\[ y = b_0 + b_1 x \]

where
\[ b_0 = \bar{y} - b_1 \bar{x} \]

and
\[ b_1 = \frac{s_{xy}}{s_x^2} \]

<table>
<thead>
<tr>
<th>Patient Number</th>
<th>WBC ((/10^3))</th>
<th>Culture Count ((/10^3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.8</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>11.0</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>17.4</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>15.9</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>18.2</td>
<td>38</td>
</tr>
</tbody>
</table>

Graph showing scatter plot with regression line.
Statistics in Medicine
DEDICATION

To my children and grandchildren and to the very many medical investigators I have been privileged to work with.

—RHR

To Lilli, Elise, Sienna, my mother Sharon, and all my collaborators over the years.

—DLG
When told, “I’m too busy treating patients to do research,” we answer:

*When you treat a patient, you treat one patient.*
*When you do research, you treat 10,000 patients.*
## Contents

*Foreword to the fourth edition*  
*Acknowledgments*  
*How to use this book?*

### 1. Planning studies: from design to publication

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Organizing a Study</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Stages of Scientific Investigation</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Science Underlying Clinical Decision-Making</td>
<td>3</td>
</tr>
<tr>
<td>1.4 Why Do We Need Statistics?</td>
<td>5</td>
</tr>
<tr>
<td>1.5 Concepts in Study Design</td>
<td>7</td>
</tr>
<tr>
<td>1.6 Study Types</td>
<td>8</td>
</tr>
<tr>
<td>1.7 Convergence With Sample Size</td>
<td>11</td>
</tr>
<tr>
<td>1.8 Sampling Schemes in Observational Studies</td>
<td>11</td>
</tr>
<tr>
<td>1.9 Sampling Bias</td>
<td>13</td>
</tr>
<tr>
<td>1.10 Randomizing a Sample</td>
<td>15</td>
</tr>
<tr>
<td>1.11 How to Plan and Conduct a Study</td>
<td>16</td>
</tr>
<tr>
<td>1.12 Mechanisms to Improve Your Study Plan</td>
<td>17</td>
</tr>
<tr>
<td>1.13 Reading Medical Articles</td>
<td>19</td>
</tr>
<tr>
<td>1.14 Where Articles May Fall Short</td>
<td>19</td>
</tr>
<tr>
<td>1.15 Writing Medical Articles</td>
<td>22</td>
</tr>
<tr>
<td>1.16 Statistical Ethics in Medical Studies</td>
<td>23</td>
</tr>
<tr>
<td>1.17 Conclusion</td>
<td>27</td>
</tr>
<tr>
<td>Appendix to Chapter 1</td>
<td>28</td>
</tr>
</tbody>
</table>

### 2. Planning analysis: how to reach my scientific objective

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 What Is in This Chapter</td>
<td>29</td>
</tr>
<tr>
<td>2.2 Notation (or Symbols)</td>
<td>29</td>
</tr>
<tr>
<td>2.3 Quantification and Accuracy</td>
<td>31</td>
</tr>
<tr>
<td>2.4 Data Types</td>
<td>32</td>
</tr>
<tr>
<td>2.5 Multivariable Concepts and Types of Adjustment Variables</td>
<td>35</td>
</tr>
<tr>
<td>2.6 How to Manage Data</td>
<td>37</td>
</tr>
<tr>
<td>2.7 Defining the Scientific Goal: Description, Association Testing, Prediction</td>
<td>40</td>
</tr>
<tr>
<td>2.8 Reporting Statistical Results</td>
<td>41</td>
</tr>
<tr>
<td>2.9 A First-Step Guide to Descriptive Statistics</td>
<td>42</td>
</tr>
</tbody>
</table>
### Contents

#### 2.10  An Overview of Association Testing 45
#### 2.11  A Brief Discussion of Prediction Modeling 50

#### 3.  Probability and relative frequency 51

- **3.1  Probability Concepts** 51
- **3.2  Probability and Relative Frequency** 55
- **3.3  Graphing Relative Frequency** 56
- **3.4  Continuous Random Variables** 57
- **3.5  Frequency Distributions for Continuous Variables** 58
- **3.6  Probability Estimates From Continuous Distributions** 62
- **3.7  Probability as Area Under the Curve** 63

References 64

#### 4.  Distributions 65

- **4.1  Characteristics of a Distribution** 65
- **4.2  Greek Versus Roman Letters** 65
- **4.3  What Is Typical** 65
- **4.4  The Spread About the Typical** 68
- **4.5  The Shape** 69
- **4.6  Sampling Distribution of a Variable Versus a Statistic** 71
- **4.7  Statistical Inference** 72
- **4.8  Distributions Commonly Used in Statistics** 75
- **4.9  Approximate Distribution of the Mean (Central Limit Theorem)** 83
- **4.10  Approximate Distribution of a Sample Quantile** 86

#### 5.  Descriptive statistics 89

- **5.1  Purpose of Descriptive Statistics** 89
- **5.2  Numerical Descriptors, One Variable** 89
- **5.3  Numerical Descriptors, Two Variables** 95
- **5.4  Numerical Descriptors, Three Variables** 100
- **5.5  Graphical Descriptors, One Variable** 101
- **5.6  Graphical Descriptors, Two Variables** 101
- **5.7  Graphical Descriptors, Three Variables** 111
- **5.8  Principles of Informative Descriptive Tables and Figures** 114

References 115

#### 6.  Finding probabilities 117

- **6.1  Probability and Area Under the Curve** 117
- **6.2  The Normal Distribution** 117
- **6.3  The t Distribution** 121
- **6.4  The Chi-Square Distribution** 124
- **6.5  The F-Distribution** 127
6.6 The Binomial Distribution 130
6.7 The Poisson Distribution 132
References 135

7. Hypothesis testing: concept and practice 137
7.1 Hypotheses in Inference 137
7.2 Error Probabilities 144
7.3 Two Policies of Testing 148
7.4 Distinguishing Between Statistical and Clinical Significance 150
7.5 Controversies Regarding the Rigid Use and Abuse of $p$-Values 151
7.6 Avoiding Multiplicity Bias 152
7.7 Organizing Data for Inference 152
7.8 Evolving a Way to Answer Your Data Question 156
Reference 159

8. Tolerance, prediction, and confidence intervals 161
8.1 Overview 161
8.2 Tolerance Intervals for Patient Measurements 163
8.3 Concept of a Confidence Interval for a Parameter 164
8.4 Confidence Interval for a Population Mean, Known Standard Deviation 166
8.5 Confidence Interval for a Population Mean, Estimated Standard Deviation 169
8.6 Confidence Interval for a Population Proportion 171
8.7 Confidence Interval for a Population Median 175
8.8 Confidence Interval for a Population Variance or Standard Deviation 176
8.9 Confidence Interval for a Population Correlation Coefficient 179
References 181

9. Tests on categorical data 183
9.1 Categorical Data Basics 183
9.2 Tests on Categorical Data: $2 \times 2$ Tables 185
9.3 The Chi-Square Test of Contingency 186
9.4 Fisher’s Exact Test of Contingency 190
9.5 Tests on $r \times c$ Contingency Tables 192
9.6 Tests of Proportion 195
9.7 Tests of Rare Events (Proportions Close to Zero) 201
9.8 McNemar’s Test: Matched Pair Test of a $2 \times 2$ Table 205
9.9 Cochran’s Q: Matched Pair Test of a $2 \times r$ Table 210
9.10 Three or More Ranked Samples With Two Outcome Categories: Royston’s $P_{trend}$ Test 211
References 213

10. Risks, odds, and receiver operating characteristic curves 215
10.1 Association Measures for Categorical Data: Risks and Odds 215
10.2 Inference for the Risk Ratio: The Log Risk Ratio Test 225
## Contents

10.3 Inference for the Odds Ratio: The Log Odds Ratio Test 227
10.4 Receiver Operating Characteristic Curves 232
10.5 Comparing Two Receiver Operating Characteristic Curves References 235 237

### 11. Tests of location with continuous outcomes

11.1 Basics of Location Testing 239
11.2 Single or Paired Means: One-Sample Normal (z) and t Tests 241
11.3 Two Means: Two-Sample Normal (z) and t Tests 245
11.4 Three or More Means: One-Factor Analysis of Variance 256
11.5 Three or More Means in Rank Order: Analysis of Variance Trend Test 265
11.6 The Basics of Nonparametric Tests 267
11.7 Single or Paired Sample Distribution(s): The Signed-Rank Test 269
11.8 Two Independent Sample Distributions: The Rank-Sum Test 272
11.9 Large Sample-Ranked Outcomes 276
11.10 Three or More Independent Sample Distributions: The Kruskal—Wallis Test 280
11.11 Three or More Matched Sample Distributions: The Friedman Test 284
11.12 Three or More Ranked Independent Samples With Ranked Outcomes: Cusick’s Nptrend Test 287
11.13 Three or More Ranked Matched Samples With Ranked Outcomes: Page’s L Test 290
11.14 Potential Drawbacks to Using Nonparametric Tests 291

### 12. Equivalence testing

12.1 Concepts and Terms 295
12.2 Basics Underlying Equivalence Testing 296
12.3 Choosing a Noninferiority or Equivalence Margin 299
12.4 Methods for Noninferiority Testing 300
12.5 Methods for Equivalence Testing 304
12.6 Joint Difference and Equivalence Testing 307

### 13. Tests on variability and distributions

13.1 Basics of Tests on Variability 311
13.2 Testing Variability on a Single Sample 312
13.3 Testing Variability Between Two Samples 315
13.4 Testing Variability Among Three or More Samples 317
13.5 Basics on Tests of Distributions 321
13.6 Test of Normality of a Distribution 322
13.7 Test of Equality of Two Distributions 331

### 14. Measuring association and agreement

14.1 What Are Association and Agreement? 337
14.2 Contingency as Association 338
14.3 Correlation as Association 342
14.4 Contingency as Agreement 344
14.5 Correlation as Agreement 345
14.6 Agreement Among Ratings: Kappa 346
14.7 Agreement Among Multiple Rankers 348
14.8 Reliability 353
14.9 Intraclass Correlation 354
References 355

15. Linear regression and correlation 357
15.1 Introduction 357
15.2 Regression Concepts and Assumptions 359
15.3 Simple Regression 363
15.4 Assessing Regression: Tests and Confidence Intervals 368
15.5 Deming Regression 376
15.6 Types of Regression 378
15.7 Correlation Concepts and Assumptions 379
15.8 Correlation Coefficients 382
15.9 Correlation as Related to Regression 384
15.10 Assessing Correlation: Tests and Confidence Intervals 385
15.11 Interpretation of Small-But-Significant Correlations 390
References 390

16. Multiple linear and curvilinear regression and multifactor analysis of variance 391
16.1 Introduction 391
16.2 Multiple Linear Regression 396
16.3 Model Diagnosis and Goodness of Fit 407
16.4 Accounting for Heteroscedasticity 412
16.5 Curvilinear Regression 414
16.6 Two-Factor Analysis of Variance 420
16.7 Analysis of Covariance 426
16.8 Three-Way and Higher Way Analysis of Variance 430
16.9 Concepts of Experimental Design 431
References 435

17. Logistic regression for binary outcomes 437
17.1 Introduction 437
17.2 Extensions of Contingency Table Analyses Simple Logistic Regression 438
17.3 Multiple Logistic Regression: Model Specification and Interpretation 441
17.4 Inference for Association Parameters 445
17.5 Model Diagnostics and Goodness-of-Fit 448
References 457
Contents

18. Poisson regression for count outcomes 459
   18.1 Introduction 459
   18.2 The Poisson Distribution 459
   18.3 Means Versus Rates 460
   18.4 Inference for the Rate of a Poisson Random Variable 461
   18.5 Comparing Poisson Rates From Two Independent Samples 463
   18.6 The Simple Poisson Regression Model 465
   18.7 Multiple Poisson Regression: Model Specification and Interpretation 469
   18.8 Obtaining Predicted Rates 472
   18.9 Inference for Association Parameters 473
   Reference 475

19. Analysis of censored time-to-event data 477
   19.1 Survival Concepts 477
   19.2 Censoring 478
   19.3 Survival Estimation: Life Table Estimates and Kaplan—Meier Curves 480
   19.4 Survival Testing: The Log-Rank Test 484
   19.5 Adjusted Comparison of Survival Times: Cox Regression 487
   References 490

20. Analysis of repeated continuous measurements over time 491
   20.1 Introduction 491
   20.2 Distinguishing Longitudinal Data From Time-Series Data 492
   20.3 Analysis of Longitudinal Data 493
   20.4 Time-Series: 514
   References 531

21. Sample size estimation 533
   21.1 Issues in Sample Size Considerations 533
   21.2 Is the Sample Size Estimate Adequate? 536
   21.3 The Concept of Power Analysis 537
   21.4 Sample Size Methods 540
   21.5 Test on One Mean (Normal Distribution) 541
   21.6 Test on Two Means (Normal Distribution) 542
   21.7 Tests When Distributions Are Nonnormal or Unknown 545
   21.8 Test With No Objective Prior Data 546
   21.9 Confidence Intervals on Means 547
   21.10 Test of One Proportion (One Rate) 549
   21.11 Test of Two Proportions (Two Rates) 551
   21.12 Confidence Intervals on Proportions (On Rates) 553
   21.13 Test on a Correlation Coefficient 555
   21.14 Tests on Ranked Data 557
   21.15 Variance Tests, Analysis of Variance, and Regression 558
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.16 Equivalence Tests</td>
<td>558</td>
</tr>
<tr>
<td>21.17 Number Needed to Treat or Benefit</td>
<td>559</td>
</tr>
<tr>
<td>References</td>
<td>562</td>
</tr>
<tr>
<td><strong>22. Clinical trials and group sequential testing</strong></td>
<td>565</td>
</tr>
<tr>
<td>22.1 Introduction</td>
<td>565</td>
</tr>
<tr>
<td>22.2 Fundamentals of Clinical Trial Design</td>
<td>568</td>
</tr>
<tr>
<td>22.3 Reducing Bias in Clinical Trials: Blinding and Randomization</td>
<td>573</td>
</tr>
<tr>
<td>22.4 Interim Analyses in Clinical Trials: Group Sequential Testing</td>
<td>577</td>
</tr>
<tr>
<td>References</td>
<td>582</td>
</tr>
<tr>
<td><strong>23. Epidemiology</strong></td>
<td>583</td>
</tr>
<tr>
<td>23.1 The Nature of Epidemiology</td>
<td>583</td>
</tr>
<tr>
<td>23.2 Some Key Stages in the History of Epidemiology</td>
<td>584</td>
</tr>
<tr>
<td>23.3 Concept of Disease Transmission</td>
<td>584</td>
</tr>
<tr>
<td>23.4 Descriptive Measures</td>
<td>586</td>
</tr>
<tr>
<td>23.5 Types of Epidemiologic Studies</td>
<td>589</td>
</tr>
<tr>
<td>23.6 Retrospective Study Designs: The Case—Control Study Design</td>
<td>591</td>
</tr>
<tr>
<td>23.7 The Nested Case—Control Study Design</td>
<td>593</td>
</tr>
<tr>
<td>23.8 The Case—Cohort Study Design</td>
<td>594</td>
</tr>
<tr>
<td>23.9 Methods to Analyze Survival and Causal Factors</td>
<td>594</td>
</tr>
<tr>
<td>23.10 A historical note</td>
<td>598</td>
</tr>
<tr>
<td>References</td>
<td>599</td>
</tr>
<tr>
<td><strong>24. Meta-analyses</strong></td>
<td>601</td>
</tr>
<tr>
<td>24.1 Introduction</td>
<td>601</td>
</tr>
<tr>
<td>24.2 Publication Bias in Meta-analyses</td>
<td>602</td>
</tr>
<tr>
<td>24.3 Fixed- and Random-Effects Estimates of the Pooled Effect</td>
<td>603</td>
</tr>
<tr>
<td>24.4 Tests for Heterogeneity of Estimated Effects Across Studies</td>
<td>606</td>
</tr>
<tr>
<td>24.5 Reporting the Results of a Meta-analysis</td>
<td>606</td>
</tr>
<tr>
<td>24.6 Further References</td>
<td>607</td>
</tr>
<tr>
<td>References</td>
<td>608</td>
</tr>
<tr>
<td><strong>25. Bayesian statistics</strong></td>
<td>609</td>
</tr>
<tr>
<td>25.1 What Is Bayesian Statistics</td>
<td>609</td>
</tr>
<tr>
<td>25.2 Bayesian Concepts</td>
<td>612</td>
</tr>
<tr>
<td>25.3 Describing and Testing Means</td>
<td>613</td>
</tr>
<tr>
<td>25.4 On Parameters Other Than Means</td>
<td>617</td>
</tr>
<tr>
<td>25.5 Describing and Testing a Rate (Proportion)</td>
<td>617</td>
</tr>
<tr>
<td>25.6 Conclusion</td>
<td>619</td>
</tr>
<tr>
<td>References</td>
<td>619</td>
</tr>
<tr>
<td><strong>26. Questionnaires and surveys</strong></td>
<td>621</td>
</tr>
<tr>
<td>26.1 Introduction</td>
<td>621</td>
</tr>
</tbody>
</table>
## Contents

26.2 Surveys 621
26.3 Questionnaires 624

### 27. Techniques to Aid Analysis 631

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.1 Interpreting Results</td>
<td>631</td>
</tr>
<tr>
<td>27.2 Significance in Interpretation</td>
<td>632</td>
</tr>
<tr>
<td>27.3 <em>Post Hoc</em> Confidence and Power</td>
<td>633</td>
</tr>
<tr>
<td>27.4 Multiple Tests and Significance</td>
<td>635</td>
</tr>
<tr>
<td>27.5 Bootstrapping, Resampling, and Simulation</td>
<td>639</td>
</tr>
<tr>
<td>27.6 Bland-Altman Plot: A Diagnostic Tool</td>
<td>642</td>
</tr>
<tr>
<td>27.7 Cost Effectiveness</td>
<td>647</td>
</tr>
</tbody>
</table>

References 649

### 28. Methods you might meet, but not every day 651

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.1 Overview</td>
<td>651</td>
</tr>
<tr>
<td>28.2 Analysis of Variance Issues</td>
<td>651</td>
</tr>
<tr>
<td>28.3 Regression Issues</td>
<td>652</td>
</tr>
<tr>
<td>28.4 Rates and Proportions Issues</td>
<td>653</td>
</tr>
<tr>
<td>28.5 Multivariate Methods</td>
<td>653</td>
</tr>
<tr>
<td>28.6 Markov Chains: Following Multiple States through Time</td>
<td>656</td>
</tr>
<tr>
<td>28.7 Markov Chain Monte Carlo: Evolving Models</td>
<td>657</td>
</tr>
<tr>
<td>28.8 Markov Chain Monte Carlo: Stationary Models</td>
<td>659</td>
</tr>
<tr>
<td>28.9 Further Nonparametric Tests</td>
<td>659</td>
</tr>
<tr>
<td>28.10 Imputation of Missing Data</td>
<td>660</td>
</tr>
<tr>
<td>28.11 Frailty Models in Survival Analysis</td>
<td>660</td>
</tr>
<tr>
<td>28.12 <em>Bonferroni</em> &quot;Correction&quot;</td>
<td>661</td>
</tr>
<tr>
<td>28.13 Logit and Probit</td>
<td>661</td>
</tr>
<tr>
<td>28.14 Adjusting for Outliers</td>
<td>661</td>
</tr>
<tr>
<td>28.15 Curve Fitting to Data</td>
<td>662</td>
</tr>
<tr>
<td>28.16 Sequential Analysis</td>
<td>663</td>
</tr>
<tr>
<td>28.17 Another Test of Normality</td>
<td>665</td>
</tr>
<tr>
<td>28.18 Data Mining</td>
<td>665</td>
</tr>
<tr>
<td>28.19 Data Science and The Relationship Among Statistics, Machine Learning, and Artificial Intelligence</td>
<td>666</td>
</tr>
</tbody>
</table>

References 667

*Appendix 1: Answers to exercises: Final* 669
*Appendix 2: Databases* 717
*Appendix 3: Tables of probability distributions* 741
*Appendix 4: Symbol index* 761
*Statistical Subject Index* 765
*Medical Subject Index* 789
Foreword to the fourth edition

The use of statistics in medical studies has revolutionized our understanding of human health in the past 70 years or so. Early and classic examples include measuring the success of the polio vaccine in the 1950s, and understanding the relationship between smoking and lung cancer that led to the Surgeon General’s warning on cigarettes starting in 1970. Since that time hundreds of thousands of medical studies have reaped the benefit of the use of statistical methods.

This update of Statistics in Medicine continues the philosophy of Riffenburgh’s first three editions, providing a nontechnical, comprehensible guide to most of the common statistical methods used to collect, analyze, and report data used in medical studies. Whether you are a medical practitioner, a student studying medicine, or a layperson who wants to understand research about your own health issues or those of a loved one, this book will guide you through what you need to know to understand medical research results. If you are a practicing medical researcher, this book will provide you with the tools, vocabulary, and understanding needed to work collaboratively with a professional statistician. While it may seem like there are too many statistical methods to cover in one book of this nature, the fact is that a large majority of medical studies use just a small handful of common methods, all covered in this book.

The fourth edition of this book brings many new features, including an additional author, Professor Daniel L. Gillen. Professor Gillen has extensive experience working with medical researchers on a range of diseases, including cancer and Alzheimer’s and has consulted extensively with the Food and Drug Administration and the biopharmaceutical industry. As a professor, he is a highly acclaimed teacher and mentor for students in both statistics and medical fields and has made substantial contributions to biostatistical methodology. Adding Professor Gillen’s expertise to that of Professor Riffenburgh, who wrote the first three editions of the book and has several decades of experience as a statistical collaborator, means that there is almost no area of medicine or biostatistics that is not familiar to one or both authors.

The book begins where you should begin if you want to understand the process of medical research—with an overview of the process, including a step-by-step guide on how to plan a successful study and avoid common mistakes. Even if you never plan or implement your own research study, this guide will help you assess strengths and weaknesses when you read the medical literature. That section is just one small example of the focus this book places on understanding best practices and detecting
common errors. The first two chapters set the stage for later chapters and include almost no formulas. There are an additional 26 chapters, for a total of 28. Chapters 3–25 cover the myriad of statistical methods most commonly used in medical research. Notable and unusual in a book of this type are the final two chapters: Chapter 27 titled “Techniques to aid analysis,” discusses subtleties that can affect the interpretation of results, and Chapter 28 has the self-explanatory title “Methods you might meet, but not every day.”

There are five new chapters in this fourth edition, as well as new subsections in existing chapters. The new chapters cover statistical methods that are occurring with increasing frequency in the medical research literature. These include logistic regression, Poisson regression, longitudinal and time-series analyses, group sequential analysis, and metaanalysis. Despite the sophistication of some of these methods, these chapters focus on conceptual understanding. The methods are illustrated with several examples, including interpretations of the models and results.

There are 17 real data sets accompanying the book, and an additional 20 data sets available from the publisher’s website, covering a wide range of medical issues. Two of the data sets are new to this edition, obtained from Professor Gillen’s work with the Alzheimer’s Disease Research Center at the UCI. All of the data sets are used in examples throughout the book, and in addition to downloadable files of the data, a background description is provided for each one.

Artificial intelligence, machine learning, and data mining are all techniques (described in Chapter 28: Methods you might meet, but not every day) being used to produce “black box” algorithms that allegedly replace human judgment for medical diagnosis and care. But these methods and algorithms are only as good as the data used to create them. For many decades, statisticians have developed and refined appropriate methods for collecting and understanding data and are aware of the dangers of overinterpreting results based on poorly designed studies. Now, more than ever, medical practitioners need a book like Statistics in Medicine to understand the importance of every stage of the design and implementation of medical studies in producing useful and accurate results.

Professor Emerita Jessica Utts*

University of California, Irvine, Irvine, CA, United States

* Dr. Jessica Utts, Professor Emerita of Statistics at the University of California, Irvine (UCI), has served as the Chair of the Department of Statistics at UCI, and Associate Vice Provost of University Outreach at UC Davis, where she was on the faculty for 30 years prior to moving to UCI. She was the President of the American Statistical Association in 2016 and President of the International Biometric Society, Western North American Region, in 1986. She served as the Chief Reader for the Advanced Placement Statistics Exam from 2014 to 2018. Professor Utts is internationally recognized for her work in parapsychology and her ability to bring statistical literacy to the masses.
## ABSTRACTS OF FOREWORDS TO PRIOR EDITIONS

(Affiliations are listed as of when the Foreword was written.)

| Foreword to the first edition: | Vice Admiral Richard A. Nelson, MD, USN  
Surgeon General of the Navy Chief, Bureau of Medicine and Surgery |
---|---|

We have a critical obligation not only to care for our patients but also to assess thoroughly our decisions on therapy, treatments effects, and outcomes. Scientifically rigorous medical studies, including sound statistical analyses are central to this process. In recent years, statistical analysis has become the hallmark of studies appearing in the leading journals.

This book represents a practical opportunity for health-care trainees and staff not only to acquaint themselves with statistics in medicine and understand statistical analysis in the medical literature but also to be guided in the application of planning and analysis in their own medical studies.

***

| Foreword to the second edition: | W.M. (Mike) O’Fallon, PhD  
Professor Emeritus of Biostatistics, Mayo Clinic; Chair,  
Department of Health Science Research Past President, American Statistical Association |
---|---|

To refer to the last century as revolutionary in the statistical world would be a gross understatement. I believe the effects of this revolution are more profound in the world of medicine and other health-related sciences than in most scientific disciplines.

Dr. Riffenburgh addresses just this problem in his successful book on statistics in medicine (presenting) statistics in language accessible to health-care practitioners so that they could effectively understand and communicate statistical information.

***

| Foreword to the third edition: | P.A. (Tony) Lachenbruch, PhD  
Professor of Public Health, Oregon State Univ. Past President, American Statistical Association Past President, International Biometric Society Elected Member, International Statistics Institute |
---|---|

The third edition of *Statistics in Medicine* provides a wonderful pathway for people to learn those methods (about 30% according to the author) used in 90% of all (medical) statistical analyses. The message is clear: most statistical applications are conducted with a relatively few procedures and these are all covered in this book.
We consistently strive to develop, communicate, and apply statistical methodologies in a pragmatic fashion in order to further progress science and public health. We are eternally grateful to the many mentors, colleagues, and students that have helped us on this journey.

We extend appreciation to all those acknowledged in prior editions. Although many of you have moved on to new challenges, your contributions to this book continue to be recognized and appreciated.

We are grateful to Dr. Jessica Utts for writing the Foreword to this fourth edition. Dr. Utts, some of whose stellar accomplishments are documented along with her Foreword has long been a friend and esteemed colleague. We are also extremely thankful to Michelle Nuño for carefully proofreading the text and providing valuable feedback.

We are grateful to Stata Corporation for providing statistical software used in much of the preparation of this book.

Finally, we thank our families for the patience and forbearance during our time of neglect while preparing this edition.
How to use this book?

1 GOALS

Purpose of the book

This book was written to teach and provide reference on statistical methods for research workers in the medical sciences. The book starts at “ground zero” in mathematics and statistics but assumes an intelligent, alert, and motivated student with ability in arithmetic and a bit of algebra. It is designed to be offered without formal university-level prerequisites.

The motivation for the book stems from our attempts to provide simple explanations to statistically naïve collaborators. Thus its level and tone are designed to be user-friendly to the health-care practitioner who does not engage in research on a regular basis. It was not written for statisticians, some of whom disapprove of its simplicity relative to many traditional statistical texts. However, we believe this is a major factor in its success.

The most used statistical methods and concepts are included

This book does not attempt to cover all biostatistical methods. Earlier editions of the book referred to an article by Emerson and Colditz¹ that analyzed the 301 articles appearing in four volumes of NEJM for statistical content. The first two editions covered 90% of the statistical methods used in these 301 articles, the remaining 10% being a scattering of infrequently used, often arcane methods. Since that time, more statistical methods have come into common use and, correspondingly, more are covered in the later editions.

Using a few data sets repeatedly allows the user to focus on the method

The book opens with a number of databases having sample sizes small enough to be given fully. These databases (DBs) are drawn from various medical areas, including urology (DB1: prostate cancer), surgery (DB2: gall bladder removal), dermatology (DB6: tattoo removal), orthopedics (DB7: femoral neck fractures), internal medicine (DB11: treating malaria), ENT-HNS (DB12: carinal resection), pulmonary medicine (DB14: bronchoconstriction), and others. Supplementary examples with larger sample sizes from other medical fields are also included, summary statistics being provided for the user. The database set is available online in MS Excel format in the Elsevier website.
2 USE AS A TEXTBOOK VERSUS AS A REFERENCE

This book is intended to serve both as a text for a course in medical statistics and as a reference to statistical methods for investigators. The first two editions started with 10 chapters for a course and 16 more for reference. It was found that the course part was too long for some users and too short for others and that the investigator had to maneuver back and forth for methods. As with the third edition, this fourth edition begins from first concepts, adding new topics that are commonly encountered in medical and public health research.

To use this edition for a course, chapters and sections can be selected according to the recommendations appearing in Table 1.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Suggested list of chapters/sections to constitute a course of indicated length.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short course</td>
</tr>
<tr>
<td>Immersion into research</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Basic tools and concepts</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6.1,2,3,4</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8.1,3,5</td>
<td>8.1,2,3,4,5,6</td>
</tr>
<tr>
<td>8.1,2,3,4</td>
<td>8</td>
</tr>
<tr>
<td>9.1,2,3,4</td>
<td>9.1,2,3,4,6,7,8</td>
</tr>
<tr>
<td>10.1</td>
<td>10.1,2,3,4</td>
</tr>
<tr>
<td>11.1,2,3,6,7,8</td>
<td>11.1,2,3,4,6,7,8,10</td>
</tr>
<tr>
<td>12.1,2,3,4,5</td>
<td>12</td>
</tr>
<tr>
<td>15.1,2,3,7,8,9,11</td>
<td>15</td>
</tr>
<tr>
<td>16.1,2,3,4,5,6,7</td>
<td>16</td>
</tr>
<tr>
<td>17.1,2,3</td>
<td>17</td>
</tr>
<tr>
<td>19.1,2,3,4</td>
<td>18</td>
</tr>
<tr>
<td>21.1,2,3,4,5,6,7,10,11</td>
<td>20</td>
</tr>
<tr>
<td>22.1,2,3</td>
<td>21.1,2,3,4,5,6,7,10,11,13,14,17</td>
</tr>
<tr>
<td>23.1,2,3,4,5,10</td>
<td>22</td>
</tr>
<tr>
<td>25</td>
<td>23.1,2,3,4,5,10</td>
</tr>
<tr>
<td>Further topics</td>
<td>As time permits</td>
</tr>
<tr>
<td>As time permits</td>
<td>As time permits</td>
</tr>
<tr>
<td>As time permits</td>
<td>As time permits</td>
</tr>
</tbody>
</table>
Students using this book are most often medical residents, fellows, and public health students encountering their first exposure to research, but the material is suited for all levels from premedical undergraduates to experienced staff by using minor alterations in speed and depth of coverage and choice of material. We have been surprised to learn that it has even been used successfully by high school students and by advanced, experienced medical investigators.

3 POSSIBLE SCHEDULES FOR COURSES

Table 1 provides a list of chapters, or sections if chapters are not to be covered completely, for three different lengths of course. Comments on the different lengths of courses are given in adjoining paragraphs.

A large number of students will already have knowledge about some pieces of Chapter 1, Planning Studies: From Design to Publication, and Chapter 2, Planning Analysis: How to Reach My Scientific Objective, but most will be missing a piece here and there. These chapters can be covered more quickly than the remainder of the book but should at the very least be skimmed to fill in this or that element of knowledge a particular student may be lacking. In general, the beginning chapters of the book can be covered relatively quickly. The coverage should slow as the material increases in complexity and reduces in familiarity.

