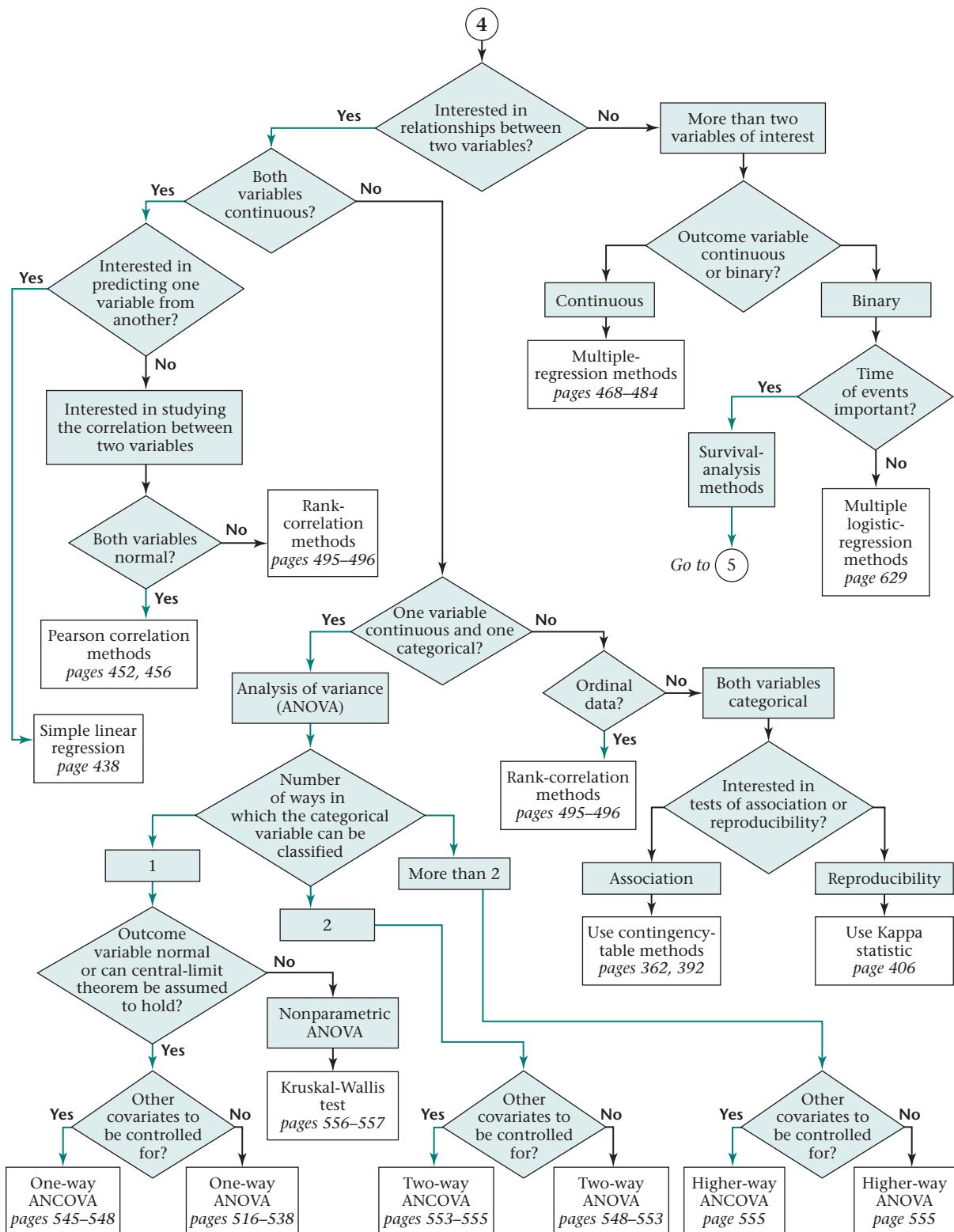


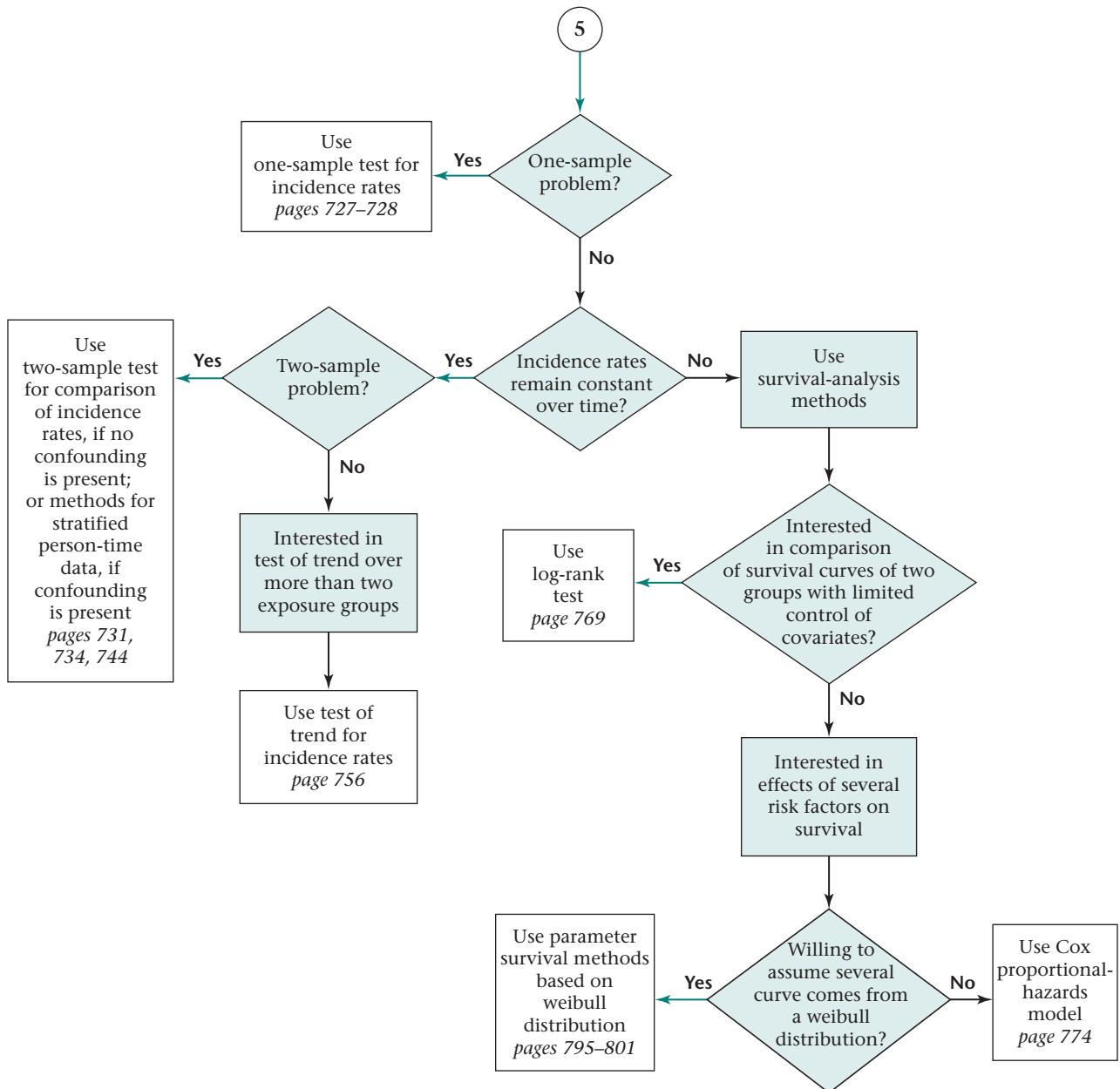
# **FLOWCHART**

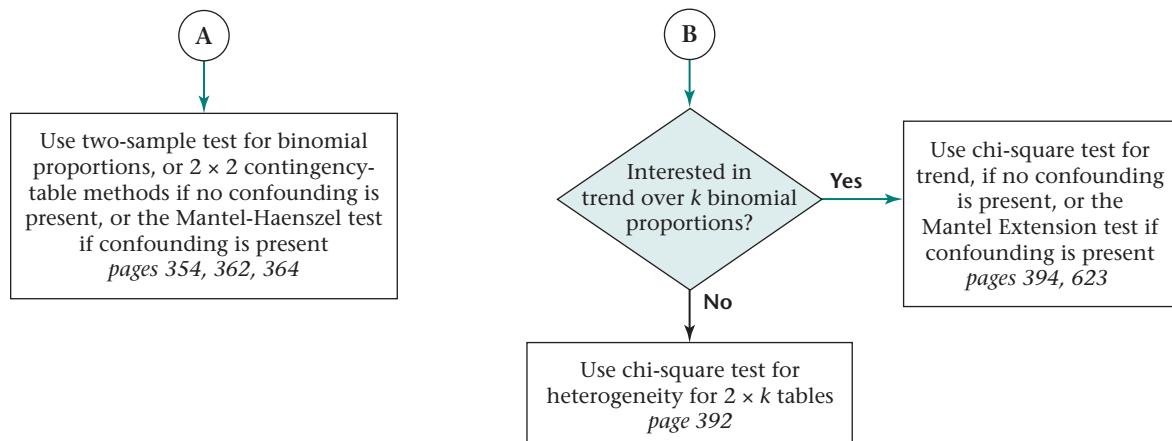
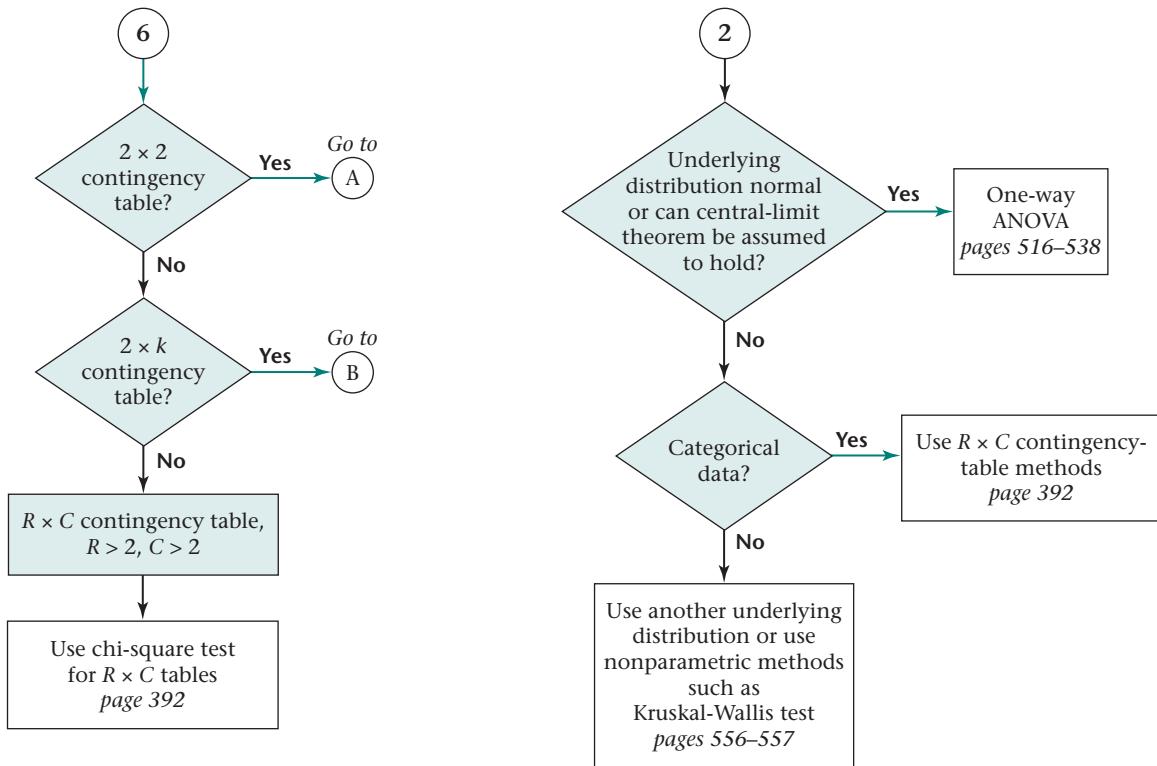
## **Methods of Statistical Inference**











# Index of Data Sets

<i>Data Set</i>	<i>Page in text</i>	<i>Data Set</i>	<i>Page in text</i>
BETACAR	579	HOSPITAL	33
BLOOD	721	INFANTBP	199
BONEDEN	30	LEAD	29
BOTOX	806	MICE	582
BREAST	808	NEPHRO	658
CORNEAL	587	NIFED	140
DIABETES	322	OTO	714
EAR	64	PIRIFORM	144
EFF	714	SEXRAT	103
ENDOCRIN	580	SMOKE	102
ESTRADL	422, 511	SWISS	319
ESTROGEN	716	TEAR	265
FEV	35	TEMPERAT	581
FIELD	810	TENNIS1	415
HEART	Study Guide	TENNIS2	316
HORMONE	315	VALID	36



# Index

- Absolute value, 218  
Accelerating failure time model, 797  
Acceptance region, 208  
Bonferroni multiple-comparisons procedure, 532  
chi-square goodness-of-fit test, 403  
chi-square test for  $R \times C$  contingency tables, 393  
chi-square test for trend in binomial proportions, 395  
Dunn procedure, 560  
 $F$  test for one-way ANOVA, 521  
 $F$  test for simple linear regression, 438  
Fisher's  $z$  test, 464  
fixed-effects one-way ANOVA (analysis of variance), 521  
hypothesis testing in multiple linear regression, 473  
kappa statistic, 406  
Kruskal-Wallis test, 557  
log-rank test, 771  
Mantel-Haenszel test, 615  
McNemar's test, 376  
multiple logistic regression, 637  
normal-theory test, 355  
one-sample binomial test (two-sided alternative), 245  
one-sample  $\chi^2$  test for variance of normal distribution, 242  
one-sample  $t$  test for correlation coefficient, 457  
one-sample  $z$  test for correlation coefficient, 459  
one-way ANOVA, 521, 525  
 $R \times C$  contingency tables, 393  
rank-correlation coefficients, 497  
sign test, 330  
sign test (normal-theory method), 330  
Spearman rank-correlation coefficient, 497  
 $t$  test for comparison of pairs of groups in one-way ANOVA, 525
- $t$  test for multiple linear regression, 474  
 $t$  test for simple linear regression, 442  
two-sample inference for incidence-rate data, 732  
two-sample test for binomial proportions (normal-theory test), 355  
two-sample test for incidence rates (normal-theory method), 732  
Yates-corrected chi-square test, 363  
 $z$  test, one-sample, 459  
 $z$  test, two-sample, 464  
Actuarial method, 766  
Addition law of probability, 44–46  
Adjusted  $R^2$ , 440  
Age-standardized risk, 611  
Analysis-of-variance estimator, 569  
Apgar score, 495, 501  
Applied statistics, 1  
Arithmetic mean, 7–9  
    vs. median, 10  
    properties of, 13–15  
    rescaled sample, 14–15  
    translated sample, 13–14  
Attributable risk, 601–606  
    estimation with multiple exposed groups, 605–606  
    interval estimation for, 603  
    and risk factor, 602
- Bar graph, 5–6, 25  
Bayes' rule, 53–56  
Bayesian inference, 56–57, 237–241  
Bayley Mental Development Index, 507–508  
Behrens-Fisher problem, 288  
Bell-shaped distribution. *See* Normal distribution  
Bernoulli trial, 131  
Between mean square, 520
- Between sum of squares, 519–520  
Between-group variability, 518–519  
Bimodal distribution, 12  
BINOMDIST function, 87, 132  
Binomial distribution, 71, 83–88  
    binomial tables, 83–86  
    electronic tables, 87–88  
    estimation for, 181–188  
    expected value of, 88–89  
    interval estimation, exact methods, 186–188  
    interval estimation, normal-theory methods, 184–186  
    maximum-likelihood estimation, 182–184  
    normal approximation to, 129–135  
    one-sample inference, 244–250  
    point estimation, 181–182  
    Poisson approximation, 96–98  
    variance of, 88–89  
Binomial tables, 83–86  
Biostatistics, 1  
Blinding, 159  
Block randomization, 158  
BONEDEN.DAT (Data Set), 29–30  
BONEDEN.DOC (Data Set), 29–30  
Bone-mineral density study, 30–31, 175–176, 256  
Bonferroni multiple-comparisons procedure, 531–534  
    acceptance and rejection regions, 532  
    experiment-wise type I error, 532  
BOTOX.DAT (Data Set), 806  
Box plots, 26–29
- Cardinal data, 327  
Carry-over effect, 667, 670–672  
Case study examples  
    confidence interval, 175–176  
    graphing of data, 29–31  
    multiple linear regression, 484–491  
nonparametric statistical methods, 344–345