Short course

Short courses on the recommended set of topics have been presented several times each at Emory University School of Medicine, Rady Childrens’ Hospital San Diego, and the Naval Medical Center San Diego. These courses were 10 clock hours equating to 12 50-minute university periods, consisting mainly of straight lecture with little student interaction. (This is a description, not a recommendation.) There is not enough time in such a course for a student to master more than the most rudimentary statistical concepts and methods; the course exposes the student to the concepts, terminology, and methods of statistics and assists the student’s ability to understand articles in medical journals.

The recommended short course covers material from the first 11 chapters of the text.

A three-credit semester course

A three-credit semester course includes 45 class periods, perhaps 42 lectures when time for examinations is removed, equating to 35 60-minute hours, or about threefold the time allowed for the short course. It includes material from additional nine chapters or about 1.5 times the amount of material. Thus the semester course can proceed
at a little slower rate than short course. The material can be studied in greater depth with time for interaction between the students and the instructor.

**A year course (six-semester credits)**

A year course covers more material (all chapters with exception to 27 and 28) than the one term course. It can proceed at about the rate of the one-term course for material occurring in both course designs but allowing for greater time spent on the more advanced material that will inevitably be more involved and less familiar.

**REFERENCE**

Planning studies: from design to publication

(A few statistical terms commonly appearing in medical articles appear in this chapter without having been previously defined. In case, a reader encounters an unfamiliar one, a glossary at the chapter’s end provides interim definitions pending formal definitions later in this book.)

1.1 ORGANIZING A STUDY

A study must be organized sooner or later. Planning in advance from an overview down to details increases efficiency, reduces false starts, reduces errors, and shortens the time spent. By definition an impactful scientific study must be credible to the scientific community. This implies that the study must meet minimal scientific standards, including valid methods, valid measurements of study outcomes, valid quantification of empirical results, and appropriate interpretations of study results. The most common pitfalls of most studies stem from a lack of a priori design and planning. If that is not enough motivation, the lazy and least stressful way to conduct a study, in the sense of least work overall, is thorough planning upfront.

1.2 STAGES OF SCIENTIFIC INVESTIGATION

Stages

We gather data because we want to know something. These data are useful only if they provide information about what we want to know. A scientist usually seeks to develop knowledge in three stages. The first stage is to describe a class of scientific events and formulate hypotheses regarding the nature of the events. The second stage is to explain these events. The third stage is to predict the occurrence of these events. The ability to predict an event implies some level of understanding of the rule of nature governing the event. The ability to predict outcomes of actions allows the scientist to make better decisions about such actions. At best, a general scientific rule may be inferred from repeated events of this type. The latter two stages of a scientific
investigation will generally involve building a statistical model. A statistical model is an abstract concept but in the most general of terms it is a mathematical model that is built on a (generally) simplifying set of assumptions that attempts to explain the mechanistic or data-generating process that gave rise to the data one has observed. In this way a statistical model allows one to infer from a sample to the larger population by relying upon the assumptions made to describe the data-generating process. While the topic may seem abstract, the reader has undoubtedly encountered and possibly utilized multiple statistical models in practice. As an example, we might look at body mass index, or BMI. BMI is an indicator of body weight adjusted for body size. The model is BMI = weight/height², where weight is measured in kilograms and height in meters. (If pounds and inches are used, the equation becomes BMI = 703 × weight/height²).

A 6-ft person (1.83 m) weighing 177 lb (80 kg) would have 80/1.83² = 23.9 BMI. To build such a model, we would start with weights and heights recorded for a representative sample of normal people. (We will ignore underweight for this example.) For a given height, there is an ideal weight and the greater the excess weight, the lower the health. But ideal weight varies with body size. If we plot weights for various heights, we find a curve that increases in slope as height increases, something akin to the way y² looks when plotted for x, so we try height². For a fixed weight the body mass measure goes down as height goes up, so the height term should be a divider of weight, not a multiplier. Thus we have the BMI formula. Of course, many influences are ignored to achieve simplicity. A better model would adjust for muscle mass, bone density, and others, but such measures are hard to come by. Height and weight are normally in every person’s medical history.

The model gives an estimate of, or approximation to, the body weight’s influence on the person’s health. More generally, a model approximates a state or condition based on measurements of influencing variables, whence its name, a model of the state, not a direct measure. The greater the predictive accuracy and reliability of a model, the more complicated the model needs to be. Usually, models are trade-offs between accessibility of measures and simplicity of interpretation versus the requirement for accuracy.

Sometimes it is necessary to formulate more complicated models in order to ensure better predictive accuracy. For example, The American College of Cardiology utilizes a model to estimate an individual’s 10-year risk of atherosclerotic cardiovascular disease (ASCVD). This model utilizes 13 variables to obtain an estimate of the probability that an individual will experience ASCVD within the next 10 years. These variables include factors such as age, sex, weight, smoking status, systolic and diastolic blood pressure, cholesterol levels, and medication use. The model then weights each of these factors in order to compute an estimate of ASCVD risk. Due to the complexity of the model, it is not easy to write down and communicate, as is the case with BMI. Instead, it is easier to produce an “online calculator” that takes in each of the influencing variables and, behind the scenes, feeds these values into the model to report a final
estimate of the probability of ASCVD. As an example, the ASCVD online calculator from The American College of Cardiology can be found at http://tools.acc.org/ASCVD-Risk-Estimator-Plus.

Following is a brief explanation of the three stages of gathering knowledge.

**THE CAUSATIVE PROCESS IS OF INTEREST, NOT THE DATA**
A process, or set of forces, generates data related to an event. It is this process, not the data per se, that interests us.

*Description:* The stage in which we seek to describe the data-generating process in cases for which we have data from that process. Description would answer questions such as: What is the range of prostate volumes for a sample of urology patients? What is the difference in average volume between patients with negative biopsy results and those with positive results?

*Explanation:* The stage in which we seek to infer characteristics of the (overall) data-generating process when we have only part (usually a small part) of the possible data. Inference would answer questions such as: Based on a sample of patients with prostate problems, are the average volumes of patients with positive biopsy results less than those of patients with negative biopsy results, for all men with prostate problems? Such inferences usually take the form of tests of hypotheses.

*Prediction:* The stage in which we seek to make predictions about a characteristic of the data-generating process on the basis of newly taken related observations. Such a prediction would answer questions such as: On the basis of a patient’s negative digital rectal examination, prostate-specific antigen of 9, and prostate volume of 30 mL, what is the probability that he has prostate cancer? Such predictions allow us to make decisions on how to treat our patients to change the chances of an event. For example, should I perform a biopsy on my patient? Predictions usually take the form of a mathematical model of the relationship between the predicted (dependent) variable and the predictor (independent) variables.

### 1.3 SCIENCE UNDERLYING CLINICAL DECISION-MAKING

**The scientific method**

*Science* is a collection of fact and theory resting on information obtained by using a particular method that is therefore called the scientific method. This method is a way of obtaining information constrained by a set of criteria. The method is required to be objective; the characteristics should be made explicit and mean the same to every user of the information. The method should be unbiased, free of personal or corporate agendas; the purpose is to investigate the truth and correctness of states and relationships, not to “prove” them. The true scientific approach allows no preference for
outcome. The method should involve the control of variables; ideally, it should elimi-
nate as far as practicable all sources of influence but one, so that the existence of and
extent of influence of that one source is undeniable. The method should be repeat-
able; other investigators should be able to repeat the experiment and come to the
same conclusion. The method should allow the accumulation of results; only by accu-
mulation does the information evolve from postulate to theory to fact. The scientific
method is the goal of good study design.

Jargon in science
Jargon may be defined as technical terminology or as pretentious language. The public
generally thinks of it as the latter. To the public, carcinoma is jargon for cancer, but to
the professional, technical connotation is required for scientific accuracy. We need to
differentiate between jargon for pomposity and jargon for accuracy, using it only for
the latter and not unnecessarily. The same process occurs in statistics. Some statistical
terms are used loosely and often erroneously by the public, who miss the technical
implications. Examples are randomness, independence, probability, and significance. Users of
statistics should be aware of the technical accuracy of statistical terms and use them
correctly.

Evidence
The accumulating information resulting from medical studies is evidence. Some types
of studies yield more credible evidence than others. Anecdotal evidence, often dis-
missed by users seeking scientific information, is the least credible, yet is still evidence.
The anecdotal information that patients with a particular disease often improve more
quickly than usual when taking a certain herb may give the rate of improvement but
not the rate of failure of the treatment. It may serve as a candle in a dark room.
However, such evidence may suggest that a credible study be done. The quality of the
study improves as we pass through registries, case—control studies, and cohort studies,
to the current gold standard of credibility, the randomized controlled prospective clin-
ical trial (RCT). (See Sections 1.5 and 1.6 for more information on types of studies.)
It is incumbent on the user of evidence to evaluate the credibility of the cumulative
evidence: number of accumulated studies, types of studies, quality of control over
influencing factors, sample sizes, and peer reviews. Evidence may be thought of as the
blocks that are combined to build the scientific edifice of theory and fact. The more
solid blocks should form the cornerstones and some blocks might well be rejected.

Evidence versus proof
The results of a single study are seldom conclusive. We seldom see true absolute proof
in science. As evidence accrues from similar investigations, confidence increases in the
correctness of the answer. The news media like to say, “The jury is still out.” In a more accurate rendition of that analogy, the jurors come in and lodge their judgment one at a time—with no set number of jurors.

**Evidence-based medicine**

Evidence-based medicine (EBM) melds the art and science of medicine. EBM is just the ideal paradigm of health-care practice, with the added requirement that updated credible evidence associated with treatment be sought, found, assessed, and incorporated into practice. It is much the way we all think we practice, but it ensures consideration of the evidence components. It could be looked at somewhat like an airliner cockpit check; even though we usually mentally tick off all the items, formal guides verify that we have not overlooked something.

One rendition of the EBM sequence might be the following: (1) we acquire the evidence: the patient’s medical history, the clinical picture, test results, and relevant published studies. (2) We update, assess, and evaluate the evidence, eliminating evidence that is not credible, weighting that remaining evidence according to its credibility, and prioritizing that remaining according to its relevance to the case at hand. (3) We integrate the evidence of different types and from different sources. (4) We add nonmedical aspects, for example, cost considerations, the likelihood of patient cooperation, and the likelihood of patient follow-up. (5) Finally, we embed the integrated totality of evidence into a decision model.

1.4 **WHY DO WE NEED STATISTICS?**

**Primary objective**

A primary objective of statistics is to make an inference about a population based on a sample from that population.

**Population versus sample**

The term *population* refers to all members of a defined group and the term *sample* to a subset of the population. As an example, patients in a hospital would constitute the entire population for a study of infection control in that hospital. However, for a study of infected patients in the nation’s hospitals, the same group of patients would be but a tiny sample. The same group can be a sample for one question about its characteristics and a population for another question.

**Objective restated**

In the context of inferring a treatment effect, the symbol $\alpha$ is assigned to the chance of concluding that a treatment difference exists when in fact it does not (otherwise known as a type I error in statistical terms). We may restate this common objective of statistics
as follows: based on a sample, we wish to bound the chance of concluding that a treatment difference exists in the population when it truly does not (a false-positive difference) by an agreed upon $\alpha$. For example, of 50 urgent care patients with dyspepsia who are given no treatment, 30 are better within an hour and of 50 given a “GI cocktail” (antacid with viscous lidocaine), 36 are better within an hour. In order to decide if the treatment is effective in the population based on this sample, in that the condition of 20% more treated than untreated patients showed improvement for these 100 patients, we calculate the probability that an improvement of this magnitude (or more) would have been observed by chance if the treatment had no effect. The question for statistics to answer is: Is it likely to work in the overall population of urgent care patients with dyspepsia, or was the result for this sample “luck of the draw”?

What statistics will not do for us

Statistics will not make uncertainty disappear. Statistics will not automatically formulate a scientific hypothesis. Statistics will not give answers without thought and effort. Statistics will not provide a credible conclusion from poor data, that is, to use an old maxim, it will not make a silk purse out of a sow’s ear. It is worth keeping in mind that putting numbers into a formula will yield an answer but the process will not inform the user whether the answer is credible. The onus is on the user researcher to apply credible data in a credible manner to obtain a credible answer.

What statistics will do for us

There is no remedy for uncertainty, but statistics allows you to measure, quantify, and account for uncertainty, and in many cases to reduce uncertainty through study design that is founded on statistical principles. This benefit is one of the most crucial and critically important bases for scientific investigation. In addition, statistics and statistical thinking can assist us to do the following:

- Refine and clarify our exact question.
- Identify the variable and the measure of that variable that will answer that question.
- Verify that the planned sample size is adequate.
- Test our sample to see if it adequately represents the population.
- Answer the question asked, while limiting bounding the risk for error in our decision.

Other benefits of statistics include the following:

- allowing us to follow strands of evidence obscured by myriad causes,
- allowing us to mine unforeseen knowledge from a mountain of data in order to generate new hypotheses,
- providing the credibility for the evidence required in EBM, and
- reducing the frequency of embarrassing mistakes in medical research.
1.5 Concepts in Study Design

Components of a study

A *medical clinical study* is an experiment or gathering of data in a designed fashion in order to answer a specific question about a population of patients. A study design may be involved, approaching the arcane. Breaking it into the components used in constructing a study will simplify it. The basic steps are as follows:

- Specify, clearly and unequivocally, a question to be answered about an explicitly defined population.
- Identify a measurable variable capable of answering the question.
- Obtain observations on this variable from a sample that represents the population.
- Analyze the data with methods that provide a valid answer to the question.
- Generalize this answer to the population, limiting the generalization by the measured probability of being correct.

Control groups and placebos

A frequent mechanism to pinpoint the effect of a treatment and to reduce bias is to provide a *control* group having all the characteristics of the experimental group except the treatment under study. For example, in an animal experiment on the removal of a generated tumor, the control animals would be surgically opened and closed without removing the tumor, so that the surgery itself will not influence the effect of removing the tumor. In the case of a drug efficacy study, a control group may be provided by introducing a *placebo*, a capsule appearing (and possibly feeling) identical to that being given to the experimental group but lacking the study drug.

Variables

A *variable* is just a term for an observation or reading giving information on the study question to be answered. Blood pressure is a variable giving information on hypertension. Blood uric acid level is a variable giving information on gout. The term *variable* may also refer to the symbol denoting this observation or reading.

In study design, it is essential to differentiate between independent and dependent variables. Let us define these terms.

An *independent* variable is a variable that, for the purposes of the study question to be answered, occurs independently of the effects being studied. A *dependent* variable is a variable that depends on, or more exactly is influenced by, the independent variable. In a study on gout, suppose we ask if blood uric acid (level) is a factor in causing pain. We record blood uric acid level as a measurable variable that occurs in the patient. Then we record pain as reported by the patient. We believe blood uric acid level is predictive of pain. In this relationship, the blood uric acid is the independent variable.
and pain is the dependent variable. While not entirely appropriate, much of the statistical literature interchanges the terms independent and dependent variable with response and predictor, respectively.

**Moving from sample to population**

We use descriptive summary statistics from the sample to estimate the characteristics of the population, and we generalize conclusions about the population on the basis of the sample, a process known as statistical inference.

**Representativeness and bias**

To make a dependable generalization about certain characteristics, the sample must represent the population in those characteristics. For example, men tend to weigh more than women because they tend to be bigger. We could be led into making wrong decisions on the basis of weight if we generalized about the weight of all humans from a sample containing only men. We would say that this sample is biased. To avoid sex bias, our sample should contain the same ratio of men to women as does the human population.

**Experimental design can reduce bias**

The crucial step that gives rise to most of the design aspects is encompassed in the phrase “a sample which represents the population.” Sampling bias can arise in many ways, some of which are addressed in Section 1.9. Clear thinking about this step avoids many of the problems. Experimental design characteristics can diminish biases.

**Exercise 1.1**

Choose a medical article from your field. Evaluate it using the guidelines given in Section 1.5.

### 1.6 STUDY TYPES

Different types of studies imply different forms of design and analysis. To evaluate an article, we need to know what sort of study was conducted.

**Registry**

A registry is an accumulation of data from an uncontrolled sample. It is not considered to be a “study.” It may start with data from past files or with newly gathered data. It is useful in planning a formal study to get a rough idea of the nature of the data: typical values to be encountered, the most effective variables to measure, the problems in sampling that may be encountered, and the sample sizes required. It does not,
however, provide definitive answers, because it is subject to many forms of bias. The very fact of needing information about the nature of the data and about sampling problems implies the inability to ensure freedom from unrepresentative sampling and unwanted influences on the question being posed.

**Cohort study**

A cohort study starts by choosing groups that have already been assigned to study categories, such as diseases or treatments, and follows these groups forward in time to assess the outcomes. To try to ensure that the groups arose from the same population and differ only in the study category, their characteristics, both medical and demographic, must be recorded and compared. This type of study is risky, because only the judgment of what characteristics are included guards against the influence of spurious causal factors. Cohort studies are useful in situations in which the proportion in one of the study categories (not in an outcome as in the case—control study) is small, which would require a prohibitively large sample size.

**Case—control study**

A case—control study is a study in which an experimental group of patients is chosen for being characterized by some outcome factors, such as having acquired a disease, and a control group lacking this factor is matched patient for patient. Control is exerted over the selection of cases but not over the acquisition of data within these cases. Sampling bias is reduced by choosing sample cases using factors independent of the variables influencing the effects under study. It still lacks evidence that chance alone selects the patients and therefore lacks assurance that the sample properly represents the population. There still is no control over how the data were acquired and how carefully they were recorded. Often, but not always, a case—control study is based on prior records and therefore is sometimes loosely termed a retrospective study. Case—control studies are useful in situations in which the outcomes being studied either have a very small incidence, which would require a vast sample, or are very long developing, which would require a prohibitively long time to gain a study result.

**Case—control contrasted with cohort studies**

The key determinant is the sequence in which the risk factor (or characteristic) and the disease (or condition) occur. In a cohort study, experimental subjects are selected for the risk factor and are examined (followed) prospectively for the disease outcome; in a case—control study, experimental subjects are selected based upon the disease outcome and are then retrospectively examined for the risk factor that, in theory, should have preceded the outcome.
Randomized controlled trial

The most sound type of study and the gold standard for establishing causal relationships is the randomized controlled trial (RCT), often called a clinical trial. An RCT is a true experiment in which patients are assigned randomly to a study category, such as clinical treatment, and are then followed forward in time (making it a prospective study) and the outcome is assessed. (A fine distinction is that, in occasional situations, the data can have been previously recorded and it is the selection of the existing record that is prospective rather than the selection of the not-yet-measured patient.) An RCT is randomized, meaning that the sample members are allocated to treatment groups by chance alone so that the choice reduces the risk of possibly biasing factors. In a randomized study the probability of influence by unanticipated biases diminishes as the sample size grows larger. An RCT should be masked or blinded when practical, meaning that the humans involved in the study do not know the allocation of the sample members, so they cannot influence measurements. Thus the investigator cannot judge (even subconsciously) a greater improvement in a patient receiving the treatment the investigator prefers. Often both the investigator and the patient are able to influence measurements, in which case both might be masked; such a study is termed double-masked or double-blinded.

Paired and crossover designs

Some studies permit a design in which the patients serve as their own controls, as in a “before-and-after” study or a comparison of two treatments in which the patient receives both in sequence. For example, to test the efficacy of drugs A and B to reduce intraocular pressure, each patient may be given one drug for a period of time and then (after a “washout” period) the other. A crossover design is a type of paired design in which patients are randomized to a given sequential administration of treatment: half the patients are given drug A followed by B and the other half is given drug B followed by A; this helps to minimize bias due to carryover effects in which the effect of the first treatment carries over into the second and contaminates the contrast between A and B. Of course, the allocation of patients to the A-first versus B-first groups must be random.

Exercise 1.2

From the discussion in Section 1.6, what sort of study gave rise to (a) DB2 (see Databases)? (b) DB14? What are the independent and dependent variables in each?

Exercise 1.3

Is the study represented by DB6 an RCT? Why or why not?
1.7 CONVERGENCE WITH SAMPLE SIZE

Another requirement for a dependable generalization about certain characteristics is that the sample must become more like the population in the relevant characteristics as the sample grows larger. Related to this, the estimator we use to estimate a population parameter should grow closer, or converge to the population parameter as the sample size grows larger. For example, we may use the sample average (the estimator) to estimate the population mean (the parameter). We would like the sample average to converge upon the population mean as the sample size grows. Such a process is illustrated in Table 1.1. The deviation from the mean can be seen to grow smaller, albeit somewhat irregularly, as the size of the sample grows closer to the population size. In a formal treatment of this, we would say that the sample average is a consistent estimator of the population mean if this holds. These issues are examined in Section 4.3.

1.8 SAMPLING SCHEMES IN OBSERVATIONAL STUDIES

Purpose of sampling schemes

The major reason for different sampling procedures is to either increase representativeness or statistical precision for estimating population parameters. Methods of sampling are myriad, but most relate to unusual designs, such as complicated mixtures of variables or designs with missing data. Four of the most basic methods, used with rather ordinary designs, are sketched here.

Simple random sampling

If a sample is drawn from the entire population so that any member of the population is as likely to be drawn as any other, that is, drawn at random, the sampling scheme is termed simple random sampling.

Table 1.1 Mean prostate-specific antigen (PSA) and its difference from the group mean as sample size grows.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Mean PSA level</th>
<th>Deviation from group mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>6.8</td>
<td>−2.0</td>
</tr>
<tr>
<td>50</td>
<td>10.7</td>
<td>1.9</td>
</tr>
<tr>
<td>100</td>
<td>11.7</td>
<td>2.9</td>
</tr>
<tr>
<td>150</td>
<td>10.0</td>
<td>1.2</td>
</tr>
<tr>
<td>200</td>
<td>9.5</td>
<td>0.7</td>
</tr>
<tr>
<td>250</td>
<td>9.4</td>
<td>0.6</td>
</tr>
<tr>
<td>260</td>
<td>9.2</td>
<td>0.4</td>
</tr>
<tr>
<td>280</td>
<td>9.0</td>
<td>0.2</td>
</tr>
<tr>
<td>290</td>
<td>8.9</td>
<td>0.1</td>
</tr>
<tr>
<td>Entire group = 301</td>
<td>8.8</td>
<td>−</td>
</tr>
</tbody>
</table>
Systematic sampling

Sometimes we are not confident that the members sampled will be drawn with truly equal chance. We need to sophisticate our sampling scheme to reduce the risk for bias. In some cases, we may draw a sample of size \( n \) by dividing the population into \( k \) equal portions and drawing \( n/k \) members equally likely from each division. For example, suppose we want 50 measurements of the heart’s electrical conductivity amplitude over a 10-second period, where recordings are available each millisecond. We could divide the 10,000 ms into 50 equal segments of 200 ms each and sample one member equilikely from each segment. Another example might be a comparison of treatments on pig skin healing. The physiologic properties of the skin vary by location on the pig’s flank. An equal number of samples of each treatment are taken from each location, but the assignments are randomized within this constraint. These schemes are named systematic sampling.

CAUTION

The term systematic sampling is sometimes used to refer to sampling from a systematic criterion, such as all patients whose name starts with G, or sampling at equal intervals, as every third patient. In the latter case the position of the patient chosen in each portion is fixed rather than random. For example, the third, sixth, and so on patients would be chosen rather than one equally likely from the first triplet, another equally likely from the second triplet. Such sampling schemes are renowned for bias and should be avoided.

Stratified sampling

In cardiac sampling, suppose a cusp (sharp peak) exists, say 100 ms in duration, occurring with each of 12 heartbeats, and it is essential to obtain samples from the cusp area. The investigator could divide the region into the 1200 ms of cusp and 8800 ms of noncusp, drawing 12% of the members from the first portion and 88% from the second. Division of the population into not-necessarily-equal subpopulations and sampling proportionally and equilikely from each is termed stratified sampling. As another example, consider a sports medicine sample of size 50 from Olympic contenders for which sex, an influential variable, is split 80% male to 20% female athletes. For our sample, we would select randomly 40 male and 10 female members.

Cluster sampling

A compromise with sampling costs, sometimes useful in epidemiology, is cluster sampling. In this case, larger components of the population are chosen equilikely (e.g., a family, a hospital ward) and then every member of each component is sampled. A larger sample to offset the reduced accuracy usually is required.
Nonuniform weighted sampling

In the case of rare prevalence in the population, a simple random sample may not yield enough subjects to estimate population parameters of interest. For example, one may wish to estimate the probability of hypertension in various age and race/ethnic subpopulations in the United States. Drawing a truly random sample would yield fewer African Americans in the sample than Caucasians, thereby leading to less precision in the estimated prevalence of hypertension among African Americans. To achieve greater precision, one may oversample African Americans. While this approach allows for greater precision of the estimated prevalence of hypertension among African Americans, an estimate of the overall prevalence of hypertension in the United States based upon the sample would be incorrect if the prevalence of hypertension among African Americans is different from other race/ethnic groups. To correct for this, statistical analyses would need to be weighted in order to account for the (intentionally) biased sampling scheme that was performed. We expand more on the concept of bias next.

1.9 SAMPLING BIAS

Bias, or lack of representativeness, has been referred to repeatedly in the previous paragraphs, for good reason. The crucial step in most design aspects lies in the phrase “a sample that represents the population.” Sampling bias can arise in many ways. Clear thinking about this step avoids most of the problems.

Although true randomness is a sampling goal, too often it is not achievable. In the spirit of “some information is better than none,” many studies are carried out on convenience samples that include biases of one sort or another. These studies cannot be considered conclusive and must be interpreted in the spirit in which they were sampled.

A pictorial example of bias

Let us, for a moment, take the patients in a certain hospital as the population. Suppose we have 250 inpatients during an influenza epidemic. We measure the white blood cell (WBC) count (in $10^9$) of all inpatients and note which arose from 30 patients in the infectious disease wards.

Fig. 1.1 shows the frequencies of WBCs for the hospital population with frequencies for those patients from the infectious disease wards appearing darker. It is clear that the distribution of readings from the infectious disease wards (i.e., the sample) is biased and does not represent the distribution for the population (i.e., the entire hospital). If we needed to learn about the characteristics of WBCs in this hospital, we should ensure a representative sample.
Increasing representativeness by random samples

The attempt to ensure representative samples is a study in itself. One important approach is to choose the sample randomly. A *random sample* is a sample of elements in which the selection is due to chance alone, with no influence by any other causal factor. Usually, but not always (e.g., in choosing two control patients per experimental patient), the random sample is chosen such that *any member of the population is as likely to be drawn as any other member*. A sample is not random if we have any advance knowledge at all of what value an element will have. If the effectiveness of two drugs is being compared, the drug allocated to be given to the next arriving patient should be chosen by chance alone, perhaps by the roll of a die or by a flip of a coin. Methods of randomizing are addressed in the next section.

Sources of bias

Let us consider biases that arise from some common design characteristics. The terms used here appear in the most common usage, although nuances occur and different investigators sometimes use the terms slightly differently.

The sources of bias in a study are myriad and no list of possible biases can be complete. Some of the more common sampling biases to be alert to are given in the following list. (Some other biases that are unique to integrative literature studies,
or meta analyses, are addressed in Chapter 24, Meta analyses.) In the end, only experience and clear thought, subjected when possible to the judgment of colleagues, can provide adequate freedom from bias.

1. **Bias resulting from method of selection.** Included would be, for example, patients referred from primary health-care sources, advertisements for patients (biased by patient awareness or interest), patients who gravitate to care facilities that have certain reputations, and assignment to clinical procedures according to therapy risks.

2. **Bias resulting from membership in certain groups.** Included would be, for example, patients in a certain geographical region, in certain cultural groups, in certain economic groups, in certain job-category groups, and in certain age groups.

3. **Bias resulting from missing data.** Included would be patients whose data are missing because of, for example, dropping out of the study because they got well, or not responding to a survey because they were too ill, too busy, or illiterate.

4. **State-of-health bias** (Berkson’s bias). Included would be patients selected from a biased pool, that is, people with atypical health.

5. **Prevalence-incidence bias** (Neyman’s bias). Included would be patients selected from a short subperiod for having a disease showing a pattern of occurrence irregular in time.

6. **Comorbidity bias.** Included would be patients selected for study who have concurrent diseases affecting their health.

7. **Reporting bias.** Some socially unacceptable diseases are underreported.

**Exercise 1.4**

*Might sex and/or age differences in the independent variables have biased the outcomes for (a) DB2? (b) DB14? What can be done to rule out such bias?*

### 1.10 RANDOMIZING A SAMPLE

#### Haphazard versus random assignment

We define haphazard assignment as selection by some occurrence unrelated to the experimental variables, for example, by the order of presentation of patients, or even by whim. Haphazard assignment may introduce bias. A haphazard assignment could be random, but there is no way to guarantee it. Sampling should be assigned by methods that guarantee randomness. It has been noted that randomization is one way to reduce bias. A more in-depth discussion of randomization strategies in the context of randomized clinical trials can be found in Chapter 22.

#### Generating random numbers

A mechanism that produces a number chosen solely by chance is a random number generator. Randomization is accomplished by using a random number generator to assign patients to groups.
1.11 HOW TO PLAN AND CONDUCT A STUDY

Planning a study is involved, sometimes seeming to approach the arcane, but need not be daunting if well organized. Total time and effort will be reduced to a minimum by spending time in organization at the beginning. An unplanned effort leads to stomach-churning uncertainty, false starts, acquisition of useless data, unrecoverable relevant data, and a sequence of text drafts destined for the wastebasket. Where does one start?

**STEPS THAT WILL AID IN PLANNING A STUDY**

1. *Start with objectives.* (Do not start by writing the abstract.) Specify, clearly and unequivocally, a question to be answered about an explicitly defined population.

2. *Develop the background and relevance.* Become familiar with related efforts made by others. Be clear about why this study will contribute to medical knowledge.

3. *Plan your materials.* From where will you obtain your equipment? Will your equipment access mesh with your patient availability? Dry run your procedures to eliminate unforeseen problems.

4. *Plan your methods and data.* Identify at least one measurable variable capable of answering your question. Define the specific data that will satisfy your objectives and verify that your methods will provide these data. Develop clearly specified null and alternate hypotheses.

5. *Plan data recording.* Develop a raw data entry sheet and a spreadsheet to transfer the raw data to that will facilitate analysis by computer software.

6. *Define the subject population* and verify that your sampling procedures will sample representatively.

7. *Ensure that your sample size will satisfy your objectives* (see Chapter 21: Sample size estimation).

8. *Anticipate what statistical analysis will yield results* that will satisfy your objectives. Dry run your analysis with fabricated data that will be similar to your eventual real data. Verify that the analysis results answer your research question.

9. *Plan analyses to investigate sampling bias.* You hope to show a future reader that your sampling was not biased.

10. *Plan the bridge from results to conclusions.* In the eventual article, this is usually termed the discussion, which also explains unusual occurrences in the scientific process.

11. *Anticipate the form* in which your conclusions will be expressed (but, of course, not what will be concluded). Verify that your answer can be generalized to the population, limiting the generalization by the measured probabilities of error.

12. *Now you can draft an abstract.* The abstract should summarize all the foregoing in half a page to a page. After drafting this terse summary, review steps (1)–(9) and revise as required.
Professional guides

A number of professional guides to study planning may be found on the Internet. As of this writing, an investigator can find a checklist for randomized controlled trials at CONSORT-statement.org, for systematic reviews at PRISMA-statement.org, for meta analysis at MOOSE within the CONSORT site, and for diagnostic accuracy studies at STARD-statement.org. Further, the International Conference on Harmonization (ICH) constitutes regulatory bodies from the United States, Europe, and Japan as well pharmaceutical professionals to provide guidance on the conduct of the drug-approval process. Most germane to this text are sections E9 and E10 of the ICH guidelines that provide statistical guidance on clinical trial design and the choice of control groups in clinical studies, respectively.

1.12 MECHANISMS TO IMPROVE YOUR STUDY PLAN

Tricks of the trade

There exist devices that might be thought of as “tricks of the trade.” The preceding section gives steps to draft a study. However, reviewers of study drafts typically see a majority of studies not yet thoroughly planned. Three devices to improve study plans, used by many investigators but seldom if ever written down, will render plans more “solid.” Using them early in the writing game often prevents a great deal of grief.

WORK BACKWARD THROUGH THE LOGICAL PROCESS

After verification that the questions to be asked of the study are written clearly and unequivocally, go to step 9 of the list in Section 1.11 and work backward. (1) What conclusions are needed to answer these questions? (A conclusion is construed as answering a question such as “Is the treatment efficacious?” rather than providing the specific conclusion the investigator desires.) (2) What data results will I need, and how many data will I need to reach these conclusions? In many cases, it is even useful to fully specify the tables and figures needed to display the relevant results that will be reported at study completion prior to any data analysis. (3) What statistical methods will I need to obtain these results? (4) What is the nature and format of the data I need to apply these statistical methods? (5) What is the design and conduct of the study I need to obtain these data? (6) Finally, what is the ambiance in the literature that leads to the need for this study in general and this design in particular? When the investigator has answered these questions satisfactorily, the study plan will flow neatly and logically from the beginning.
ANALYZE DUMMY DATA
If you had your data at this stage, you could analyze them to see if you had chosen the appropriate data and the right recording format needed for that analysis. However, although you do not have the data per se, you have a good idea what they will look like. You have seen numbers of that sort in the literature, in pilot studies, or in your clinical experience. Use a little imagination and make up representative numbers of the sort you will encounter in your study. Then subject them to your planned analysis. You do not need more than a few; you can test out your planned analysis with 20 patients rather than 200. You do not need to have the data in the relative magnitudes you would like to see; a correlation coefficient even very different from the one that will appear in your later study will still tell you whether or not the data form can be used to calculate a legitimate correlation coefficient. This is not lost time, because not only will you learn how to perform any analyses with which you are not intimately familiar, but also when you obtain your actual data, your analysis will be much faster and more efficient. This step is worth its time spent to avoid that sinking feeling experienced when you realize that your study will not answer the question because the hematocrits from 200 patients were recorded as low, normal, or high rather than as percentages.

PLAY THE ROLE OF DEVIL’S ADVOCATE
A device that is useful at the planning stage, but perhaps more so when the finished study is drafted, is to “put on the hat” of a reviewer and criticize your own work. This is not an easy challenge. It requires a complete mental reset followed by self-disciplined focus and rigid adherence to that mindset. Indeed, it requires the investigator to use a bit of acting talent. Many recognized actors achieve success by momentarily believing that they are the characters they are playing; in this case the character is a demanding, “I’ve-seen-it-all,” somewhat cynical reviewer. A number of little mechanisms can help. Note everything that can be construed as negative, however trivial. Examine each paragraph to see if you can find some fault in it. The object is to try really hard to discover something subject to criticism. When you have finished, and only then, go back and consider which of the criticisms are valid and rewrite to preempt a reviewer’s criticism. Remember that you would rather be criticized by a friend than by an enemy, and, if you carry off this acting job properly, you are being your own best friend. The most difficult problem to recognize and repair in a study draft is lack of clarity. As is often true of computer manual writers, if you know enough to explain, you know too much to explain clearly. The author once had a colleague who advised, “Explain it like you would to your mother.” Find a patient person who knows nothing about the subject and explain the study, paragraph by paragraph. This technique often uncovers arcane or confusing passages and suggests wordings that can clarify such passages.
1.13 READING MEDICAL ARTICLES

Two primary goals

The two primary goals in reading the medical literature are keeping up with new developments and searching for specific information to answer a clinical question. The mechanisms to satisfy these goals are not very different. Keeping up is accruing general information on a specialty or subspecialty, whereas searching is accruing information about a particular question; the distinction is merely one of focus.

Ways to improve efficiency in reading medical articles

There are several ways to improve efficiency in reading medical articles:

1. Allow enough time to think about the article. A fast scan will not ferret out crucial subtleties. It is what a charlatan author would wish the reader to do, but it disappoints the true scientist author.

2. From the title and beginning lines of the abstract, identify the central question about the subject matter being asked by the author. Read with this in mind, searching for the answer; this will focus and motivate your reading.

3. Ask yourself, if I were to do a study to answer the author’s question, how would I do it? Comparing your plan with the author’s will improve your experiment planning if the author’s plan is better than yours, or it will show up weaknesses in the article if it is worse than yours.

4. A step process in answering a study question was posed in Section 1.11. Verify that the author has taken these steps.

5. Read the article repeatedly. Each time, new subtleties will be discovered and new understanding reached. Many times we read an article that appears solid on a first perusal only to discover feet of clay by the third reading.

6. When seeming flaws are discovered, ask yourself: Could I do it better? Many times we read an article that appears to be flawed on a first perusal only to find on study and reflection that it is done the best way possible under difficult conditions.

The reader may rightly protest that there is not sufficient time for these steps for each of the myriad articles appearing periodically in that reader’s field. A fast perusal of articles of minor importance is unavoidable. Our advice is to weigh the importance of articles appearing in a journal and select those, if any, for solid study that may augment basic knowledge in the field or change the way medicine is practiced. There will not be many.

1.14 WHERE ARTICLES MAY FALL SHORT

Apart from bias, there are several statistical areas in which journal articles may fall short. Some of those giving the most frequent problems are addressed here.
Confusing statistical versus clinical significance

Statistical significance implies that an event is unlikely to have occurred by chance; clinical significance implies that the event is useful in health care. These are different and must be distinguished. A new type of thermometer may measure body temperature so accurately and precisely that a difference of 1/100 degree is detectable and statistically significant; but it is certainly not clinically important. In contrast a new treatment that increases recovery rate from 60% to 70% may be very significant clinically but associated with a level of variability that prevents statistical significance from appearing. When “significance” is used, its meaning should be designated explicitly if not totally obvious by context. Indeed, we might better use the designation clinically important or clinically relevant.

Violating assumptions underlying statistical methods

The making of assumptions, more often implicit than explicit, will be discussed in Sections 4.6 and 7.1. If data are in the correct format, a numerical solution to an equation will always emerge, leading to an apparent statistical answer. The issue is whether or not the answer can be believed. If the assumptions are violated, the answer is spurious, but there is no label on it to say so. It is important for the reader to note whether or not the author has verified crucial assumptions.

Generalizing from poorly behaved data

How would we interpret the mean human height? By using a bimodal distribution? If we made decisions based on the average of such a bimodal distribution, we would judge a typical man to be abnormally tall and a typical woman to be abnormally short. Authors who use descriptors (e.g., mean, standard deviation) from a distribution of sample values of one shape to generalize to a theoretical distribution of another shape mislead the reader. The most frequently seen error is using the mean and standard deviation of an asymmetric (skewed) distribution to generate confidence intervals that assume a symmetric (normal) distribution. We see a bar chart of means with little standard error whiskers extending above, implying that they would extend below symmetrically; but do they? For example, the preoperative plasma silicon level in DB5 is skewed (stretched out) to the right; it has a mean of about 0.23 with standard deviation 0.10. Suppose for clinical use, we want to know above what level the upper quarter of the patients is? From the data, the 75th percentile is 0.27, but generalization from a normal (bell-shaped) distribution with mean 0.23 and standard deviation 0.10 claims the 75th percentile to be 0.35. The risk-of-danger level starts much higher using the normal assumption than is shown by the data. The author should specify for the reader the shape of the sample distributions that were used for generalization.
Failure to define data formats, symbols, or statistical terms

An author labels figures in a table as means but adds a plus/minus (±) symbol after the values, for example, 5.7 ± 1.2. Are the 1.2 units a standard deviation, a standard error of the mean, a confidence interval, or something else? Beware of the author who does not define formats and symbols. A further problem is the use of statistical terms. A paper might describe data samples as “compared for significance using the general linear model procedure.” The general linear model is quite general and includes a host of specific tests. Examination of the data and results would lead a statistician to conclude that the authors compared pairs of group means using an F test, which is the square of the t test for two means. The text could have said, “Pairs of means were compared using the t test.” However, this may be a wrong conclusion. Are we at fault for failing to understand the jargon used? No. It is incumbent on authors to make their methodology clear to the reader. Beware of authors who try to convince rather than inform.

Using multiple related tests that inflate the probability of false results

The means of a treatment group are compared with those of a placebo group for systolic blood pressure, diastolic blood pressure, heart rate, and WBC count using four t tests, each bounding the probability of a false-positive result (i.e., concluding that the treatment affects the mean of the response when it, in fact, does not) at 5% (the α level). In this case the risk of a false positive accumulates to approximately 20% if the tests were independent of one another. (Four such tests yield a risk = 1 − (1 − α)^4 = 0.185.) If we performed 20 such tests, we would be almost sure that at least one positive result is spurious. The solution to this problem is to use a multivariate test that tests the means of several variables simultaneously. Can t or other tests ever be used multiple times? Of course. When independent variables do not influence the same dependent variable being tested, they may be tested separately. (This issue is further discussed in Chapter 7: Hypothesis testing: concept and practice and Chapter 11: Tests of location with continuous outcomes.)

Failure to clearly distinguish a priori analyses from exploratory data-driven analyses

As we have discussed, a good scientific study begins with a well-defined objective that is generally framed in terms of one or more hypotheses. After addressing these hypotheses, it is often useful to perform exploratory analyses to determine if something was “missed” in the data or in order to generate new hypotheses to be validated in future studies. A common pitfall of many studies is that statistical inferences (p-values, confidence intervals, etc.) are presented for these exploratory analyses as if
they were \textit{a priori} stated. Given their data-driven nature, however, these inferences do not control error rates and hence must be interpreted with caution. As a reader, it is critical to distinguish analyses that were \textit{a priori} planned from those found by searching through the data as one will have far more confidence in the inference provided for the former than the later. This point is so critical that some journals, particularly those in the field of psychology, have begun requiring authors to preregister their analysis plan prior to actually conducting their study and reporting this plan with their final results.

\textbf{Choosing inappropriate statistical tests}

Too many investigators choose the wrong method to test a hypothesis, usually through ignorance, occasionally by machination. This alters the risk associated with a conclusion (e.g., the distribution of the $p$-value when no association exists) and may even change the conclusion of the investigation. Although some involved studies require the judgment of an experienced professional statistician for design and analysis, many studies may be assessed for appropriate choice of test using Table 2.1 and its associated explanation.

\section*{1.15 WRITING MEDICAL ARTICLES}

\textbf{Format}

As a first draft, merely write down in a straightforward manner why the study needed to be done, what you did, what you found, and what these findings mean to medicine. Assuming you planned your study well according to earlier sections of this chapter, a good start on background, objectives, methods and materials, and statistical plan may follow the study plan by changing the tense from future to past. After drafting results, discussion, and conclusions, reformat your text in the pattern used by the journal to which you plan to submit. Use other articles in that journal as a guide to mode of expression, amount of explanation, and general mood. Journals provide “Instructions to Authors” that will give detailed guidance. Once you have a solid draft, rewrite with attention to detail.

\textbf{Level of composition}

Editors are very jealous of the space in their journals and readers have more to read than they have time or patience for. Your article has a higher chance of being accepted, and then being read, if it is short and succinct. Provide only enough detail to make your study credible.

Remember that few readers are as interested in your topic as you are and no reader will have your insight accrued from months or years of daily involvement with your
data. Therefore do not take for granted their understanding. Write and rewrite and rerewrite until the rendition is simple and clear. As advised in Section 1.12, write it so your mother could understand it.

**Selecting content**

Of 1000 people who read your title, how many will read your abstract? The title is crucial and must embody an unspoken invitation to be read. It must “grab” the reader. Spend a great deal of time and thought on the title.

Given a good title, you will be lucky if 100 of the 1000 scan the abstract and most will go no further. It is not foolish to spend nearly as much time on composing the abstract as in composing the remainder of the article.

Given a good abstract, perhaps 40 will thumb through looking at tables and charts, 20 will read the article, and 4 will study and evaluate the article carefully. Therefore your charts should be able to stand alone with complete captions so that a reader can get much of your message from them without working through the text.

Most readers will “turn off” and go to another article if yours is overloaded with many charts, many tables, or copious detail. Unless you are writing a seminal paper that blankets an entire field of study (very rare), focus on the one main lesson you want to convey to your reader and stick to that, avoiding excursions into interesting side issues. Do not fall into the trap of trying to pass on every detail you observed.

Continually ask yourself what you want your reader to take away from your article. The best paper will present that idea in one dramatic, simple graph or table and leave the presentation at that. Next week, your reader will remember one dramatic graph; your reader will not remember anything about it after wading through 10 graphs.

### 1.16 STATISTICAL ETHICS IN MEDICAL STUDIES

**Ethics in the conduct of medical studies is a broad topic**

Ethics in medical research covers inter alia the contrast between research and practice; informed consent issues; the organization, responsibility, and conduct of institutional review boards and animal care committees; adherence to the World Medical Association’s Helsinki Declaration and its revisions; issues in the trade-off between care for populations (epidemic containment, immunization, placebo groups in clinical trials) and care for individual patients; and the interaction between the practices of statistics and medicine. The latter issue includes inter alia integrity of statistical methods chosen, documentation and availability of statistical methods used, qualifications of a data analyst, patient data privacy, rules for halting a study when efficacy is shown prematurely,
random patient allocation to treatment groups in the presence of uncertainty about equal treatment preference, and the issue that poorly done statistics can lead to poorer patient care. The resolution of many of these statistical issues is obvious or has been well treated. The latter issue is the principal focus of this section.

**Patient protection requirements**

Although patient protection in medical research is not directly a statistical issue, the statistical designer in medical research must remain aware of its implications. It often affects statistical design, for example, forbidding certain control arms, constraining sample sizes in some studies, and requiring early termination when shown to benefit patients. For example, it does not permit denying treatment to cancer patients even though knowing the progress of untreated cancer would provide a baseline against which to evaluate cancer treatments. Different nations have different standards, but most developed nations have similar standards. In the United States as of 2004, the basic documents specifying standards are the following: the Nuremberg Code, 1949 (developed as a result of the Nuremberg war crimes trials at the end of World War II); the Belmont Report, 1979 (on protection of human subjects); United States Code 10 USC 980 Humans as Experimental Subjects, 1999; Food and Drug Administration (FDA) 21 CFR (Code of Federal Regulations) 312 Investigational New Drugs; FDA 21 CFR 812 Investigational Device Exemptions; Department of Defense 32 CFR 219 Protection of Human Subjects; Department of Health and Human Services 45 CFR 46 Protection of Human Subjects, 1999; and finally the 2003 implementation of 45 CFR 160/164 Health Insurance Portability and Accountability Act of 1996 (HIPAA) that establishes patient-protected health information. New standards documents appear sporadically; therefore the ethical investigator must stay updated and informed regarding standards.

**Patient identifiers in data sheets**

Health Insurance Portability and Accountability Act of 1996 (HIPAA) rules require safeguards to patient identity. In setting up a spreadsheet to record data for research, plan to identify patients with a code, not with names or personal numbers that an unscrupulous person could use to obtain medical information about the patient. If there are no further data to be obtained, all patient identifiers should be shredded or deleted electronically. If future access to patient identity may be needed, keep a master code separate from the data in a locked file or safe.

**Statistical control parameters and sample sizes are at issue**

If the statistical design and analysis of a study leads to erroneous conclusions, patients’ health care resulting from that study will be degraded. The choice of the appropriate
statistical method is a topic throughout this book; the smaller topic that this section addresses is that of the choice of and interaction among error risks (α and β), test sidedness, and sample size.

Examples of ethical considerations in specifying sample size test parameters

This section pursues two examples of ethical considerations. (1) In a trial investigating the effect of a new muscarinic agent on return of salivary function after head and neck irradiation, how many patients are required for the trial when randomized to control versus pilocarpine? (2) In an investigation of a recursive-partitioning model in early-stage breast cancer to determine the need for axillary sampling, based on historical control data, how many patients are required for statistical validity?

RELATIONSHIP AMONG THE STATISTICAL PARAMETERS

Required sample size $n$ may often be estimated by methods of Chapter 21, Sample size estimation. An estimate of the standard deviation of the response variable being used is obtained and α, the bound we wish to place on the risk of falsely concluding an association when no association exists, and β, the risk of failing to conclude an association exists under a given effect size, δ. The remaining choice is for δ itself, the size of the difference between treatments that will answer the clinical question being posed (often based on the investigator’s clinical experience). Medical journal reviewers generally seem to expect choices of $\alpha = 0.05$ and $\beta = 0.20$ (power $= 0.80$), although power $= 0.90$ is being seen increasingly. The error risk associated with concluding a treatment effect exists when it does not will be $\alpha$ or $\alpha/2$, depending on whether a one- or two-sided test is being used. In most cases a computer operation using these parameters as inputs will provide $n$.

IMPLICATIONS OF $\alpha$ AND $\beta$

$\alpha$ is the upperbound we wish to place on risk of inferring a difference between experimental groups when there is in fact no such difference and $\beta$ is the risk of inferring no difference between groups when there is in fact a difference of a given magnitude, say δ. Choosing $\alpha = 0.05$ and $\beta = 0.20$ (power $= 0.80$) implies setting the rate of false negatives at four times the rate of false positives.

EFFECT ON PATIENTS FROM THE XEROSTOMIA STUDY

A false-positive outcome from the study implies inferring the new treatment to be better when it is not; using it unnecessarily exposes patients to the risk for possible serious negative side effects. A false-negative outcome implies inferring the new treatment to be no better when, in fact, it is, thereby failing to palliate xerostomia. The false-positive outcome is worse for the patient than is the false-negative outcome. The $\alpha/\beta$ ratio choice of 0.05/0.20 (or 1/4) is justifiable.
EFFECT ON PATIENTS FROM THE BREAST CANCER STUDY
In the other trial, however, the false-positive outcome implies axillary metastases that are not there and unnecessarily doing an axillary sampling. The false-negative outcome implies the absence of cancer when it is present and subsequent failure to offer appropriate therapy. In this case the false-positive outcome represents less loss to the patient than does a false-negative outcome. To set the probability of missing a cancer at four times the probability of an unnecessary sampling may not serve patients well. The investigator should take a higher $\alpha$ and lower $\beta$. Would this require an untenable increase in sample size required? If the error rates were reversed to $\alpha/\beta = 0.20/0.05$ (four unnecessary samplings expected per cancer missed), the sample size would increase no more than about 10%. However, if $\alpha$ should be mandated at 0.05 as is usual in medical research (based more on tradition than good reason), the 4/1 ratio would require $\beta$ to be 0.0125 and the required sample size would increase by about 1.25-fold.

THE CHOICE OF $\alpha$ AND $\beta$ SHOULD INVOLVE CLINICAL IMPLICATIONS
The choice of $\alpha$ and $\beta$, often a hard choice, follows as a matter of judgment. Thus this decision is partly clinical, partly statistical. Furthermore, although the sample size selected will not directly affect patients not participating in the study, the selection of $\alpha$ and $\beta$ must be carried through into the statistical testing after acquiring the data, and test results may affect patients in greater generality.

EFFECT OF TEST SIEDNESS ON THE PATIENT
After head-and-neck irradiation, can the muscarinic agent inhibit or enhance the return of salivary function, or can it solely enhance it? The choice of a one-sided versus two-sided test should be made before gathering data. The statistical reason is that an investigator must be a stern self-disciplinarian to choose a two-sided test once the data show on which side of a hypothesized mean the sample mean lies. However, there is an ethical reason as well. When a two-sided test is appropriate, a one-sided test doubles the error rate assigned to the chosen tail, which gives rise to two results. The first result is a statistical result of benefit to the investigator: a smaller $\delta$ is required to obtain significance in a given sample. The second result is a clinical outcome to the detriment of the patient: too large a number of healthy patients will be treated as ill, and too small a number of ill patients will not be treated. Choosing a two-sided test when a one-sided test is appropriate creates the same classes of mistake but with opposite results.

CHOOSING SIEDNESS
Often the choice of sidedness is obvious: if we are subjecting the patient to treatments A and B and have no idea which one will be better, two-sidedness is appropriate. If, however, we expect the result associated with one side to be more likely (especially if we prefer that result to be chosen), sidedness should be thoughtfully selected. In this
case, sidedness should be chosen when the study is initially conceived and should be chosen in answer to the following questions: can a result on the nonexpected side possibly occur physically? If not, select a one-sided test. If so, could a result on the nonexpected (perhaps, nonpreferred) side affect the patient? If not, select a one-sided test; if so, select a two-sided test. Note that the choice here, although affecting $\alpha$ and $\beta$, is made on other grounds, which is why $\alpha$ and $\beta$ selection and sidedness selection are discussed separately.

**SELECTION OF THE CLINICAL DIFFERENCE $\delta$**

Another potential problem arises when $\delta$ is selected. When a study is being developed, a larger $\delta$ will require a smaller $n$. The bigger a difference between two group averages is, the easier it is to detect it. Therefore, for protocols that are expected to accrue slowly, or for which a low $n$ is expected, the temptation exists to maximize $\delta$ to allow for a small enrollment with subsequent early closure. This is clearly statistical tampering, but it obscures an ethical question: if the proposed new therapy is really so much better than the current one, how may the researcher in good faith not offer the patient the superior course? This manipulation of $\delta$ poses an ethical dilemma, the answer to which is that we be honest.

**EFFECT OF THE CLINICAL DIFFERENCE $\delta$ ON THE PATIENT**

Even despite an honest selection of $\delta$, the issue persists of its influence on patients affected by the study results. The smaller the actual $\delta$, the less sensitive the analysis will be in finding a true treatment effect, even with a larger $n$; the larger the actual $\delta$, the more sensitive the analysis will be. In the crush for funding and in a forest of competing alternative treatments, it is easier to try many small trials than an enormous one. Are we then reducing our likelihood of improving the lot of all patients to reduce the risk to the small group we include in our trial? We must assure ourselves in trial design that such a trade-off is based solely on the potential patient benefit involved, without influence of personal benefit to the investigator.

**1.17 CONCLUSION**

A medical study holds the potential, whether explicitly or implicitly, of a deleterious outcome for all the patients treated in the future as a result of the information emerging from that study. Ethical soundness in statistical design currently is rarely discussed in the context of study planning. In clinical studies, statistical ethical soundness must be sought side-by-side with clinical ethical soundness.

**Exercise 1.5**

For the question of DB5, should a one- or two-sided test be used?
APPENDIX TO CHAPTER 1

Glossary of statistical terms used in Chapters 1 and 2

A few statistical terms commonly appearing in medical articles appear incidentally in these chapters. In case, a reader encounters an unfamiliar one, this glossary provides interim definitions pending formal definitions later in this book.

\( \alpha \) the upper bound placed on the risk of a false-positive outcome when planning a study.

\( \beta \) the risk of a false-negative outcome under an assumed effect when planning a study.

\( \delta \) the strength or level of an outcome required to be clinically relevant.

**Interquartile range (IQR)** a measure of how spread out a data set is; distance from first quartile to third quartile; includes middle half of data.

**Mean** the most commonly used average; the sum of values divided by the number of values.

**Median** another common average; the middle point of a data set where half of the data lies to its left and half to its right.

**p-value** given observed data from a study, the *posthoc* probability of observing data as or more indicative of an alternative hypothesis (e.g., a treatment effect) than we did when the null hypothesis (e.g., no treatment effect) is true.

**Parameter** a population-level characteristic of the distribution of response variable, for example, a population mean. We use a sample statistic (e.g., the sample mean) to estimate a parameter (e.g., the population mean).

**Percentile** a value of a sample below which that percent of the sample lies. For example, the 50th percentile is the sample value below which half of the sample values lie (the median).

**Quartile** a value of a sample below which that quarter of the sample lies. For example, the second quartile is the sample value below which half of the sample values lie (the median).

**Standard deviation (SD)** a measure of how spread out a data set is. When the SD about a parameter, for example, the mean, is small, we may be confident that the data estimate is close to its population value.

**Standard error of the mean (SEM)** if repeated samples are taken from the same data source and a mean calculated on each, the SEM is the SD of that set of means.

**Statistical model** a statistical model is a mathematical model that is built on a simplifying set of assumptions that attempts to explain the mechanistic or data-generating process that gave rise to the data one has observed. A model approximates a state or condition based on measurements of influencing variables, whence its name, a model of the state, not a direct measure.

**Test** a procedure to find the likelihood that a hypothesis about an outcome is correct. The type of test is often named for the type of probability distribution of the outcome, giving rise to names like a normal test, a \( t \) test, or an \( F \) test.

**Type I error** concluding a result to be significant when it is not; a false-positive result.

**Type II error** concluding a result to be not significant when it is; a false-negative result.
Planning analysis: how to reach my scientific objective

2.1 WHAT IS IN THIS CHAPTER

Chapter 1, Planning studies: from design to publication, looked at how to plan and conduct a study, including defining the data that will be recorded. What will you do with those data? Some investigators are new to the game and are not used to symbols and rigorous treatment of quantities. For folk at the beginning of the exciting adventure into research, Sections 2.2–2.5 provide some concepts and rules for symbols and quantities. Then the management of data is addressed and finally setting data up for descriptive statistics and statistical tests.

2.2 NOTATION (OR SYMBOLS)

An index of symbols used in this chapter appears as an appendix at the end of Chapter 1, Planning studies: from design to publication. Any reader at ease with formulas and symbols, for example, Σ, may not need to refer to this section. A more complete such index appears toward the back of the book.

Purpose of symbols

Many people have some difficulty with mathematical symbols. They are, after all, a cross between shorthand and a foreign language. They are awkward to use at first, because users must translate before they get a “gut feel” for the significance of the symbols. After some use a μ or an s² takes on an intuitive meaning to the user and the awkwardness fades. Indeed, symbols are intended to avoid awkwardness, not to create it. If we were to write out concepts and their relationships in words, we would soon be so overcome by the verbosity that we would lose track of what we were trying to do.

Categories of symbols

Most symbols arise from one of three categories: names, operators, or relationships.

1. Name symbols, such as x or μ, may be thought of as families. x may denote prostate volume measures all taken together as a family; x is the “family name.” μ may
denote the mean prostate volume for a population of men. If we need to refer to members of the family, we can affix a “first name” to the family name, usually (but not always) in the form of a subscript. \( x_1 \) is the name of the first prostate volume listed, \( x_2 \) the second, and so on. \( \mu_1 \) may denote the mean of the population of American men, \( \mu_2 \) of Japanese men, and so forth. If we think of name symbols in this fashion, for example, \( y_7 \) denoting the seventh member of the \( y \) family, the mystique of symbols reduces. Most symbols are name symbols.

2. **Operator (or command) symbols** represent an act rather than a thing. There are not many operator symbols. Operator symbols, such as \( \div \) or \( \Sigma \), say “do this thing.” \( \div \) says “divide the quantity before me by the quantity after me,” and \( \Sigma \) says “add together all members of the family that follow me.” There is no indicator of which symbols are name and which are operator, in the same way that there are none in grammar that distinguish nouns from verbs. They become clear by context after some use.

3. **Relationship symbols**, such as \( = \) or \( > \), express some relation between two families or two members of a family. We are all familiar with statements such as \( 2 + 2 = 4 \) or \( 6 > 5 \). The characters 2, 4, 5, and 6 are name symbols for members of the family of integers, \( + \) is an operator symbol, and \( = \) and \( > \) are relationship symbols.

**Formulas**

Putting these ideas together, we can see the meaning of the “formula” for a sample mean, in which we add together all \( n \) values of observations named \( x \) in the sample and divide by \( n \): \((\Sigma x)/n\). The operator symbol \( \Sigma \) says “add together what follows.” The name symbol of what follows is \( x \), the members of the sample of prostate volumes. Thus we add together all the members of the sample, that is, all prostate volumes. The parentheses say “we do what is indicated within before taking the next action.” The operator symbol \( / \) says “we divide what comes before \([\Sigma x]\) by what comes after \((n)\).”

**Becoming familiar with symbols**

The foregoing statement is a tortuously long expression of a simple idea. The user having difficulty with symbols and formulas who goes through such a mental process on each formula for *only a few times, however, will soon find that formulas become a natural way to express relationships, much easier than trying to put them into words. The advantage of symbols over words increases as the relationship being expressed becomes more complicated.

**A first formula**

Now, we can express symbolically the definition of a mean \( m \): \( m = (\Sigma x)/n \).
**Indicator symbols**

If we want to indicate a member of a family, but not wish to specify which member at the moment, we can use what might be called an *indicator symbol*. Any symbol can be used, but the most common in statistics are *i*, *j*, and *k*. If *x_1*, *x_2*, …, *x_5* are the members of the *x* family of prostate volumes put in order from smallest to largest, we can say *x_i* < *x_{i+1}* 1, which is shorthand for the relationship that any member of the *x* family (any prostate volume) is smaller than the member (volume) to its right.

**Symbols for ranks**

We need a way to differentiate ranked data from unranked data. In Table DB1.1, *x_1* is 32.3; when ranked, the first *x* is 16.2, the old *x_3*. A prime (′) is a common way to indicate an observation after it has been ranked. Thus *x'_1* would be the smallest, *x'_2* the next to smallest, and *x'_n* the largest. In Table DB1.1 the first three volumes are *x_1* = 32.3, *x_2* = 27.0, and *x_3* = 16.2. If we ranked the 10 values from smallest to largest, we would have *x'_1* = 16.2, *x'_2* = 16.4, and *x'_3* = 27.0.

**Exercise 2.1**

In the equation \( s^2 = \frac{\sum x_i^2 - nm^2}{n} \), what category of symbol is (a) “n”? (b) “m”? (c) “\( \Sigma \)”? (d) “/”? (e) “=”? (f) “i”?

---

**2.3 QUANTIFICATION AND ACCURACY**

**Statistics and quantification**

Knowledge gained from data is usually more informative and more accurate if the data are *quantitative*. Whereas certain quantities, such as the acceleration of an object in a vacuum, may be expressed with certainty and therefore lie within the realm of (deterministic) mathematics, most quantities that we deal with in life, and particularly in medicine, exhibit some uncertainty and lie within the realm of probability. *Statistics* deals with the development of probabilistic knowledge using observed quantities. These quantities may be as simple as counting the number of patients with flu symptoms or measuring the blood pressure of patients with heart disease, but they should be quantities that shed light on what we want to learn. Statistics as a discipline is interested not in the data per se but rather in the process that has generated these data.

**Quantifying data**

Often the information being gathered arises in quantitative form naturally, such as the counts or measures just mentioned. In other cases, the information is given in words or pictures (true/false, “patient is in severe pain,” X-ray findings) and must be
converted to quantities. A great deal of time and work may be saved, indeed sometimes, whole studies salvaged, if the form in which the data are needed is planned in advance and the data are recorded in a form subject to analysis.

Quantities should be clearly defined and should arise from a common measurement base. Counts are straightforward. True—false, presence—absence, or male—female may be converted to 0—1. Verbal descriptions and visual images are more difficult. Cancer stages may be rated as A, B, C, or D and in turn converted to 1, 2, 3, or 4; however, in order to be compared among patients or among raters, they must be based on clearly defined clinical findings. On X-ray film, the diameter of an arterial stenosis may be measured with a ruler. However, because the scale (enlargement) of radiographs and the size of patients may vary, the ruler-measured stenosis is better given in ratio to a nonstenotic diameter of the same artery at a standardized location.