- one-sample hypothesis testing, 256  
 one-way ANOVA, 538–548  
 two-sample hypothesis testing,  
   293–295  
 Case-control study, 589  
 Categorical data, measures of effect for,  
   591–601  
 Causal pathway, 609–610  
 Censored data, 761–763  
   interval, 763  
   left, 763  
   right, 763  
   treatment of, 761–763  
 Censored observations, 763  
 Central-limit theorem, 166–168  
 CHIDIST function, 244  
 Childhood Respiratory Diseases (CRD)  
   Study, 35, 512  
 Chinese Mini-Mental Status Test  
   (CMMS), 65  
 Chi-square distribution, 177–179, 824  
 Chi-square test  
   acceptance and rejection regions,  
 393, 395  
   goodness-of-fit test, 401–404  
   for homogeneity of odds ratios,  
 619–620  
   for homogeneity of rate ratios,  
 748–749  
   p-value, 393, 396–397  
   for  $R \times C$  contingency tables,  
 392–393  
   for trend in binomial proportions,  
 394–396  
   for trend-multiple strata, 623–624  
   and Wilcoxon rank-sum test,  
 397–400  
*Civil Action, A* (book and movie), 189  
 Clinical trials, 386–389  
 Cluster sampling, 151  
 Clustered binary data, 674–687  
   generalized estimated equations,  
 681  
   hypothesis testing, 674–675  
   power estimation for, 679–680  
   regression models, 681–686  
   sample-size estimation for, 679–680  
   two-sample test for binomial  
 proportions, 675–676, 679  
 CMMS (Chinese Mini-Mental Status  
   Test), 65  
 Coefficient of variation (CV), 20–21,  
   566–567  
 Cohort study, 589  
 Column effect, 550  
 Column margins, 358  
 Combinations, 79–83  
 Complement, 42  
 Complete-case method, 706  
 Compound symmetry correlation  
   structure, 681–682, 687  
 Concordant pair, 374  
 Conditional logistic regression,  
   649–651  
 Conditional probability, 46–51  
   relative risk, 47  
   total-probability rule, 48–51  
 Confidence interval. *See also* Interval  
   estimation  
   comparison of means from two  
 paired samples, 275–276  
   factors affecting length of, 174  
   and hypothesis testing, 235–237  
   for mean of normal distribution,  
 171–175  
   one-sided, 193–195  
   sample-size determination, 233–234  
 Confidence limits  
   for expectation of Poisson variable,  
 827  
   for incidence rates, 729–730  
 Confounding, 607–610  
   causal pathway, 609–610  
   negative confounder, 608  
   positive confounder, 608  
   stratification, 608  
   variable, 607  
 Contingency-table method, 357–359  
   2 x 2 contingency table, 357  
   expected table, 359–362  
   and multiple logistic regression,  
 632–633  
   observed table, 359  
   significance testing, 359–364  
   Yates-corrected chi-square test,  
 362–367  
 Continuity correction, 123  
 Continuous distribution, 3  
 Continuous probability distributions,  
   108–137  
   conversion from  $N(\mu, \sigma^2)$  to  $N(0, 1)$   
 distribution, 120–124  
   general concepts, 108–111  
   linear combinations of random  
 variables, 124–129  
   normal approximation to binomial  
 distribution, 129–135  
   normal approximation to Poisson  
 distribution, 135–137  
   normal distribution, 111–114  
   standard normal distribution,  
 114–120  
 Continuous random variable, 72  
   expected value of, 110  
   variance of, 111  
 Corrected sum of cross products, 432  
 Corrected sum of squares, 432  
 Correlation coefficient, 127–129,  
   452–455  
   dependent, comparison of, 465–467  
   interval estimation for, 460–462  
   multiple correlation, 493–494  
   one-sample  $t$  test, 455–457  
   one-sample  $z$  test, 457–460  
   overview, 452  
   partial correlation, 491–492  
   power and sample-size estimation,  
 463  
 rank correlation, 494–499  
 sample (Pearson) correlation  
   coefficient, 452–453  
 sample regression vs. sample  
   correlation coefficients, 453–455  
 sample vs. population correlation  
   coefficients, 453  
 sample-size estimation for, 462–463  
 statistical inference for, 455–467  
 two-sample  $z$  test, 463–465  
 Covariance, 126  
 Covariates, 546  
 Cox proportional-hazards model,  
   774–783  
   assumptions testing, 781–783  
   hazard ratio estimation for  
 continuous independent  
 variables, 775  
   hazard ratio estimation for  
 dichotomous independent  
 variables, 775  
   power estimation for, 781–783  
   sample-size estimation for, 786–787  
 CRD (Childhood Respiratory Diseases)  
   Study, 35, 512  
 Critical values, 209  
   Kruskal-Wallis test, 835  
   Spearman rank-correlation  
 coefficient, 834  
   standardized statistic, 836  
   Wilcoxon rank-sum test, 831–832  
   Wilcoxon signed-rank test, 830  
 Critical-value method, 209, 212, 836  
 Cross-over design, 666–674  
   assessment of treatment effects,  
 667–670  
   carry-over effect, 667, 670–672  
   definition, 666  
   sample-size estimation for, 672–674  
   washout period, 667  
 Cross-sectional study, 270, 589  
 Cumulative incidence, 59  
 Cumulative incidence rates, 725  
 Cumulative odds ordinal logistic  
   regression model, 656  
 Cumulative-distribution function, 78,  
   110, 115–116  
 CV (coefficient of variation), 20–21,  
   566–567  
 Data  
   cardinal, 327  
   interval scale, 327  
   nominal scale, 328  
   ordinal, 328  
   person-time. *See* Person-time data  
   ratio scale, 327  
 DBP (diastolic blood pressure), 571  
 Deductive reasoning, 149  
 Degrees of freedom, 169  
 Delta method, 593–594  
 Denominator degrees of freedom, 282  
 Dependent correlation coefficient,  
   465–467

- Dependent events, 43–44  
 Dependent random variables, 126–129  
 Dependent variable, 429  
 Descriptive statistics, 5–31  
     arithmetic mean, 13–15  
     case studies, 29–31  
     coefficient of variation, 20–21  
     computer packages, 31  
     graphic methods, 24–29  
     grouped data, 22–24  
     measures of location, 5–13  
     measures of spread, 15–18  
 Diabetes Prevention Study, 320  
 DIABETES.DAT (Data Set), 322, 349, 511  
 Diastolic blood pressure (DBP), 571  
 DIFCHISQ, 647  
 Direct standardization, 611–612  
 Discordant pair, 374–375  
 Discrete distribution, 3  
 Discrete probability distributions, 71–98  
     binomial distribution, 83–88  
     combinations, 79–83  
     cumulative-distribution function of discrete random variable, 78  
     expected value of discrete random variable, 75–76  
     permutations, 79–83  
 Poisson approximation to binomial distribution, 96–98  
 Poisson distribution, 90–93  
 Poisson probabilities, 93–94  
 probability-mass function, 73–75  
 random variable, 72–73  
 variance of discrete random variable, 76–77  
 Discrete random variable, 72  
     cumulative-distribution function, 78  
     expected value of, 75–76  
     probability-mass function for, 73–75  
     variance of, 76–77  
 Disease variable, 589  
 Disease-odds ratio, 596  
 Distribution  
     binomial, 71, 83–89  
     chi-square, 177–179  
     continuous, 3  
     discrete, 3  
     frequency, 3, 22–24, 73–75  
     mode, 11–12  
     negatively skewed, 10–11  
     Poisson, 71  
     positively skewed, 10–11  
     probability, 73–75  
     sampling, 160–161  
     symmetric, 10–11  
 Double blind, 159  
 Drop-in rate, 386  
 Dropout rate, 386  
 Dummy variable, 485, 541–542  
 Dunn procedure, 559–560  
 Effect modification, 618–620  
 El Paso Lead Study (LEAD.DAT), 29–30  
 ENDOCRIT.DAT, 580  
 EPESE (Established Populations for Epidemiologic Studies of the Elderly), 706  
 Epidemiologic studies  
     attributable risk, 601–606  
     clustered binary data, 674–687  
     confounding, 607–610  
     cross-over design, 666–674  
     equivalence studies, 663–666  
     extensions to logistic regression, 649–657  
     longitudinal data analysis, 687–696  
     Mantel-Haenszel test, 612–624  
     measures of effect for categorical data, 591–601  
     meta-analysis, 658–663  
     missing data, 706–711  
     multiple logistic regression, 628–649  
     power estimation for stratified categorical data, 625–628  
     sample-size estimation for stratified categorical data, 628  
     standardization, 610–612  
     study design, 588–591  
 Equal variances, 281–287  
     F distribution, 282–284  
     F test, 284–287  
 Equivalence studies, 663–666  
     definition, 663  
     inference based on confidence-interval estimation, 663–664  
     sample-size estimation for, 664–666  
 Error, in estimate, 4  
 Error mean square, 565  
 Error term, 550  
 ESD (Extreme Studentized Deviate) statistic, 296–300, 830  
 Established Populations for Epidemiologic Studies of the Elderly (EPESE), 706  
 Estimate, 4  
     error in, 4  
 Estimated mean difference, 4  
 Estimated regression line, 431, 433  
 Estimation, 149–195. *See also* Interval estimation; Point estimation; Power estimation; Sample-size estimation  
     binomial distribution, 181–188  
     central-limit theorem, 166–168  
     chi-square distribution, 177–179  
     interval, 168–169, 179–181, 184–188, 190–193  
     maximum-likelihood, 182–184  
     mean of distribution, 160–175  
     one-sided confidence intervals, 193–195  
     point, 160–162, 176–177, 181–182, 189–190  
     Poisson distribution, 189–193  
     population, 150–152  
     of power, 303–304  
     randomized clinical trials, 156–160  
 random-number tables, 152–156  
 sample, 150–152  
 standard error of the mean, 162–165  
 t distribution, 169–175  
 variance of distribution, 176–181  
 Estimator, 161  
 ESTRADL.DAT (Data Set), 422, 511  
 ESTROGEN.DAT (Data Set), 716  
 Events, 39  
     complement, 42  
     dependent, 43–44  
     exhaustive, 49  
     independent, 43  
     mutually exclusive, 40, 49  
     simultaneous, 41  
     symbol for, 40  
 Exact binomial probabilities, 811–815  
 Exact methods. *See also* Normal-theory methods  
     comparison of incidence rates, 734–736  
     McNemar's test, 377–380  
     one-sample binomial test (two-sided alternative), 247–249  
 Exact Poisson probabilities, 815–817  
 Excel, 4  
 Excel statistical package procedures  
     BINOMDIST, 87, 132  
     CHIDIST, 244  
     HYPGEOMDIST, 370–371  
     NORMDIST, 122–123  
     NORMINV, 122  
     NORMSINV, 501–502  
     POISSON, 192  
     TDIST, 211–212  
     TINV, 174, 209–210  
 Exchangeable correlation structure, 681–682, 687  
 Exhaustive events, 49  
 Expected mean square, 566  
 Expected table, 359–362, 391–392  
 Expected value  
     for 2 x 2 contingency tables, 360–362  
     of binomial distribution, 88–89  
     of continuous random variable, 110  
     of discrete random variable, 75–76  
     of hypergeometric distribution, 369–370  
     of linear combinations of random variables, 125–126  
     of Poisson distribution, 95–96  
 Exposure variable, 589  
 Exposure-odds ratio, 596–599  
 Externally Studentized residual, 478  
 Extreme outlying value, 28  
 Extreme Studentized Deviate (ESD) statistic, 296–300, 830  
 F distribution, 282–284  
     denominator degrees of freedom, 282  
     lower percentiles of, 284  
     numerator degrees of freedom, 282

- percentage of, 828–830  
 $p$ th percentile, 283  
*F* test, 284–287  
 acceptance and rejection regions, 438  
 for fixed-effects one-way ANOVA, 518–521  
 for multiple regression, 472  
 $p$ -value, 438  
 for simple linear regression, 437–441
- Factorial, 80–81  
 False negative, 52  
 False positive, 52  
 False-discovery rate, 536–538  
 FEF (forced mid-expiratory flow), 516  
 FEV.DAT (Data Set), 35, 315–316, 506, 512–513  
 Fisher's exact test, 367–373  
 exact probability of observing table with cells, 369  
 general layout of data, 368  
 general procedure, 371–372  
 hypergeometric distribution, 369–371  
 $p$ -value, 371  
 Fisher's *z* test, 463–465  
 acceptance and rejection regions, 464  
 for comparing two correlation coefficients, 464  
 $p$ -value, 465  
 Fisher's *z* transformation, 458, 833  
 Fitted regression lines, 448–451  
 assumptions, 448  
 influential points, 450  
 outliers, 450  
 standard deviation of residuals, 448–449  
 Studentized residuals, 448–449  
 Fixed-effects one-way ANOVA (analysis of variance), 516–518, 563. *See also* One-way ANOVA (analysis of variance)  
 acceptance and rejection regions, 521  
 between-group variability, 518–519  
*F* test, 520–521  
*F* test for group means comparison, 518  
 hypothesis testing, 518–522  
 interpretation of parameters, 517–518  
 $p$ -value, 521  
 within-group variability, 518–519
- Flat distribution, 238  
 Flowcharts  
 categorical data method for statistical inference, 409  
 methods for statistical inference, 503  
 person-time data methods, 803  
 two-sample statistical inference, 308
- Follow-up study, 270  
 Forced mid-expiratory flow (FEF), 516
- Framingham Eye Study, 50  
 Framingham Heart Study, 638–642  
 Frequency definition of probability, 56  
 Frequency distribution, 3, 22–24, 73–75
- Gaussian distribution, 108, 111–114  
 Generalized estimated equations (GEE), 681  
 Geometric mean, 12–13  
 Goodness of fit  
 of logistic-regression models, 643–649  
 regression lines, 435–436, 448–451  
 of Weibull survival model, 794  
 Goodness-of-fit test, 74, 401–404  
 Gossett, William, 169  
 Grand total, 358  
 Graphic methods, 24–29  
 bar graph, 25  
 box plots, 26–29  
 stem-and-leaf plots, 25–26  
 Greene-Touchstone study, 428–429  
 Grouped data, 22–24
- Harvard Medical Study, 415, 425  
 Harvard Pilgrim Health Care, 200  
 Hazard function, 760, 766, 789  
 Hazard rates, 759  
 Hazard ratio, 775  
 Heavy smokers, 516  
 Homogeneity of binomial proportions, 358  
 HORMONE.DAT (Data Set), 315, 349, 413, 506, 580, 716  
 HOSPITAL.DAT (Data Set), 33, 555  
 Hypergeometric distribution, 369–371  
 HYPGEOMDIST function, 370–371  
 Hypothesis testing, 149–150  
 acceptance and rejection regions, 473  
 Bayesian inference, 237–241  
 case study, 256, 293–295  
 chi-square goodness-of-fit test, 401–404  
 clustered binary data, 674–675  
 comparison of means from two paired samples, 275–276  
 and confidence intervals, 235–237  
 critical-value method, 209, 212  
 equality of two variances, 281–287  
*F* test, 472  
 Fisher's exact test, 367–373  
 fixed-effects one-way ANOVA, 518–522  
 general concepts, 204–207  
 interval estimation, 280–281  
 multiple logistic regression, 636–642  
 multiple-regression analysis, 472–476  
 null hypothesis, 205–206  
 one-sample  $\chi^2$  test for variance of normal distribution, 241–244  
 one-sample inference for binomial distribution, 244–250  
 one-sample problem, 204  
 one-sample test for mean of normal distribution, 207–221  
 one-sided alternatives, 207–215  
 outliers, 295–301  
 paired *t* test, 271–274  
 power estimation for comparing binomial proportions, 381–389  
 power of test, 221–226  
 $p$ -value, 473  
*R*  $\times$  *C* contingency tables, 390–400  
 sample-size determination, 228–234  
 sample-size estimation for comparing binomial proportions, 381–389  
 stratified person-time data, 742–745  
*t* test, 474  
 true state of nature, 207  
 two-sample inference, 269–307  
 two-sample problem, 204  
 two-sample *t* test for independent samples with equal variances, 276–279  
 two-sample *t* test for independent samples with unequal variances, 287–293  
 two-sample test for binomial proportions, 353–367  
 two-way ANOVA, 550–553  
 type I error, 206–207  
 type II error, 206–207
- Imputation, 707  
 Incidence, 59  
 Incidence density, 59, 726  
 Incidence rates, 725  
 confidence limits, 729–730  
 and cumulative incidence, 726  
 exact test, 734–736  
 interval estimation, 729–730  
 log-rank test, 768–769  
 normal-theory test, 731–733  
 one-sample inference, 727–729  
 point estimation, 729–730  
 power estimation for, 753–754  
 rate ratio, 736–737  
 sample-size estimation, 750–751  
 trend testing, 755–758  
 two-sample inference, 730–737
- Independent events, 43  
 Independent samples, 381–383, 387–388  
 Independent variable, 429  
 Independent-sample design, 270  
 Inductive reasoning, 149  
 INFANTBP.DAT (Data Set), 199, 315, 465–467, 507  
 Inferential statistics, 3  
 Influential points, 450  
 Interaction effect, 550  
 Internally Studentized residual, 478  
 Interval censoring, 763