**Accuracy versus precision**

Quantification per se is not enough; the quantities must be sufficiently accurate and precise. **Accuracy** refers to how well the data-gathering intent is satisfied; that is, how close the arrow comes to the bull’s-eye. **Precision** refers to the consistency of measurement; that is, how tightly the arrows cluster together, unrelated to the distance of the cluster from the bull’s-eye.

**How much precision?**

To decide how much precision is enough, identify the smallest unit that provides useful clinical information. (Just to be safe, one unit farther might be recorded and then rounded off for clinical reporting.) Body temperature of a patient recorded in 10-degree units is useless. Tenths-of-a-degree units are most useful. Carrying temperature to hundredths-of-a-degree adds no clinical benefit, because too many unknown variables influence such minute temperature differences.

**Exercise 2.2**

Do the following estimators tend to measure accuracy or precision: mean, standard deviation (SD)?

**2.4 DATA TYPES**

**Types of data**

Quantitative data may be of three major types: continuous (also called interval), rank order (also called ordinal), or categorical (also called nominal).

1. **Continuous** data are positions on a scale. (A common ruler is such a scale.) In continuous data, these positions may be as close to one another as the user wishes to
discern and record. Prostate volumes and PSA level form examples of this type. (See DB1 from the Database collection.) One patient’s volume was measured as 32.3 mL. If we needed more accuracy for some reason and had a sufficiently accurate measuring device, we might have recorded it as 32.34 or 32.3387. Discrete data are a subset of continuous data that are recorded only as distinct values; there is a required distance between adjacent data. Age recorded in integral years is an example of this type. The age could be 62 or 63, but we agree not to record 62.491.

2. **Rank-order** data are indicators of some ordering characteristic of the subject, such as ranking smallest to largest, most likely to survive to least likely, and so on. There are two classes of data for which ranks are needed. First, ranking is the only data form available for judgment cases in which the investigator cannot measure the variable but can judge the patients’ order. An example is triaging the treatment order of patients after a disaster or military battle. Second, ranking forces equal distances between adjacent continuous data where the distances are disparate or carry disparate clinical implications. An example is platelet count in thrombocytopenia; the clinical implication of a difference from 20,000 to 80,000 is far different from that of 220,000–280,000. The continuous data average of counts of 30,000, 90,000, and 400,000 for three patients are 170,000, implying that the group is healthy. Ranking the patients as 1, 2, and 3 removes the misinformation arising from the first minus second difference being 70,000 and the second minus third difference being 310,000. When continuous data are ranked, ties are recorded as the average of the ranks they would have had had they not been tied. For example, let us rank, from smallest to largest, the first eight preoperative serum silicon levels of DB5. The ranks would be as follows: first, 0.13; second, 0.15; third, 0.18; fourth, 0.20; and fifth, 0.24. The sixth and seventh are tied, both being 0.39; the rank of both is recorded as 6.5. Last, the eighth is 0.42. Notably, ranking retains some, but not all, of the information of continuous data.

3. **Categorical** data are indicators of type or category and may be thought of as counts. We often see 1–0 (or 1–2) for male–female, true–false, or healthy–diseased. In such cases, the names of the categories may occur in any sequence and are not orderable; nonorderable categorical data are sometimes called *nominal* data. In Table DB1.1 the 0–1 indicators of negative–positive biopsy results were categorical data; seven of the patients had negative biopsy results (a proportion of 70%) and three (30%) had positive biopsy results. *Ratios* are occasionally named as a fourth type of data. For the purpose of using data in statistical methods, ratios generally behave like continuous data and thus are not separated out in this book.

**Distinguishing between types**

Sometimes, an investigator may choose the data type in which numerical outcomes from a study or experiment are recorded. PSA values are basically
continuous data. However, they may also be ranked: smallest, next smallest, . . . , largest. If PSA is categorized into three groups, <4, 4–10, and >10, the values are still ordered, but we have lost a lot of information. So many ties will arise that analysis methods will be sensitive only to three ranks, one for each category. Although we could analyze them as unranked categories A, B, and C, the methods treating them as ranks first, second, and third are still “stronger” methods. When rendering data into categories, note whether or not the categories fall into a natural ordering. If they do, we can treat them as ranks. For example, categories of ethnic groups do not fall into a natural order, but the pain categories of severe, moderate, small, and absent do.

Note that we can always change data from higher to lower type, that is, continuous to discrete to ranked to categorical, but the reverse is not possible. Thus it is always preferable to record data as high up on this sequence as possible; it can always be dropped lower.

Rounding

Rounding, most often a rather benign convenience, at times can change data type. The continuous observations 1/3 and 2/3, if rounded to two decimal places, would become 0.33 and 0.67, now discrete observations. If rounded to integers, they would become 0 and 1, which might lose too much accuracy or be mistaken for categorical data. We should carry full accuracy in calculations and then round to the accuracy that has clinical relevance. For example, if four readings of intraocular pressure (by an applanation tonometer) were 15.96, 17.32, 22.61, and 19.87, then the mean would be 18.94. In clinical use the portion after the decimal is not used and would only add distracting detail, so the mean would be reported as 19.

Ratings

Ratings form a class of data all their own, in that they may be any type. There is a great deal of confusion between ratings and rankings in medical literature, even in some published textbooks on biostatistics, and the user must be careful to distinguish between them. In contrast to ranks, which are judgments about one patient or event relative to others, ratings are judgments about a patient or event on that patient’s own merits alone, regardless of others, as in the rating of a tumor as 1 of 4 cancer stages. In contrast to ranks, which should be all different except perhaps for occasional ties, ratings quite properly could be all the same. Ratings behave like continuous data if there are many categories, and they may be analyzed as such if the samples are of fair size. Ratings behave like categorical data if there are few categories. Regardless of the number of categories, ratings may be ranked and rank methods used, but there are usually so many ties that they weaken rank methods.
String (or alphabetic) data

This book is restricted to quantitative data. Often data are recorded as words, called strings in computer terminology, rather than as numbers, for example, “male—female” or “side effects—no side effects”. Most often, these verbal data can be converted to numeric data for statistical analysis. Indeed, if statistical analysis is intended, time and effort will be saved by recording them as numeric data from the outset. If the investigator cannot or will not convert the verbal data to numeric data, some qualitative methods exist to analyze them, but the analyses are more complicated and less effective than quantitative analyses.

Exercise 2.3
Identify the data type from the following data sources: (a) cancer stage 1, 2, 3, or 4; (b) employment class of patient as production worker, service worker, and homemaker; and (c) SBP of the patient.

2.5 MULTIVARIABLE CONCEPTS AND TYPES OF ADJUSTMENT VARIABLES

Univariate versus multivariate measures

The ideas of dependent and independent variables were seen in Section 1.5. We measure a variable, say migraine headache pain, give the patient a controlled dose of a nonsteroidal antiinflammatory drug (NSAID), and measure the pain again in an hour. The dose is the independent variable; it is controlled to be the same on all patients in the experiment and is therefore independent of the patients’ response. The change in pain level is the dependent variable; it depends on the treatment. If there is only one variable of each type, the analysis is called univariate. Sometimes, we have more than one independent variable. We might introduce a different treatment in consort with the NSAID, perhaps an ice pack, providing multiple independent variables. Sometimes, we have more than one dependent variable. We might measure duration of pain as well as level of pain, providing multiple dependent variables.

Multiple, multivariate, and multivariable

“Multi” connotes several or many, but technical uses vary. “Multiple” usually just implies “more than one.” Sometimes, it connotes multivariate (when the adjective “multivariate” coupled with the noun it modifies generates too many syllables), as in “multiple regression.” “Multiple regression” or “multivariable analysis” implies one dependent variable depending on more than one independent variable. Methods satisfying that definition are collectively called multivariable analysis. Multiple regression is a form
of multivariable analysis. In simple regression a dependent variable (e.g., heart rate) is predicted from an independent variable (perhaps minutes of stair climbing). In multiple regression the dependent variable (heart rate) is predicted from several independent variables simultaneously (minutes of stair climbing, patient age, number days per week patient exercises). One step further is the case of more than one dependent variable, which is termed multivariate analysis, above the level of this book. If, in DB3, we wish to predict the dependent variables 5- and 10-day serum theophylline levels simultaneously from the independent variables baseline serum level and age, we would use a form of multivariate analysis. The several terms met in this paragraph are arcane enough to be used without proper care in some articles, and so the reader must be the careful one.

**Types of adjustment variables in multiple regression analyses**

A common scientific objective of many studies is to estimate the association between a specific independent variable (often termed the *predictor of interest*) and a dependent variable, with the hope of isolating the causal association between the two. We have previously noted that randomized experiments represent the gold standard design for establishing causation. This is because randomization, at least in large sample sizes, will tend to yield balance between any other variables that might impact the association between the predictor of interest and the dependent variable. However, in observational (nonrandomized) experiments, such control is not available by design. As a classic example, if one considered observational data aimed at assessing the association between alcohol consumption and the risk of lung cancer, a simple unadjusted estimate of the association between these variables would likely yield a positive association. However, this association is *confounded* by cigarette smoking. That is, individuals that smoke are more likely to drink alcohol, and as we know smoking is related to an increase in the risk of lung cancer. Throughout the remainder of the text, we define a *confounder* as a variable that is causally related to the predictor of interest and also associated with the dependent variable of interest. Many of the regression techniques to be discussed in later chapters were developed to allow for the adjustment of confounding variables, which results in artificially forcing comparison groups to have an equal propensity for the confounding factor and thereby eliminating the spurious association that one might otherwise observe if a simple association was estimated.

Another common type of adjustment variable is termed an *effect modifier*. An effect modifier is a variable for which the association between the predictor of interest and the dependent variable changes depending upon the value of this variable. For example, consider an estrogen blocking treatment for breast cancer. Such a treatment would likely have a differential effect on the ability to treat breast cancer in females when compared to males. This is because of the differential hormone levels in males compared to females. In this case, it would be important to estimate and report the
association between the treatment and progression of disease separately for females and males. In later chapters, we will see that in the context of a regression model, effect modifiers can be incorporated into an analysis through the use of an interaction term. As such, effect modifiers are often referred to as interacting variables in the literature.

2.6 HOW TO MANAGE DATA

Make a plan

As with other aspects of experimental design, a plan reduces errors, work, and time spent. Assess what you have for raw data and what you want to finish with, and plan how to get from the former to the latter.

Raw data

Raw data may be in patients’ charts or even scattered among charts, lab reports, and survey sheets. Data are often in words that must be coded into numbers. Data may be in different units. Numbers may be recorded to different accuracies. In short, raw data can be messy. The investigator must account for the discrepancies and coordinate the data to a common format. The raw data must be assessed for quality and data lacking credibility eliminated. The first step is to assess the data in its several forms and plan a recording sheet that will amalgamate and coordinate the data. The data may be assessed and quality assured while being entered into the recording form so that a finished dependable product emerges ready for transfer to a format amenable to analysis.

Data in format for analysis

The goal is a quantified spreadsheet, most often with cases (such as patients) down the columns and variables (such as age, lab test result, and survival) across rows. The entries for all variables to be used in the analysis should be numerical. For example, use 1 for male and 2 for female, not the letters M and F. For most analyses, all data should be lodged in a single spreadsheet, at least to start. If control data, for example, are placed in one spreadsheet and experimental data in another, the investigator will just have to go to the work of merging them and reverifying correctness in order to carry out a statistical test of difference between the groups. Different methods of statistical analysis may require different formats of data for the analysis to run. Unfortunately, the required format is seldom obvious, requiring the user to go to some effort to ferret it out. The format noted above with cases down the side and variables across the top is the safest to start with. When two groups A and B of patients are being compared, place all patients down the side and form an additional column to designate group, 1 for group A, say, and 2 for B.
Data quality

It is essential to verify the correctness of data at entry, at each change of data format (which, of course, implies making the fewest changes possible to avoid unnecessary work and errors), and again before beginning analysis. Verifying data is tedious and boring but essential. A busy, sometimes tired investigator might read a raw datum as 29 but type 92 into the spreadsheet. Failing to discover and correct such little errors could well weaken or even reverse the conclusion of the study, in some ways a worse outcome than losing the data, time, and effort altogether. All too often, an investigator finds a datum to be erroneous near the end of a project and must redo the entire analysis, interpretation, and reporting text.

Missing data

If a small number of observations are missing, they may be estimated one by one using one of several technically defined methods, denoted imputation. Such a case might occur if an animal dies or a patient is lost to follow up. Some imputation methods are: estimate the mean from readings of other patients on the same variable, use the last reading for that patient if the readings occur in a time series, or estimate the missing datum from a multiple prediction scheme based on a number of variables. Statistical software exists for imputation. Lacking software, the first approach (usually not the best) might be used, substituting the average readings of the other patients on that variable.

Keeping records

Absolutely essential is keeping a complete record of original data plus changes (altering, adding, deleting) in data, along with the reason for change. This process may seem obvious, but all too often steps are omitted in the pressured environment that often accompanies research. Make a disciplined habit of keeping complete records to avoid that sinking feeling leading to desperate panic that arises when an investigator cannot justify data in composing an article or cannot answer a reviewer’s criticism.

Software for data management

A number of data management software packages exist, but Microsoft Excel is by far the most commonly used in medical research. In addition, all statistical analysis software packages have their own spreadsheet capabilities. Data management packages allow ease of data manipulation, such as repositioning data subsets or filling in columns of repetitive numbers, but are limited in their statistical analysis capability. Statistical packages have less capability and are more awkward in data manipulation but allow a nearly complete range of statistical analysis. Investigators having complicated data but very simple analysis may stick with a data management package for the entire project.
Those that have simple data but involved analyses often go directly to a statistical package. The majority prepare their data with data management software and transfer to statistical software for analysis.

**Software for statistical analysis**

There are many statistical software packages. Only a sample of packages will be named here. Booths at a recent Joint Statistical Meetings providing packages for general statistical analysis were, in alphabetical order, JMP, Minitab, NCSS, R, SAS, SPSS, Stata, Statistica, and Systat. Several other good packages exist. All of these packages will perform the more common tasks required by the clinical investigator.

The technical differences lie mostly in access and display formats and in the choice of less frequently used methods. Other differences lie in cost and technical support availability. The packages vary considerably in cost, ranging from a free (R) to several thousand US dollars. Most come in a single comprehensive package, but some, notably SAS and SPSS, are sold in separate subpackages, each containing capabilities for a particular class of methods. Most software packages are sold with perpetual licenses, in which a particular version remains usable indefinitely. The maker profits from providing upgrades that many users buy. However, some makers, such as SPSS, SAS, S-Plus, and JMP, charge yearly fees for the package to remain active. Also, different makers provide different levels of access to customer support, varying from a quick phone call when “stumped” to difficult and remote access. Some are free, some charge a yearly fee, and some charge by the question.

In addition to commercially available statistical software, there are many options available on the web called “freeware.” Much used is R, but one has to learn how to use it (program it) before it becomes practical. Often an individual test or estimation technique may be found with a search engine, but an investigator should verify that the technique provides a correct computation before relying on it.

**Choosing software**

The choice of a package is important. Some time and effort is required to become facile with it, and a user does not want to spend the time and effort to change packages after developing some mastery. In a sense, the user is “stuck with” the package first bought. However, it is almost impossible to know how comfortable a user will become with a package from a demonstration; the choice usually comes from other indicators. Our suggestion to the clinical investigator who is not using the software daily is to choose a single component, perpetual package with free and easily accessible user support. The investigator starting to use a new package will need a lot of support. This need tails off to occasional as facility with the software increases.
Specialty software

In addition to general analysis software, specialty analysis packages exist to handle missing data imputation, equivalence testing, sample size estimation, simulation, data mining, meta analysis, etc.

2.7 DEFINING THE SCIENTIFIC GOAL: DESCRIPTION, ASSOCIATION TESTING, PREDICTION

Throughout this text, commonly used statistical methods are presented, and the appropriateness of such methods is introduced relative to the types of data one seeks to analyze (e.g., binary vs continuous outcomes and independent vs correlated responses). While it is critical to use appropriate statistical methods for the type of data being considered, it is just as critical to choose an analysis strategy that is specific to the scientific goal one seeks to address. While not exhaustive, scientific goals in medical statistics can generally be classified into three broad areas: description or hypothesis generation, association estimation and testing, and prediction of a future observation.

Descriptive analyses are just that, they seek to describe the relationships between multiple variables using a set of observed data. The function of a descriptive analysis is critical to the scientific approach as we often generate hypotheses by first observing the state of nature then conjecturing mechanisms that define the relationships we observe. A data-driven descriptive analysis is not a confirmatory analysis in any way but seeks to identify potential associations that are to be replicated and validated in new experiments or studies. In these future studies, one’s goal is to estimate a specific association of interest and draw inference on it. The association of interest is most commonly a comparison of some statistical summary measure across subpopulations. For example, we might postulate that individuals with low levels of the protein amyloid beta in their cerebrospinal fluid are at higher risk for Alzheimer’s disease when compared to individuals with high levels of amyloid beta. Thus the association of interest is a comparison of the probability of Alzheimer’s disease (the dependent variable) comparing subpopulations with low versus high amyloid-beta levels (the predictor of interest). When the goal is estimating and testing a specific association, study design becomes critical in order to avoid biases regarding the inferred association. Specifically, one must carefully consider how the data have been sampled, decide what confounding factors are to be controlled if the data do not arise from a randomized experiment, and avoid performing many tests in order to avoid inflating the probability of observing a given association by chance.

Prediction is a common goal of many statistical analyses and is often the first goal that comes to mind when a researcher thinks of statistics and model building. In this case, one seeks to build a statistical model using observed independent variables in order to predict a new future response. For example, we may wish to predict an individual’s 5-year probability of a myocardial infarction given a set of variables, including a
family history of disease, past disease experience of the individual, anthropometric measures, vital signs, and serum biomarkers such as cholesterol, triglycerides, and C-reactive protein. The role of statistics in this case is to provide a framework to estimate a fairly complicated model that can be optimized for predicting the dependent variable using an observed set of data. We generally refer to the data used to estimate the model as training data. This is because our goal is not to predict the data we already have (we know what those responses are) but to predict the outcome of a new observation in the future given their covariate values. As such, when building a prediction model, it is critical to explore many possible models and to compare them based upon how well they predict results that are outside of the training data in order to obtain an honest assessment of how well each model will perform when applied to new data. Section 2.11 further elaborates on the criteria needed for choosing a good predictive model.

What many researchers find confusing is that the statistical methods used to address each of the abovementioned goals are often the same, but the analysis strategy used for choosing a given model is quite different in each case. For example, one might reasonably use a logistic regression model (see Chapter 17: Logistic regression for binary outcomes) to estimate and test the association between amyloid-beta levels and the risk of Alzheimer’s disease, as well as to build a prediction model for the 5-year probability of a myocardial infarction. However, the approach to selecting a final logistic regression model is quite different given the differing scientific objectives. In the case of estimating the association between amyloid-beta levels and the risk of Alzheimer’s disease, one should a priori consider the potential confounding factors that should be adjusted for in order to make a “fair” comparison between populations with low and high amyloid-beta levels (e.g., age and sex would surely need to be adjusted for). Defining the model and adjustment covariates a priori is essential in this case because the goal is to test the association, and we wish to avoid performing multiple tests on various models as this will increase the probability of false-positive findings, thereby leaving standard p-values and confidence intervals meaningless. On the other hand, in order to choose a good predictive model, it will be necessary to consider the prediction performance of many potential models (including and excluding particular variables). In this case a p-value and control of the false-positive rate is not the objective. Instead, having an honest assessment of the out-of-sample predictive performance of each model is the criterion one need consider.

2.8 REPORTING STATISTICAL RESULTS

Advice on reporting is scattered throughout this book, imbedded in methods to develop the statistics to be reported. Among the more important suggestions are as follows:

• Long imposing tables of results and rows of figures will be read by few and remembered the next day by none. Select the core of what you want the reader to take away from your article and show it in one dramatic chart or table so it will be
remembered. If ancillary or supporting tables or figures are necessary, place them in a subordinate section or appendix for those few who will actually study your article.

- No one is offended by clear simple exposition in plane language, so long as it is technically correct. Allow jargon and abstruse mathematics only when absolutely necessary.
- Stay focused on your objective, present the development of your analysis in a logical fashion leading from objective to result, and avoid side excursions irrelevant to that objective.
- In reporting a statistical test, be sure that all descriptive statistics inputs have been reported and that the name of the test and its underlying assumptions and parameters are clear.
- Carry as much accuracy as you have available in computation and round the finished product to the accuracy of clinical interest.


### 2.9 A FIRST-STEP GUIDE TO DESCRIPTIVE STATISTICS

#### Statistical terms

In the remainder of this chapter a number of statistical terms are used. Most medical investigators will have seen these terms in journal articles frequently enough to be at least somewhat familiar with them. They are all defined and explained more fully at the appropriate places in this book, but some terms have to be used in order to present first-step guides. A reader encountering an unfamiliar term may find it explained in Section 1.18 on a shallow level that will serve in the interim until reaching later chapters.

#### Accessing computer help

Requesting descriptive statistics from a software program is usually simple when data are formatted as recommended in Section 2.6. Most statistical software packages have a menu tab at the top of the screen labeled “statistics” or “analysis.” In that menu, there is a choice labeled “descriptive statistics” or something similar. A click on that yields a display of descriptors or summaries of the selected data. Usually given are sample size, mean, standard deviation, and range, and often medians, rates of occurrence (e.g., %), and other values. Any common descriptor is usually found in the menu choices.
Univariate descriptive statistics

Chapter 5, Descriptive statistics, describes the more common descriptive statistics, with confidence intervals in Chapter 8, Tolerance, prediction, and confidence intervals. Familiarity with them will help select which are appropriate to use in an application. As a first step, data type and shape of distribution suggest the most common descriptors, as can be seen in Table 2.1.

Exercise 2.4

From Table 2.1, what description would you use to convey information on the following patient samples? (a) Average systolic blood pressure (SBP). (b) How spread out SBP is about the mean. (c) Stenosis of an artery or not. (d) Posttreatment psoriasis as absent (0), mild (1), moderate (2), or severe (3). (e) Similarity of occurrence patterns between pulse oximetry and minutes of strenuous exercise. (f) Relative frequency of occurrence of HIV inception by sex.

Bivariate and trivariate descriptive statistics for confounding and effect modification

As previously noted, in most observational studies where we seek to better understand the possible causal pathway between a predictor of interest and some outcome, it will be necessary to adjust for factors that may confound the relationship of interest. Recall that, by definition, a confounder must be causally associated with both the predictor of interest and the outcome. As such, it is useful to provide descriptive statistics that highlight these pairwise associations. Bivariate descriptive statistics are geared for this purpose. For example, consider again observational data aimed at assessing the association between alcohol consumption and the risk of lung cancer. We previously noted that this association is likely to be confounded by smoking status since individuals who smoke are more likely to drink alcohol, and as we know smoking is related to an increase in the risk of lung cancer. Bivariate descriptive statistics can be useful for highlighting this. One possibility would be to describe the observed proportion of alcohol consumed among smokers and nonsmokers. A difference in these two proportions would suggest that alcohol use is associated with smoking status. Similarly, one could then describe the observed proportion of lung cancer cases by smoking status. Indications of both associations would then suggest the role of smoking status as a confounder in the analysis. We will see in later chapters that for continuous variables, scatterplots, wherein one plots the value of one variable versus the other, will be a useful tool for describing bivariate relationships.

We have also noted that the assessment of effect modification is also a common objective in statistical analyses. Recall that an effect modifier is a variable for which the association between the predictor of interest and the dependent variable changes
Table 2.1 Type of descriptors most commonly used, providing a starting point for selection.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Numerical description&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Association</th>
<th>Visual description&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical (nominal) variables</td>
<td>Count or rate</td>
<td>Tetrachoric correlation; Φ, V, γ, κ</td>
<td>Bar chart</td>
</tr>
<tr>
<td>Rank-order or nonnormal</td>
<td>Median IQR</td>
<td>Rank (Spearman) correlation</td>
<td>Box plot</td>
</tr>
<tr>
<td>continuous variables</td>
<td>Mean SD</td>
<td>Pearson correlation</td>
<td>Bar chart with whiskers</td>
</tr>
<tr>
<td>Normal (bell-shaped) variables</td>
<td>Mean SD</td>
<td>Confidence intervals on a mean</td>
<td>Pie chart</td>
</tr>
</tbody>
</table>

After developing a descriptor, decide if it conveys the information you want the reader to take away from your data. If not, go to more advanced sections of the book for alternative options. CI, Confidence intervals; IQR, interquartile range; SD, standard deviation.

<sup>a</sup>For more detail on numerical description, see Chapter 4, Distributions; on center, spread, correlation, and charts, see Chapter 5, Descriptive statistics; on categorical association measures Φ, V, γ, κ, see Chapter 14, Measuring association and agreement; and on confidence intervals, see Chapter 8, Tolerance, prediction, and confidence intervals.
depending upon the value of this variable. Trivariate descriptive statistics seek to assess the bivariate association between two variables conditional upon the value of third variable. As such, that can be useful for describing potential effect modification. Consider again the example of an estrogen blocking treatment for breast cancer. We previously noted that the treatment would likely have a different effect in males than in females, implying that sex would modify the effect of treatment. To describe this, it would be important to produce bivariate descriptive statistics summarizing the time to progression of disease by treatment arm separately between males and females.

### 2.10 AN OVERVIEW OF ASSOCIATION TESTING

#### Setting up a test within a study

Since biological events tend to include imperfectly controlled variables, we can describe, explain, and predict events only probabilistically. For example, infection with a particular agent will likely but not certainly cause a particular symptom. Three stages were described in Section 1.2. We first seek to describe events. For this, we use descriptive statistics, as in Section 2.9. We sometimes use statistical tests to decide which groupings should be presented descriptively. Much of the mechanics of tests, although pursuing different goals, are used in confidence intervals as part of description. Confidence intervals are examined in Chapter 8, Tolerance, prediction, and confidence intervals. When we are acquainted with the nature of biological events, we seek to explain the events, inferring causal factors for these events. Tests to include or exclude causal factors compose much of the explanation. They begin here, with more specific guidance seen in many of the remaining chapters. When we have identified the primary causes, we seek to predict these events from specific values and combinations of the causes, pursued further in Section 2.11 and detailed in Chapters 15—20.

We use statistical tests to identify the significant variables that are associated with the outcome of interest. The largest part of clinical decision-making is based on tests. This section discusses setting up and conducting a test, and the next few sections provide a start in the selection of an appropriate test to use in a variety of circumstances.

#### A priori specification of a hypothesis

The goal of statistical testing is to generalize one’s statement regarding the plausibility of an association of interest outside of an observed sample of data to the broader population. Statistical tests rely upon the rules of probability to assess how uncommon an observed association would have been if, in the population, no association truly exists. Because we can never absolutely rule out the presence or absence of an association in the population using only a sample of observations, probability is used to bound the probability of a false-positive result (the Type I error of our test). That is, we seek to
limit the proportion of times we would have erroneously concluded that an association truly exists in the population when it in fact does not. Most statistical testing methods assume that only a single hypothesis is being tested when controlling the false-positive rate. However, common sense tells us that if we test many hypotheses, then the probability that we conclude an association exists on any one of them will increase as the number of tested hypotheses increases. This will happen by chance, even if none of the associations being tested truly exist. This phenomenon is often described as *multiple testing bias* or *multiplicity bias*. While there are some statistical methods that try to adjust for multiple testing bias, in most cases, they are far from perfect fixes. Further, it is common for investigators to fail to adjust for multiple tests and even more common to fail to report just how many tests they performed. The result is that published scientific findings contain many more false-positive results than would be anticipated in the absence of multiple testing. In light of this, perhaps the most important aspect of statistical testing is to first (i.e., prior to analyzing any data) think critically about what hypothesis is to be tested and what potential adjustment variables (confounding and effect modifying variables) are to be included in the analysis. If this is carefully considered and a priori specified, a single test can be performed and the false-positive rate can be validly controlled.

### STEPS IN SETTING UP A TEST

A sequence of steps in test development where a reader judging articles or an investigator planning research should look for will be examined in some detail in Section 7.8. A simplified sequence is given below to be of help when investigators set up their tests.

1. Specify clearly, completely, and unequivocally the question that you will be asking of your data.
2. Identify, specify in detail, and plan how to measure the variable(s) to answer that question.
3. Review your definitions of sample and population, and verify the appropriateness of generalization.
4. Review the sampling scheme used (to be used) to obtain your data.
5. Specify exactly the null and alternative hypotheses.
6. Select risks of a false-positive result (Type I error of size $\alpha$) and of a false-negative result (Type II error of size $\beta$).
7. Choose the form of the test (see the following subsection).
8. Verify that your sample size is adequate to achieve the proposed power.
9. At this point, obtain your data.
10. Organize your data, transfer them to a spreadsheet in the format required for analysis, and verify their correctness to avoid errors.
11. Identify and test possible biases.
12. Carry out the calculations of your test statistics and form your conclusions.
Choosing the right test

WHAT DO YOU WANT YOUR DATA TO TELL YOU?
You might ask if a rate of illness has risen from a historical prevalence of 0.2%, if the means of WBC from treatment and placebo groups are different, if the variability differs in drug content of two brands, or if the time to recover from an anesthetic, follow a normal distribution.

FIRST-STEP GUIDES
The following three sections provide a first step in choosing an appropriate test. Sometimes, subtleties in design or data availability will require more sophisticated tests, but the tests listed here will serve the majority of cases. It would do well to verify test appropriateness with an experienced biostatistician.

WHAT TO EXPECT FROM THESE GUIDES
These guides are presented before theory or methods, as an aid to getting started. The user should become familiar with statistical basics from the early chapters, after which the step may be made from the guide to the indicated sections later in the book.

A first-step guide to tests of rates or averages

THE TABLE FORMAT
Table 2.2 provides a first-step guide to choosing univariate tests of rates or averages. Back pain is one of the most frequent complaints in American society. In the civilian population, work-related cases result in over 1,000,000 lost workdays per year and nearly 15,000,000 visits to physicians. Back pain and simple treatments familiar to everyone were chosen for all illustrations. Variations of the same simple example, back pain as responding to treatment by aspirin (ASA) or ibuprofen (IPA), are posed to illustrate the distinctions among the data formats and analysis patterns addressed by different tests. Using the same variables for each test allows the different test results to be contrasted. Detailed examples may be found in the third edition of this book as Appendix to Chapter 2.

INFORMATION REQUIRED TO CHOOSE A TEST
In order to select a test appropriate to your question and your data, at least the following information is required:
1. What is your independent variable? What is your dependent variable? (If you are concerned only with level of association or correlation, it does not matter.)
2. What class of data do you have (categorical, rank order, continuous)?
3. Are your measurements independent (e.g., on patients in treatment versus placebo groups) or paired/matched (e.g., pretreatment, posttreatment, and 6-month follow-up on same patient)?
Table 2.2 A first-step guide to choosing statistical tests.