- Interval estimation, 150, 168–169, 179–181. *See also* Confidence interval; Estimation; Sample-size estimation  
 for attributable risk, 603  
 comparison of means from two independent samples, 280–281  
 comparison of means from two paired samples, 275–276  
 for correlation coefficients, 460–462  
 exact methods, 186–188, 190  
 for incidence rates, 729–730  
 intraclass correlation coefficient, 569  
 for linear regression, 443–447  
 multiple logistic regression, 642–643  
 normal-theory methods, 184–186  
 for Poisson distribution, 190–193  
 for predictions from regression lines, 445–447  
 rank-correlation coefficients, 499–503  
 of rate ratio, 736–737  
 rate ratio, 746–747  
 for regression parameters, 443–445  
 risk difference, 592–593  
 for risk ratio, 594–595, 599–601  
 of survival probabilities, 764–765
- Interval scale, 327
- Intraclass correlation coefficient, 568–571. *See also* Correlation coefficient  
 definition, 568  
 interpretation of, 569  
 interval estimation, 569  
 as measure of reliability, 571  
 point estimation, 569
- Inverse normal function, 119
- Kaplan-Meier estimator, 760–767  
 estimation of hazard function, 766  
 interval estimation of survival probabilities, 764–765  
 treatment of censored data, 761–763
- Kappa statistic, 404–408  
 acceptance and rejection regions, 406  
 guidelines for evaluating kappa, 408  
*p*-value, 407
- Kruskal-Wallis test, 555–561  
 acceptance and rejection regions, 557  
 comparison of specific groups, 559–560  
 critical values for, 835  
*p*-value, 558  
 procedure, 556–557  
 rank assignment, 558
- Large-sample test, 727–728
- LEAD.DAT (Data Set), 29, 344
- LEAD.DOC (Data Set), 29
- Least significant difference (LSD), 524–528
- Least-squares line, 431, 433
- Left censoring, 763
- Light smokers, 516
- Likelihood, 182
- Linear combinations of random variables, 124–129  
 dependent random variables, 126–129  
 expected value of, 125–126  
 variance of, 125–126, 129
- Linear contrast, 528–531  
 definition, 125  
 multiple-comparisons procedure, 534–536  
*t* test, 529  
 variance of, 129
- Linear regression  
 $F$  test for, 437–441  
 interval estimation for, 443–447  
 simple, 437–445  
 standard deviation of residuals, 448–449  
 standard errors for estimated parameters, 444–445  
*t* test for, 441–443
- Linear-regression methods, 427–451  
 assumptions, 448  
 dependent variable, 429  
 $F$  test for simple linear regression, 437–441  
 independent variable, 429  
 interval estimates for regression parameters, 443–445  
 interval estimation, 443–447  
 interval estimation for predictions, 445–447  
 method of least squares, 431–435  
 overview, 427  
 regression line, 428, 430  
 standard errors for estimated parameters, 444–445  
*t* test for simple linear regression, 441–443
- Logistic regression  
 conditional, 649–651  
 interval estimation, 642–643  
 matched, 649–653  
 ordinal, 656–657  
 point estimation, 642–643  
 polychotomous, 653–656  
 residuals in, 643–647
- Logit transformation logit ( $p$ ), 629
- Log-rank test, 767–773  
 acceptance and rejection regions, 771  
 incidence rates, 768–769  
 procedure, 769–770  
*p*-value, 771
- Longitudinal data analysis, 687–696  
 in clinical trial setting, 688–691  
 interpretation of parameters, 687–688  
 measurement-error methods, 696–706
- Longitudinal study, 270, 304–307
- LSD (least significant difference), 524–528
- Mann-Whitney U test. *See* Wilcoxon rank-sum test
- Mantel extension test, 622–624
- Mantel-Haenszel test, 612–624  
 acceptance and rejection regions, 615  
 effect modification, 618–620  
 estimation in matched-pair studies, 620–621  
 estimation of odds ratio for stratified data, 616–618  
 procedure, 614–615  
*p*-value, 615  
 testing for trends in presence of confounding, 622–624
- Masking problem, 298
- Matched logistic regression, 649–653
- Matched pair, 373–380  
 concordant, 374  
 discordant, 374  
 estimation of odds ratio, 620–621  
 two-sample test for binomial proportions, 373–380  
 type A discordant pair, 374  
 type B discordant pair, 374–375
- Matched-pair design, 79
- Mathematical statistics, 1
- MAXFWT (mean finger-wrist tapping score), 484–491, 538–541
- Maximum-likelihood estimation, 182–184
- McNemar's test, 373–380  
 acceptance and rejection regions, 376  
 for correlated proportions, 375, 377  
 exact test, 377–380  
 normal-theory test, 375–377  
*p*-value, 376, 378
- Mean  
 arithmetic, 7–9, 13–15  
 geometric, 12–13  
 standard error of, 162–165
- Mean deviation, 17
- Mean finger-wrist tapping score (MAXFWT), 484–491, 538–541
- Mean of distribution, 160–175  
 central-limit theorem, 166–168  
 confidence interval, 171–176  
 estimation of, 160–175  
 interval estimation, 168–169  
 one-sample test, 207–215  
 point estimation, 160–162  
 standard error of the mean, 162–165  
 $t$  distribution, 169–175
- Measurement-error methods, 696–706  
 measurement-error correction with gold-standard exposure, 697–701  
 measurement-error correction without gold-standard exposure, 701–703

- regression-calibration approach, 699–700, 703–704
- Measures of effect for categorical data, 591–601
- odds ratio, 595–601
  - risk difference, 592–593
  - risk ratio, 593–595
- Measures of location, 5–13
- arithmetic mean, 7–9
  - geometric mean, 12–13
  - median, 9–10
  - mode, 11–12
- Measures of spread, 15–18
- quantiles (percentiles), 16–17
  - range, 15–16
  - standard deviation, 17–20
  - variance, 17–20
- Median, 9–10
- Meta-analysis, 658–663
- models, 662
  - random-effects model, 659–660
  - tests of homogeneity of odds ratios, 661–663
- Method of least squares, 431–435
- corrected sum of cross products, 432
  - corrected sum of squares, 432
  - estimation of least-squares lines, 433
  - least-squares line, 431
  - raw sum of cross products, 432
  - raw sum of squares, 432
- MICE.DAT (Data Set), 582
- Minimum variance unbiased estimator, 161
- MINITAB package, 4, 22–24, 133, 154, 798–799
- Minnesota Heart Study, 151
- Missing data, 706–711
- Mode, 11–12
- Moderate smokers, 516
- Multiple correlation, 493–494
- Multiple imputation, 707
- Multiple linear regression, 468–491.
- See also* Simple linear regression
  - case study, 484–491
  - estimation of regression equation, 468–471
  - goodness of fit, 476–482
  - hypothesis testing, 472–476
  - multiple correlation, 493–494
  - and one-way ANOVA, 541–545
  - partial correlation, 491–493
  - partial F test, 476
  - partial-regression coefficients, 476
  - partial-residual plot, 479–483
  - rank correlation, 494–499
  - standardized regression coefficient, 471
- Multiple logistic regression, 628–649
- acceptance and rejection regions, 637
  - and contingency-table analysis, 632–633
  - estimation of odds ratio for continuous independent variables, 635–636
- estimation of odds ratio for dichotomous independent variables, 631–632
- goodness of fit, 643–649
- hypothesis testing, 636–642
- interval estimation, 642–643
- model, 628–629
- point estimation, 642–643
- prediction with, 642–643
- p-value, 637
- regression parameters, 629–631
- residuals in, 643–647
- Multiplication law of probability, 42–44
- Multisample inference, 516–577
- intraclass correlation coefficient, 568–571
  - Kruskal-Wallis test, 555–561
  - mixed models, 572–576
  - one-way ANOVA, 516–548
  - random-effect one-way ANOVA, 562–568
- Mutually exclusive events, 40, 49
- Negative cofounder, 748
- Negatively skewed distribution, 10–11
- NEPHRO.DAT (Data Set), 658, 714–715
- NIFED.DAT (Data Set), 200, 262, 313
- Nominal scale, 328
- Noninformative prior distribution, 238
- Noninhaling smokers, 516
- Nonparametric statistical methods, 327–345
- case study, 344–345
  - chi-square goodness-of-fit test, 401–404
  - Fisher's exact test, 367–373
  - kappa statistic, 404–408
  - McNemar's test, 373–380
  - R x C contingency tables, 390–400
  - sample size and power, 381–389
  - sign test, 329–333
  - Wilcoxon rank-sum test, 339–343
  - Wilcoxon signed-rank test, 332–339
- Nonsmokers, 516
- Normal approximation
- binomial distribution, 129–135
  - Poisson distribution, 135–137
- Normal distribution, 108, 111–114. *See also* Standard normal distribution
- electronic tables, 118–119
- NORMDIST function of Excel, 122–123
- NORMINV function of Excel, 122
- one-sample test, 207–215
- probability-density function of, 115
- p<sup>th</sup> percentile, 123
- standard, 114–120
- table, 818–821
- Normal range, 118
- Normal variable, standardization of, 121
- Normal-theory methods, 184–186, 244–247. *See also* Exact methods
- McNemar's test, 375–377
- sign test, 329–332
- two-sample test for binomial proportions, 353–356
- Normal-theory test, 731–733
- Normative Aging Study, 63
- NORMDIST function, 122–123
- NORMINV function, 122
- NORMSINV function, 501–502
- Null hypothesis, 205–206
- Numerator degrees of freedom, 282
- Nurses' Health Study, 151, 562, 625
- Observed contingency table, 359
- Odds ratio, 595–601
- chi-square test for homogeneity of, 619–620
  - disease-odds ratio, 596
  - exposure-odds ratio, 596–599
  - interval estimation for, 599–601
  - in meta-analysis, 661–663
  - and multiple logistic regression, 631–632
  - odds in favor of success, 595
  - point estimation for, 600
  - probability of success, 595
  - test of homogeneity, 661–663
- One-sample  $\chi^2$  test for variance of normal distribution, 241–244
- One-sample inference, 204–256
- Bayesian inference, 237–241
  - for binomial distribution, 244–250
  - exact methods, 247–249
  - general concept, 204–207
  - normal-theory methods, 244–247
  - one-sample test for mean of normal distribution, 207–221
  - one-sample test for variance of normal distribution, 241–244
  - for Poisson distribution, 251–256
  - power and sample-size estimation, 249–250
  - power of test, 221–228
  - sample-size determination, 228–234
  - two-sided alternatives, 215–221
- One-sample inference for incidence-rate data, 727–730
- exact test, 728–729
  - large-sample test, 727–728
- One-sample problem, 204
- One-sample *t* test, 455–457
- acceptance and rejection regions, 457
  - for correlation coefficient, 456
  - p-value, 457
- One-sample test
- one-sided alternatives, 207–215
  - two-sided alternatives, 215–221
  - z* test, 220–221
- One-sample *z* test
- acceptance and rejection regions, 459
  - p-value, 460
  - z* transformation of *r*, 458