<table>
<thead>
<tr>
<th>Dependent variable (outcome; result of experiment) example: back pain readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Categories</td>
</tr>
<tr>
<td>0,1 (absent, present)</td>
</tr>
</tbody>
</table>

### Independent variable (fixed by experimental design or by nature)

<table>
<thead>
<tr>
<th>1 or 2 Categories</th>
<th>2 Categories</th>
<th>&gt;2 Categories</th>
<th>Ranked</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or 2 Categories</td>
<td>Not matched (2 samples)</td>
<td>Placebo (0) vs ASA (1)</td>
<td>Fisher’s exact test of contingency Section 9.4</td>
<td>Fisher’s exact test or χ² test of contingency Section 9.3–9.5</td>
</tr>
<tr>
<td></td>
<td>Matched (repeated) or single sample</td>
<td>Before vs after ASA</td>
<td>Binomial’s (or Poisson) test of proportion or McNemar’s test: Section 9.6–9.8</td>
<td>χ² Goodness of fit test against uniform distribution Section 13.6</td>
</tr>
<tr>
<td>&gt;2 Categories</td>
<td>Not matched</td>
<td>Group: Plac (0), IPA (1), ASA (2)</td>
<td>Fisher’s exact test of contingency Section 9.4</td>
<td>Fisher’s exact test or χ² test of contingency Section 9.5</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>Sequence: Plac (0), IPA (1), ASA (2)</td>
<td>Cochran’s Q test Section 9.9</td>
<td>χ² Goodness of fit test against uniform distribution Section 13.6</td>
</tr>
<tr>
<td>Ranked</td>
<td>Not matched</td>
<td>ASA: 0, 5, 10 Grains</td>
<td>Royston’s P Trend test Section 9.10</td>
<td>Kruskal–Wallis test Section 11.10</td>
</tr>
<tr>
<td></td>
<td>Matched</td>
<td>ASA at 0, 2, 4 h</td>
<td>Cochran’s Q test Section 9.9</td>
<td>χ² Goodness of fit test against uniform distribution Section 13.6</td>
</tr>
<tr>
<td>Continuous</td>
<td></td>
<td>Grains of ASA</td>
<td>Logistic regression Section 17.2</td>
<td>Multinomial regression Section 13.6</td>
</tr>
</tbody>
</table>

Variations of a simple example, back pain as responding to treatment by aspirin (ASA) or ibuprofen (IPA), are posed to illustrate the data and analysis pattern addressed by each test. In column 2, pain types are categorized as “distributed pain” (1), “sharp local pain” (2), or “severe ache” (3). In some cases, the test shown is not the only test that could be used and other tests might be preferable, depending on data definitions and objectives. In the Continuous column, if assumptions, for example, normal distribution, and equal variances, are badly violated, use the Ranked column entries.
AN EXAMPLE FROM PROSTATE CANCER
An example might help the reader to use Table 2.2. Is the PSA level for 15 cancer patients different after hormone therapy than before? We want to test the mean within-subject change in PSA level from before to after therapy.

The independent variable is the time of measurement: before therapy and after therapy; this time is the same for all patients and is not affected by the outcome of the therapy. The time points *pre* and *post* are categorical; every recording falls into one or the other. The time points are paired; each patient has two matched records. Starting with the side of Table 2.2, two categorical matched variables lead us to row 2 of the matrix.

The dependent variable is the difference in PSA levels at those times; the level *depends* on the therapy. Each PSA level is a continuous measurement, so the difference is continuous. From the top of Table 2.2 a continuous dependent variable leads us to column 4 of the matrix. The intersection of row 2 and column 4 suggests a paired *t*-test. We proceed to Section 11.2 and conduct our test as directed.

LIMITATIONS OF TABLE
This table addresses tests for only one independent and one dependent variables. Cases with multiple independent variables are addressed in Chapter 11, Tests of location with continuous outcomes, and subsequent chapters. Cases with multiple dependent variables are beyond the scope of this book.

This table addresses only tests for rates and averages. Tests for variability or for distributions are treated in Chapter 13, Tests on variability and distributions, and for correlation in Chapter 14, Measuring association and agreement, and Chapter 15, Linear regression and correlation.

LARGE SAMPLE TESTS
Prior to wide availability of computers and statistical software, several tests were prohibitively manpower intensive when samples were large. “Large sample” methods were developed, particularly for normal, Poisson, signed-rank, and rank-sum tests. These are seldom used today in light of modern computational power. The reader should be aware of their availability.

In addition, the chi-square test of contingency (only; not other uses of the chi-square distribution), developed prior to Fisher’s exact test, is really a large sample alternative that needs to be used only in cases of multiple categories with large samples.

Exercise 2.5
*From Table 2.2, what test on data from DB14 would you choose for the following questions.*

(a) What will identify or negate age bias? (b) What will identify or negate sex bias? (c) Is eNO from before to after 5 minutes of exercise different between EIB and non-EIB patients?
2.11 A BRIEF DISCUSSION OF PREDICTION MODELING

In Section 2.7, we noted that a common scientific goal is to be able to use an existing sample of data to provide information for predicting the response of a new observation if it were drawn from the population. For example, we might use data from the National Health and Nutrition Study (NHANES) to characterize the height of children based upon their age, sex, race/ethnicity, etc. Thus the goal in this case is not to test equality of the distributions (or some characteristic of the distributions) between subpopulations but instead to use our existing data to build a model that relates the response to the other predictor variables we have observed. We can almost certainly guarantee that no model we build will be perfect, in the sense that it will yield perfect predictions for all future observations. As such, the key to effectively building a prediction model is to try and minimize the error when predicting new observations in the population.

A structured approach to forming a prediction tool

A sound approach to achieving the goal of prediction can be thought of as three steps:

1. **Training**: In the *training step*, we use existing data to estimate how the predictors that have been measured best associate with the response we intend to predict. In the training step, we often consider many different prediction rules. For example, in the case of predicting height, we might consider one model that includes only age and sex, and in another model, we might consider age, sex, and race/ethnicity.

2. **Validation**: In the *validation step*, we use another set of data to estimate prediction error, or how far a prediction rule is from the actual value on measured observations, and to choose the top-performing prediction rule among a set of possible prediction rules that were formulated in the training step.

3. **Test**: Once we have decided upon a final model, it is necessary to test the model in an independent sample of data to assess whether the predictive ability of the model replicates in the population. This is known as the *test step* and is only performed for one model (the model chosen in the validation step). As an analogy to hypothesis testing, the test step can be thought of as replicating the test of an association in another sample to reaffirm confidence in our scientific conclusions.
Probability and relative frequency

3.1 PROBABILITY CONCEPTS

Probability defined

We should not confuse probability with possibility, an error frequently seen among the lay public and in the media. Possibility is 1 or 0: an event can occur, or it cannot. In its simplest form, probability may be considered the likelihood that a possible event occurs on a random opportunity to occur. On a flip of a coin the possible events are a head (H) or a tail (T). Which occurs on a flip is a random variable with value 0 or 1. The probability of a 1 is the likelihood of a head on a flip, a half if the coin is unbiased. We denote this probability as \( P(H) = 0.5 \).

How did we find 0.5 for \( P(H) \)? Suppose, you have an infectious disease ward with 25 patients, 10 of whom are in their 20s–30s (denoted as Y for young), 5 in their 40s–50s (M for middle), and 10 aged 60 or older (S for senior). If you draw the name of one patient at random, what is the probability of selecting a Y patient?

The basic definition of probability of a randomly chosen event is the number of equally likely ways that the event can occur divided by the number of ways any event can occur.

In this case a Y patient can be drawn 10 ways and any patient’s name can be drawn 25 ways, so the probability of a younger patient being selected is \( 10/25 \), or \( P(Y) = 0.4 \). Similarly, on the unbiased coin flip, there is 1 way for H to occur and 2 equally likely ways for either side to occur, so \( P(H) = 0.5 \).

Probability can vary from 0 to 1

If an event cannot occur, the number of ways it can occur is 0. From the definition the probability of this event must be 0, defining the lower bound of probability. If an event is certain to occur, the number of ways it can occur must include all the ways any event can occur, forming a fraction with top and bottom the same. From the definition the probability must be 1, defining the upper bound of probability. Thus all probability lies within the limits of 0, an impossible event, and 1, a certain event.
The additive rule of probability

There exist some rules for multiple events. To return to the infectious disease ward, if a name is drawn at random, what is the probability that it is either a Y or an M? The additive rule says that the probability of one or another event occurring is the sum of the individual probabilities, or

\[ P(Y \text{ or } M) = P(Y) + P(M) \]  \hfill (3.1)

if the two events are mutually exclusive, that is, cannot occur together. In this case, \( P(M) = 5/25 = 0.2 \), so \( P(Y \text{ or } M) = P(Y) + P(M) = 0.4 + 0.2 = 0.6 \).

Suppose, events can occur together. You have 20 patients with liver disease, including four patients with hepatitis B, denoted as \( H_b \), and six with cirrhosis of the liver, denoted as \( C \). Two patients have both \( H_b \) and \( C \). If a patient is selected randomly, what is the probability that the disease will be \( H_b \) or \( C \)? \( P(H_b) = 4/20 = 0.2 \), \( P(C) = 6/20 = 0.3 \), and \( P(H_b \text{ and } C) = 2/20 = 0.1 \). Because the two patients that have both \( H_b \) and \( C \) were counted in both \( P(H_b) \) and \( P(C) \), one of these counts must be subtracted, so

\[ P(H_b \text{ or } C) = P(H_b) + P(C) - P(H_b \text{ and } C), \]  \hfill (3.2)

if \( H_b \) and \( C \) may occur together. In this case, \( P(H_b \text{ or } C) = 0.2 + 0.3 - 0.1 = 0.4 \).

The multiplicative rule of probability

If two names from the infectious disease ward are drawn at random, what is the probability that one will be \( Y \) and the other, \( M \)? The multiplicative rule says that the probability of both the events occurring is the product of the individual probabilities:

\[ P(Y \text{ and } M) = P(Y) \times P(M), \]  \hfill (3.3)

if the events are independent. In this case, because the names are simultaneously drawn at random, the events are independent (the choice of one does not depend upon the value of the other). Thus \( P(Y \text{ and } M) = 0.4 \times 0.2 = 0.08 \).

If the characteristics are not independent, as in the liver disease case, we have to consider conditional probability, that is, the probability of an event given that a related event has occurred. A vertical bar is used to denote conditionality, so that the probability of \( C \) given that \( H_b \) occurs is written as \( P(C|H_b) \). If we know \( H_b \) is a property of the draw, and there are four \( H_b \) patients, two of whom are also \( C \), then the probability of a \( C \) out of the \( H_b \) patients is \( P(C|H_b) = 2/4 = 0.5 \). For dependent events the multiplicative rule is written

\[ P(H_b \text{ and } C) = P(H_b) \times P(C|H_b) \]  \hfill (3.4)
if the events are not independent. In this case, \( P(H_b \text{ and } C) = 0.16 \times 0.5 = 0.08 \), agreeing with the direct calculation of the probability of two patients with both \( H_b \) and \( C \) out of 25.

**Bayes’ rule**

Bayes’ rule, also known as Bayes’ formula or theorem, appears at this point in the book as a quaint relationship in conditional probability. However, it has profound implications, as can be seen in Chapter 25, Bayesian statistics. This rule was noted by the English clergyman Thomas Bayes (1702–61). It has become so important that the Rev. Bayes house in Tunbridge Wells, Kent, has become a national monument labeled with a brass plaque.

Suppose a patient exhibits some symptom \( S \) (perhaps fever) of a disease \( D \) (perhaps flu) during the winter season. Epidemiologists tell us that 6% of the populace exhibit fever, the prevalence of flu is 3%, and 90% of patients with flu have a fever. Thus we know \( P(S) = 0.06 \), \( P(D) = 0.03 \), and \( P(S|D) = 0.90 \). What we do not know is the chance that a patient with fever has the flu, or \( P(D|S) \). From Eq. (3.4), \( P(D \text{ and } S) = P(D) \cdot P(S|D) \). But we could also write \( P(D \text{ and } S) = P(S) \cdot P(D|S) \). Setting these two equal to each other and solving for \( P(D|S) \), we obtain Eq. (3.5), the simplest form of Bayes rule.

\[
P(D|S) = \frac{P(D)P(S|D)}{P(S)}
\]  

(3.5)

If we substitute our probability estimates in Eq. (3.5), we can obtain an estimate of the probability that a patient with fever will have the flu: \( (0.03 \times 0.90)/(0.06) = 0.45 \). There are many more sophisticated forms of Bayes’ rule that are used in various applications.

**Combinations**

Suppose, you inject an anesthetic into a broken wrist joint after repair and want to know how long pain relief lasts. You measure pain pre-repair, immediately post-repair, at 1 week, and at 1 month (times 1, 2, 3, and 4, respectively). You want to compare all possible pairs, that is, pre- to post-repair, pre-repair to a week, \ldots, a week to a month. How many pairs must you compare?
This falls in the realm of combinatorics, said “the combination of \( n \) things taken \( k \) at a time,” symbolized \( \binom{n}{k} \). In our case, we have four things taken two at a time, or \( \binom{4}{2} \). A required definition is a factorial, symbolized by a number followed by a “!”.

It represents the multiplication of the number times, the number minus 1, times the number minus 2, etc., as \( 3! = 3 \cdot 2 \cdot 1 = 6 \). The number of combinations is calculated as

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\ldots(n-k+1)}{k!}.
\]

In our case of four time points to measure pain, the number of possible pairs to compare is given by

\[
\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4 \times 3}{2 \times 1} = 6.
\]

If we are dealing only with pairs, the number of possible pairs simplifies to

\[
\binom{n}{2} = \frac{n(n-1)}{2}.
\]

**Exercise 3.1**

Suppose, you have 10 young (Y), 5 middle aged (M), and 10 senior (S) patients. (a) Find the probability that a randomly selected patient is not young. (b) Find the probability that neither of two randomly drawn patients is young.

**Exercise 3.2**

Suppose, you have 25 liver disease patients. Five have Hepatitis C (Hc). Ten each have hepatitis B (Hb) and cirrhosis (C), six of which have both Hb and C. (a) Find the probability that a randomly selected patient will have Hb or C. (b) Find the probability that in two randomly selected patients, you will find both Hb and C.

**Exercise 3.3**

Suppose, in our liver disease ward from Exercise 3.2, 12 patients display jaundice. We know that 90% of hepatitis B patients in our catchment display jaundice. If a patient selected at random displays jaundice, what is the probability that this patient has hepatitis B?

**Exercise 3.4**

You measure the healing effect of a placebo and four different antibiotics on patients with a particular disease. To compare each treatment with each other, how many pairings do you have?
3.2 PROBABILITY AND RELATIVE FREQUENCY

A graph of probability

We could represent a probability by a graph. Let the horizontal (x) axis represent the events that are possible to occur and the vertical (y) axis, a height proportional to the probability of occurrence of each possible event. The vertical axis will, of course, lie from 0 to 1.

Consider the roll of a fair die, originally white but with two faces colored red. The possible outcomes are red (R) or white (W). The x-axis holds the two positions R and W. When the die is rolled, there are six ways for a face to land upward. The probability of a red is $\frac{2}{6} = \frac{1}{3}$ and that of white, $\frac{4}{6} = \frac{2}{3}$. We can graph the probabilities as in Fig. 3.1.

Relative frequencies estimate probabilities

Suppose, we do not know these probabilities. We roll the die a number of times and count the frequency with which red occurs. As an example, the die was rolled 20 times and six red faces appeared. The relative frequency of an event’s occurrence is defined as the frequency with which that event occurs divided by the frequency with which it is possible to occur. This definition is the same as that for probability, except that we are counting observed occurrences rather than known possibilities. In this case the six occurrences of red divided by the 20 opportunities to occur yields a relative frequency of 0.3. Note the similarity to probability. If we did not know the $\frac{1}{3}$ probability from theory, we could get an idea of it from our experiment. This correspondence is one of the basic relationships of statistical theory: relative frequencies estimate probabilities and relative frequency distributions approximate probability distributions.

Figure 3.1 A probability graph of the throw of a white die with two red faces.
Relative frequency in medicine

The relative frequency of a die is a dull little exercise. However, in medicine we often want to know how likely an event is but have no probabilities. Indeed, often probabilities could not even be calculated at all, because, for example, the probability denominator (the number of ways any event could occur) might have to include all people of the defined class in the world. Finding relative frequencies of occurrence estimates these probabilities for us.

Suppose, we wanted the probability of a stroke as a side effect of a certain drug. We estimate it by dividing the number of strokes experienced (say four) by the number of administrations of the drug recorded (say 9483). We find a relative frequency of 0.00042 as an estimate of the probability.

We also note that a relative frequency is a rate of occurrence, a concept used daily in medicine. Probabilities and relative frequencies are very relevant and important concepts in both medical practice and medical research.

The accuracy of estimation increases as the sample size grows larger

An experiment was performed, rolling the white die with two red faces. \( P(R) = 1/3 \). How good is an experimental estimate? We roll the die 20 times, yielding 6 red faces. The estimate of \( P(R) \) is 0.300. The error rate is \((1/3 - 3/10) / (1/3)\), giving us an error of 10%. The exercise was extended, and the die rolled 100 times, yielding 31 red faces. This estimates \( P(R) \) as 0.310, reducing the error to 6%. As the sample size grows larger, the relative frequency estimate can be expected to grow closer to the probability.

We say “expected” because it does so on average. Since the occurrences are random, we occasionally get a run of occurrences that take the relative frequency away from rather than toward the probability. However, in the long run (quite literally a “long run” of occurrences), the relative frequency will converge on the probability. This fact is a result of what is known as the weak law of large numbers, a result discussed in further detail in Section 4.3.

Exercise 3.5

Returning to our liver disease ward of Exercise 3.2, we look at six randomly chosen charts and find two patients with cirrhosis. What is our relative frequency and how does it compare with the probability of cirrhosis for the ward?

3.3 GRAPHING RELATIVE FREQUENCY

When a variety of outcomes occur, the relative frequency of each can be graphed to show the distribution of relative frequencies. The very simple distribution of red and
white from the experimental 20 rolls of the die with two red faces would contain two bars, similar to Fig. 3.1, but with the bar for red outcome of height 0.3 and the bar for white outcome of height 0.7. In general, for each possible category, we count the occurrences and divide this count by the total sample size to find that category’s relative frequency. The vertical axis ranges from 0 to 1 and the relative frequency is represented by the bar’s height.

3.4 CONTINUOUS RANDOM VARIABLES

Relative frequencies and probabilities for categorical variables are straightforward. Those for continuous variables have an added complication: there are no clear delineations of categories.

As an example in medicine, consider the maximal diameter of tumor size (cm) for 115 liver cancer patients. We must choose intervals along the x-axis, after which each interval may be treated like a category for computation. Of course, choosing a different interval set yields different computations. The distribution ranges from 0.7 to 5.4 cm with mean 2.74 cm and standard deviation (see the Appendix of Chapter 1) 1.03 cm. Let us choose intervals of 1 cm from 0 to 6 cm. Relative frequencies may be plotted as appear in Fig. 3.2.

Suppose, we had a vast number of tumor size recordings, perhaps 1000 times the 115. We could increase the number of intervals, reducing the interval size from 1 cm to 1 mm, even to 0.1 mm. As we went through this process, the ragged top of the graph would become less and less ragged until it approached a smooth curve.

Figure 3.2 Relative frequencies of tumor size in 115 liver cancer patients.
The resulting curve would approximate the probability distribution of tumor sizes more and more closely. A continuous random variable has no natural intervals, allowing its probability distribution, conceptually represented by an infinite number of intervals each of zero width, to be fully smooth. The distribution will have a measure of average to represent its center and a measure of variability to represent its width. Some continuous distributions have other measures to represent shape characteristics.

### 3.5 FREQUENCY DISTRIBUTIONS FOR CONTINUOUS VARIABLES

A frequency distribution is nothing more than the way the data are distributed along the scale (or axis) of the variable of interest, but it is a concept through which most of the essential elements of medical statistics can be accessed.

**Start with a tally**

How is a frequency distribution for a continuous variable formed? Let us look at the prostate volumes in Database DB1. The variable of interest is volume in milliliters (mL). We scan the (entire set of) data to ascertain the range of volumes, which extends from about 3 to 114 mL. We list 5-mL intervals (choosing intervals is discussed later), starting with 0 to <5 (expressed this way to indicate “up to but not including 5” and done to avoid a possible overlap) and ending with 110 to <115. We go through the data sequentially, making a tick mark for each datum. A space, a crossover, or some other indicator might be placed every five tick marks to facilitate counting. Such a tally sheet for the data from Table DB1.1 for 10 patients is shown as Fig. 3.3A. If we rotate Fig. 3.3A a quarter turn counterclockwise and enclose the tick marks into boxes, towers or bars are formed, as shown in Fig. 3.3B. Continuing this procedure for the data for the first 50 patients, we produce Fig. 3.3C. The pattern of the distribution is beginning to emerge. Fig. 3.3D represents the distribution of data for 100 patients, Fig. 3.3E represents that for 200 patients, and Fig. 3.3F shows the distribution of data for all 301 patients. Fig. 3.3F is the frequency distribution for the volume data.

**Frequencies expressed as a histogram**

A certain number of values have fallen into each defined interval, or bin. This is the frequency for that bin. There are 40 patients with volumes between 20 and 25 mL. Bins may be combined: for example, the frequency for volumes less than 30 mL is 122. This type of chart, representing frequencies by the heights of bars, is called a histogram.
Figure 3.3 Evolution of a bar chart showing the frequency distribution of prostate volume on 301 patients: (A) tally, (B) first 10 data, (C) 50 data, (D) 100 data, (E) 200 data, and (F) 301 data.
Figure 3.3 (Continued)
Relative frequencies in continuous distributions

As with rates the frequency itself does not always tell us what we want to know. Usually of more interest is the relative frequency, which is the proportion of the sample in a bin. The proportion of patients with volumes between 20 and 25 mL is $40/301 = 0.133$. The proportion of patients with volumes less than 30 mL is $122/301 = 0.405$. We have been trained by news reports and such sources to think in terms of parts per hundred, or percent, the proportion multiplied by 100. In the preceding sample, 40.5% of the prostate volumes are less than 30 mL. We noted in Section 3.4 that the relative frequency distribution of a variable estimates the probability distribution for that variable.

Effect of increasing sample size

We can observe the convergence phenomenon by comparing Fig. 3.3B–F. The distribution patterns are rather different for the earlier figures but become increasingly stable as we add more and more data. The increasing stability of statistical behavior with increasing sample size is an important property of sampling, noted in Section 1.7, about which more will be seen later.

Choosing intervals

Consider the prostate-specific antigen data of Table DB1.1. What is the range (i.e., largest minus smallest values)? The smallest value is 4.1 and the largest 9.0. The range is $9.0 - 4.1 = 4.9$. What intervals should we choose for a tally? With only 10 observations, a small number of intervals are appropriate. Interval widths that are easy to plot and easy to interpret are sensible, such as a width of 1.0 for this example. With a range of about 5, we are led to 5 intervals. If we had had a large number of data, we might have chosen 10 or even 15 intervals. Our intervals become $>4-5$, $>5-6$, $>6-7$, $>7-8$, $>8-9$. The student can construct the tally and the histogram.

Exercise 3.6

For extent of carinal resection in DB12, (a) select intervals for a tally, (b) tally the data, (c) find the median, and (d) convert the tally into a relative frequency distribution, (e) comment on whether the distribution appears by eye to be near normal. If not, in what characteristics does it differ? (f) give a visual estimate (without calculation) of the mean and mode. How do these compare with the median found in (c)?

Exercise 3.7

For patient age in DB12, (a) select intervals for a tally and (b) tally the data. In tallying, (c) record the median when 25 data have been tallied, then at 50, at 75, at 100, and at the end. Plot the median depending on the number of data sampled similar to the plot of Fig. 3.2.
Does the median converge to the final median? (d) Convert the tally into a relative frequency distribution. (e) Comment on whether the distribution appears by eye to be near normal. If not, in what characteristics does it differ? (f) Give a visual estimate (without calculation) of the mean and mode. How do these compare with the final median found in (c)?

Exercise 3.8
For International Normalized Ratio (INR) readings from the clinic in DB13, (a) select intervals for a tally, (b) tally the data, (c) find the median, and (d) convert the tally into a relative frequency distribution. (e) What does the median compared with the mean = 2.40 say about symmetry?

Exercise 3.9
For INR readings from the lab in DB13, (a) select intervals for a tally, (b) tally the data, (c) find the median, and (d) convert the tally into a relative frequency distribution. (e) What does the median compared with the mean = 2.28 say about symmetry?

Exercise 3.10
The squares of the 60 plasma silicone levels in DB5 are ($\times 10^4$): 225, 169, 1521, 400, 1521, 1764, 576, 324, 144, 676, 100, 121, 225, 361, 729, 784, 121, 121, 324, 324, 576, 2304, 484, 729, 361, 324, 1024, 961, 361, 441, 441, 576, 100, 144, 784, 625, 484, 441, 529, 484, 484, 576, 1444, 2025, 529, 484, 324, 225, 16, 196, 576, 400, 324, 576, 225, 361, 676, 900, 484, 576. Plot a relative frequency distribution of these squares using intervals: 0—< 0.02, 0.02—< 0.04, . . . , 0.22—0.24. Look ahead a chapter to Fig. 4.3. Does it appear more similar to the normal or the chi-square probability distribution?

Exercise 3.11
For INR readings from the clinic in DB13, examine the plot of the frequency distribution in intervals of width 0.25 done in Exercise 3.8(d). Again, look ahead to Fig. 4.3. Is the plot more similar to a normal or a chi-square probability distribution?

3.6 PROBABILITY ESTIMATES FROM CONTINUOUS DISTRIBUTIONS

The relationship
Suppose, a prostate volume $\geq 65$ mL is taken as suggestive of benign prostatic hyper trophy (BPH). From our sample of 301 prostate patients, 28 readings equal or exceed 65 mL; the proportion $28/301 = 0.093$ shows evidence of BPH. This is the relative frequency of BPH suspects in our sample. If $r$ patients out of $n$ show a certain characteristic, $r/n$ is the relative frequency of that characteristic. However, we are less
interested in the characteristics of the sample of older male patients at our hospital than of the population of older male patients in the United States. We want to know what proportion of the population is suspect for BPH. If we could measure the prostate volume of every older man in the United States, we would be able to find the relative frequency distribution for the population, which would be the probability distribution. We could then find the probability that any randomly chosen older male in America shows indications of BPH. This probability would be represented as the area under the probability distribution $\geq 65$ mL divided by the total area. This calculation is similar to measuring the area under the bars in the sample distribution $\geq 65$ mL and dividing by the area under all the bars, yielding 0.093 (9.3%).

**Estimated probability and the term “parameter”**

Because we cannot measure prostate volume for the entire population, we use what information we can obtain from the sample. The 0.093 relative frequency (9.3%) does not tell us the exact probability of equaling or exceeding 65 mL in the population, but it estimates this probability. If we took another similar sample from another hospital, the relative frequency of that sample would likely be different, but we would expect it to be somewhere in the same vicinity. The term parameter is used to distinguish an unchanging characteristic of a population from an equivalent characteristic of a sample, an estimate of this parameter, which varies from sample to sample. How well does the estimate approximate the parameter? We are able to measure the confidence in this estimate by methods that are discussed later.

**3.7 PROBABILITY AS AREA UNDER THE CURVE**

**Concept**

The $x$-axis of a discrete probability graph has one position for each possibility, with a height over that position between 0 and 1 representing the probability. The sum of probabilities over all possibilities equals 1. The graph of a continuous probability is a smooth curve having height between 0 and 1 and an area over the entire range of possibility equal to 1.

To find a probability associated with a continuous variable, we must define an interval. The area under the curve in that interval represents the probability that a random event will fall into that interval.

For example, the normal (bell) curve is symmetric so that half the area under the curve lies over the interval to the left of the curve’s center. Thus the probability that a randomly chosen value from that probability distribution will fall in the left half is $1/2$. 

The graphical relationship of relative frequency and probability

Looking at Fig. 3.2, relative frequency distributions of a sample of tumor sizes, we see that it is not far from symmetric and bell-shaped. Does it arise from a normal probability distribution? Fig. 3.4 repeats the frequency distribution with a normal probability distribution superposed. While clearly not exactly in agreement with the normal shape, the frequency distribution is not far off. Later chapters will address ways to test whether the frequency distribution either is normal (and slightly off due to the variation of randomness in the sampling) or probably did not arise from a normal distribution.

Chapter 4, Distributions, will examine properties of probability distributions and present the most commonly seen types that are used extensively in statistical methods.

REFERENCES

Distributions

4.1 CHARACTERISTICS OF A DISTRIBUTION

The fundamental statistical information is the distribution of data, because it contains all the information we need for our statistical methods. From a sample distribution, we can learn what sample proportion of individuals is part of an interval of interest. From a population distribution, we can learn the probability that a randomly chosen individual will arise from that interval of interest. From distributions, we can learn what is typical or characteristic of a distribution and how closely the distribution clusters about this typical value. We can learn about the regularity or “bumpiness” of a distribution and how close it is to symmetrical. We can learn about classes or types of distributions that are fundamental to statistical theory and methods. These characteristics of distributions are addressed conceptually in the next few sections. Formulas, expressing the nature of the distributions in symbols, are given in Chapter 5, Descriptive statistics, on summary statistics and later chapters on individual methods.

4.2 GREEK VERSUS ROMAN LETTERS

It is common among statisticians to use Greek letters to represent population names and Roman letters to represent sample names. Thus to represent a mean and variance, introduced in the next two sections, we use $\mu$ and $\sigma^2$ for the population parameters and $m$ and $s^2$ for the sample estimates. Some users of statistics have not adopted this convention, but it helps keep the important distinction between population and sample straight and will be used in this text. In our view the use of $\bar{x}$ to denote the mean is an inconsistent historical remnant.

4.3 WHAT IS TYPICAL

Averages

If we could have only one quantity to give us information about a distribution, it would usually be the typical, central, or “average” value. There are several measures of central tendency, three of which are mentioned in this section.
The mean, or average, is used most often. The population mean is denoted \( \mu \). We will denote the sample mean \( m \), although (as noted earlier) many books denote it by the letter representing the variable being averaged with a bar over it (e.g., \( \bar{x} \)). To obtain the sample mean, we add all values and divide by the number of values added. The mean prostate volume from Table DB1.1 is 32.73 mL. The mean of a probability distribution is the “center of gravity” of that distribution. If the picture of the distribution were cut from a material of uniform density, say wood, the mean would be the position on the horizontal axis at which the distribution would balance. Thus values far from the center carry more influence, or “moment,” than values near the center.

If we are sampling from two populations, say \( x \) and \( y \), we can use subscripts (previously mentioned as the “member-of-the-family” indicators) to keep them straight, such as \( \mu_x \) and \( \mu_y \) for the population means, respectively, and \( m_x \) and \( m_y \) for the sample means, respectively.

The median of a sample is the position on the horizontal axis at which half the observations fall on either side. If we order the prostate volumes in Table DB1.1, that is, put them in sequence of increasing value, the median would be the midpoint between the fifth and sixth volumes. The fifth and sixth volumes are 30.5 and 30.9 mL, respectively, so the median is 30.7 mL. For a distribution with some extreme observations, for example, income of HMO executives in the distribution of HMO employee incomes, or unusually long-lived patients in the distribution of survival times of cancer patients, the median is a better descriptor of the typical than is the mean. For a probability distribution the population median is the position on the horizontal axis at which half the area under the curve falls on either side.

The mode of a sample is the most frequently occurring value, which is the position on the horizontal axis centered under the tallest bar. The mode of the prostate volumes in Table DB1.1 can be seen in Fig. 3.3B as 32.5. The mode is not very dependable for small samples and should be used only for large samples. The position of the mode depends on the bar width and starting position. Also, the mode is not necessarily unique. An example of a distribution with two modes might be the heights of a mixed sample of men and women, because men, on average, tend to be taller than women. For a probability distribution the mode is the horizontal axis value under the highest point on the curve. For a multimodal probability distribution the position of each height on the curve having a lesser height to both left and right is a mode.

Other indicators of central tendency exist, such as the midrange and geometric mean, but both have poor estimating qualities or are less generally understood and hence are used only for specialized situations.