- One-sided alternatives, 207–215  
 power of test, 221–226  
 sample-size determination, 228–232
- One-sided confidence intervals, 193–195
- One-tailed test, 208
- One-way ANCOVA (analysis of covariance), 545–548
- One-way ANOVA (analysis of variance), 517, 525. *See also* Two-way ANOVA (analysis of variance)  
 acceptance and rejection regions, 521, 525
- Bonferroni multiple-comparisons procedure, 531–534
- case study, 538–548
- comparison of specific groups, 521–538
- dummy variable, 541–542
- F* test, 518–521
- false-discovery rate, 536–538
- linear contrast, 528–531
- LSD procedure, 524–528
- and multiple regression, 541–545
- multiple-comparisons procedure for linear contrasts, 534–536
- pooled estimate of variance, 523–524
- p*-value, 521, 525
- random-effects model, 563–568
- t*-test based on pairs of groups, 522–528
- Ordinal data, 328
- Ordinal logistic regression, 656–657
- OTO.DAT (Data Set), 714–715
- Outliers, 295–300, 450
- Outlying value, 28
- p* (Logit transformation logit), 629
- Paired samples, 383–386
- Paired *t* test, 271–274
- Paired-sample design, 270
- Parametric statistical methods, 327
- Parametric survival analysis, 787–795
- Partial correlation, 491–493
- partial-regression coefficients, 476
- Partial-residual plot, 479–483
- Passive smoking, 516
- PC-SAS TTEST, 294
- Pearson correlation coefficient, 452–455  
 definition, 452  
 interpretation, 452–453  
 vs. population correlation coefficient, 453  
 vs. sample regression coefficient, 453–455
- Pearson residual, 643–647
- Percentiles, 16–17, 793–794, 799–802
- Permutations, 79–83
- Person-time data, 725–802  
 cumulative incidence, 725–726  
 incidence density, 726  
 for incidence-rate data, 730–737
- Kaplan-Meier estimator, 760–767  
 log-rank test, 767–773
- one-sample inference for incidence-rate data, 727–730
- parametric regression models for survival data, 795–802
- parametric survival analysis, 787–795
- power estimation for, 738–740
- proportional-hazards model, 774–783
- sample-size estimation, 740–742
- stratified, 742–750
- survival analysis, 758–760
- trend testing, 755–758
- Physician's Health Study, 389, 590–591
- Point estimates, 150
- Point estimation. *See also* Estimation; Interval estimation  
 binomial distribution, 181–182  
 for incidence rates, 729–730  
 intraclass correlation coefficient, 569  
 mean of distribution, 160–162  
 multiple logistic regression, 642–643  
 for odds ratio, 600  
 Poisson distribution, 189–190  
 rate ratio, 736–737, 746–747  
 risk difference, 592–593  
 for risk ratio, 594–595  
 variance of distribution, 176–177
- Poisson approximation, 96–98
- Poisson distribution, 71, 90–93  
 electronic tables, 94  
 estimation for, 189–193  
 expected value of, 95–96  
 interval estimation, 190–193  
 normal approximation to, 135–137  
 one-sample inference, 251–256  
 point estimation, 189–190  
 Poisson tables, 93–94  
 variance of, 95–96
- POISSON function, 192
- Poisson probabilities, 93–94
- Poisson tables, 93–94
- Poisson variable, 827
- Polychotomous logistic regression, 653–656
- Population, 150–152
- Population correlation coefficient, 453
- Population variance, 76–77
- Positive confounder, 608
- Positively skewed distribution, 10–11
- Posterior distribution, 237
- Posterior predictive interval, 239
- Posterior probability, 57
- Power estimation. *See also* Estimation  
 in clinical trial setting, 386–389  
 for clustered binary data, 679–680  
 for comparing two binomial proportions, 381–389  
 comparing two means, 303–304  
 for comparison of two incidence rates, 739–740
- for correlation coefficients, 463
- for incidence rates, 753–754
- for person-time data, 738–742
- for proportional-hazards model, 783–785
- for stratified categorical data, 625–628
- for stratified person-time data, 753–754
- Power of test, 221–226
- Predictive value negative (PV<sup>−</sup>), 51
- Predictive value positive (PV<sup>+</sup>), 51
- Prevalence, 59
- Prevalence study, 589
- Prior distribution, 237
- Prior probability, 56
- Probability, 38–59  
 addition law of, 44–46  
 Bayes' rule, 53–56  
 Bayesian inference, 56–57  
 conditional, 46–51  
 definition, 39  
 event, 39  
 frequency definition of, 56–57  
 incidence, 59  
 multiplication law of, 42–44  
 mutually exclusive events, 40  
 notation, 40–42  
 posterior, 57  
 prevalence, 59  
 prior, 56  
 receiver operating characteristic curve, 57–59  
 sample space, 39  
 screening tests, 51–52  
 total-probability rule, 48–51
- Probability distribution, 73–75
- Probability model, 4
- Probability-density function, 109
- Probability-mass function, 73–75
- PROC GENMOD, 682–686
- PROC MI, 708
- PROC MIANALYZE, 708
- PROC MIXED, 688–691
- PROC TTEST, 291–292
- Product-limit estimator, 761
- Product-limit method, 766
- Proportional Hazards Regression (SAS PHREG), 776
- Proportional odds ordinal logistic regression model, 656
- Proportional-hazards model, 774–783  
 assumptions testing, 781–783  
 hazard ratio estimation for continuous independent variables, 775  
 hazard ratio estimation for dichotomous independent variables, 775  
 power estimation for, 781–783  
 sample-size estimation for, 786–787
- Weibull survival model, 797
- Proportional-mortality study, 248
- Prospective study, 589

- Pseudorandom numbers, 153  
 $PV^-$  (predictive value negative), 51  
 $PV^+$  (predictive value positive), 51  
 $p$ -value, 210–215
  - chi-square goodness-of-fit test, 403
  - chi-square test for trend in binomial proportions, 396–397
  - exact method, 332–333
  - F* test for equality of two variances, 284–287
  - F* test for one-way ANOVA, 521
  - Fisher's exact test, 371
  - hypothesis testing in multiple linear regression, 473
  - kappa statistic, 407
  - log-rank test, 771
  - Mantel-Haenszel test, 615
  - McNemar's test, 376, 378
  - multiple logistic regression, 637
  - normal-theory method, 329–332
  - normal-theory test, 356
  - one-sample binomial test (exact method), 247
  - one-sample *t* test for correlation coefficient, 457
  - one-sample *t* test for mean of normal distribution, 218
  - paired *t* test, 272
  - R x C* contingency tables, 393
  - sign test (exact method), 332–333
  - sign test (normal-theory method), 330–331
  - Spearman rank-correlation coefficient, 497
  - statistical significance of, 212
  - t* test for comparison of pairs of groups in one-way ANOVA, 525
  - two-sample inference for incidence-rate data, 735–736
  - two-sample *t* test for independent samples with unequal variances, 289–291
  - two-sample test for binomial proportions, 356
  - two-sample test for incidence rates (normal-theory method), 733
  - Yates-corrected chi-square test, 363
- Quantiles, 16–17
- R x C* contingency tables, 390–400
  - acceptance and rejection regions, 393
  - chi-square test, 392–393
  - chi-square test for trend in binomial proportions, 394–396
  - definition, 390
  - expected table, 391–392
  - $p$ -value, 393
  - test for association, 390–394
  - Wilcoxon rank-sum test, 397–400
- $R^2$ , 439–440
- Random assignment, 154
- Random digits, 822
- Random effects one-way ANOVA (analysis of variance), 562–568
- Random numbers, 152
- Random sample, 150
- Random selection, 153
- Random variables, 72–73
  - continuous, 72
  - correlation coefficient, 127–129
  - cumulative-distribution function, 110
  - dependent, 126–129
  - discrete, 72
  - linear combinations of, 124–129
  - probability-density function, 109
  - standard deviation, 76
- Random-effects one-way ANOVA (analysis of variance)
  - balanced case, 563
  - unbalanced case, 563
- Randomization, 156
- Randomized clinical trials, 156–160
  - design features, 158–159
  - double blind, 159
  - single blind, 159
  - stratification, 159
  - unblinded, 159
- Random-number tables, 152–156
- Range, 15–16
- Rank sum, 336
- Rank-correlation coefficients, 494–499
  - acceptance and rejection regions, 497
  - interval estimation for, 499–503
  - $p$ -value, 498
  - Spearman, 495–499
  - t* test for, 496
- Ranking procedure, 335
- Rate ratio, 736–737
  - chi-square test for homogeneity of, 748–749
  - interval estimation, 736–737, 746–747
  - point estimation, 736–737, 746–747
  - stratified person-time data, 745–748
- Ratio scale, 327
- Raw sum of cross products, 432
- Raw sum of squares, 432
- Receiver operating characteristic (ROC) curve, 57–59
- Recidivism rate, 188
- Reference population, 151
- Regression coefficient from first imputed data set, 710
- Regression component, 435
- Regression line, 428–430
  - estimated, 431, 433
  - fitting, 431–435
  - goodness of fit, 435–436, 448–451
  - inferences about parameters from, 435–443
  - interval estimates for regression parameters, 443–445
  - interval estimation for predictions, 445–447
- least-squares line, 431
- method of least squares, 431–435
- predicted value, 434
- predictions for individual observations, 445
- regression component, 435
- regression sum of squares, 436
- residual component, 435
- residual sum of squares, 437
- slope and intercept of, 432
- standard deviation of residuals, 448–449
- total sum of squares, 436
- two-sided 100%  $x (1-\alpha)$  confidence intervals, 444
- Regression mean square, 437
- Regression parameters, interval estimation for, 443–445
- Regression sum of squares, 436
- Regression to the mean, 241
- Rejection region, 208
  - Bonferroni multiple-comparisons procedure, 532
  - chi-square goodness-of-fit test, 403
  - chi-square test for  $R \times C$  contingency tables, 393
  - chi-square test for trend in binomial proportions, 395
  - Dunn procedure, 560
  - F* test for one-way ANOVA, 521
  - F* test for simple linear regression, 438
  - Fisher's *z* test, 464
  - fixed-effects one-way ANOVA (analysis of variance), 521
  - hypothesis testing in multiple linear regression, 473
  - kappa statistic, 406
  - Kruskal-Wallis test, 557
  - log-rank test, 771
  - Mantel-Haenszel test, 615
  - McNemar's test, 376
  - multiple logistic regression, 637
  - one-sample binomial test (normal-theory method), 245
  - one-sample  $\chi^2$  test for variance of normal distribution, 242
  - one-sample *t* test for correlation coefficient, 457
  - one-sample *z* test for correlation coefficient, 459
  - one-way ANOVA, 521, 525
  - R x C* contingency tables, 393
  - rank-correlation coefficients, 497
  - sign test, 330
  - Spearman rank-correlation coefficient, 497
  - t* test for comparison of pairs of groups in one-way ANOVA, 525
  - t* test for multiple linear regression, 474
  - t* test for simple linear regression, 442
  - two-sample inference for incidence-rate data, 732

- two-sample test for binomial proportions (normal-theory test), 355
- two-sample test for incidence rates (normal-theory method), 732
- Yates-corrected chi-square test, 363
- z* test, one-sample, 459
- z* test, two-sample, 464
- Relative risk (*RR*), 47, 592
- Reliability coefficient, 571
- Reproducibility studies, 566
- Rescaled sample, 14–15
- Residual component, 435
- Residual mean square, 437
- Residual sum of squares, 437
- Retrospective study, 589
- Right censored data, 763
- Risk difference, 592
- interval estimation, 592–593
  - point estimation, 592–593
- Risk ratio, 593–595
- definition, 592
  - delta method, 593–594
  - estimation for case-control studies, 598
  - interval estimation, 594–595
  - point estimation, 594–595
- ROC (receiver operating characteristic) curve, 57–59
- Row effect, 550
- Row margins, 358
- RR* (relative risk), 47, 592
- Sample, 150–152
- independent, 381–383
  - median, 9–10
  - paired, 383–386
  - random, 150
  - regression coefficient, 453–455
  - space, 39
  - standard deviation. *See* Standard deviation
  - variance. *See* Variance
- Sample (Pearson) correlation coefficient, 452–455. *See also* Correlation coefficient
- definition, 452
  - interpretation, 452–453
  - vs. population correlation coefficient, 453
  - vs. sample regression coefficient, 453–455
- Sample-size estimation, 228–234. *See also* Interval estimation
- based on CI width, 233–234
  - in clinical trial setting, 386–389
  - for clustered binary data, 679–680
  - for comparing two binomial proportions, 381–389
  - for comparison of two incidence rates, 740–742
  - for correlation coefficients, 462–463
  - for cross-over design, 672–674
- for equivalence studies, 664–666
- for incidence-rate data, 750–751
- independent samples, 381–383, 387–388
- for longitudinal studies, 304–307
- one-sided alternatives, 228–234
- paired samples, 383–386
- and power, 249–250
- for proportional-hazards model, 786–787
- for stratified categorical data, 628
- for stratified person-time data, 750–753
- two-sample inference, 301–303
- two-sided alternatives, 232–233
- Sampling distribution, 160
- SAS, 4
- SAS General Linear Model procedure, 550–552, 565–566
- SAS PHREG (Proportional Hazards Regression), 776
- SAS PROC GENMOD program, 682–686
- SAS PROC LOGISTIC program, 633–634
- SAS PROC PHREG program, 651–653
- SAS PROC REG program, 469–470
- Satterwaite's method, 288–289
- Scatter plot, 5, 7
- Scheffé's multiple-comparison procedure, 534–536
- Screening tests, 51–52
- false negative, 52
  - false positive, 52
  - predictive value, 51
  - sensitivity of symptom, 51
  - specificity of symptom, 52
- SEER Tumor Registry, 40
- Sensitivity of symptom, 51
- SEXRAT.DAT (Data Set), 103, 199, 262, 418
- SHEP (Systolic Hypertension in the Elderly Program), 156–157
- Sign test, 329–333
- acceptance and rejection regions, 330
  - exact method, 332–333
  - normal-theory method, 329–332
  - p*-value, 330–333
- Significance level, 206
- Simple linear regression. *See also* Multiple linear regression
- F* test for, 437–441
  - standard errors for estimated parameters, 444–445
  - t* test for, 441–443
- Simple random sample, 150
- Single blind, 159
- SMOKE.DAT (Data Set), 102–103, 348, 508, 758–760, 776, 804
- Spearman rank-correlation coefficient, 495–499
- acceptance and rejection regions, 497
- interval estimation for, 500
- p*-value, 498
- t* test for, 496
- two-tailed critical values for, 834
- Specificity of symptom, 52
- Spread, 15
- SPSS, 4
- SPSS<sup>x</sup>/PC CROSSTABS program, 366, 380
- SPSS<sup>x</sup>/PC McNemar's test program, 380
- Standard deviation, 17–20
- properties of, 18–20
  - of random variable, 76
- Standard error of the mean, 162–165
- Standard normal distribution, 114–120.
- See also* Normal distribution
  - (100 *x u*)th percentile, 119
  - cumulative-distribution function, 115–116
  - electronic tables, 118–120
  - normal tables, 115–118
  - p*th percentile, 119
  - symmetry properties, 116–117
- Standardization, 610–612
- age-standardized risk, 611
  - direct, 611–612
  - of normal variable, 121
- Standardized morbidity ratio, 192, 253
- Standardized mortality ratio, 253–254
- Standardized regression coefficient, 471
- Statistics, 1, 3
- Steam-and-leaf plots, 25–26
- Step function, 78
- Strata, 4, 608
- Stratification, 159, 608
- Stratified person-time data
- estimation of rate ratio, 745–748
  - homogeneity of rate ratio, 748–749
  - hypothesis testing, 742–745
  - inference for, 742–750
  - power estimation for, 753–754
  - sample-size determination, 750–753
- Studentized residuals, 448–449, 478–479
- Student's *t* distribution, 169–175
- Study design, 588–591
- case-control study, 589
  - cohort study, 589
  - cross-sectional study, 589
  - prevalence study, 589
  - prospective study, 589
  - retrospective study, 589
- Study population, 151
- Sufficient statistic, 238
- Survival analysis, 758–760
- hazard function, 760
  - hazard rates, 759
- Survival function, 760, 766
- Survival probability, 760
- SWISS.DAT (Data Set), 319, 582, 720, 805
- Symmetric distribution, 10–11
- Systolic Hypertension in the Elderly Program (SHEP), 156–157