### Convergence with increasing sample size

In Section 1.7 the increasing stability of a sample with increasing sample size was mentioned. The mean of a sample gets closer to, that is converges on, the population
mean as the sample size grows larger. This property is known as the Weak Law of Large Numbers or the Bienaymé–Tchebycheff Inequality (also Tchebycheff alone and using various spellings). Although we need not remember the name, this relationship is essential to expressing confidence in our sample estimates of population means and to finding the sample size required for studies and experiments. To illustrate this convergence, sample means of differences between preexercise and 20-minute postexercise nitric oxide levels from DB14 are shown in Fig. 4.1 for the first 5, first 10, first 15, …, first 35, and all 38 patients seen. (They are listed in the order they were presented.) The means can be seen to converge on the 1.176 mean of all 38 patients as the sample size increases. Estimates are over, under, under, over, over, over, and over, respectively, whereas the size of error tends to diminish with each sample size increase, deviation from the mean grows slightly larger at sample sizes 20 and 30 because some patients with greater variability appeared. The convergence occurs in probability and numerically in the long run, not necessarily with every additional datum.

Exercise 4.1
For computational ease, use the glycosaminoglycans (GAG) levels for type I assay in DB8 rounded to 1 decimal place: 0.7, 0.4, 0.3, 0.2. Calculate (a) the mean and (b) the median.

Exercise 4.2
For computational ease, use the glycosaminoglycans (GAG) levels for type II assay in DB8 rounded to 1 decimal place: 0.5, 0.5, 0.4, 0.3. Calculate (a) the mean and (b) the median.
4.4 THE SPREAD ABOUT THE TYPICAL

Types of spread indicators

After we know the typical value of a distribution, it follows naturally to want to know how closely the values cluster about this average. As with the average, there are different measures of variability, but the variance and its square root, the standard deviation, are the most frequently used. In cases of an asymmetric or irregular distribution, the interquartile range (IQR) is usually a better indicator of spread.

The *range* (*highest minus lowest* values) gives us a hint, but it uses the information from only two of our observations, and those are the most unstable values in the sample. Although the range is not a good measure of spread, it and the two extreme values that define it are often of clinical interest in themselves.

The *variance* of a set of values is the *average of squared deviations from the mean*. Just for computational illustration, consider a population composed of the values 1, 2, and 3, which have the mean 2. First, the mean is subtracted from each observation and each difference is squared, yielding \((1 - 2)^2\), \((2 - 2)^2\), and \((3 - 2)^2\). Squaring these values makes them all non-negative so they do not cancel each other out, and it allows the more deviant observations to carry more weight. These squared differences are then averaged, that is, added and divided by their number, resulting in the population variance: \(\frac{(1 + 0 + 1)}{3} = 0.666\ldots\). The population variance is denoted \(\sigma^2\). A sample variance, commonly denoted \(s^2\), is calculated in the same way if \(\mu\) is known. However, if \(\mu\) is estimated by \(m\), it has been found that adjusting the divisor to be the sample size less one (as \(n - 1\)) provides an estimate of the variance that is closer, on average, to the population variance.

The *standard deviation is the square root of the variance*, denoted \(\sigma\) for populations and \(s\) for samples. The variance is often used in statistical methods because square roots are difficult to work with mathematically. However, the standard deviation typically is used in describing and interpreting results because it can represent distances along the axis of the variable of interest, whereas the variance is expressed in squared units. The meaning of squared volumes of the prostate would be somewhat obscure.

The quartiles of a distribution are the values at which 25%, 50%, and 75% of data are less than the respective values. The *IQR* is the width of the central half of a data set, extending from the point at which 25% of the data fall to the left up to the point at which 25% fall to the right. If a dataset contains some outliers, that is, data falling far from the apparent bulk of the distribution, squaring their distance from the mean to compute the variance enlarges the variance so much that it does not represent the spread very well. In such a case the IQR represents the spread better. The two, however, do not correspond to each other in terms of the portion of distribution covered, so that they must be interpreted differently.
Let us calculate the variance and standard deviation of prostate volumes from Table DB1.1. Using the sample mean $m = 32.73$ from the previous section, the observations minus $m$ are $-0.43, -5.73, \ldots, 4.07, -16.33$. Their squares are $0.1849, 32.8329, \ldots, 16.5649, 266.6689$, and the sum of these squares is $2281.401$. Dividing the sum of squares by 9 gives the variance, often called the mean square, as $s^2 = 253.409$. The standard deviation is just the square root of the variance, or $s = 15.92$. Engineers often call the standard deviation the root-mean-square, which may help keep its meaning in mind. The computation used here follows the conceptual method of calculation. A simpler computational formula, plus examples, will be given in Chapter 5, Descriptive statistics, on summary statistics.

**Exercise 4.3**
For computational ease, use the glycosaminoglycans (GAG) levels for type I assay in DB8 rounded to 1 decimal place: 0.7, 0.4, 0.3, 0.2. Calculate (a) the variance and (b) the standard deviation.

**Exercise 4.4**
For computational ease, use the glycosaminoglycans (GAG) levels for type II assay in DB8 rounded to 1 decimal place: 0.5, 0.5, 0.4, 0.3. Calculate (a) the variance and (b) the standard deviation.

**4.5 THE SHAPE**
A distribution may have almost any shape (provided it has areas only above the horizontal axis). Fortunately, only a few shapes are common, especially for large samples and populations where the laws of large numbers hold. Sample distributions, although different from sample to sample, often approximate probability distributions. When they do, this correspondence allows informative conclusions about the population to be inferred from a sample.

**Most common shapes**
The two most famous distribution shapes are the uniform and the normal curve, sometimes loosely termed a bell curve. The uniform distribution is an equal likelihood case, in which all events (such as choosing patients for a certain treatment in a study) have equal chance of occurring, so that all probabilities achieve the same height on the graph of the distribution. The distribution is just a rectangle. The bell curve has a graph showing a central hump (mode) with observations tailing off symmetrically, or approximately so, on either side. The normal is a bell curve with certain additional characteristics that are noted in Sections 4.7 and 6.2.
In describing the shape of a distribution, we look for what might be called smoothness. A “well behaved” distribution has a single clearly seen mode, tails off in not too jagged a fashion on either side, and relates in a meaningful way to the variable of interest. Then we look for symmetry. If one tail is “dragged out” more than the other, we say that the distribution is skewed in that direction. We look for the approximate center of the distribution. In a unimodal unskewed (symmetric) curve the mode, median, and mean will coincide. If it is right skewed, the median is to the right of the mode and the mean is to the right of both. Finally, we look to see how much it is spread out about its center, that is, its approximate standard deviation. By examining the distribution in this fashion, we can get a much clearer picture of what forces are at work in the generation of the data than through knowing only the mean and the standard deviation.

**EXAMPLE: PROSTATE VOLUMES FROM DB1**

Fig. 4.2 shows the frequency distribution of 301 prostate volumes from Fig. 3.3F with the population probability distribution approximated by a freehand fit. Note that it is skewed to the right. In agreement with this, the mean of 36.5 mL is to the right of both the median (32.4 mL) and the mode (32.5 mL). Before seeing any data, we might have expected a symmetric distribution. Some of the skew seems to arise from the presence of benign prostatic hypertrophy (BPH), which causes a large increase in prostate volume. However, we suspect the main cause is the limitation on the left by zero, whereas there is no limitation to the right. The mean $\pm 1$ standard deviation includes the interval on the volume scale of about $21-52$ mL; probabilistic interpretations of this are better.
discussed later, but it seems safe to say that we would not consider a prostate volume in this interval to be clinically abnormal.

**Standardizing a distribution**

Distributions come with various scales on the variable of interest, that is, with varying widths and locations. However, the scale differences can be overcome by standardizing the distribution. A distribution is *standardized* by subtracting the mean from every observation and dividing the result by the standard deviation. This forces the distribution to have a mean of 0, a standard deviation of 1, and a scale measured in standard deviations. This transformation allows shapes of different distributions to be compared, uncluttered by scale differences, and it allows a single table of areas under a curve of a particular type to be made rather than a new table for every scale. The *sample distribution* of a variable, say $x$, is standardized when $x$ is transformed into $y$ by using

$$y = \frac{x - m}{s}$$

and the *probability distribution* of $x$ is transformed by using

$$y = \frac{x - \mu}{\sigma}$$

Probability distributions are transformed further so that their total area under the curve is always 1.

**4.6 SAMPLING DISTRIBUTION OF A VARIABLE VERSUS A STATISTIC**

To this point, we have focused on the sampling distribution of a measurement, for example, the distribution of blood pressure values on the members of a defined population. In statistics, we commonly wish to make statements about a population parameter, for example, the population mean. We do so by considering an estimate of the mean using a sample of data. An estimate calculated from sample data is called a *statistic*. The population or probability distribution value it estimates is the corresponding *parameter*.

Just as there is a sampling distribution for individual measurements such as blood pressure values, there is (at least in theory) a sampling distribution for a statistic that could be formed on any given sample of data. For example, we could consider taking multiple independent samples of 100 subjects each, measuring the systolic blood pressure of each subject, and computing the sample mean of systolic blood pressure in each sample. Just as we can graph the approximate distribution of systolic blood
pressure values in each sample, we can also graph the computed means across the multiple samples. The latter would be an approximation to the sampling distribution of the sample mean.

Many statistical methods rely upon assumptions regarding either the distribution of the individual observations and/or the sampling distribution of a particular statistic such as the sample mean. The two are commonly confused by new practitioners of statistics, but the two concepts and the ability to satisfy assumptions regarding them are quite different. Importantly, the distribution of a statistic generally does not have the same shape as the distribution of individual observations. Further, we cannot make precise statements about the true distribution of observations in the population. However, as we will see in Section 4.9, we will often be able to make concrete statements about the sampling distribution of a statistic when we have sampled a relatively large amount of data.

4.7 Statistical Inference

Making inferences about a population on the basis of a sample from that population is a major task in statistics. A statistical inference is a conclusion about a state or process in nature drawn from quantitative, variable evidence in a way that specifies the risk for error about such a conclusion. The quality of the inference is formed by the degree to which the sample represents the population. Although we can never eliminate the risk of error from inference, we can quantify and limit its probability.

Desirable properties of a sample statistic

There are several indicators of the quality of a statistic. Stated conceptually rather than mathematically, a statistic should be the following:

Unbiased. If the statistic were calculated for all possible samples of size \( n \) from the population, the mean of these statistics would equal the parameter.

Sufficient. No other statistic calculated from that sample provides more information.

Efficient. The statistic estimates its parameter with minimum variance.

Consistent. The estimate converges on its parameter as the sample size approaches the population size.

Inference via a confidence interval

One important form of inference is the generation of confidence intervals. The goal is to measure and quantify the uncertainty in a sample’s estimate of a population parameter. Given a sample mean on some variable \( x \), for example, what can be said about the mean of the population? Mathematically, \( m \) has been shown to be the best estimate of \( \mu \), but how good is it in a particular case? By following a number
of steps, we can place a confidence interval about this mean, saying “We are 95% confident that this interval encloses the population mean.” (For medical purposes, 95% confidence is used most often, more by habit than because a 5% risk is better than some other risk. We can never be 100% sure.) This procedure infers that an interval is likely to contain the population mean by yielding a set of plausible values of the population mean. What defines “plausible”? This is defined by the set of hypothesized mean values for which, given our observed data, we would fail to reject the hypothesis that the population mean is equal to those values when using a level 0.05 test.

**Inference about a difference**

A decision is a variation on inferring a conclusion. An important form of inference is to decide whether a characteristic, such as a mean, is the same or different between two samples, or between a sample and a population. We state that these two population means are the same by means of a null hypothesis, so termed because it postulates a null difference between the population means. Then we specify an upper bound on the probability that we conclude there is a difference when in fact there is not. This may be thought of as a bound on the probability of a false-positive result, the same \( \alpha \), frequently taken to be 5%. Finally, by comparing our estimated probability of observing data as or more indicative of a difference than we did in our sample (the \( p \)-value) and comparing this to \( \alpha \), we either do or do not reject the null hypothesis, which is equivalent to inferring that a difference between the true means has or has not been established.

**Inference about equivalence**

A parallel goal, less commonly encountered, is the need to show whether or not two characteristics, such as means, are the same. For example, a new drug is developed that is much cheaper than an existing drug. Does it perform as well? In this case the error in making a wrong decision may be thought of as a false-negative result: that is, inferring equivalence when, in fact, there is a difference. The null hypothesis we must reject or not becomes the statement that the two mean performances are different. A difficulty not appearing in a difference test is that we must specify how far different the means are, say a difference of \( d \) units. The inference becomes the decision that the means are not different by as much as \( d \) units, and therefore are equivalent for clinical purposes, or that that equivalence has not been shown. The value \( d \) is usually chosen on clinical grounds so that the decision includes some arbitrary judgment not present in a difference test. More about equivalence testing can be found in Chapter 12, Equivalence testing.
Steps in inference

In making statistical inferences about a population parameter, we follow several steps: (1) we assume the nature of the probability distribution of the estimate of the parameter, based on the data from which it is drawn. For example, “The distribution of the sample mean is normal (a particular type of bell shape).” Many statistical inferences are based on more than one assumption, for example, that the underlying probability distribution is normal and that each sample value drawn cannot be predicted by other sample values (the observations are independent). (2) We arbitrarily choose a bound for the probability, often but not necessarily 5%, of falsely rejecting a null hypothesis when the null hypothesis is true. (3) We find an interval about our estimator of the population parameter on the x-axis designated by those values outside of which lies 5% (or other risk defined by $\alpha$) of the area under the probability distribution of the estimator that is derived under the assumption that the null hypothesis holds (calculated or looked up in tables). (4) If we are testing a null hypothesis, we reject the hypothesis if our observed estimate lies outside this interval and we fail to reject it if it lies inside. If we want a confidence interval, this is obtained by considering all possible theoretical values of the population parameter for which we would fail to reject these hypothesized values. The reader can see that there is a logical relationship between confidence intervals and hypothesis testing.

The logic in the preceding paragraph does not follow the way we usually think. It will require intense concentration on initial exposure. Students seldom comprehend it fully at first. Return to it repeatedly, preferably by using it when working through the methods given in later chapters and give it time to “sink in.”

One admonition should be noted. The assumptions made about distributions (step 1) are seldom actually stated, but they are made whether explicitly or implicitly and we must accept their consequences whether they were stated or not.

Effect of violated assumptions; robustness

What happens when an assumption is violated? The computations can be made in any case and there is no flag to alert the user to the violation. When assumptions presume erroneous characteristics of probability distributions, the areas under the curves are computed incorrectly and decisions are made with erroneous confidence. For example, in finding a confidence interval, we may believe that $\alpha$ is 0.05 when in fact it is much greater or smaller. We must give careful attention to assumptions. Fortunately, the commonly used statistical methods often are robust, meaning that they are not very sensitive to moderate violations of assumptions.

More detail on confidence intervals and hypothesis testing are given in Chapter 7, Hypothesis testing: concept and practice, and Chapter 8, Tolerance, prediction, and confidence intervals, respectively, after additional foundation is provided; additional attention to assumptions appears at the end of Section 8.1.
Exercise 4.5
As an investigator, you wish to make an inference from DB3 about whether or not the drug affects serum theophylline level (5 days). From Section 4.7, what steps would you go through?

Exercise 4.6
As an investigator, you wish to make an inference from DB14 about whether or not the difference in eNO from before exercise to 20 min after is 0. From Section 4.7, what steps would you go through?

4.8 DISTRIBUTIONS COMMONLY USED IN STATISTICS

Section 3.2 explains how frequency distributions arise from sample data. Section 3.5 demonstrates that the population distribution, arising from sampling the entire population, becomes the probability distribution. Section 4.7 shows that this probability distribution is used in the process of making statistical inferences about population characteristics on the basis of sample information. There are, of course, endless types of possible probability distributions. However, luckily, most statistical methods use only six probability distributions.

The six distributions commonly used in statistics are the normal, t, χ² (chi-square), F, binomial, and Poisson. Continuous data depend mostly on the first four distributions. Rank-order methods depend on distributions of ranks rather than continuous data, but several of them can be transformed to calculate probabilities from the normal or chi-square distributions. Categorical data depend mostly on the chi-square, binomial, and Poisson distributions, with larger samples transformed to normal. We need to become familiar with only these six distributions to understand most of the methods given in this text. Fig. 4.3 shows examples of these six types of distributions. The following paragraphs describe these distributions and some of their properties, needed to use and interpret statistical methods.

Normal distribution
The normal distribution, sometimes called Gaussian after the mathematician Carl Friedrich Gauss, is the perfect case of the famous bell curve. We standardize a normal (Gaussian) variable, x, transforming it to z by subtracting the mean and dividing by the standard deviation, for example,

$$ z = \frac{x - \mu}{\sigma} $$ (4.3)

The normal distribution then becomes the standard normal, which has mean 0 and standard deviation 1. This transformation is usually made in practice because the
Figure 4.3 The six common probability distributions used in elementary statistical methods. The normal (A) and \( t \) (B) are used with inferences about means of continuous data, the \( \chi^2 \) (C) and \( F \) (D) about standard deviations (more exactly, variances) of continuous data, and the binomial (E) about proportions of categorical data. The Poisson (F) and \( \chi^2 \) are used to approximate the binomial for larger sample sizes.
probability tables available are usually of the standard normal curve. The statistical “normal” must not be confused with the medical “normal,” meaning nonpathologic.

**SHORTHAND FOR THE NORMAL**

The normal distribution is used so much that a shorthand symbol is helpful. The common symbol is $N$ for “normal” with the mean and variance (square of the standard deviation) following in parentheses. Thus $N(5,4)$ indicates a normal distribution with mean of 5 and variance of 4 (standard deviation = 2). $N(0,1)$ indicates the standard normal with mean of 0 and standard deviation of 1, used for normal tables.

**t Distribution**

The $t$ distribution answers the same sorts of questions about the mean as does the normal distribution. It arises when we must use the sample standard deviation $s$ to estimate an unknown population standard deviation $\sigma$.

**THE STANDARD T**

Standardizing the mean includes dividing by the standard deviation. The known $\sigma$ is a constant, so the division just changes the scale. However, when $\sigma$ is unknown, which happens in most cases of clinical studies, we must divide by $s$. Instead of Eq. (4.3) as in the normal, we use

$$t = \frac{x - \mu}{s} \quad \text{or} \quad t = \frac{x - m}{s}$$

(4.4)

depending on whether $\mu$ is known or estimated by $m$. The sample $s$ is not a constant like the population $\sigma$, but it is composed of observations and, therefore, follows a probability distribution. This division by $s$ introduces $t$ as a new variable: one drawn from a normal distribution divided by a variable drawn from a more difficult distribution, that for the root of a sum of squares of normally distributed random variables. The probability distribution for this variable was published in 1908 by W. S. Gosset, who named it $t$. (As a historical note, he published under the pseudonym “Student,” because the policy of his employer, Guinness Brewery, was to use pseudonyms.)

The $t$ looks like the normal curve, as seen in Fig. 4.3A. However, it is a little fatter because it uses $s$, which is less accurate than $\sigma$. Whereas the normal is a single distribution, $t$ is a family of curves. In Fig. 4.4, two standard $t$ distributions are superposed on a standard normal. The particular member of the $t$ family depends on the sample size, or more exactly, on the degrees of freedom.

**Degrees of freedom**

Degrees of freedom, often abbreviated $df$, is a concept that may be thought of as *that part of the sample size $n$ not otherwise allocated*. This concept relates to quite a number of
aspects of statistical methods; thus \( df \) may be explained in a number of ways. Some of these aspects are more difficult than others, and even experienced users find some of them very challenging. Do not expect to understand \( df \) fully at once. Comprehension usually starts at a rudimentary level and sophisticates slowly with use. \( df \) is related to the sample number, usually to the number of observations for continuous data methods and to the number of categories for categorical data methods. As a start, it will be enough to conceive of \( df \) as a sample number adjusted for other sources of information, more specifically, the number of unrestricted and independent data entering into a calculated statistic. In the \( t \) distribution, we might think informally of \( n \) “pieces” of information available. Consider the form of \( t \) in which we use the sample mean \( m \). Once we know \( m \), we have \( n - 1 \) pieces of information remaining which can be selected by the sampling procedure; when we have obtained \( n - 1 \) observations, the \( n \)th one may be found by subtraction. (This is, of course, not how the data are obtained, but rather a more abstract mathematical allocation of information.) Because \( df \) is tied to \( n \), the sample values converge to the population values as both \( n \) and \( df \) increase. \( t \) Converges on the normal as \( df \) increases. Fig. 4.4 shows the standard normal curve with two \( t \) curves superposed, one with 10 \( df \) and the other with 5 \( df \). The fewer the \( df \), the less accurate an estimate \( s \) is of \( \sigma \), and thus the greater is the standard deviation of \( t \) (the “fatter” the \( t \) curve).

Chi-square (\( \chi^2 \)) distribution
We often want to make decisions, or other inferences, about a standard deviation. For computational ease, we usually make the decision about its square, the variance, because
any decision about one implies the equivalent decision about the other. Basically, a variance is a sum of squares of values. If each comes from a normal curve, its mathematical pattern has a right-skewed shape like that in Fig. 4.3C. That distribution is called the chi-square distribution (\( \text{ch} \) pronounced like \( k \)). It is obtained by multiplying the calculated sample variance \( s^2 \) by the constant \( df/\sigma^2 \), where \( df = n - 1 \) and \( \sigma^2 \) is the population variance. Often the Greek symbol for chi-square (\( \chi^2 \)) is used. Because all elements are squares, a chi-square cannot be negative. It rises from zero to a mode and then tails off in a skew to the right. As in the normal and the \( t \), we use areas under the curve taken from tables or computer calculation routines to make inferences.

**CHI-SQUARE IN A HYPOTHESIS TEST: CRITICAL VALUE**

As a further example of hypothesis testing, introduced in Section 4.7, consider the days to heal to a certain standard after a new surgical procedure as compared with that after an older established one. Say we know the new procedure is slightly better on average, but could it be so much more variable as to nullify its advantage? The heal-time variance of the old procedure is known from such a large number of cases that it can be taken as having converged to the population variance \( \sigma^2 \). Our estimate of the heal-time variance of the new procedure from a sample of patients is \( s^2 \). We are asking if \( s^2 \) is probably larger than \( \sigma^2 \). We test the null hypothesis that \( s^2 \) is no different from \( \sigma^2 \) by using the ratio

\[
\frac{s^2 \times df}{\sigma^2}
\]

that can be shown to follow the chi-square distribution. If \( s^2 \) is much greater than \( \sigma^2 \), the ratio will be large. We choose our \( \alpha \), the risk of being wrong if we say there is a difference when there is not. This \( \alpha \) is associated with a certain value of chi-square, which we term the critical value, because it is the value that separates the acceptance from the rejection of the null hypothesis. If the calculated value of \( (s^2 \times df)/\sigma^2 \) is greater than the critical value, we reject the null hypothesis and conclude that there is statistical evidence that \( s^2 \) is larger. Otherwise, we conclude that \( s^2 \) has not been shown to be larger.

**F Distribution**

In the example given in the previous paragraph, suppose we were comparing heal times for two new surgical procedures rather than a new one against an established one. The variances for both are estimated from small samples. We test the two-sample variances \( s_1^2 \) and \( s_2^2 \) to see if, in the populations being sampled, one is bigger in probability. To do this, we use their ratio, dividing the bigger by the smaller value. (The conclusion may also be extended to standard deviations: is one standard deviation
bigger in probability than the other?) The probability distribution for this ratio is called
$F$, named (by George Snedecor) after Sir Ronald Fisher, the greatest producer of prac-
tical theories in the field of statistics. The $F$ distribution looks similar to the chi-square
(inded, it is the ratio of two independent chi-square—distributed variables), as can be
seen in Fig. 4.3D. $F$ has one complication not appearing in $\chi^2$: an additional $df$ value.
Because $F$ involves two samples, the $df$ for each sample must be used, making
table reference a bit more involved.

**Rank-order methods**

If we ask statistical questions about data recorded as ranks rather than measurements
on intervals, we are dealing with fewer assumptions about the distributions of the data
being measured. Statistical methods have been developed for rank-type data, often
called *nonparametric* methods. In addition, continuous-type data can always be ranked.
The advantage of using rank-order methods for continuous data is that we do not need
to consider their sampling distributions. In the case of data following well-defined distri-
butions, such as the normal, rank-order methods are not as efficient as methods assuming
distributions. However, if the sampling distributions of continuous data are “poorly
behaved,” assumptions are violated, and results of parametric statistical methods can be
wrong. In such cases, rank-order methods may give more dependable results.

**Binomial distribution**

When using *categorical data*, we are usually asking questions about counts or proportions
(which are based on the same theory because one can be converted into the other). The
binomial distribution has a proportion as its parameter. This proportion, denoted $\pi$ for a
population and $p$ for a sample, and the number in the count, denoted $n$, fully characterize
the binomial. The binomial shown in Fig. 4.3E is the distribution with $\pi = 1/3$ and
$n = 12$. If our population were evenly split by sex, for example, the probability that a
randomly selected patient is male would be 0.5, the binomial parameter. Suppose we
treated 20 patients for some disease and found that 15 were male. Is the disease sex
related? If not, this mix of male and female patients would have arisen by chance alone
from our binomial with parameter 0.5. We can calculate this probability and find that
the chance of 15 or more male patients from an equal population is only 2% (0.021).
We have strong evidence that the disease is more prevalent in males.

**LARGE-SAMPLE APPROXIMATIONS TO THE BINOMIAL**

When we have more than a few observations, the binomial becomes difficult to
calculate. Fortunately, the binomial can be approximated fairly well by one of two dis-
butions, the normal or the Poisson. If the population proportion (the binomial
parameter) is toward the middle of its 0–1 range, say, between 0.1 and 0.9, the
normal distribution approximates large-sample binomials fairly well. In this case the mean of the population being sampled is \( \pi \) and the variance is \( \pi(1 - \pi)/n \) (with its square root being the standard deviation). As the population proportion moves farther toward either 0 or 1, the binomial becomes too skewed to use the normal. The binomial distribution shown in Fig. 4.3E has a parameter of 1/3, and already we can see the beginning of a skew.

Poisson distribution

In the case of a rare event the population parameter will be very near 0 or 1; the large-sample binomial can be adequately approximated by the Poisson distribution, named after the French mathematician, Siméon Denis Poisson, who published a theory on it in 1837. The Poisson distribution arises in cases in which there are many opportunities for an event to occur, but a very small chance of occurrence on any one trial. Exposure to a common but not very infectious virus should result in an incidence of illness that follows the Poisson. For quite large samples, again there exists a normal approximation. In this case the mean of the population being sampled remains \( \pi \) but the variance becomes \( \pi/n \) (with its square root being the standard deviation).

Exercise 4.7

If we were to make an inference about the mean of differences between pre-op and post-op plasma silicone levels in DB5, what would be the df? To make this mean into a t statistic, we subtract what value and then divide by what statistic? To decide if \( t \) is significantly larger than 0 or not, we choose a “cut point” greater than which \( t \) is significant and smaller than which, not. What do we call this cut point?

Exercise 4.8

If we were to make an inference about the mean of differences between clinic and lab INR readings in DB13, what would be the df?

Exercise 4.9

If we wished to make an inference about the proportion of patients experiencing any nausea in DB2, what distribution would we be using in this inference?

Exercise 4.10

If we wished to make an inference about proportion of patients in DB12 who died after 2 or less cm of carinal resection, what distribution would we be using?

Joint distributions of two variables

So far, we have looked at distributions of one variable. Can we deal with the distribution of two variables at the same time? Yes, indeed. We can also look at the
distribution of more than two variables, but that will be met in a later chapter. Let us look at it through an example.

**EXAMPLE: EFFECT OF IMMUNOGLOBULIN ON ITP**

Suppose we had platelet counts, say \( x \), on patients with idiopathic thrombocytopenic purpura prior to treatment and then a second count, say \( y \), 24 hours after administering immunoglobulin. We wonder if there is any relationship between \( x \) and \( y \), that is, might it be that the lower the pretreatment level, the less effect there is from the immunoglobulin, or vice versa? If so, a study might be carried out to learn if we can predict the required dosage of immunoglobulin based on the pretreatment platelet count. Or is there no relation?

**Joint frequency distribution**

If we plot the two measures, \( x \) on one axis and \( y \) on the other, we are likely to find points more concentrated in some areas than others. Imagine drawing a grid on the plot, which is lying flat on a table, and making vertical columns proportional in height to the number of points in each grid square. This *joint frequency distribution* is a three-dimensional analog to a two-dimensional bar chart. Where the points are concentrated, we would find a “hill,” sloping off to areas in which the points are sparse. If the hill is symmetrically round (or elliptical with the long axis of the ellipse parallel to one of the axes), any value of \( x \) would lead to a single \( y \) value as the best prediction; that \( y \) value would be the highest point of the hill for that \( x \) and would be the same \( y \) for all \( x \). We would think that there is no predictive capability. In contrast, if the hill is a ridge with a peak extending from the bottom left to the top right of the graph, the high point of \( y \) associated with a value of \( x \) would be different for each \( x \), and we would think that some predictive relationship exists; for example, \( y \) increases as \( x \) increases.

**Relationship between two variables**

There exist statistical methods to assess the relationship between two variables, mostly falling under the topics *correlation* and *regression* (see Chapter 5: Descriptive statistics, Chapter 15: Linear regression and correlation, and Chapter 16: Multiple linear and curvilinear regression and multifactor ANOVA).

The concept of *covariance* is fundamental to all treatments of the relation between two variables. It is an extension of the ideas of variance and standard deviation. Let us think of sampling from the two variables \( x \) and \( y \); we might keep their standard deviations straight by suffixing an indicator name on the appropriate \( \sigma \) or \( s \), for example, \( \sigma_x \) or \( s_x \). The covariance of two variables measures *how they jointly vary, one with respect to the other*. The common symbol is similar to the standard deviation symbol, but with two subscripts indicating which variables are being included. Thus the population and sample covariances would be \( \sigma_{xy} \) and \( s_{xy} \), respectively. Formulas for these concepts and examples appear in Sections 5.1 and 5.3.
Exercise 4.11
In DB3 the two related variables baseline and 5-day theophylline levels vary jointly. If we were to calculate a measure of how one varies relative to the other, it would be named a what?

4.9 APPROXIMATE DISTRIBUTION OF THE MEAN (CENTRAL LIMIT THEOREM)

Often, we want to make inferences about a population mean based on a sample mean. For example, consider our prostate volume data from 301 patients. Each observation is drawn from a population that has a given probability distribution. If we wanted to make conclusions about a single patient, we would use this distribution. However, suppose we want to use the sample data to make conclusions about the average prostate volume of American men. The sample mean is composed of individual volume readings, each of which has a probability distribution. As noted in Section 4.7, the sample mean must also have a probability distribution, but it will be different.

Sample means of continuous data are distributed normal

If the sample in question is drawn from a normal population (e.g., if prostate volumes were distributed normal), the probability distribution of the sample mean is exactly normal. If the sample is drawn from any other distribution (e.g., the prostate volumes are from a skewed distribution), the probability distribution of the sample mean is still approximately normal and converges on normal as the sample size increases. This remarkable result is due to a famous mathematical relationship, named the central limit theorem. Because much of our attention to statistical results is focused on means, this theorem is frequently applied.

The central limit theorem is dramatically illustrated in Figs. 4.5 and 4.6. In Fig. 4.5, we observe how the frequency distribution of means becomes closer to normal as we take means first of 5 observations, then 10, and finally 20. Fig. 4.5A shows 117 prostate-specific antigen (PSA) readings of patients whose PSA leaves little doubt about a biopsy decision—that is, PSA < 4 (no biopsy) or PSA > 10 (definite biopsy). A normal curve with the mean and standard deviation of the PSA data is superposed. Fig. 4.5B shows the frequency distribution of 200 means of 5 observations each drawn randomly (using a computer random number generator) from the 117 PSA readings, with a normal curve superposed. Fig. 4.5C and D shows the same sort of display for means of samples of 10 and 20, respectively. The convergence to the normal can be seen clearly.