- t* distribution, 169–175, 823  
*t* test  
 acceptance and rejection regions, 442, 474  
 for comparison of pairs of groups, 522–528  
 hypothesis testing in multiple linear regression, 474  
 for linear contrasts in one-way ANOVA, 529  
 for multiple linear regression, 474  
 one-sample, for correlation coefficients, 455–457  
 for simple linear regression, 441–443  
 Spearman rank-correlation coefficient, 496
- Tables  
 2 x 2 contingency tables, 360–362  
 binomial, 83–86  
 binomial distribution, 87–88  
 chi-square distribution, 824  
 confidence limits for expectation of Poisson variable, 827  
 critical values for standardized statistic, 836  
 exact binomial probabilities, 811–815  
 exact Poisson probabilities, 815–817  
 expected, 359, 391–392  
 Extreme Studentized Deviate outlier statistic, 830  
 Fisher's *z* transformation, 833  
 Kruskal-Wallis test, 835  
 normal distribution, 115–119, 818–821  
 observed, 359  
 percentage of *F* distribution, 828–830  
 Poisson, 93–94  
 Poisson distribution, 94  
 Poisson variable, 827  
*R* x *C* contingency tables, 390–400  
 random digits, 822  
 random-number, 152–156  
 standard normal distribution, 118–120  
*t* distribution, 823  
 two-tailed critical values for Spearman rank-correlation coefficient, 834  
 two-tailed critical values for Wilcoxon rank-sum test, 831–832  
 two-tailed critical values for Wilcoxon signed-rank test, 830  
 Target population, 151  
 TDIST function, 211–212  
 TEAR.DAT (Data Set), 265, 350, 584  
 TENNIS1.DAT (Data Set), 415, 656, 715  
 TENNIS2.DAT (Data Set), 316, 346, 666, 670, 716  
 Test for homogeneity of binomial proportions, 358  
 Test of association, 359  
 Test of independence, 359  
 Test statistic, 209  
 TINV function, 174, 209–210  
 Total sum of squares, 436, 518  
 Total-probability rule, 48–51  
 Translated sample, 13–14  
 Treatment efficacy, assessment of, 667–670  
 Trend, testing for, 755–758  
 Trimodal distribution, 12  
 True mean difference, 4  
 True state of nature, 207  
 T-TEST program, 273–274  
 Tukey approach, 575  
 Two-sample hypothesis-testing problem, 269  
 Two-sample inference, 269–307  
 case study, 293–295  
 comparison of means from two independent samples, 280–281  
 equality of two variances, 281–287  
 estimation of power, 303–304  
 estimation of sample size, 301–303  
*F* test, 284–287  
 interval estimation, 275–276, 280–281  
 outliers, 295–301  
 paired *t* test, 271–274  
 two-sample *t* test for independent samples with equal variances, 276–279  
 two-sample *t* test for independent samples with unequal variances, 287–293  
 Two-sample inference for incidence-rate data, 730–737  
 acceptance and rejection regions, 732  
 exact test, 734–736  
 hypothesis testing, 730–731  
 normal-theory test, 731–734  
*p*-value, 733, 735–736  
 rate ratio, 736–737  
 Two-sample problem, 204  
 Two-sample test for binomial proportions, 353–367  
 acceptance and rejection regions, 355  
 clustered data, 675–676  
 contingency-table method, 357–367  
 equal number of sites per individual, 679  
 for matched-pair data, 373–380  
 normal-theory method, 353–356  
*p*-value, 356  
 Two-sided alternatives, 215–221  
 confidence interval, 235–237  
 power of test, 226–228  
 sample-size determination, 232–234  
 Two-tailed critical values  
 Spearman rank-correlation coefficient, 834  
 Wilcoxon rank-sum test, 831–832  
 Wilcoxon signed-rank test, 830  
 Two-tailed test, 216  
 Two-way ANCOVA (analysis of covariance), 553–555  
 Two-way ANOVA (analysis of variance), 548–553. *See also* One-way ANOVA (analysis of variance)  
 column effect, 550  
 definition, 548  
 error term, 550  
 fixed effect, 573  
 general model, 549  
 hypothesis testing, 550–553  
 interaction between two variables, 549  
 interaction effect, 550  
 random effect, 573  
 row effect, 550  
 Type I error, 206–207  
 Type II error, 206–207
- Unbiased estimator, 176  
 Unblinded clinical trial, 159  
 Unimodal distribution, 12  
 Unmatched study design, 81
- VALID.DAT (Data Set), 36, 199–200, 262, 506, 508
- Variables  
 confounding, 607  
 continuous random, 72, 111  
 dependent, 126–129, 429  
 discrete random, 72, 76–77  
 disease, 589  
 dummy, 485, 541–542  
 exposure, 589  
 independent, 429  
 normal, 121  
 Poisson, 827  
 random, 72–73, 76, 109–110, 124–129
- Variance, 17–20  
 of binomial distribution, 88–89  
 of continuous random variable, 111  
 of discrete random variable, 76–77  
 estimation of, 176–181  
 of hypergeometric distribution, 369–370  
 of linear combinations of random variables, 125–126, 129  
 of Poisson distribution, 95–96  
 population, 77  
 properties of, 18–20
- Variance of distribution, 176–181  
 chi-square distribution, 177–179  
 interval estimation, 179–181  
 point estimation, 176–177
- Variance-covariance matrix, 544
- Variance-stabilizing transformation, 450

Washout period, 667  
 Weibull survival model  
   definition, 789  
   estimation of parameters, 790–793  
   estimation of percentiles, 793–794,  
     799–802  
   goodness of fit, 794  
   hazard functions of Weibull  
     distribution, 789  
   proportional hazards model, 797  
 Wilcoxon rank-sum test, 339–343  
   and chi-square test for trend,  
     397–400  
 $p$ -value, 342  
 ranking procedure, 340–341  
 two-tailed critical values for,  
   831–832

Wilcoxon signed-rank test, 333–339  
    $p$ -value, 337  
   ranking procedure, 335–336  
   two-tailed critical values for, 830  
 Within mean square, 520, 565  
 Within sum of squares, 519–520  
 Within-group variability, 518–519  
 Woburn study, 189–192  
 Wolfe's test, 466  
 Women's Health Initiative, 742  
 Woolf procedure, 600  
 $\chi^2$ . *See* Chi-square test  
 Yates-corrected chi-square test,  
   362–367  
   for 2 x 2 contingency tables, 362

acceptance and rejection  
   regions, 363  
 $p$ -value, 363  
 short computational form, 364–367  
 z test, one-sample, 218  
   acceptance and rejection  
     regions, 459  
   for correlation coefficient, 457–460  
    $p$ -value, 460  
 z test, two-sample, 464–465  
   acceptance and rejection regions,  
     464  
   for comparing two correlation  
     coefficients, 464  
    $p$ -value, 465  
 z transformation, 458



## INDEX OF APPLICATIONS (continued)

### DENTISTRY

clinical trial of two treatments for periodontal disease: Example 13.63, **680**; Example 13.64, **681**; Example 13.65, **682**; Example 13.66, **684**  
efficacy of a dental-education program in preventing the progression of periodontal disease: Problems 9.1–9.3, **346**  
estimation of the frequency of tooth loss among male health professionals: Problems 4.91–4.93, **105**  
longitudinal study of caries lesions on the exposed roots of teeth: Example 13.62, **676**

### DERMATOLOGY

comparison of two ointments in preventing redness on exposure to sunlight: Example 9.7, **329**; Example 9.8, **332**; Example 9.10, **333**; Example 9.11, **335**; Example 9.12, **337**; Example 9.13, **338**  
comparison of vidabrine vs. placebo in treating recurrent herpes labialis: Example 13.61, **674**

### DIABETES

association between ethnicity and diabetes: Review Question 3B.2, **51**  
association between ethnicity and HgbA1c among diabetes patients: Review Question 12B.3, **548**  
clinical trial among subjects with impaired glucose tolerance in the Diabetes Prevention Study: Problems 8.133–8.136, **320**  
effect of compliance with insulin on growth in boys with type I diabetes: Problems 5.86–5.88, **145**; Problems 8.152–8.154, **322**; Problems 9.47–9.50, **349**; Problems 11.88–11.91, **510**; Problems 11.97–11.99, **511**  
effect of the insulin pump on HgA1c levels among diabetics: Problems 8.163–8.166, **323**  
genetic profile of patients with type I diabetes: Problems 12.68–12.69, **585**  
incidence rates of blindness among insulin-dependent diabetics: Problems 4.63–4.67, **103**  
long-term trends in incidence of type II diabetes in Rochester, Minnesota: Problems 11.85–11.87, **510**  
plasma-glucose levels in sedentary people: Problems 7.7–7.9, **259**  
results from a weight loss trial among diabetics: Problems 5.114–5.116, **147**  
selection of patients for a treatment trial comparing an oral hypoglycemic agent with standard insulin therapy: Example 6.15, **153**  
side effects of insulin-pump therapy: Problems 10.26–10.27, **412**

### ENDOCRINOLOGY

age at onset of spermatozoa in urine samples of pre-adolescent boys: Problems 10.47–10.50, **413**  
change in bone density over 7 years after treatment with alendronate: Problems 11.75–11.80, **509**  
clinical trial for the prevention of fractures: Problems 7.99–7.101, **266**  
comparison of bone loss between alendronate- and placebo-treated patients: Problems 5.65–5.67, **142–143**  
effect of calcium and estrogen supplementation on bone loss: Problems 12.31–12.34, **580**  
effect of cod liver oil supplementation in childhood to bone density in middle age: Problems 12.74–12.77, **585–586**  
effect of low-fat diet on estrogen metabolism: Problems 9.61–9.64, **350–351**  
effect of low-fat diet on hormone levels in postmenopausal women: Problems 8.107–8.110, **317**  
effect of obesity on hormonal profile and impact on breast-cancer risk in women: Problems 11.92–11.96, **510–511**  
effect of raloxifene in preventing fractures among postmenopausal women: Problems 13.85–13.89, **719**  
effects of tobacco use on bone density in middle-aged women: Figure 2.11, **32**; Problems 2.38–2.46, **37**; Problems 4.79–4.83, **104**; Problems 6.86–6.87, **201**; Review Question 7B.1, **234**; Example 7.58, **256**; Problems 7.73–7.74, **263–264**; Problems 9.41–9.42, **349**; Problems 11.57–11.64, **508**; Review Question 13B.1, **612**  
hypothyroxinemia as a cause of subsequent motor and cognitive abnormalities in premature infants: Problems 11.48–11.52, **507–508**  
plasma hormones as risk factors for postmenopausal breast cancer: Example 13.72, **701**; Example 13.73, **704**; Problems 13.78–13.81, **718**  
relationship between calcium content of drinking water and the rate of fractures: Problems 13.15–13.18, **713**  
reproducibility of plasma hormones in split blood samples: Examples 12.27–12.30, **562–565**; Example 12.32, **568**; Example 12.33, **569**; Table 12.37, **580**; Problem 12.55, **582**

### ENVIRONMENTAL HEALTH

association between selenium level and cognitive function: Problems 10.127–10.130, **423**  
effect of exposure to anesthetic gases on cancer incidence: Problems 6.33–6.35, **198**  
effect of nuclear-power plants on birth defects: Problems 4.52–4.54, **102**  
effect of occupational exposure to 2,4,5-T herbicide on pulmonary function: Problems 7.82–7.84, **264**  
incidence of childhood leukemia in Woburn, Massachusetts: Examples 6.53–6.55, **189–190**; Examples 6.57–6.58, **191**  
measurement of exposure to low levels of radiation among shipyard workers: Example 4.5, **72**  
projected health effects of chronic exposure to low levels of lead in young children: Figures 2.9–2.10, **30**; Problems 2.31–2.32, **36**; Problems 6.67–6.69, **200**; Example 8.20, **293**; Examples 8.23–8.26, **297–299**; Problems 8.93–8.95, **316**; Table 9.5, **345**; Example 11.50, **485**; Problems 11.44–11.47, **507**; Tables 12.7–12.13, **538–548**; Problems 12.44–12.46, **581**; Problems 13.82–13.84, **718**  
rate of congenital malformations in offspring of Vietnam-veteran fathers: Problem 4.32, **101**  
relationship between daily particulate air pollution and mortality in Steubenville, Ohio: Problem 5.55, **142**  
relationship between emergency-room admissions and level of pollution: Problems 4.71–4.74, **104**  
relationship between pollution levels and heart-attack rates: Problems 5.89–5.91, **145**  
variation in temperature within a household: Problems 12.37–12.39, **581**; Problem 12.56, **582**

## INDEX OF APPLICATIONS (continued)

### EPIDEMIOLOGY

selection of random samples for serum testing from participants in the Nurses' Health Study: Example 6.10, **151**

### GASTROENTEROLOGY

comparison of two treatments for duodenal ulcer: Problems 10.8–10.12, **410**

random assignment of treatments for a clinical trial for duodenal ulcer: Problems 6.1–6.4, **196**

relationship between protein concentration of duodenal secretions and pancreatic function in cystic fibrosis:

Problems 12.42–12.43, **581**

### GENETICS

correlation between body weights of fathers and first-born sons: Example 11.29, **458**; Example 11.31, **459**; Example 11.32, **461**

dominant, recessive, and sex-linked mode of inheritance: Problems 3.30–3.47, **61–62**; Problems 4.94–4.96, **105–106**

genetic counseling for families with dominant disease: Problems 3.100–3.103, **67**

genetic effects on cholesterol levels: Table 8.21, **313**

genetic factors modulating the effect of cigarette smoking on renal-cell carcinoma: Problems 10.102–10.104, **421**

genetic marker for coronary heart disease: Review Question 14B.2, **738**

genetic markers for breast cancer: Example 14.4, **727**; Example 14.5–14.6, **728**; Example 14.7, **729**; Example 14.8, **730**

genetic profile of patients with type I diabetes: Problems 12.68–12.69, **585**

genetics of macular degeneration: Table 5.8, **147**

genetics of phenylketonuria: Problems 6.111–6.115, **202–203**

nested case-control study to assess SNPs associated with cardiovascular disease: Examples 12.16–12.17, **536–537**

patterns in sex-ratio data: Problem 4.57, **103**; Problem 6.59, **199**

prevalence of birth defects in a population: Problems 3.111–3.113, **67**

sequencing of ribosomal 5S RNA: Problems 7.19–7.20, **259**

### GYNECOLOGY

effect of contraceptive method on fertility: Example 8.6, **274**; Example 8.8, **275**

relationship between IUD use and infertility: Problems 13.1–13.7, **713**

sample of time intervals between successive menstrual periods: Table 2.3, **11**

test of a home pregnancy test-kit: Problems 3.79–3.82, **65**

use of basal body temperature to estimate the exact day of ovulation: Example 6.24, **165**; Example 6.33, **173**; Examples 6.35–6.36, **174**

### HEALTH PROMOTION

effect of fiber intake on weight gain in women: Review Question 10D.1, **390**

effect of quitting smoking on weight gain in middle-aged women: Problems 8.141–8.145, **320–321**

effect of walking on a treadmill on heart rate: Review Question 8A.2, **275**; Problems 8.167–8.170, **323–324**

gender differences in weight perception among adolescents: Problems 10.100–10.101, **420**

influence of retirement on level of physical activity among elderly women in the ARIC study: Problems 8.137–8.140, **320**

obesity among high school students: Review Question 7C.1, **250**

relationship between education and obesity: Problems 10.82–10.85, **419**

relationship between ethnicity and obesity among women: Problems 10.112–10.113, **422**

relationship between time to quitting smoking and total mortality: Problems 14.48–14.52, **806–807**

risk factors influencing success of smoking-cessation programs: Problems 4.55–4.56, **102**; Problems 9.28–9.32, **348**; Problems 11.65–11.67, **508**; Example 14.26, **758**; Example 14.27, **759**; Example 14.29, **761**; Example 14.34, **768**; Example 14.35, **771**; Example 14.37, **776**; Problems 14.7–14.11, **803**

smoking cessation as a preventive measure for heart disease: Example 6.52, **188**

smoking-cessation strategies for heavy-smoking teenagers: Problems 8.127–8.128, **319**

### HEALTH SERVICES ADMINISTRATION

comparison of length of stay in two different hospitals for patients with the same diagnosis: Problems 9.9–9.10, **346**

reproducibility of designation of medical malpractice: Problems 10.51–10.52, **415**

### HEMATOLOGY

hematologic data for patients with aplastic anemia: Problems 11.1–11.7, **504**

### HEPATIC DISEASE

effect of different hormones on pancreatic and biliary secretions in laying hens: Problems 8.82–8.85, **314–315**;