Fig. 4.6 illustrates a similar concept but when the observed values are actually binary. In this case, we consider the proportion of individuals over age 65 with a history of coronary heart disease. Clearly, each observation does not follow a normal distribution as it takes on only one of two values (“yes” or “no,” 1 or 0). Fig. 4.6A depicts the observed data when
a sample of size $N = 100$ is taken. Parts (B)—(D) depict the sampling distribution of the sample proportion (mean) if a sample of size $N = 10, 100,$ or $1000$ were drawn from the population, respectively. On each plot, we overlay the normal approximation to the sampling distribution of the mean from the central limit theorem. Notice two things: as the sample size grows, (1) the range of observed sample means narrows, indicating greater precision in our estimate of the true proportion of individuals with coronary heart disease and (2) the normal approximation to the sampling distribution becomes more accurate.

**Measure of variability in the sample mean**

Before we meet the concepts of hypothesis testing, it will be convenient to become familiar with the *standard deviation of the sample mean*. This quantity measures the variability of the sample mean $m$. It has been called the *standard error of the mean* (SEM) for historical reasons that do not have much meaning today.
4.9 Approximate Distribution of the Mean (Central Limit Theorem)

Figure 4.6 Example of the central limit theorem for a binary response variable (an indicator of coronary heart disease in this case). Part (A) depicts the observed numbers of individuals without and with CHD in a sample of size $N = 100$. Parts (B)–(D) depict the sampling distribution of the sample proportion (mean) if a sample of size $N = 10, 100, \text{ or } 1000$ were drawn from the population, respectively. On each plot, we overlay the normal approximation to the sampling distribution of the mean from the central limit theorem. Note that (A) is purposefully shown in a different color to highlight the difference in the observed frequency for the raw PSA values versus the frequency of the sample mean from repeated samples in (B–D). Also notice two things: as the sample size grows, (1) the range of observed sample means narrows, indicating greater precision in our estimate of the true proportion of individuals with CHD and (2) the normal approximation to the sampling distribution becomes more accurate. CHD, Coronary heart disease.

Population standard error of the mean

Let us consider our 301 prostate volumes as a population. The mean of the 301 volumes becomes the population mean $\mu = 36.47 \text{ mL}$ and the standard deviation $\sigma = 18.04 \text{ mL}$. If we randomly draw prostate volumes, they will generally be somewhat different from one another, and these differences will follow some frequency distribution; $\sigma$ is the standard deviation of this distribution. Similarly, if we calculate means $m_1, m_2, \ldots$, from repeated samples, they will generally be somewhat different from one another and will follow some frequency distribution; the SEM is the standard deviation...
of this distribution. It is often symbolized as $\sigma_m$. It will be smaller than $\sigma$, because it reflects the behavior of 301 observations rather than 1. It turns out that

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

(4.6)

the population standard deviation divided by the square root of the population size. In the case of the 301 volumes, $\sigma_m = 18.04/\sqrt{301} = 1.0398$ mL. Because we know from the central limit theorem that the probability distribution of $m$ is normal, we have the distribution of $m$ as $N(\mu, \sigma_m^2) = N(36.47, 1.0812)$.

**Sample standard error of the mean**

In the same way that we can estimate $\sigma$ by $s$ when we do not know $\sigma$, we can estimate $\sigma_m$ by $s_m$, using the standard deviation of a sample. The sample SEM, $s_m$, is the sample standard deviation divided by the square root of the sample size, or

$$s_m = \frac{s}{\sqrt{n}}$$

(4.7)

Let us take the 10 prostate volumes given in Table DB1.1 as our sample. We can calculate $m = 32.73$ mL and $s = 15.92$ mL. Then $s_m = 15.92/\sqrt{10} = 5.3067$ mL.

**Exercise 4.12**

Calculate the sample SEM of the rounded GAG levels (0.7, 0.4, 0.3, 0.2) for type I assay in DB8.

**Exercise 4.13**

In DB10 the difference in time to perform between legs has mean $= 0.16$ and standard deviation $= 0.2868$. Calculate the SEM.

### 4.10 APPROXIMATE DISTRIBUTION OF A SAMPLE QUANTILE

We are all familiar with percentiles. A set of values is divided into 100 parts; the 90th percentile, for example, is the value below which 90% of the data fall. Other quantile types are deciles, in which the set of values is divided into 10 parts, and quartiles, 4 parts. Quantiles are useful descriptors of how a data distribution is shaped. $Q_1$, $Q_2$, and $Q_3$ are the first, second, and third quartiles (one quarter, half, three-quarters of the data fall below). It can be seen that, if $Q_2 - Q_1 \approx Q_3 - Q_2$ (“$\approx$” means “approximately equal to”), the distribution is approximately symmetric; if $Q_2 - Q_1$ and $Q_3 - Q_2$ are quite different, the distribution is skewed.

A *sample quantile* is computed by ordering the observations in a sample from smallest to largest and selecting the value that corresponds to the specified proportion of
observations coming before it. For example, if considering the median, this is the 0.5 quantile or 50th percentile. In this case the sample median would be the “midpoint” of the ordered observations. Hence, we would compute the sample median by choosing the value for which half of the ordered observations precede this value and half of the observations are greater than this value. If the true quantile falls between two observations, a weighted average of those observations used as the estimate of the quantile. In the case of the median falling between two observations, the estimate is just the midpoint between them.

Just as in the case of the sample mean, the sampling distribution of a sample quantile can also be approximated by the normal distribution when the sample size is large. This makes it possible to characterize the precision in our estimate of a population quantile. However, it should be noted that the approximation to the distribution of a sample quantile is not trivial to implement in practice because it relies upon knowledge, or the assumption, of the underlying distribution of the observations. This, in general, is not the case as with the sample mean. As such, approximating the distribution of a sample quantile typically either requires the researcher to make strong and typically untestable assumptions about the distribution that gave rise to the data or to turn to simulation or resampling–based approaches to estimate the sampling distribution of the quantile.
Descriptive statistics

5.1 PURPOSE OF DESCRIPTIVE STATISTICS

Descriptive statistics are numerical summaries of data sets. Concepts for some descriptors were given in Chapter 4, Distributions.

The interpretation of descriptive statistics is too often neglected in favor of the more dramatic statistical testing. The very important purpose of descriptive statistics is to perceive the forces driving the data. They are sort of a statistical X-ray: they allow the user to see some of what is giving rise to the data, that is, what states or events underlie the obscuring variability. They also provide important information regarding the population that has been sampled, and hence what population a statistical analysis will be able to generalize to.

Good descriptive statistics will always relate to the scientific goal of one’s analysis. For example, if the goal is to compare responses between groups, descriptive statistics will summarize important covariates between the groups. This provides the analyst with a means to assess differences that may confound the relationship of interest. If present, such confounding factors would need to be adjusted for in a comparison of responses between groups in order to make a fair comparison. Descriptive statistics can also highlight potential effect modification, a concept discussed in Chapter 2, Planning analysis: how to reach my scientific objective. Last, but certainly not least, descriptive statistics can reveal problems with the data that may otherwise go unnoticed. A good descriptive analysis will present the range and distribution of covariates in the data set, highlighting irregular, improbable, or impossible values. In the latter case, further investigation of the data-collection procedure would be warranted.

5.2 NUMERICAL DESCRIPTORS, ONE VARIABLE

Section format

In this section, formulas for these summary statistics are given so that the user will know just how they are calculated and will have a convenient reference to verify the calculation method used. The symbol representing a concept is given, followed by the formula.

When the number $n$ of observations is indicated, $n$ refers to the sample size when the formula is calculated from a sample, and to the population size when the formula
is calculated from a population. The distinction between samples and populations is usually clear by context and is stated if not clear.

**Quantiles**
Quantiles were met in Section 4.10. A quantile gives the proportion of data lying below it, estimating the corresponding proportion of the underlying probability distribution. A set of quantiles provides some picture of the shape of a distribution. The most frequently used quantiles are quartiles, deciles, and percentiles. In medicine, PPT, that is, parts per thousand, is sometimes seen. It might be thought of as a “kilotile.”

**Rate**
Perhaps, the most frequently used descriptor is the *rate*, the relative frequency of occurrence of an event. As we saw in Chapter 3, Probability and relative frequency, the rate of an event is found as the number of occurrences of that event divided by the number of occurrences of any event. If you have 20 patients with migraine headaches and 12 experience relief when treated with ibuprofen, the rate of relief from ibuprofen is 0.60.

**Mean**
μ denotes a population’s (arithmetic) mean, whereas *m* denotes a sample’s (arithmetic) mean. The two have the same calculation, so the difference is whether *n* represents the size of a population or a sample. The sample mean *m* sometimes appears in the literature represented as the variable with an overbar, such as \( \bar{x} \). Recall that “\( \Sigma \)” means “add together all data elements whose symbol follows me.” Thus if the variable “\( x \)” contains the three data elements, 1, 2, and 4, then “\( \Sigma x \)” implies “1 + 2 + 4.”

\[
\mu = \frac{\sum x}{n} \quad \text{or} \quad m = \frac{\sum x}{n},
\]

depending on whether *n* denotes the number in a population or in a sample, respectively. (The usual mean is called arithmetic because the values are added, in contrast to the geometric mean, a form used only in some special cases, in which they are multiplied.)

**EXAMPLE: PROTEIN-SPECIFIC ANTIGEN FROM TABLE DB1.1**
Consider the mean protein-specific antigen (PSA) value from Table DB1.1:

\[
m = \frac{(7.6 + 4.1 + \cdots + 7.9)}{10} = \frac{67.5}{10} = 6.75.
\]
Median
The median, having no standard symbol (sometimes \( m_d \) is used, and, rarely, \( \tilde{m} \)), may be found for either a population or a sample. The following formula is an algorithm in words rather than in symbols:
1. Put the \( n \) observations in order of size.
2. Median is the middle observation if \( n \) is odd.
3. Median is the halfway between the two middle observations if \( n \) is even.

It will be apparent that the median is also the second quartile (\( Q_2 \)) and 50th percentile. The median frequently is used to represent the average of asymmetric data sets, such as survival data, where occasional long survivors skew the data, rendering the mean poorly descriptive of the typical patient.

**EXAMPLE: PROTEIN-SPECIFIC ANTIGEN FROM TABLE DB1.1**
Let us put the 10 PSA observations from Table DB1.1 in order of size: 4.1, 4.4, 5.9, 6.1, 6.8, 7.6, 7.7, 7.9, 8.0, and 9.0. \( n \) is even. The two middle observations are 6.8 and 7.6. The median is halfway between them: 7.2.

Mode
The mode, also having no standard symbol (sometimes \( m_o \) is used), may be found for either a population or a sample, provided that \( n \) is large. The mode can be read from a histogram but will be approximate, depending on the choice of the histogram's interval widths and starting point.
1. Make a histogram of the data.
2. Mode is the center value of the highest bar.

**EXAMPLE: PROTEIN-SPECIFIC ANTIGEN FROM TABLE DB1.1**
We can again use PSA values from Table DB1.1 as an example, although a mode from 10 observations would not be used in practice. The mode requires a large number of observations to have any accuracy and is not often used in clinical studies. A histogram of the observations with a bin width of 1 would appear as in Fig. 5.1. The mode is the center of the tallest bar, 7.5.

Range
The range is given by the largest value minus the smallest value, telling the region of the \( x \)-axis covered by the data. For example, the range of temperatures (°F) in a sample of patients might be \( 105^\circ - 97^\circ = 8^\circ \). The smallest (97°) and largest (105°) values are often of clinical interest in themselves. However, the range is not a good measure of the spread of a distribution because it consists of the two most extreme and
therefore usually unstable values in the sample. We might have one very sick patient, while the others are contained in a well-behaved bell curve extending from 97\(^\circ\) to 100\(^\circ\) with a range of 3\(^\circ\).

**Variance and standard deviation**

The variance of a population is denoted \( \sigma^2 \). Conceptually, it is the average of the squares of differences between the observations and their mean: \( \sigma^2 = \Sigma (x - \mu)^2/n \). It may be calculated more easily by the equivalent form

\[
\sigma^2 = \frac{\Sigma x^2 - n\mu^2}{n}.
\]  

The variance of a sample is denoted \( s^2 \). It is similar, with \( m \) replacing \( \mu \) and using \( n - 1 \) in place of \( n \) to avoid a theoretical bias. The sample form equivalent to Eq. (5.2) is

\[
s^2 = \frac{\Sigma x^2 - nm^2}{n - 1}.
\]  

The standard deviations \( \sigma \) and \( s \) are just the square roots of the respective variances.

**EXAMPLE: PROTEIN-SPECIFIC ANTIGEN FROM TABLE DB1.1**

The mean of the PSA values from Table DB1.1 was \( m = 6.75 \). The variance may be found by
\[ s^2 = \left( \frac{7.6^2 + 4.1^2 + \cdots + 7.9^2 - 10 \times 6.75^2}{9} \right) \]
\[ = \left( \frac{57.76 + 16.82 + \cdots 62.41 - 10 \times 45.5625}{9} \right) \]
\[ = \left( \frac{478.890 - 455.625}{9} \right) = \frac{23.265}{9} = 2.585. \]

The standard deviation is just the square root of the variance, or
\[ s = \sqrt{2.585} = 1.6078. \]

**Standard error of the mean**

The standard error of the mean (SEM) was introduced in Section 4.9. It is simply the standard deviation of the value \( m \), the mean of \( n \) observations. If \( n \) is the population size, the SEM is symbolized as \( \sigma_m \); if \( n \) is the sample size, the (estimated) SEM is symbolized as \( s_m \). They are calculated using the standard deviations of the observations, \( \sigma \) or \( s \), as appropriate.

\[ \sigma_m = \frac{\sigma}{\sqrt{n}} \text{ or } s_m = \frac{s}{\sqrt{n}}. \]  \hspace{1cm} (5.4)

**EXAMPLE: PROTEIN-SPECIFIC ANTIGEN FROM TABLE DB1.1**

From the 10 PSA values of Table DB1.1, \( s \) was found to be 1.6078. The estimated SEM is \( 1.6078/\sqrt{10} = 1.6078/3.1623 = 0.5084 \).

**Standard error of the mean for two independent samples**

In many cases, we are faced with two samples that may be drawn from the same population, and we want to know if their means are different. To assess this, we need to use the standard deviation of the mean (SEM) using the information from both samples. If the standard deviation of observations from that population is known, say \( \sigma \), the SEM is simply \( \sigma \) divided by a sample size term, for example,

\[ \sigma_\mu = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \]  \hspace{1cm} (5.5)

If, however, we do not know the population standard deviation, which is usually the case, we must estimate it from \( s_1 \) and \( s_2 \), the sample standard deviations calculated from the two sets of observations. This requires two steps: (1) use \( s_1 \) and \( s_2 \) to find the standard deviation of the pooled observations, say \( s_p \), and (2) then find the SEM,
CHAPTER 5 Descriptive statistics

say \(s_m\), in a form similar to Eq. (5.5). The algebra is worked backward from the formulas for the two-sample standard deviations to find \(s_p\), as we would have calculated it if we had pooled all the observations at the outset. It turns out to be

\[
s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{5.6}
\]

Then, \(s_m\) is

\[
s_m = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5.7}
\]

EXAMPLE: PROTEIN-SPECIFIC ANTIGEN FROM TABLE DB1.1

Consider again the PSA values from Table DB1.1. Suppose, we had taken two samples: the first four \((n_1 = 4)\) and the remaining six \((n_2 = 6)\) observations. Their standard deviations are \(s_1 = 2.1205\) and \(s_2 = 1.3934\). Substituting these values into Eq. (5.6), we obtain

\[
s_p = \sqrt{\frac{3 \times 2.1205^2 + 5 \times 1.3934^2}{8}} = \sqrt{2.8997} = 1.7029.
\]

The 1.7029 estimate is slightly larger than the 1.6078, we obtained from the original 10 observations because we had to allocate a degree of freedom for each of two means (one appearing in \(s_1\) and the other in \(s_2\)) and had to divide by \(n - 2\) rather than \(n - 1\).

\[
s_m = 1.7029 \times 0.6455 = 1.0992.
\]

Interquartile range

If the distribution of interest is notably asymmetric, the standard deviation loses much of its desirable characteristics as a measure of spread. The interquartile range \((IQR)\) is an alternative, not tied to symmetry. The IQR is \(Q_3 - Q_1\), the central half of data whatever shape the distribution may have. There is no direct relationship between the standard deviation and the IQR; it will be different for every distribution. In the normal probability distribution, \(m \pm 1 \sigma\) encloses 68% of the area under the curve while the IQR encloses 50%. Half of the IQR equals \(0.675\sigma\). This is not helpful because we would be unlikely to use the IQR if the distribution were normal. The IQR must be interpreted in its own right for each distribution.
EXAMPLE: EFFECT OF EXERCISE ON ENO
Examine the difference between eNO before exercise and 5 minutes after in DB14. The mean is 4.2 while the median (Q₂) is 1.8. Q₁ = -0.8 and Q₃ = 7.1, so IQR = 7.9. Compare this to m - 1σ = -3.0 and m + 1σ = 11.3. 8% of the data lie to the left, but 16% to the right of the m ± 1σ interval, showing the effect of asymmetry. More on this topic will be discussed in Chapter 7, Hypothesis testing: concept and practice.

Exercise 5.1
From DB7, find the (a) mean, (b) median, (c) variance, (d) standard deviation, (e) first quartile, (f) third quartile, and (g) SEM of the bone density. Round all answers to one decimal place after calculation.

Exercise 5.2
From DB10, find the (a) mean, (b) median, (c) variance, (d) standard deviation, (e) first quartile, (f) third quartile, and (g) SEM of the distance covered in a triple hop on the operated leg.

Exercise 5.3
From DB15, find the (a) mean, (b) median, (c) variance, (d) standard deviation, (e) first quartile, (f) third quartile, and (g) SEM of temperatures at depth 1 for each treatment.

Exercise 5.4
From DB12, find the mode age.

Exercise 5.5
From DB10, find the two-sample SEM of distance covered using the operated leg sample and the nonoperated leg sample.

5.3 NUMERICAL DESCRIPTORS, TWO VARIABLES

Covariance
When we consider two measures that vary simultaneously, each one alone can be described by the methods of the preceding section, but we are often interested in how they vary jointly, or “covary,” one relative to the other. This covariance, introduced in Section 4.8 in the context of joint frequency distributions, requires paired recordings, that is, a reading on y for each reading on x. The calculation of the covariance is similar to the variance, except that instead of squaring the x – mean term, we use the x – mean term times its paired y – mean term, as \( \sigma_{xy} = \sum (x - \mu_x)(y - \mu_y)/n \). These forms, for the population and sample covariances, respectively, become
\[
\sigma_{xy} = \frac{\sum xy - n\mu_x \mu_y}{n}
\]  
(5.8)

and

\[
s_{xy} = \frac{\sum xy - nm_x m_y}{n - 1}
\]  
(5.9)

**EXAMPLE: PROSTATE VOLUME WITH AGE**

Let age from Table DB1.1, take the \( x \) position in Eq. (5.9) and prostate volume take the \( y \) position.

\[
s_{xy} = \frac{(75 \times 32.3 + 68 \times 27.0 + \cdots + 74 \times 16.4 - 10 \times 65.1 \times 32.73)}{9}
\]

\[
= \frac{(2422.50 + 1836.00 + \cdots + 1213.60 - 10 \times 2130.723)}{9}
\]

\[
= \frac{(21,211.40 - 21,307.23)}{9} = \frac{-95.83}{9} = -10.6478.
\]

**INTERPRETATION**

If one variable tends to increase as the other increases, such as systolic and diastolic blood pressure, the covariance is positive and large; large values of one are multiplied by large values of the other, which makes a very large sum. If one tends to decrease as the other increases, as with PSA and prostate density, the covariance is negative and large; large positive values of one multiply large negative values of the other. Conversely, if increases and decreases in one variable are unrelated to those of the other, the covariance tends to be small.

**Correlation coefficient for continuous variables**

The covariance could be very useful in indicating a shared behavior or independence between the two variables, but there is no standard for interpreting it. The covariance can be standardized by dividing by the product of standard deviations of the two variables. It is then called the (Pearson) **correlation coefficient**, designated \( \rho \) (Greek rho, rhymes with snow) for a population correlation and \( r \) for a sample correlation. Thus the formulas for calculating correlation coefficients \( \rho \) (population) or \( s \) (sample), respectively, are
5.3 Numerical Descriptors, Two Variables

\[ \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \text{ or } r_{xy} = \frac{s_{xy}}{s_x s_y} \]  

(5.10)

**INTERPRETATION**

This standardized covariance, the correlation coefficient, may be easily interpreted. If either variable is perfectly predictable from the other, the correlation coefficient is 1.00 when they both increase together and −1.00 when one increases as the other decreases. If the two variables are independent, that is, a value of one provides no information about the value of the other, the correlation coefficient is 0. A correlation coefficient of 0.10 is low, showing little predictable relationship, whereas 0.90 is quite high, showing that one increases rather predictably as the other increases. However, we have no standard to determine what is considered “small” or “large.”

**EXAMPLE: PROSTATE VOLUME WITH AGE**

Continuing the age and prostate volume example from before, we note that the standard deviation of age is \( s_x = 6.9992 \) and that of volume is \( s_y = 15.9351 \). The covariance was calculated as \( s_{xy} = -10.6478 \). Then

\[ r_{xy} = -\frac{10.6478}{(6.9992 \times 15.9351)} = -0.0956. \]

This result would tell us that, for our 10 patients in Table DB1.1, volume tends to decrease as age increases, but in a very weak relationship. We should note that \( n = 10 \) is too small a sample to provide us with much statistical confidence in our result.

**Caution**

We must remember that correlation methods measure relationship only along a straight line. If one variable increases when the other increases but not in a straight line the correlation may not be high despite good predictability. Some examples of nonlinear relationships include when weight is predicted by height (weight is a power of height) or when teenage growth depends on time (growth is a logarithm of time). More sophisticated prediction methods are needed for curvilinear relationships. Elementary information on this topic appears in Chapter 16, Multiple linear and curvilinear regression and multifactor analysis of variance.

Furthermore, this linear correlation measures only the pattern of data behavior, not interchangeability of data. For example, temperature measured in the same patients using one thermometer in degrees Celsius and another in degrees Fahrenheit would have an almost perfect correlation of 1.0, but the readings could not be intermingled.
Exercise 5.6
From DB10, find (a) the covariance and (b) the correlation coefficient of distance covered between the operated and nonoperated legs.

Exercise 5.7
From DB10, find (a) the covariance and (b) the correlation coefficient of seconds to perform the hops between the operated and nonoperated legs.

Exercise 5.8
From DB14, find the correlation coefficient of eNO between before exercise and 20 minutes after.

Correlation coefficient for rank-order variables
Do two rank-ordered variables have a correlation coefficient? Yes, indeed. The concept and interpretation are not remarkably different from that for continuous variables, so we may think about it in much the same way. The calculation is different. We start with two columns of paired readings, say \( x_i \) and \( y_i \). We replace each column with the ranks of its readings and make a third column \( d_i^2 = (\text{rank of } i\text{th } x - \text{rank of } i\text{th } y)^2 \) and sum the \( d_i^2 \). The rank correlation formula appears as Eq. (5.11). The subscript \( s \) stands for its developer, Charles Spearman.

\[
rs = 1 - 6\sum d_i^2 \over n(n^2 - 1).
\]  

(5.11)

EXAMPLE: PERFORMANCE MEASURE ON AN OPERATED LEG
As an example, consider the time (second) to perform a triple hop between the operated and nonoperated legs from DB10. The data appear as Table 5.1.

Table 5.1 Ranks of hop times on operated and nonoperated legs from DB10, with rank difference and rank difference squared.

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Operated rank</th>
<th>Nonoperated rank</th>
<th>Rank difference</th>
<th>Rank difference squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>8</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>6</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The sum of \( d_i^2 \) = 6. By substituting in Eq. (5.11), we find \( r_s = 0.929 \). The ability to perform is highly correlated between legs.

Rank correlation is examined further in Chapter 14, Measuring association and agreement.

**Exercise 5.9**

Calculate the rank correlation coefficient for hop distances between the two legs in DB10.

**Correlation coefficient for categorical variables: tetrachoric correlation**

Measuring correlation for categorical variables is quite different in both concept and calculation. Given two variables with joint frequencies of occurrence, what is an estimate of the association between the two variables? A measure named tetrachoric correlation, existing for many decades, was fully developed theoretically in 1984.\(^1\) The tetrachoric correlation may be calculated by statistical software when available, but not all software packages provide it. The formula will be given here in case it is needed.

**EXAMPLE POSED: EFFECT OF DRUG ON POSTOPERATIVE NAUSEA**

As an example, consider the second table in DB2: frequency of occurrence of no nausea versus nausea following gall bladder surgery against frequency of use of Drug versus Placebo. Is there a correlation between drug and freedom from nausea?

**A 2 × 2 CONTINGENCY TABLE**

A table showing the frequency of occurrence of possibilities in one variable as contingent upon having occurred in the other variable is called a contingency table. These tables are considered at greater length in Section 9.1. At present, we just need to know what one is and, to permit calculation formulas, what we name the entries in the table. **Table 5.2** shows a typical contingency table with treatment or not on the left and desired result or not at the top. The entries of the cells in the table are numbers of occurrence, denoted \( n \) with subscripts showing location, the first subscript denoting row number and the second, column number. Thus the number of occurrences in row 1, column 2 is \( n_{12} \). The marginal numbers, or totals, have the number

**Table 5.2** A sample contingency table with notation for cell entries.

<table>
<thead>
<tr>
<th></th>
<th>Desired result</th>
<th>Undesired result</th>
<th>Marginal total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>( n_{11} )</td>
<td>( n_{12} )</td>
<td>( n_{1} ).</td>
</tr>
<tr>
<td>Placebo</td>
<td>( n_{21} )</td>
<td>( n_{22} )</td>
<td>( n_{2} ).</td>
</tr>
<tr>
<td>Marginal total</td>
<td>( n_{.1} )</td>
<td>( n_{.2} )</td>
<td>( n_{..} )</td>
</tr>
</tbody>
</table>
over which the row or column is totaled replaced by a dot, as the sum of the first column is \( n_{11} \). The overall total would then be \( \sum n_{ij} \) or just \( n \).

A basic component of the calculation is the upper left to lower right product divided by the lower left to upper right product. (This happens to be the odds ratio for the relationship, further discussed in Section 10.1.) Let us define a calculation

\[
C = \left( \frac{n_{11}n_{22}}{n_{12}n_{21}} \right)^{\pi/4},
\]

where \( \pi \) is the circular constant 3.14159\ldots and \( \pi/4 \approx 0.7854 \). Then the tetrachoric correlation \( r_t \) between treatment and result is given by

\[
r_t = \frac{C - 1}{C + 1} \quad (5.12)
\]

and its variance by

\[
\text{var}(r_t) = \left[ \frac{\pi C}{2(1+C)^2} \right]^2 \left( \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right) \quad (5.13)
\]

(Limiting cases: If \( n_{11} \) or \( n_{22} \) = 0, \( r_t \) is taken as \(-1\); \( n_{12} \) or \( n_{21} \) = 0, \( r_t \) is taken as \(+1\).)

**EXAMPLE CONTINUED: EFFECT OF DRUG ON POSTOPERATIVE NAUSEA**

The gall bladder surgery nausea data from DB2 form Table 5.3. \( C = \left( \frac{34 \times 19}{(22 \times 6)} \right)^{\pi/4} = 3.4806 \). \( r_t = 0.55 \). We note that there is a substantial correlation between using the drug and relief from nausea. The variance of \( r_t \) may be found as 0.02182, with its square root 0.1477 as standard deviation.

**Exercise 5.10**

For DB6, calculate the tetrachoric correlation between type of tattoo ink and ease of removal.

### 5.4 NUMERICAL DESCRIPTORS, THREE VARIABLES

In Chapter 2, Planning analysis: how to reach my scientific objective, we discussed the concept of effect modification. Recall that if the relationship between a predictor of interest and an outcome varies by the level of a third variable, then the third variable

<table>
<thead>
<tr>
<th>Drug</th>
<th>No nausea</th>
<th>Nausea</th>
<th>Marginal total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>34</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>Placebo</td>
<td>22</td>
<td>19</td>
<td>41</td>
</tr>
<tr>
<td>Marginal total</td>
<td>56</td>
<td>25</td>
<td>81</td>
</tr>
</tbody>
</table>

Table 5.3 The contingency table for the gall bladder surgery nausea data.
is termed an effect modifier. The purpose of numerical descriptors across three variables is generally to investigate and highlight potential effect modification. This is most commonly performed by stratifying or separating the data by the levels of the potential effect modifier. For example, we previously looked at the correlation between prostate volume and age. One may wish to assess whether the association between prostate volume and age changes depending upon the ethnicity of the patient. In this case a stratified descriptive analysis would separately report the correlation estimate between prostate volume and age in each ethnicity group. A similar approach can be used when the predictor of interest and the response are categorical. We previously discussed a $2 \times 2$ contingency table, which depicts the frequency of occurrence of possibilities in one variable as contingent upon having occurred in the other variable. Stratified contingency tables mean that a separate contingency table is constructed for each level of a third covariate, thus allowing one to assess whether the frequency of events in the table varies across the third covariate.

In later chapters, we will assess ways to formally test whether an association changes depending upon the level of a third variable. In regression models, this will be accomplished by including what are termed interaction effects. In contingency tables, tests for a common association across the stratified tables will be constructed.

### 5.5 Graphical Descriptors, One Variable

**Common types**

Graphs and charts allow us to visualize distributions and other properties of data. From a chart, we can often get a rough idea of a mean, a standard deviation, or a proportion. Although there are several types of charts, the most common are the histogram (seen in Chapter 3: Probability and relative frequency and Fig. 5.1), the pie chart, the line chart, and, for two variables, the scattergram. Currently, the mean-and-standard-error chart and the box-and-whisker chart are also seen frequently in medical literature.

**Making a histogram**

When forming a histogram, the choice of intervals is important. It is, to some extent, an art. An unfortunate choice of intervals can change the apparent pattern of the distribution. Enough intervals should be used so that the pattern will be affected minimally by altering the beginning and ending positions of an interval. The beginning and ending positions should be chosen to be convenient in reading the histogram and should relate to the meaning of the variable being charted. Recall the prostate volume data depicted in Fig. 3.3F. **Fig. 5.2A** shows the prostate volume data allocated to 6 intervals rather than the 24 shown in Fig. 3.3F; useful information is obscured by the
lack of detail. Fig. 5.2B shows the data allocated to 48 intervals; the viewer is distracted by too much detail. Furthermore, if we had fewer data, say only 50 or so values, the excess number of intervals would obscure the distribution pattern as badly as too few. The choice of number, width, and starting points of intervals arises from the user’s judgment. They should be considered carefully before forming a histogram.

**Histograms with differing interval widths**

The histogram is constructed so that the number of observations lying in an interval is represented by the *area* of a rectangle (or bar) rather than its height. If all intervals are of equal width, the histogram is no different from a bar chart (except perhaps cosmetically) which simply displays the frequency of counts across discrete groups. However, if one interval had been chosen that is twice the width of the others, the rectangle height over that interval must be half the height it would be in a bar chart. This area-in-lieu-of-height presentation avoids a possible misinterpretation by the viewer. For example, if the prevalence of a disease in an epidemic is displayed monthly for several months, but the last rectangle represents only 15 days, its height per infected patient would be doubled in a histogram to convey the true pattern of prevalence.