Problems 9.34–9.37, **348–349**; Problems 10.37–10.41, **413**; Problems 11.35–11.36, **506**; Problems 12.27–12.30, **580**; Problems 13.48–13.51, **716**

relationship of hepatoma to cirrhosis of the liver: Problems 5.50–5.52, **141**

### HOSPITAL EPIDEMIOLOGY

association between amount of sleep among medical house staff and medical errors: Problems 7.90–7.92, **265**

distribution of number of admissions to the emergency room: Problems 4.97–4.99, **106**

relationship between adverse events and mortality during hospital stay: Problems 10.59–10.60, **416**

## INDEX OF APPLICATIONS (continued)

### HYPERTENSION

active-control designs for testing new anti-hypertensive agents: Problems 13.41–13.43, **715**  
assessment of anti-hypertension drug treatment to reduce stroke risk among elderly people with isolated systolic hypertension (SHEP study): Example 6.17, **156**; Example 6.19, **159**  
association between glaucoma and hypertension: Problems 6.25–6.26, **197**  
association between left ventricular hypertrophy and hypertension: Review Question 13B.2, **612**  
cardiovascular-reactivity measures: Table 3.10, **64**; Problems 11.30–11.32, **506**  
comparison of an arteriosonde blood-pressure machine with the standard cuff: Example 6.39, **177**; Example 6.41, **181**; Example 7.45, **241**; Example 7.46, **243**  
comparison of blood pressure between Caucasian and African-American girls in the Bogalusa Heart Study: Review Questions 8B.2–8B.3, **293**  
comparison of the blood-pressure levels of vegetarians and non-vegetarians: Examples 12.20–12.22, **548–553**; Problems 12.58–12.59, **582**  
comparison of body-mass index between hypertensive and normotensive subjects: Problems 8.78–8.81, **314**  
comparison of plasma aldosterone levels between black and white children: Problems 5.47–5.49, **141**  
comparison of the random-zero machine and the standard cuff: Tables 9.11–9.12, **348**  
contribution of endothelin to blood-pressure regulation: Problems 11.53–11.56, **508**  
difference in prevalence of hypertension by ethnic group: Problems 5.120–5.122, **148**; Review Question 10E.3, **400**  
distribution of blood pressure among Samoans: Example 6.5, **150**  
distribution of diastolic blood pressure in 35 to 44 year old men: Example 5.2, **108**; Example 5.3, **109**; Example 5.6, **111**; Example 5.10, **113**; Example 5.20, **120**; Example 5.24, **124**; Examples 6.3–6.4, **149**  
distribution of diastolic blood pressure in the HDPP program: Examples 10.39–10.40, **401**; Example 10.41, **403**  
effect of dietary pattern on blood pressure: Problems 8.129–8.132, **319–320**  
effect of position on level of blood pressure: Table 2.14, **35**  
effect of post-menopausal hormones on blood pressure: Example 13.60, **673**; Problems 13.60–13.66, **716–717**  
effectiveness of hypertension-treatment programs: Problems 4.34–4.36, **101**; Example 6.12, **152**; Problems 6.40–6.46, **198**  
effectiveness of ingestion of linoleic acid on blood pressure: Table 9.10, **347**  
effectiveness of stress management in reducing blood pressure: Problems 8.102–8.106, **317**  
efficacy of treatment for hypertension based on home blood-pressure readings: Table 6.12, **201**; Table 6.13, **202**  
ethnic differences in rates of hypertension among children: Problems 13.90–13.97, **719–720**  
evaluation of an automated blood-pressure machine: Table 1.1, **3**; Example 3.26, **54**; Example 10.25, **378**; Problems 12.9–12.11, **578**  
familial blood-pressure relationships: Problems 5.17–5.20, **138–139**; Example 11.2, **427**; Example 11.35, **463**; Example 11.36, **464**  
hypertension screening in the home: Example 3.12, **42**  
inclusion of birthweight and body length in a multiple regression model: Example 11.46, **475**  
judging the effectiveness of anti-hypertensive medication: Problems 5.62–5.64, **142**  
non-pharmacologic therapies for hypertension: Problems 12.14–12.17, **579**  
norms for high blood pressure in children: Problems 5.53–5.54, **141–142**; Problems 11.18–11.23, **505**  
oscillometric devices for detecting high blood pressure in children: Problems 7.104–7.107, **266–267**  
prevalence of hypertension in the U.S. population: Example 3.35, **59**  
relationship between birthweight and infant blood pressure: Example 13.3, **590**  
relationship between blood lead and blood pressure: Problems 13.31–13.33, **714**  
relationship between obesity and hypertension: Example 11.47, **475**; Review Question 13A.2, **601**; Problems 13.90–13.92, **719**  
relationship between salt-taste and sugar-taste response to blood pressure in children: Problems 6.56–6.58, **199**; Problems 8.87–8.88, **315**; Example 11.37, **465**; Problems 11.42–11.43, **507**  
relationship between the use of oral contraceptives and level of blood pressure in women: Example 8.2, **269**; Table 8.1, **271**; Example 8.5, **272**; Example 8.7, **275**; Example 8.9, **276**; Example 8.10, **278**; Example 8.11, **281**; Example 8.16, **286**; Examples 8.28–8.32, **301–303**  
reliability of blood-pressure measurements: Example 12.31, **567**; Example 12.34, **570**; Table 12.34, **579**  
risk factors for newborn blood-pressure measurements: Example 11.38, **468**; Example 11.39, **469**; Examples 11.40–11.41, **470–471**; Examples 11.42–11.44, **471–472**; Example 11.45, **475**; Example 11.48, **476**; Example 11.49, **481**; Example 11.51, **492**; Example 11.52, **493**  
screening for high-blood pressure in children: Problems 5.53–5.54, **151**  
screening procedures for detecting hypertension: Example 7.44, **239**; Problems 7.97–7.98, **266**  
side effects of anti-hypertensive agents: Problems 7.85–7.87, **265**  
testing of new anti-hypertensive agents: Example 4.4, **72**, Examples 4.6–4.7, **73**; Table 4.2, **74**; Example 4.8, **74**; Example 4.11, **76**; Example 7.23, **219**; Problems 7.26–7.27, **260**; Example 8.33, **304**; Example 8.34, **306**  
use of both weight and BMI in the same regression model: Example 11.47, **475**

### INFECTIOUS DISEASE

allergic-reaction rates: Problems 14.22–14.24, **804**  
clustering of gonorrhea cases in central cities: Problem 4.25, **100**  
comparability of infectious-disease diagnoses as reported by the attending physician and by chart review: Problem 10.17, **410**  
comparison of efficacy of 2 antibiotics: Problems 3.104–3.106, **67**  
comparison of vidarabine vs. placebo in treatment of recurrent herpes labialis: Example 13.61, **674**  
differences in effectiveness and toxicity of aminoglycoside antibiotics: Example 6.18, **158**; Problems 13.34–13.36, **715**

## INDEX OF APPLICATIONS (continued)

- distribution of annual number of polio deaths: Example 4.37, **95**; Table 4.9, **95**; Example 10.3, **352**  
distribution of annual number of typhoid-fever deaths: Example 4.31, **90**; Example 4.33, **92**; Figure 4.5, **93**; Example 4.36, **94**  
distribution of MIC of penicillin G for *N. gonorrhoeae*: Table 2.4, **12**  
distribution of number of eosinophils in 100 white-blood cells: Problem 5.25, **139**  
distribution of number of lymphocytes in 100 white-blood cells: Example 4.26, **85**; Example 5.9, **112**; Problems 5.26–5.28, **139**  
distribution of number of neutrophils in 100 white-blood cells: Example 4.15, **79**; Examples 4.23–4.24, **83**; Example 4.28, **86**; Example 5.1, **108**; Examples 5.34–5.35, **132–133**; Problems 5.29–5.30, **139**; Example 6.1, **149**  
distribution of time to onset of AIDS following seroconversion among hemophiliacs: Problems 6.64–6.66, **199**  
effect of aspirin in preventing ototoxicity among patients receiving gentamicin: Problems 10.117–10.120, **422**  
effectiveness of different smallpox vaccines: Problems 10.89–10.92, **419–420**  
incidence of H1N1 influenza in Australia and New Zealand: Problems 14.72–14.75, **811**  
incidence rate of influenza among high-school students: Problems 14.32–14.36, **805**  
prevalence of HIV-positive people in a low-income census tract: Example 6.6, **150**  
red wine intake to prevent the common cold: Review Question 13E.2, **649**  
relationship between the use of oral contraceptives and bacteruria: Table 13.8, **610**; Example 13.21, **610**; Example 13.22, **611**; Problems 13.9–13.14, **713**  
reproducibility of assessment of generalized lymphadenopathy among people at high risk for AIDS: Table 8.18, **312**; Problem 9.43, **349**  
retrospective chart review of patients in a Pennsylvania hospital: Table 2.11, **33**; Examples 8.18–8.19, **291–292**; Problems 8.7–8.11, **309**; Problems 9.11–9.12, **346**; Problems 10.6–10.7, **410**; Review Question 11B.2, **455**; Review Questions 11C.2–3, **467–468**; Review Question 11D.3, **484**; Review Question 12C.4, **555**  
risk factors for *Chlamydia trachomatis*: Example 13.34, **628**; Example 13.36, **631**; Example 13.38, **636**; Example 13.39, **637**  
risk factors for HIV infection among intravenous drug users: Problems 4.58–4.62, **103**; Problems 10.53–10.54, **415**  
sample of admission white-blood counts in a Pennsylvania hospital: Table 2.2, **10**  
screening of newborns for HIV virus in five Massachusetts hospitals: Table 4.14, **100**  
side effects of a flu vaccine: Example 4.30, **87**  
side effects of a polio-immunization campaign in Finland: Example 4.40, **98**; Review Question 10F.2, **404**  
surveillance methods for detecting infection following caesarean-section birth: Problems 10.93–10.97, **420**  
validation study of accuracy of assessment of hospital-acquired infection among coronary-bypass patients: Problems 3.107–3.110, **67**

### MENTAL HEALTH

- Alzheimer's-disease prevalence: Table 3.5, **61**; Problems 3.16–3.27, **61**  
association between selenium levels and cognitive function: Problems 10.127–10.130, **423**  
comparison of physician and spouse reports for diagnosing schizophrenia: Problems 10.108–10.111, **421–422**  
effect of widowhood on mortality: Problems 10.33–10.36, **412**; Problems 13.29–13.30, **714**; Problem 13.108, **720**  
evaluation of a Mental Function Index to identify people with early signs of senile dementia: Problems 12.12–12.13, **578**  
matched pair study for schizophrenia: Examples 4.16–4.17, **79–80**  
use of APOE gene to diagnose Alzheimer's disease: Table 3.15, **66**  
use of Chinese Mini-Mental Status Test to identify people with dementia in China: Table 3.12, **65**  
use of vitamin E supplementation to prevent Alzheimer's disease: Review Question 9B.2, **344**

### MICROBIOLOGY

- pod weight of plants inoculated with nitrogen-fixing bacteria vs. uninoculated plants: Table 2.18, **37**; Problems 8.116–8.120, **318**; Problems 9.44–9.46, **349**  
quality control for susceptibility testing: Table 6.10, **197**

### NEUROLOGY

- changes in symptoms in clinical trial of Parkinson's disease patients: Problems 8.175–8.177, **324**  
risk of cancer among cystic-fibrosis patients: Problems 5.59–5.61, **142**

### NUTRITION

- association between high salt intake and cause of death: Example 10.16–10.17, **367**; Example 10.19, **370**; Example 10.20, **372**  
calcium intake in low-income populations: Problems 8.2–8.6, **309**; Problems 8.12–8.18, **309**  
comparison of blood-pressure levels between vegetarians and non-vegetarians: Examples 12.20–12.22, **548–553**  
comparison of dietary vitamin C intake between smokers and non-smokers: Problems 8.182–8.186, **325**  
comparison of protein intake among vegetarians and non-vegetarians: Problems 12.1–12.5, **577**  
distribution of total carbohydrate intake in children: Problems 5.6–5.9, **138**  
effect of cod liver oil supplementation in childhood on bone density in middle age: Problems 12.74–12.77, **585–586**  
effect of oat bran intake on cholesterol levels: Problems 7.54–7.58, **262**  
effectiveness of dietary counseling in achieving sodium restriction: Table 8.20, **313**; Problem 9.33, **348**  
iron-deficiency anemia in low-income populations: Problems 7.33–7.40, **260–261**  
measuring compliance with a sodium-restricted diet: Problem 9.33, **348**  
prevalence of bladder cancer in rats fed a high-saccharin diet: Example 6.50, **186**; Example 6.51, **187**  
protective effect of vitamins A and E vs. cancer: Example 2.1, **5**; Example 10.26–10.27, **381**  
recall of pre-school diet of their children by 70–79 year-old women: Problems 11.118–11.121, **574–575**

## INDEX OF APPLICATIONS (continued)

relationship between breast-cancer incidence and dietary-fat intake: Example 13.70, **697**; Example 13.71, **701**; Problems 13.73–13.77, **717–718**  
relationship between dietary and plasma vitamin C in the EPIC-Norfolk Study: Problems 11.109–11.112, **512**  
relationship of dietary intake assessed by food-frequency questionnaire vs. the diet record: Table 2.16, **36**; Problems 5.56–5.58, **142**; Problems 7.59–7.60, **262**; Problem 11.33, **506**; Problems 11.68–11.72, **509–510**  
reproducibility of a food-frequency questionnaire: Example 10.9, **359**; Example 10.15, **365**; Example 10.42–10.44, **404–406**  
serum cholesterol levels in vegetarians: Problems 7.26–7.28, **260**  
validation of a dietary questionnaire administered over the Internet: Example 11.33, **462**; Example 11.34, **463**

### OBSTETRICS

accuracy of daughter's report of maternal smoking during pregnancy: Table 3.19, **69**  
Apgar score: Examples 11.53–11.55, **494–496**; Example 11.57, **501**  
association between socioeconomic status and birth defects: Problems 4.100–4.103, **106**  
cigarette smoking and low-birthweight deliveries: Table 5.5, **146**  
distribution of birthweights in the general population: Example 5.5, **110**; Example 6.8, **150**  
drug therapy for preventing low-birthweight deliveries: Problem 7.22, **260**; Problems 8.37–8.40, **311**; Problems 9.7–9.8, **346**  
estriol levels in pregnant women as an indicator of a low-birthweight fetus: Example 11.1, **427**; Example 11.3, **428**; Examples 11.4–11.5, **429**; Example 11.8–11.11, **433–434**; Example 11.12, **439**; Example 11.15, **443**; Example 11.16, **444**; Example 11.24, **453**; Example 11.26, **455**  
infant-mortality rates in the U.S., 1960–2005: Problems 11.13–11.17, **505**  
probability of a male live childbirth: Example 3.2, **39**  
rate of congenital malformations in offspring of Vietnam-veteran fathers: Problem 4.32, **101**  
relationship between birthweight and gestational age: Problems 3.48–3.50, **62**  
relationship between low socioeconomic status and low birthweight: Examples 7.2–7.3, **204–205**; Example 7.4, **206**; Example 7.6, **206**; Examples 7.10–7.11, **209–210**; Example 7.12, **211**; Example 7.14, **212**; Example 7.16–7.17, **213**; Example 7.19, **215**; Example 7.29, **225**; Example 7.31, **226**; Example 7.33–7.34, **230**; Example 7.36, **231**; Example 8.1, **269**  
sample of birthweights from 100 consecutive deliveries: Table 2.7, **22**  
sample of birthweights from 1000 consecutive deliveries: Table 6.2, **155**; Example 6.16, **154**; Example 6.22, **163**; Example 6.23, **164**; Problems 6.52–6.55, **198**  
sample of birthweights from a San Diego hospital: Table 2.1, **8**  
screening tests for Down's syndrome: Problems 10.121–10.123, **422–423**  
surveillance methods for detecting infection following caesarean-section birth: Problems 10.93–10.97, **420**  
variability of an assay for *M. hominis* mycoplasma: Problems 6.36–6.39, **198**

### OCCUPATIONAL HEALTH

excess cancer deaths in nuclear-power-plant workers: Example 7.49, **248**  
incidence of bladder cancer and Hodgkin's disease among rubber workers: Examples 7.52–7.53, **251–252**; Example 7.54–7.57, **252–255**  
incidence of bladder cancer among workers in the tire industry: Problems 4.23, **100**; Problem 6.75, **200**  
incidence of stomach cancer among workers in the tire industry: Problem 4.24, **100**; Problem 6.76, **200**  
mortality experience of workers exposed to waste disposal during the Manhattan Project: Problems 7.43–7.46, **261**  
mortality experience of workers in a chemical plant: Problems 3.28–3.29, **61**  
mortality experience of workers with exposure to EDB: Example 4.38, **95**; Example 6.59, **192**  
proportion of lung-cancer deaths in chemical-plant workers: Problems 7.28–7.32, **260**

### OPHTHALMOLOGY

association between body mass index and AMD: Problems 13.111–13.112, **721**  
association between cigarette smoking and glaucoma: Review Question 14C.1, **749**; Review Question 14D.12, **754–755**  
association between ethnic origin and genetic type in retinitis pigmentosa: Problem 10.15, **410**  
association between glaucoma and hypertension: Problems 6.25–6.26, **197**  
change in electroretinogram (ERG) amplitude following surgery for patients with retinitis pigmentosa: Review Question 9A.5, **339**  
change in serum retinol and serum triglycerides after taking high doses of vitamin A among retinitis-pigmentosa patients: Problems 10.64–10.69, **416–417**  
change in visual field in retinitis pigmentosa patients: Problems 11.100–11.103, **511–512**; Problems 14.68–14.71, **810–811**  
comparison of eye drops in preventing redness and itching in people with hay fever: Example 9.9, **333**  
comparison of lens photographs of cataractous and normal eyes: Table 8.16, **310**  
comparison of mean ERG amplitude among patients with different genetic types of retinitis pigmentosa: Problems 8.97–8.101, **316**  
comparison of Sorbinil vs. placebo for the prevention of diabetic retinopathy: Problems 10.143–10.146, **424–425**  
comparison of the ocular anti-inflammatory properties of four different drugs in albino rabbits: Example 12.23, **556**; Example 12.24, **557**; Example 12.26, **560**  
comparison of visual acuity in people with the dominant and sex-linked forms of retinitis pigmentosa: Example 9.15, **339**; Example 9.16, **341**; Example 9.17, **342**; Example 10.38, **399**  
compliance with lutein supplement tablets among macular-degeneration patients: Problems 5.92–5.95, **145–146**; Problems 11.81–11.84, **509–510**

# Index of Applications

## ACCIDENT EPIDEMIOLOGY

cumulative incidence of automobile accidents among medical interns: Problems 5.99–5.102, **146**

## AGING

risk factors predicting survival among subjects in the EPESE study: Example 13.74, **706**; Example 13.75, **707**; Example 13.76, **710**  
use of vitamin E supplementation to prevent Alzheimer's disease: Review Question 9B.2, **344**

## BACTERIOLOGY

distribution of the number of bacterial colonies on a petri or agar plate: Example 4.32, **91**; Example 4.34, **92**; Example 5.36, **136**; Example 6.2, **149**; Example 6.60, **193**

## BIOAVAILABILITY

comparison of bioavailability of four different beta-carotene preparations: Problems 12.20–12.26, **579–580**; Problems 12.52–12.54, **582**; Problems 14.12–14.15, **803**

## BLOOD CHEMISTRY

monitoring clinical-chemistry measurements in pharmacologic research: Problems 5.31–5.35, **139**

## BOTANY

distribution of tree diameters: Example 5.21, **122**

## CANCER

abortion as a risk factor for breast cancer: Problems 10.61–10.63, **416**

age at first birth as a risk factor for breast cancer: Example 3.1, **38**; Example 10.4, **353**; Example 10.5, **355**; Example 10.7, **357**; Example 10.10, **360**; Example 10.13, **363**; Example 10.33, **390**; Example 10.35, **393**; Example 10.36, **394**; Example 10.37, **396**; Example 13.2, **590**; Example 13.10, **598**; Example 13.11, **600**; Example 13.33, **673**

age at menarche as a risk factor for breast cancer: Problems 7.66–7.68, **263**

age at menarche as a risk factor for ovarian cancer: Problems 13.120–13.123, **721–722**

age at surgery for undescended testis as a risk factor for testicular cancer: Problems 7.112–7.116, **267**

arsenic exposure as a risk for non-melanoma skin cancer: Review Question 13C.2, **624**; Review Question 13D.1, **628**

association between age at menarche and ovarian cancer: Review Question 10A.2, **367**

association between aspirin intake and colon cancer: Problems 14.43–14.47, **806**; Problems 14.61–14.62, **807–808**

association between oral-contraceptive use and breast cancer: Example 13.32, **625**; Example 13.33, **626**; Example 14.1, **725**; Example 14.2, **726**; Example 14.9, **731**; Example 14.10, **732**; Example 14.11, **734**; Example 14.13, **737**; Problems 14.1–14.6, **802–803**

association between oral-contraceptive use and endometrial cancer: Problem 10.14, **410**

association between oral-contraceptive use and ovarian cancer: Example 10.1–10.2, **352**; Problems 13.109–13.110, **721**

association between parity and ovarian-cancer incidence: Review Question 7A.3, **215**

association between postmenopausal hormone use and breast cancer: Problems 10.105–10.107, **421**; Example 14.17, **742**; Example 14.18, **745**; Example 14.19, **747**; Example 14.20, **749**; Example 14.21, **750**; Example 14.22, **752**; Example 14.23, **754**; Problems 14.53–14.56, **807**; Problems 14.63–14.67, **808–809**

cardiovascular disease mortality among women with breast cancer: Problems 7.102–7.103, **266**

cigarette smoking as a risk factor for lung cancer: Example 3.24, **52**; Problems 11.24–11.29, **505–506**; Review Question 13A.3, **601**; Examples 13.12–13.13, **602–603**; Examples 13.14–13.15, **604–605**; Review Question 14H.4, **783**

clinical trial of efficacy of maintenance chemotherapy for leukemia patients: Example 14.45, **787**; Example 14.46, **793**; Example 14.47, **794**; Example 14.48, **797**; Example 14.49, **799**

cluster of leukemia cases in Woburn, Massachusetts: Example 4.2, **71**

## INDEX OF APPLICATIONS (continued)

comparison of exemestane versus tamoxifen for treatment of women with breast cancer: Review Question 14F.3, **767**; Review Question 14G.2, **774**  
comparison of 5-year survival for breast cancer between two different chemotherapy regimens: Example 10.21, **373**; Example 10.24, **376**; Example 10.29, **384**; Example 10.30, **385**; Example 13.29, **621**  
comparison of a new treatment and a standard treatment for cancer – the case for a one-sided confidence interval: Example 6.61, **193**; Example 6.63, **194**  
comparison of incidence rates of breast cancer between Chinese and American women: Problems 5.110–5.113, **147**  
comparison of risk factors for different types of breast cancer according to ER/PR status: Example 13.46, **654**  
comparison of stage of breast cancer by ethnic group: Problems 10.22–10.23, **411**  
comparison of two active treatments for early-stage breast cancer: Example 13.53, **663**; Example 13.54, **664**; Example 13.55, **665**  
cumulative incidence of breast cancer: Example 14.3, **727**  
effect of beta-carotene on cancer incidence: Problems 10.74–10.76, **418**  
effect of ethnicity on serum estradiol and body-mass index in premenopausal women: Table 9.14, **350**  
effect of obesity on hormonal profile and impact on breast-cancer risk in women: Problems 11.92–11.96, **510–511**  
effect of passive smoking on cancer risk: Example 13.23, **612**; Example 13.24, **616**; Examples 13.25–13.26, **617**; Example 13.27, **619**  
effect of PUVA treatment for psoriasis on malignant melanoma: Problems 7.63–7.65, **262–263**  
excess cancer deaths in nuclear-power-plant workers: Example 7.49, **248**  
family history of breast cancer as a risk factor for breast cancer: Example 3.25, **52**; Example 4.39, **96**; Example 6.48, **184**; Example 6.49, **186**; Example 7.47, **244**; Example 7.48, **247**; Example 7.50, **249**; Example 7.51, **250**; Problem 7.21, **260**; Example 14.12, **736**  
genetic factors modulating the effect of cigarette smoking on renal-cell carcinoma: Problems 10.102–10.104, **421**  
genetic markers for breast cancer: Example 14.4, **727**; Examples 14.5–14.6, **728**; Example 14.7, **729**; Example 14.8, **730**  
incidence of breast cancer: Example 3.4, **39**; Example 3.36, **59**  
incidence of childhood leukemia in Woburn, Massachusetts: Examples 6.53–6.55, **189–190**; Examples 6.57–6.58, **191**  
incidence of colon cancer based on SEER rates: Problems 7.117–7.119, **267–268**  
incidence of ovarian cancer: Review Question 14A.2–3, **730**  
incidence of stomach cancer: Example 3.5, **40**  
influence of social class on age at menarche and impact on breast-cancer risk: Problems 8.149–8.151, **322**  
lifetime risk of breast cancer: Review Question 4C.1, **89**  
lung-cancer incidence for men vs. women: Review Question 10B.3, **373**  
mammography as a screening test for breast cancer: Example 3.18, **47**; Example 3.21, **49**; Problems 6.77–6.80, **200**  
neuroblastoma screening in young children: Problems 7.79–7.81, **264**  
plasma hormones as risk factors for postmenopausal breast cancer: Example 13.44, **650**; Example 13.45, **651**; Example 13.72, **701**; Example 13.73, **704**; Problems 13.78–13.81, **718**; Problems 13.117–13.119, **721**  
prevalence of bladder cancer in rats fed a high-saccharin diet: Example 6.50, **186**; Example 6.51, **187**  
prevalence of malignant melanoma among 45- to 54-year-old women in the U.S.: Example 6.42, **181**; Example 6.43, **182**  
prostate-specific antigen (PSA) test as a screening test for prostate cancer: Review Question 3C.3, **52**  
protective effect of vitamin A or vitamin E vs. cancer: Example 2.1, **5**; Examples 10.26–10.27, **381**  
randomized clinical trial testing the effect of estrogen-replacement therapy vs. placebo on breast-cancer incidence: Example 14.14, **738**; Example 14.15, **740**; Example 14.16, **742**; Example 14.44, **787**  
relationship between breast-cancer incidence and dietary-fat intake: Example 13.70, **697**; Example 13.71, **700**; Problems 13.73–13.77, **717–718**  
relationship between breast-cancer incidence and menopausal status: Problems 14.25–14.26, **805**  
relationship between breast-cancer incidence and parity: Example 14.24, **755**; Example 14.25, **757**  
relationship between lung-cancer incidence and heavy drinking: Example 13.16, **607**; Example 13.17, **608**; Example 13.18, **609**  
relationship between plasma vitamin-A concentration and stomach-cancer risk: Problems 7.51–7.53, **261–262**  
relationship of dietary factors to colon cancer: Example 13.9, **598**  
risk of cancer among patients with cystic fibrosis: Problems 5.59–5.61, **142**  
screening techniques for esophageal cancer: Problems 6.109–6.110, **202**  
serum estradiol as a screening test for breast cancer: Table 3.17, **68**  
survival of women with breast cancer undergoing radical mastectomy: Example 6.11, **152**  
two-stage model of breast-cancer carcinogenesis: Problems 4.87–4.90, **105**  
vitamin E as a preventive agent for cancer: Table 5.2, **139**

## CARDIOLOGY

association between coronary flow reserve and myocardial velocity ratio in hypertensive patients: Problems 11.104–11.108, **512**  
comparison of angioplasty (PTCA) with medical therapy for treating single-vessel coronary disease: Problems 13.37–13.38, **715**  
comparison of duration of exercise for coronary-artery disease patients randomized to medical therapy or PTCA: Table 8.24, **314**; Problems 13.45–13.46, **736**  
effect of calcium-channel blockers on heart rate and blood pressure for patients with unstable angina: Problems 6.70–6.74, **200**; Example 7.32, **228**; Example 7.37, **233**; Examples 7.38–7.39, **234**; Problem 7.62, **262**  
glucose level as a risk factor for carotid-artery stenosis: Problems 8.111–8.112, **317**  
reduction of infarct size in patients with myocardial infarction: Example 7.13, **211**; Example 7.15, **212**; Example 7.30, **225**

## CARDIOVASCULAR DISEASE

association between ankle-arm index and S-T segment depression: Table 3.18, **69**  
association between childhood SES and subclinical markers of atherosclerosis: Problems 8.146–8.148, **321–322**

## INDEX OF APPLICATIONS (continued)

association between cholesterol levels in spouse pairs: Example 11.22, **452**; Example 11.26, **455**; Example 11.27, **455**; Example 11.28, **456**  
association between high salt intake and cause of death: Examples 10.16–10.17, **367**; Example 10.19, **370**; Example 10.20, **372**  
association between obesity and coronary disease: Example 13.20, **609**  
baldness pattern as a risk factor for MI: Problems 13.98–13.101, **720**  
change in hematocrit in a patient with intermittent claudication: Review Question 11A.2, **447**  
changes in the incidence and case-fatality rate of myocardial infarction 1990–2000: Problems 7.10–7.14, **259**  
cholesterol levels before and after adopting a vegetarian diet: Table 2.13, **34**  
cigarette smoking as a risk factor for MI in women: Problems 10.18–10.21, **410–411**  
clinical trial of lipid-lowering agents and antioxidants to prevent progression of atherosclerosis among patients with clinical coronary disease: Problems 12.60–12.62, **583**  
comparison between antithrombotic drug regimens after coronary stenting to prevent stent thrombosis: Problems 10.131–10.134, **423–424**  
comparison between treatment groups in a study of in-hospital mortality among CABG patients: Problems 8.172–8.175, **324**  
comparison of aspirin vs. placebo in the Physicians' Health Study: Example 6.21, **160**; Example 10.31, **386**; Example 10.32, **388**; Problems 10.1–10.5, **409–410**; Example 12.14, **534**; Example 13.4, **590**; Example 14.40, **781**  
comparison of aspirin vs. placebo in the Women's Health Study: Problems 13.113–13.116, **721**  
comparison of drugs for easing pain in unstable angina: Problems 5.45–5.46, **140**  
comparison of HDL cholesterol levels between Caucasian and African-American adults in the ARIC study: Problems 8.159–8.162, **323**  
deaths due to heart failure: Review Question 5C.4, **137**  
different methods of measuring cholesterol: Figure 2.4, **15**  
distribution of duration of cigarette smoking: Problems 5.12–5.13, **138**  
distribution of serum cholesterol: Review Question 5B.3, **124**; Problems 5.1–5.5, **138**; Problems 5.14–5.16, **138**  
distribution of serum triglycerides: Example 5.4, **109**; Example 5.8, **111**; Example 6.26, **166**  
effect of calcium-channel blockers on blood pressure and heart rate: Problem 7.72, **290**  
effect of oat bran on serum cholesterol: Problems 7.54–7.58, **262**  
effect of obesity on hypertension: Problems 5.68–5.71, **143**  
excess cardiac deaths attributable to an earthquake: Problems 4.68–4.70, **104**  
genetic markers for coronary heart disease: Review Question 14B.2, **738**  
Hispanic paradox: Problems 3.97–3.99, **66**  
hyperinsulinemia as an independent risk factor for ischemic heart disease: Problems 8.121–8.124, **318–319**  
microenzymatic vs. autoanalyzer method of cholesterol measurement: Figure 2.4, **15**  
nested case-control study to assess SNPs associated with cardiovascular disease: Examples 12.16–12.17, **536–537**  
predicting the incidence of coronary heart disease as a function of several risk factors: the Framingham Heart Study: Example 13.40, **638**; Example 13.41, **643**; Example 13.42, **645**; Example 13.43, **648**  
prevalence and incidence of different types of cardiovascular morbidity in Minnesota: Example 6.9, **151**  
racial trends in heart rate among children: Problems 8.23–8.24, **309–310**  
relationship between LDL cholesterol and obesity in children: Problems 11.37–11.41, **506**  
relationship between the use of oral contraceptives and heart disease in women: Example 10.6, **356**; Example 10.8, **358**; Example 10.11, **361**; Example 10.14, **364**; Example 13.1, **589**; Example 13.5, **593**; Examples 13.7–13.8, **595–596**; Review Question 13A.1, **601**; Example 13.15, **647**; Problems 13.104–13.107, **720**  
reproducibility of activated Protein C (APC), a serum marker of thrombosis: Table 2.17, **36**; Review Question 12E.2, **567–568**; Review Question 12F.2, **571**  
reproducibility of cardiovascular risk factors in children: Table 2.6, **21**  
risk factors for sudden death in the Framingham, Massachusetts population: Problems 13.44–13.47, **716**  
secondary prevention trial of lipid lowering in patients with previous MI: Problems 10.42–10.46, **413**  
serum-cholesterol levels in children of men with heart disease: Example 6.37, **175**; Example 7.1, **204**; Example 7.5, **206**; Example 7.9, **208**; Example 7.18, **214**; Examples 7.27–7.28, **224–225**; Example 7.35, **230**; Examples 7.40–7.41, **236**; Example 7.42, **237**; Example 7.43, **239**; Example 8.12, **282**; Example 8.15, **285**; Example 8.17, **289**  
serum-cholesterol levels in recent Asian immigrants to the U.S.: Example 7.20, **216**; Example 7.21, **217**; Example 7.22, **218**; Example 7.24, **220**  
serum-cholesterol levels in vegetarians: Problems 7.23–7.25, **260**  
testing for genes associated with cardiovascular disease: Review Question 14B.2, **758**  
trends in coronary heart disease mortality in Olmstead County, Minnesota, over 20 years: Problems 3.121–3.123, **68**  
validation study of accuracy of assessment of hospital-acquired infection among coronary-bypass patients: Problems 3.107–3.110, **67**  
variability of cholesterol measurements in children: Problems 6.84–6.85, **200–201**

## CEREBROVASCULAR DISEASE

clinical trial of Warfarin to prevent stroke in patients with atrial fibrillation: Example 6.20, **159**  
distribution of cerebral blood flow (CBF) in normals: Example 5.22, **122**

## DEMOGRAPHY

mortality among Americans of Chinese descent: Problems 7.41–7.42, **261**  
relationship of fertility rates to survival outcomes of previous births in Norway: Table 3.13, **66**  
sex-ratio data in humans: Problem 7.61, **262**; Problems 10.77–10.79, **418**

## INDEX OF APPLICATIONS (continued)

contralateral design for assessing the effect of an eye drop in lowering intraocular pressure in glaucoma patients: Review Question 10D.3, **390**  
contralateral design for evaluation of effectiveness of an eye drop in preventing itching: Problems 7.69–7.72, **263**; Problems 9.38–9.40, **349**  
distribution of astigmatism in 1033 Army recruits: Table 2.12, **34**  
distribution of intraocular pressure: Example 5.23, **123**; Problems 7.110–7.111, **267**  
effect of an eye drop in increasing tear break-up time among patients with dry eye: Problems 7.88–7.89, **265**; Problems 9.55–9.60, **350**; Example 12.35, **572**; Example 12.36, **574**; Problems 12.63–12.67, **584–585**  
effect of different fluoroquinolones on corneal sensitivity: Problems 12.82–12.84, **586–587**  
effect of diflunisal on intraocular pressure: Problems 8.19–8.22, **309**  
effect of dose of Botox on eye pain: Problems 8.178–8.181, **324–325**  
effect of medication regimen on intraocular pressure among glaucoma patients: Table 7.9, **266**  
effect of sunlight on mice with retinitis pigmentosa: Problem 12.57, **582**  
effectiveness of an eye drop in preventing dry eye symptoms: Problems 10.124–10.126, **423**  
effectiveness of a topical antiallergic eye drop in preventing the signs and symptoms of allergic conjunctivitis: Problems 8.113–8.115, **317–318**  
genetic forms of retinitis pigmentosa: Example 4.1, **71**; Problems 4.94–4.96, **105–106**  
genetics of macular degeneration: Table 5.8, **147**; Problems 10.135–10.138, **424**  
incidence of cataract among people 65+ years of age: Example 3.22, **50**  
incidence rates of blindness among insulin-dependent diabetics: Problems 4.63–4.67, **103**  
incidence rates of cataract among people with excessive exposure to sunlight: Problems 7.47–7.50, **261**  
prevalence of glaucoma among the elderly as determined by Eyemobile screening: Review Question 5C.3, **137**  
prevalence of low vision among the elderly: Example 6.7, **150**  
rate of field loss in retinitis-pigmentosa patients: Problems 5.83–5.85, **144–145**  
reproducibility of tear break up time among dry eye patients: Problems 12.70–12.73, **585**  
testing of a drug on ocular-hypertensive patients to prevent glaucoma: Example 7.25, **221**  
treatment trial of vitamin supplements for retinitis pigmentosa: Review Question 12D.3, **561**; Example 13.67, **687**; Example 13.68, **688**; Example 13.69, **691**; Example 14.30, **762**; Example 14.31, **764**; Example 14.32, **765**; Example 14.33, **766**; Example 14.36, **772**; Example 14.38, **776**; Example 14.39, **778**; Example 14.41, **783**; Example 14.42, **785**; Example 14.43, **786**; Problems 14.16–14.21, **803–804**  
twin study design for macular degeneration: Review Question 10C.3, **380**  
visual acuity as an ordinal variable: Example 9.4, **328**

### ORNITHOLOGY

comparison of wing length of different species of stonechats: Problems 12.78–12.81, **586**

### ORTHOPEDICS

accuracy of the FAIR test in diagnosing piriformis syndrome: Problems 5.78–5.82, **144**  
comparison of the FAIR test with patients' self-report for assessment of severity of piriformis syndrome: Table 3.16, **68**  
treatment trial for piriformis syndrome: Problems 14.37–14.42, **806**

### OTOLARYNGOLOGY

comparison of antibiotics for treating acute otitis media: Problems 3.75–3.78, **64**; Problems 10.56–10.58, **416**; Problems 10.86–10.88, **419**; Problems 13.52–13.53, **716**; Problems 13.70–13.72, **717**  
comparison of medical and surgical treatment for children with chronic otitis media: Example 10.28, **383**  
comparison of the ototoxicity of different aminoglycosides: Problem 13.35, **715**  
duration of middle-ear effusion in breast-fed and bottle-fed babies: Table 9.9, **347**  
number of episodes of otitis media in the first 2 years of life: Example 4.3, **72**; Table 4.3, **75**; Example 4.10, **75**; Examples 4.12–4.13, **76**; Example 4.14, **78**; Problems 4.26–4.31, **100–101**  
sibling history of ear infection as a risk factor for otitis media in the first year of life: Problems 13.67–13.69, **717**

### PEDIATRICS

age at onset of night-time bladder control: Problems 10.80–10.81, **418**  
age at onset of spermatozoa in urine samples of pre-adolescent boys: Problems 10.47–10.50, **413**  
age trends in pulse rates in children: Example 11.6, **430**  
association between climate conditions in infancy and adult height: Problems 8.155–8.158, **322–323**  
duration of middle ear effusion in breast and bottle fed babies: Table 9.9, **347**  
effect of in-vitro fertilization on birth defects: Problems 5.96–5.98, **146**  
hypothyroxinemia as a cause of subsequent motor and cognitive abnormalities in premature infants: Problems 11.48–11.52, **507–508**  
inclusion of birthweight and body length in a multiple regression model: Example 11.46, **475**  
number of episodes of otitis media in the first 2 years of life: Example 4.3, **72**; Table 4.3, **75**; Example 4.10, **75**; Examples 4.12–4.13, **76**; Problems 4.42–4.45, **101–102**  
projected health effects of chronic exposure to low levels of lead in young children: Figures 2.9–2.10, **30**; Problems 2.31–2.32, **36**; Problems 6.67–6.69, **200**; Example 8.20, **293**; Examples 8.23–8.26, **297–299**; Problems 8.93–8.95, **316**; Table 9.5, **345**; Example 11.50, **485**; Problems 11.44–11.47, **507**; Tables 12.7–12.13, **538–548**; Problems 12.44–12.46, **581**; Problems 13.82–13.84, **718**

## INDEX OF APPLICATIONS (continued)

- racial trends in heart rate among children: Problems 8.23–8.24, **309–310**  
relationship between LDL cholesterol and obesity in children: Problems 11.37–11.41, **506**  
relationship between salt-taste and sugar-taste response to blood pressure in children: Problems 6.56–6.58, **199**; Problems 11.42–11.43, **507**  
risk factors for newborn blood-pressure measurements: Example 11.38, **468**; Example 11.39, **469**; Examples 11.40–11.41, **470–471**; Examples 11.42–11.44, **471–472**; Example 11.45, **475**; Example 11.48, **476**; Example 11.49, **481**; Example 11.51, **492**; Example 11.52, **493**  
serum-cholesterol levels in children of men with heart disease: Example 6.37, **175**; Example 7.1, **204**; Example 7.5, **206**; Example 7.9, **208**; Example 7.18, **214**; Examples 7.27–7.28, **224–225**; Example 7.35, **230**; Example 8.12, **282**; Example 8.15, **285**; Example 8.17, **289**

### PHARMACOLOGY

- clinical pharmacology of ampicillin: Problems 6.30–6.32, **197**  
concentration of aspirin in urine samples: Table 8.15, **310**

### PULMONARY DISEASE

- asthma incidence in children: Problems 7.15–7.18, **259**  
comparison of triceps skin-fold thickness in normal men and men with chronic airflow limitation: Table 6.8, **196**  
decline in pulmonary function over time: Problems 5.21–5.24, **139**  
differential diagnosis for lung cancer vs. sarcoidosis: Example 3.27, **55**  
distribution of duration of cigarette smoking: Problems 5.12–5.13, **138**  
distribution of forced vital capacity in grade-school children: Examples 5.14–5.15, **117–118**  
effect of occupational exposure to 2,4,5-T herbicide on pulmonary function: Problems 7.82–7.84, **264**  
effect of ozone exposure on pulmonary function: Table 8.22, **313**  
establishing standards for FEV for children in Tecumseh, Michigan: Example 11.14, **440**; Examples 11.17–11.18, **445**; Example 11.19, **446**; Examples 11.20–11.21, **447**; Example 11.25, **454**; Example 11.26, **455**  
familial aggregation of chronic bronchitis: Example 4.27, **85**; Example 4.29, **87**  
genetic determinants of FEV: Problems 11.9–11.12, **505**  
influence of passive smoking and parental phlegm on pneumonia and bronchitis in early childhood: Problems 13.19–13.28, **713–714**  
interventions to improve compliance with asthma medication among inner-city children: Problems 10.139–10.142, **424**  
parental smoking as a determinant of lung function in children: Problems 8.41–8.48, **311**  
relationship between asbestos exposure and death due to chronic obstructive pulmonary disease: Problems 5.10–5.11, **138**  
relationship between parental smoking and respiratory disease in childhood: Problems 3.51–3.61, **62–63**  
relationship of age, sex, height, and smoking to pulmonary function in children: Table 2.15, **35**; Problems 5.42–5.44, **140**; Problems 8.85–8.86, **315**; Problem 8.96, **316**; Problem 11.34, **506**; Problems 11.113–11.117, **513–514**  
reproducibility of dyspnea diagnoses: Problem 10.16, **410**  
risk factors for COPD death in a Chinese population: Problems 14.57–14.60, **807**  
short-term effects of sulfur-dioxide exposure in young asthmatics: Problems 12.6–12.8, **577–578**  
temporal trends in smoking-cessation rates: Table 3.8, **63**  
tuberculin skin test as a screening test for tuberculosis: Example 3.3, **39**  
use of saliva thiocyanate as an objective indicator of cigarette smoking: Table 3.9, **63**  
working environment of passive smokers and relationship to pulmonary function: Example 2.2, **5**; Problems 8.31–8.33, **310**; Examples 12.1–12.13, **516–532**; Example 12.15, **535**

### RADIOLOGY

- evaluation of accuracy of ratings in CT images by a radiologist: Example 3.32, **57**  
screening techniques for esophageal cancer: Problems 6.109–6.110, **202**

### RENAL DISEASE

- analgesic abuse and kidney disorder: Problems 8.125–8.126, **319**; Problems 10.28–10.31, **412**; Problems 12.47–12.51, **582**; Problem 13.8, **713**; Problems 13.102–13.103, **720**; Problems 14.27–14.31, **805**  
cause of death among patients with analgesic abuse: Example 9.6, **328**  
comparison of nephrotoxicity of different aminoglycosides: Example 13.50, **658**; Example 13.51, **660**; Example 13.52, **661**; Problem 13.34, **715**  
effect of protein on course of kidney disease among diabetic patients: Example 5.29, **126**; Example 5.31, **128**  
physiological and psychological changes in patients with end-stage renal disease: Table 6.10, **197**  
prevalence of bacteriuria over 2 surveys: Problems 4.37–4.41, **101**  
serum-creatinine levels: Examples 5.25–5.28, **125–126**; Problem 7.1, **259**  
treatment of patients with diabetic nephropathy with captopril: Table 8.19, **312**

### RHEUMATOLOGY

- comparison of muscle function between patients with rheumatoid arthritis and osteoarthritis: Table 8.23, **314**  
effectiveness of ibuprofen in preventing inflammation among osteoarthritis patients: Review Question 13G.2, **666**  
testing a new drug for relief of pain from osteoarthritis: Example 7.7, **207**

## INDEX OF APPLICATIONS (continued)

### SEROLOGY

variability of an assay for *M. hominis* mycoplasma: Problems 6.36–6.39, **198**

### SEXUALLY TRANSMITTED DISEASE

association between type of STD and previous episodes of urethritis: Problem 10.13, **410**

comparison of spectinomycin and penicillin G in treating patients with gonorrhea: Problems 6.27–6.29, **197**; Problems 10.24–10.25, **412**  
diagnosing patients for syphilis: Example 3.15, **44**; Example 3.20, **48**

### SLEEP DISORDERS

change in prevalence of sleep-disordered breathing by age and sex: Example 13.30, **622**; Example 13.31, **623**

### SPORTS MEDICINE

comparison of Motrin and placebo in the treatment of tennis elbow: Problems 8.89–8.92, **316**; Problems 9.13–9.14, **346**; Example 13.56, **666**; Example 13.57, **669**; Examples 13.58–13.59, **671**; Problems 13.54–13.59, **716**

effect of playing surface on the rate of Canadian football injuries: Problems 6.102–6.103, **201–202**

risk factors for tennis elbow: Problem 10.55, **416**; Examples 13.48–13.49, **656**; Problems 13.39–13.40, **715**

### UROLOGY

age at onset of night-time bladder control: Problems 10.80–10.81, **418**

### WOMEN'S HEALTH

relationship between abortion and breast cancer: Problems 4.75–4.77, **104**

### ZOOLOGY

preference of different bird species for different types of sunflower seeds: Problems 10.70–10.73, **417–418**