**EXAMPLE: PROSTATE VOLUMES**

Fig. 5.3 illustrates the requirement for using areas in lieu of heights for unequal intervals. Suppose in recording prostate volumes, the intervals had been organized by 5-mL increments until 65 mL, and a final interval of 65–115 was used. Then the bar chart of volumes (see Fig. 3.3F) would be amended so that the last bar would have height 28, as in Fig. 5.3A. If we adjust by converting the height to area, we obtain Fig. 5.3B, which gives a more accurate depiction.
Pie chart

A pie chart represents proportions rather than amounts. Its main use is to visualize the relative prevalence of a phenomenon rather than its absolute prevalence. It also has the advantage of avoiding the illusion of sequence that sometimes is implied by the order of bars in a bar chart, regardless of whether it was intended or not. To draw a pie chart, the user must allocate the 360 degrees of a circle to the components in proportion to their prevalence. A prevalence of 20% is shown by $0.2 \times 360$ degrees = 72 degrees of angle about the center of the circle.

**EXAMPLE: ACCURACY OF DIGITAL RECTAL EXAMINATION**

Let us look at the prediction of biopsy result by digital rectal examination (DRE) from Table DB1.1. There are 30% true negatives (DRE $-$, BIOP $-$), 20% false negatives (DRE $-$, BIOP $+$), 10% true positives (DRE $+$, BIOP $+$), and 40% false positives (DRE $+$, BIOP $-$) results. Fig. 5.4 visually conveys these percentages of results in a pie chart. The first piece of pie includes 30% of 360 degrees = 108 degrees.

Line chart

In a bar chart, if we connected the center of the bar tops by line segments and then erased the bars, we would have a form of line chart. The main use of a line chart is to convey information similar to a bar chart but for intervals that form a sequence of time or order of events from left to right. In Fig. 3.3F, we intend no progression of frequency in logical sequence as prostate volumes increase. In contrast, the frequencies of patients per age follow an interesting pattern as age progresses.

**EXAMPLE: PROSTATE ABNORMALITIES BY AGE**

Fig. 5.5 depicts frequencies for our 301 DB1 patients for 5-year age intervals. We note that the frequencies increase successively to about the mid-60 and then decrease.
successively thereafter. It is not difficult to conjecture the forces causing this data pattern in these patients who have presented for possible prostate problems. Prostate problems are rare in young men and increase in frequency with age. However, starting in the late 60s, men are dying of other causes in increasing numbers; therefore the number of survivors available to present is decreasing with age.

**Figure 5.4** A pie chart representing proportions of biopsy results as predicted by the DRE. *DRE*, Digital rectal examination.

**Figure 5.5** A line chart representing frequencies of men presenting with prostate problems according to age. The points (circles) showing frequencies are positioned over the centers of the defined age intervals and are then connected by straight-line segments.
RELATION OF A LINE CHART TO A PROBABILITY DISTRIBUTION

Conceptually, it is important to observe that as the sample size increases and the width of the intervals decreases, the line chart of a sample distribution approaches the picture of its probability distribution.

Mean-and-standard-error chart

A diagram showing a set of means to be compared, augmented by an indication of the size of uncertainty associated with each mean, is appearing more and more frequently in medical articles.

SHOWING THE MEANS

When a relationship among several groups is of interest, a lot of information is given by a plot in which groups are given positions on the horizontal axis and means are shown by vertical height above each position. Increasing time or increasing amount of a drug might appear on the horizontal axis. The clinical response to this sequence is shown by the mean. For example, postoperative pain rated by patients on a visual analog scale may be related to the amount of pain-relieving drug used. Four standard levels of drug dose, starting with zero, could be positioned on the horizontal scale and mean pain rating could be shown vertically over each respective dose. In this case, we would expect the means to be decreasing with increasing dose. In other cases, they could be increasing or going up then down or even scattered in no perceptible pattern. We can tell a great deal about the process going on from the pattern.

SHOWING THE VARIABILITY ABOUT THE MEANS

We must ask whether, for example, the downward pattern of pain with increasing drug is meaningful, because we can make it look huge or minuscule by altering the vertical scale. A crucial part of the information is how different the means are relative to the variability in the data. The means may be decreasing, but with such small decrements relative to the data variability that it might have happened by chance and we do not accept the decreasing pattern as shown to be accurate. A useful solution is to show the associated uncertainty as “whiskers” on the means, that is, as lines up and down from the mean indicating variability by their lengths. This variability depicted may be standard deviation, standard error, or some related measure.

EXAMPLE: DISTRIBUTION OF PROSTATE VOLUMES

Fig. 5.6 shows prostate volume for 291 patients in the 50- to 89-year age range separated into decades of age: 50s, 60s, 70s, and 80s. The means are shown by solid circles. The whiskers indicate about two standard errors above and below the mean, which includes 95% of the data on an idealized distribution (see Chapter 4: Distributions, for a more detailed discussion of this topic). We can see by inspection that the mean
volumes appear to increase somewhat by age, but that there is so much overlap in the variability that we are not sure this increase is a dependably real phenomenon from decade to decade. However, we would take a small risk in being wrong by concluding that the increase from the youngest (50s) to the oldest (80s) is a real change.

**EFFECT OF IRREGULAR DATA**

Does this chart tell the full story? If the data per group are distributed in a fairly symmetric and smooth bell-type curve, most of the relevant pattern may be discerned. However, if the data are distributed irregularly and/or asymmetrically, this chart actually covers up important relationships and may lead to false conclusions. That is because the assumptions of regularity and symmetry are made for this chart and, as usual, relationships are distorted when assumptions are violated. Charts that are “data dependent” rather than “assumptions dependent,” such as the box-and-whisker charts discussed next, often will provide a better understanding of the data.

**Box-and-whisker chart**

A practical way to explore and understand available data is to diagram them such that they display not only the typical aspects (e.g., distribution center and spread) but also atypical characteristics (e.g., asymmetry, data clumps, and outlying values). The box-and-whisker chart does this rather well, although no technique displays all the vagaries of data. Fig. 5.7 shows such a diagram, representing aspects of the distribution of age for the 301 urologic patients. As indicated in the diagram, the box includes the 25th to the 75th percentiles of data, with the median (the 50th percentile) as an intermediate
line. If the median is in the center of the box, the middle half of the data is nearly symmetric with the median not far different from the mean. An off-center median indicates asymmetry. The whiskers extend the plot out to another indicator of spread, marked by “fences” at the end of the whiskers, somewhat closer to the tails of the data distribution. This indicator might be range minus outliers or a defined percentile enclosed. In Fig. 5.7, it is 3/2 of the box height above the top of the box and below the bottom of the box, which would include about 99% of the data in a large normal sample. Whisker lengths that are similar and are about half the semi box length are further evidence of symmetry and a near-normal distribution. Unequal whisker lengths indicate asymmetry in the outer parts of the data distribution. Whisker lengths shorter or longer than 3/2 the box length indicate tendency toward a “flat-topped” or “peaky” distribution. A short whisker attached to a long box portion or vice versa shows evidence of “lumpy” data. Finally, the presence of data far out in the tails, as well as their distance out, is shown by dots above and below the whisker ends in Fig. 5.7. Other characteristics can be detected with a little experience in using the box-and-whisker chart.

**EFFECT OF NO ASSUMPTIONS ABOUT THE DISTRIBUTION**

Charts such as the box-and-whisker with no distributional assumptions, such as was used in the mean-and-standard-error chart, are completely dependent on the data available at that moment, and general distributional characters cannot be easily inferred; they are designed to describe the sample, not the population. This is at once their great strength and their great weakness. They must be used with care.
EXAMPLE: DISTRIBUTION OF PROTEIN-SPECIFIC ANTIGEN LEVELS

Consider the sample of 291 urologic patients with PSA levels of 50 ng/mL or less included in the 50- to 89-year age range separated into decades of age: 50s, 60s, 70s, and 80s. Readings above the line drawn at PSA level 10 show a high risk of prostate cancer. PSA, Protein-specific antigen.

SHOWING SAMPLE SIZE

One piece of information lacking from the chart in Fig. 5.8 is sample size. Might some of the differences be attributable to a small number of patients in the group? A variation on the box-and-whisker chart is to draw the width of the box proportional to sample size so that relative sample sizes (not actual numbers) can be compared. Fig. 5.9 is a redrawing of Fig. 5.8 with box widths proportional to sample size. The smaller size in the 50s and especially in the 80s lends less credence to the results for these age decades.

Exercise 5.11

From DB7, construct a histogram of patient ages.
Exercise 5.12
From DB7, construct a histogram of bone density in the intervals 80—< 140, 140—< 160, 160—< 180, and 180—< 200.

Exercise 5.13
From DB7, construct pie charts of patient (a) sexes and (b) ages (grouped 17—19, 20—22, 23—25, >25).

Exercise 5.14
From DB14, construct a histogram of the 20-minute eNO change.

Exercise 5.15
From DB14, construct a histogram of the 20-minute eNO change with all values greater than 5 in one interval.

Exercise 5.16
From DB14, construct a pie chart of the four categories: male with EIB, female with EIB, male without EIB, and female without EIB.

Exercise 5.17
From DB11, construct a line chart of number rats surviving malaria by day number for the three treatments.

Exercise 5.18
From DB14, construct a line chart for eNO means over time for EIB and not-EIB groups.
CHAPTER 5 Descriptive statistics

Exercise 5.19
From DB3, construct (a) a mean-and-standard-error chart (± 1.96 SEM) and (b) a box-and-whisker chart for serum theophylline levels at baseline, 5 days and 10 days.

Exercise 5.20
From DB14, construct (a) a mean-and-standard-error chart (± 1.96 SEM) and (b) a box-and-whisker chart of eNO for EIB patients across time.

5.6 GRAPHICAL DESCRIPTORS, TWO VARIABLES
Depicting the relationship between variables
Suppose we have two types of readings on each patient, perhaps heart rate and blood oxygen level. We may examine either one alone by single variable methods. What interests us here is to explore how the data behave (i.e., how they are distributed) in the two dimensions simultaneously. Are these variables independent or correlated? If they are not independent, what is the nature of their relationship? Graphical representation will often display subtleties not apparent from summary statistics.

Scatterplot
The simplest depiction is to plot the pair of readings for each patient on perpendicular axes. We can see whether points appear to be randomly scattered or clustered. If clustered, we can see the locations and shape of these clusters.

EXAMPLE: PROSTATE VOLUME BY AGE
We might ask whether there is a relationship between age and prostate volume for patients at risk for prostate cancer. (Let us omit volumes greater than 65 mL, because many of the larger glands are due to benign prostate hypertrophy, biasing the cancer risk group.) In Fig. 5.10, volumes are plotted on the vertical axis and age on the horizontal axis. A pattern of some sort might have suggested a relationship, for example, if prostate volumes tended to reduce with age. However, no obvious pattern appears. If you were given an age, would this plot allow you to suggest an average volume different from the overall average? No. We would say the two variables appear independent.

Exercise 5.21
From DB3, construct a scattergram of serum theophylline level at 10 days depending on level at baseline.

Exercise 5.22
From DB14, construct a scattergram of 20-minute eNO change by age.
5.7 GRAPHICAL DESCRIPTORS, THREE VARIABLES

Two-dimensional frequency distribution

For one variable, we provided its frequency distribution by dividing the axis into intervals and showing the number of cases in each interval (histogram, line chart, etc.). For two variables, we could divide the space into rectangles and count the cases in each rectangle. However, to show the number of cases by heights, we need a third dimension. Although a three-dimensional (3D) solid model could be developed out of plaster, wood, or some other material, it is not very practical; we would like to show the 3D image in two dimensions, that is, on a page. This can be done and there exist computer software packages to assist.

CHALLENGES IN THREE-DIMENSIONAL IMAGING

The user is challenged to manage several characteristics at once, including scaling and aspect. The viewer should be able to read the variable or frequency values for any point in the three dimensions. It is especially difficult to show the height of a column, because most columns are not flush against the scale depiction. The viewer must extrapolate the height from a scale somewhat removed that depends on the aspect angles. In addition, the extrapolation may be distorted by perspective. (Distances appear smaller as they recede from the viewer in a 3D model, so they may be drawn smaller than in a 2D image.) Another issue is the ability to see data farther back in the image that are obscured by data in front. It may be possible to swap ends on one or both axes or to change the aspect. When the data are “bumpy,” it is usually impossible to render an image that shows all the subtleties.

Figure 5.10  A scattergram depicting prostate volumes below 65 mL as related to age of patient.
EXAMPLE OF LOW CORRELATION: PROSTATE VOLUME BY AGE

Fig. 5.11 shows a 3D bar chart of the data from Fig. 5.10. It gives the viewer some perception of the joint frequency distribution of age and prostate volume. Age intervals are chosen as decades, and volume is measured in 10-mL intervals. What is the height of the tallest column? We cannot even see the bottom. A viewer could approximate the height by adroit use of dividers and straight edge. What we can see is that the volume distribution patterns (not the actual column heights) for the respective age decades are not that different. This implies a low correlation between age and volume and, indeed, the calculation of the correlation coefficient yields $r = 0.05$.

EXAMPLE OF HIGH CORRELATION: PROTEIN-SPECIFIC ANTIGEN BY PROSTATE-SPECIFIC ANTIGEN DENSITY

Fig. 5.12 shows a 3D bar chart of PSA against prostate-specific antigen density (PSAD). The plot was limited to PSA levels less than 20 ng/mL, because the few larger PSA values would have compressed the graphical area of interest too much for characteristics to be discernible. The direction of the PSA axis was reversed to avoid obscuring much of the result behind higher bars. Because PSA is a major part of the calculation of PSAD, we would expect a high correlation; the calculation of the
correlation coefficient yields $r = 0.70$. It can be seen that data frequency is high when the two variables lie along the diagonal line from left to right and it is sparse when one is large and the other is small, which is evidence of an association.

**Statistical graphs in exploring data**

The use of pictorial descriptors has two main purposes. First, as descriptive statistics, they present a visual image of the data all at once, which not only helps to describe the interrelationships among the data but also allows the viewer to retain this image. A visual image in memory often provides a more complete picture than does remembering a smattering of table entries, even when the table contains the same information. Second, pictorial summaries may suggest the forces giving rise to patterns of data. These perceived patterns are not sufficient evidence in themselves for scientific conclusions, but they suggest hypotheses to be posed that can then be tested with quantitative results. Graphs and charts are formidable tools in *data exploration*.

**Exercise 5.23**

*From DB14, construct a 3D bar chart of 20-minute eNO change by age using 5-year intervals for age and 10-unit intervals for eNO change.*
5.8 PRINCIPLES OF INFORMATIVE DESCRIPTIVE TABLES AND FIGURES

Good table formatting practices

If organized correctly, statistical tables can be the most informative component of a scientific report. They allow for easy comparison of summary statistics across comparison groups and across multiple variables. However, if not constructed thoughtfully, statistical tables can be overwhelming for a reading leading to confusion and misinterpretation of results.

A few minimum guidelines should be utilized when constructing statistical tables. First, if a table is meant to convey similarities or differences in the distribution of covariates across groups, the comparison groups should be placed across the columns and the covariates to be compared should be placed in the rows. A reader generally finds it more natural to compare values that are side-by-side as opposed to stacked on top of one another. Columns of tables should also be uniformly formatted. Decimal points should line up in all columns and each column should be left- or right-justified. Simple practices like these allow a reader to take in and compare information quickly and efficiently.

Another common problem that often arises in statistical tables is in regard to the number of significant digits used for reporting summary statistics. Far too often, writers use more significant digits than are warranted. For example, consider a study that has collected the age of patients in years. When computing the mean of age in the sample, it would not be appropriate to report the mean age up to 4 significant digits (e.g., 43.6524 years). The data-collection procedure did not afford such precision and hence it is unwarranted and unhelpful. Instead, rounding to a single significant digit (e.g., 43.7 years) would be more appropriate. When reporting a summary statistic for multiple groups, it is also important to always keep the number of significant digits the same in each group.

Proper annotation of tables is also critical. What summary statistic is being reported should be clearly delineated. For example, if the mean $+/-$ the standard deviation is being reported it is critical to clearly state this in the table. Also, units for all variables should be shown. If the mean of systolic blood pressure is reported, one should indicate that the units of measurement are mmHg.

Good graphing practices

Although graphing practices is not a central focus of this book, a few comments may be in order. Books having such a focus are available and the reader is encouraged to seek more complete advice from these sources.

The goal of graphing is to transmit information. When constructing graphs, we are well advised to concentrate on that goal. We should avoid whatever does not
contribute to that goal, because it is likely to distract the viewer from the central message. Data should be presented in a way that relates to clinical understanding, including format and data intervals. The size and darkness of images and labels should be consistent and balanced, with the major components of the message slightly, not greatly, larger and/or darker. Be sure that all components are labeled—axes, numbers on axes, responses from different groups, and so on; if the viewer is left wondering what an image is or means, the overall message is degraded. In his first grand book of several works, Edward Tufte\textsuperscript{2} encouraged minimizing the amount of ink. Avoid displaying two dimensions in a 3D graph, despite such a practice being default in certain software packages. Avoid “cutesy” axes relating to the topic at hand that contribute confusion but no information. For example, the author has seen airline economic data drawn with the axis in the shape of an airplane fuselage and tail, preventing the viewer from reading data off the graph. Graphing practices seen weekly in news magazines and national newspapers tend to attempt drama at the expense of accuracy and are seldom models to emulate. In another grand book, Darrell Huff\textsuperscript{3} showed the use of two dimensions to distort a one-dimension comparison: the height of money bag images represented relative average worker income in two nations, but the viewer perceives the relative area of the bags, roughly the square of height.

REFERENCES


Finding probabilities

6.1 PROBABILITY AND AREA UNDER THE CURVE

Chapter 3, Probability and relative frequency, addressed probability and Chapter 4, Distributions, addressed the common probability distributions. We know that the area under a probability distribution over the entire range of possible values is 1. This chapter tells us how to find the probability associated with a portion of the distribution. As an overly simple example, let us look at the probabilities of the white die with two red faces (Fig. 3.1) from the viewpoint of its distribution rather than the probability definitions. The distribution consisted of the two bars of heights 1/3 \[P(R)\] and 2/3 \[P(W)\], summing to 1. The probability of \(R\) is the portion of the distribution associated with red, which is 1/3 of the area under the distribution. Indeed, \textit{probabilities of occurrences correspond to areas under portions of probability distributions.}

This relationship applies directly to the \textit{risks of error in a clinical decision}. The probability of inferring an occurrence when it is absent, often the probability of a false-positive result on a medical test, is usually denoted by \(\alpha\). The probability of inferring the absence of an occurrence when it is present, often the probability of a false negative, is usually denoted by \(\beta\). Rarely must we calculate these probabilities directly, because we may use computers and/or tables to find them. This chapter deals with finding these probabilities of error.

6.2 THE NORMAL DISTRIBUTION

The standard normal

Recall from Section 4.8 that the normal distribution is a perfect case of the famous bell curve. Although an infinite number of cases of the normal exists (arising from various means and standard deviations), we need deal with only one, the standard normal. Any normally distributed variable or sample of observations becomes a standard normal variable with mean 0 and standard deviation (and variance) 1, symbolized by \(z\), by subtracting the mean from each value and dividing by the standard deviation. Thus \(z\) \textit{represents the number of standard deviations away from the mean}, with positive values to the right of the mean and negative values to the left of the mean. Fig. 6.1 shows a
standard normal distribution with \( z = 1.96 \) (frequently seen in practice) and the corresponding area \( \alpha = 0.025 \) under the curve to the right of that \( z \). (Addition of a similar area in the left tail would provide \( \alpha = 0.05 \) in total.)

Table of the standard normal

Table I (see Tables of Probability Distributions at the back of the book) contains selected values of \( z \) with four areas that are often used: (1) the area under the curve in the positive tail for the given \( z \), that is, one-tailed \( \alpha \); (2) the area under all except that tail, that is, \( 1 - \alpha \); (3) the areas combined for both positive and negative tails, that is, two-tailed \( \alpha \); and (4) the area under all except the two tails, that is, \( 1 - \alpha \). Table 6.1 shows a segment of Table I.

Probability of certain ranges occurring

Fig. 3.4 shows the relative frequency distribution of tumor sizes of 115 liver cancers. A normal curve with the same mean (2.77 cm) and standard deviation (1.01 cm) superposed. Let us denote mean by \( \mu \) and standard deviation by \( \sigma \) for shorthand. We ask what percent of tumors are larger than 5 cm. If we are willing to accept the normal curve as the probability distribution of liver tumor sizes, the probability of a tumor larger than 5 cm is the proportion of curve greater than 5. The value 5 cm lies \( 5 - 2.77 = 2.23 \) cm to the right of \( \mu \). How many \( \sigma \) is that? \( 2.23/1.01 = 2.21 \) \( \sigma \) to the right of \( \mu \). Table 6.1 shows the area in the right tail to the right of 2.20 is 0.014. About 1.4% of tumors are larger than 5 cm.
As further illustration, what percent of tumors are less than 1 cm? 1 cm lies 2.77 − 1 = 1.77 cm or 1.77/1.01 = 1.75 σ to the left of μ. Since the normal curve is symmetric, the area under the curve to the left of μ − 1.75σ is the same as the area to the right of μ + 1.75σ. (This provides the information in a table half as long.) From Table 6.1, 1.75σ yields an area halfway between 0.045 and 0.036, which tells us that about 4% of tumors are less than 1 cm.

Fig. 6.2 shows Fig. 3.4 with the 4% area to the left of 1 cm shaded black. Similarly, we could calculate that 1.4% of the tumors exceed 5 cm.

Using the standard normal

Another way to look at the same information is to transform the data to the standard normal by the relationship $z = (x - \mu)/\sigma$, saving a step in calculation. Because $\mu = 2.77$ cm and $\sigma = 1.01$ cm, tumors greater than 5 cm become $(5 - 2.77)/1.01 = 2.21$, which refers to the same position in the table. A standard normal $z$-value represents the number of standard deviations from the mean.
Critical value

If we wanted to select all patients with tumor sizes in the 5% upper tail, that is, to identify the upper 5%, what would be our separating critical value? By looking at Table I, we find that \( \alpha = 0.05 \) is paired with \( z = 1.645 \). This indicates all tumors of size more than \( \mu + 1.645\sigma = 2.77 + 1.645 \times 1.01 = 4.43 \) cm lie in the upper 5% of the probability curve.

**EXAMPLE: PROPORTION OF CARINAL RESECTION PATIENTS UNDER 30 YEARS OF AGE**

As another example of how to find normal probabilities, consider the age of patients undergoing resection of the tracheal carina (DB12). Of 134 patients, the mean age is about 48 years with a standard deviation (taken as \( \sigma \) for this large a sample) of about 16 years. (Values are rounded to make the calculation easier.) What percentage of patients presenting for carinal resection are younger than 30 years? Age 30 is 18 years less than (i.e., to the left of) the mean, or \( 18/16 = 1.125 \) standard deviations below the mean. Our question becomes what proportion of the curve is more than \( 1.125\sigma \) below the mean? Because the distribution is symmetric, the result will be the same as the proportion more than \( 1.125\sigma \) above the mean. By using this property, we need to have only one tail of the distribution tabulated. We look up \( z = 1.125 \) in Table I or Table 6.1. \( z \) lies one-fourth of the way from 1.10 to 1.20, which yields a probability approximately one-fourth of the way from 0.136 to 0.115, or about 0.136 – 0.005 = 0.131. Thus the chance of a patient younger than 30 years presenting for carinal resection is about 13 out of 100.
Exercise 6.1
For a certain population of young healthy adults, diastolic blood pressure (DBP) follows a normal distribution with $\mu = 120$ and $\sigma = 5$ mmHg. You have a patient with DBP = 126 mmHg. What percentage of the population has higher DBP measurements?

6.3 THE $t$ DISTRIBUTION

Why we need $t$

Often we need to look at a confidence interval on a mean when we do not know the population standard deviation and must estimate it from a small sample. In this case the normal distribution does not apply. We need the $t$ distribution, introduced in Section 4.8. This is a distribution similar in appearance to the normal distribution, but with a slightly wider shape to compensate for lack of accuracy due to estimating rather than knowing the standard deviation. Recall that the smaller the sample, the wider the distribution, so the particular member of the family of $t$ distributions to be used depends on the degrees of freedom ($df$), a variation of the sample size, explained in Section 4.8.

The nature of $t$

The $t$ distribution is a symmetric, bell-shaped distribution that looks and behaves very much like the normal. It is used where one would choose to use a normal, but where the standard deviation is estimated by $s$ rather than being the known $\sigma$. Even the calculations are similar. For example, $(x - \mu)/s$ provides a standard $t$-value. The difference is that using the estimated rather than the known standard deviation yields a less confident and therefore more spread out distribution, with the spread depending on the $df$, a variation of the sample size. Each $df$ yields a member (a curve) of the $t$ family. When using a $t$ arising from one sample, $df = n - 1$; for a $t$ arising from two samples, $df = n - 2$. As $df$ grows large, the $t$ distribution gets closer to the normal distribution. For infinite $df$, the $t$ becomes the normal.

We note that in the rare case where $\mu$ is known but $\sigma$ is estimated by $s$, $t = (x - \mu)/s$.

The $t$ pictured

Fig. 6.3 shows a $t$ distribution for 9 $df$ with 2.5% of the area under the curve’s right tail shaded. Note that it is similar in appearance to the normal distribution depicted in Fig. 6.1, except that the critical value corresponding to $\alpha = .025$ (or 0.05 for two-tailed tests) lies $2.262 s$ (sample standard deviations) to the right of the mean rather than $1.96\sigma$ (population standard deviations).
Table II (see Tables of Probability Distributions) contains selected distances \( t \) away from the mean for the most commonly used one- and two-tailed \( \alpha \) and \( 1 - \alpha \) areas under the curve for various \( df \). These \( t \)-values correspond to italicized \( z \)-values in Table I. Table 6.2 shows a portion of Table II that may be used to follow the examples.

**EXAMPLE OF A TOLERANCE INTERVAL USING \( t \)**

Let us follow the prostate volume example as in the preceding section on the standard normal distribution, except that we shall use \( m \) and \( s \) from the small sample of Table DB1.1 rather than \( \mu \) and \( \sigma \). (Actually, a small skewed sample such as this may be treated better by the rank-order methods of Chapter 11, Tests of location with continuous outcomes, but the example serves to illustrate \( t \).) For the 10 volumes, \( m = 32.7 \text{ mL} \) and \( s = 15.9 \text{ mL} \). Because we are dealing with only one sample, \( df = n - 1 = 9 \). We, therefore, need look only at the \( df = 9 \) row in Table 6.2. The prostate volume 59 mL yields \( t = (59 - 32.7)/15.9 = 1.65 \), which places it 1.65 standard deviations above the mean. This lies between 1.38 which falls under \( \alpha = 0.10 \) on the 9 \( df \) row and 1.83 which falls under \( \alpha = 0.05 \). We can conclude that between 5% and 10% of patients will have volumes greater than this patient. Similarly, the 83-mL prostate yields \( t = 3.16 \), which is between \( \alpha = 0.01 \) and \( \alpha = 0.005 \); less than 1% of patients will have volumes this large. We do not tabulate \( t \) for as many possible values as we do the standard normal, because a full table would be required for every possible
We can calculate on a computer \( \alpha = 0.067 \) for \( t = 1.65 \) and \( \alpha = 0.006 \) for \( t = 3.16 \) if we need them. Tolerance intervals are treated in some detail in Chapter 8, Tolerance, prediction, and confidence intervals.

**Critical Value for t**

The critical value above which 1% of prostate volumes occur is found by using the 9 \( df \) \( t \)-value in the one-tailed \( \alpha = 0.01 \) column, namely, \( t = 2.82 \). We calculate the volume at 2.82 standard deviations above the mean, that is, \( m + 2.82 \sigma = 32.7 + 2.82 \times 15.9 = 77.54 \). We expect that no more than 1% of volumes will exceed 77.5 mL.

**Example: Proportion of Leg Surgery Patients (DB10) Who Can Perform**

DB10 provides data on performance of hamstring or quadriceps surgery patients. What proportion can perform a triple hop in less than 1.95 seconds? By calculating from the database, we find \( m = 2.70 \) and \( \sigma = 0.53 \). With \( n = 8 \) from one sample, \( df = 7 \). The postulated 1.95 seconds is \( x \), so \( t = (1.95 - 2.70)/0.53 = -1.415 \). Because the \( t \)
distribution is symmetric, only the right tail has been tabulated; the area to the left of
\(-1.415\) is the same as the area to the right of \(1.415\). By looking in the row for 7 \(df\) in
Table 6.2, we find 1.415 in the first column under the heading for one-tailed \(\alpha = 0.10\), implying that 10% of the area under the \(t\) curve falls to the right of 1.415. By using the symmetry, we conclude that 10% of the patients will have hop times less
than 1.95 seconds.

**Exercise 6.2**
In DB10, the number of centimeters covered in the hop test using the operated leg has
\(m = 452.8\) and \(s = 91.7\). What interval will include 95% of such patients (i.e., 2.5% in each
tail)?

### 6.4 THE CHI-SQUARE DISTRIBUTION

**Why we need chi-square**
A random variable that follows a chi-square distribution \(\chi^2\) is composed of a sum of squares. This distribution arises in a test of a variance (the averaged sum of squares of deviations from a mean) and in a test of contingency (the weighted sum of squares of observed deviations from expected). Other test statistics have also been found to follow a chi-square distribution, for example, the log-rank test of two survival curves. Statistical applications involving \(\chi^2\), being composed of squares, must always have a positive value and can grow to any large size.

Also of interest at this point in the book is the confidence interval. In this case, we need to focus on a measure of variability as a statistic. Suppose the amount of active ingredient in a medicinal capsule is crucial: less than \(\mu - \epsilon\) mg fails to work and more than \(\mu + \epsilon\) mg damages the patient. We need a confidence interval on \(\mu\) to be confident that the probable variability in content is not too large. Because the standard deviation is a square root, which is difficult to work with mathematically, confidence intervals on variability are developed using the sample variance \(s^2\) and afterward are converted to standard deviation units. As noted in Section 4.8, the sample variance \(s^2\), drawn randomly from a normal population, when multiplied by the constant \(df/\sigma^2\) follows a chi-square distribution. Thus we need chi-square to develop confidence intervals on the variance. (\(df\) for simple confidence intervals will be \(n - 1\).)

**Chi-square pictured**
Remember that the chi-square, being composed of squares, cannot be negative; however, it can grow to any large size. Thus it is like a lopsided bell curve with the right tail stretched out—right skewed. Fig. 6.4 shows the chi-square distribution for 9 \(df\).
6.4 The Chi-Square Distribution

Tables of chi-square

Table III (see Tables of Probability Distributions) provides the chi-square values that yield commonly used values of $\alpha$, that is, the probability that a randomly drawn value from the distribution lies in the tail demarked by the tabulated chi-square value. Table 6.3 shows a segment of Table III, which may be used to follow the examples. Because the chi-square distribution is asymmetric, we cannot take an area in one tail and expect the other tail to be the same. Finding areas in a tail is much the same as in the $t$: The desired area in the tail specifies the column, the $df$ specifies the row, and the critical chi-square value (the value demarking the tail area) lies at the row-column intersection. Table III provides chi-square values for the more commonly used right tail; Table IV (see Tables of Probability Distributions) provides values for the left tail.

**EXAMPLE: VARIABILITY OF PROSTATE VOLUMES**

In DB1, we saw that $\sigma = 16.35$ mL for the population of prostate volumes (excluding known BPH). How large a standard deviation $s$ would we have to observe in a sample size of 10 to have a 0.05 or less probability that it could have been so large by chance alone? From either Table 6.3 or Fig. 6.4, the critical chi-square values (the values demarking the tail area) for $\alpha = 0.05$ with $n - 1 = 9$ df is $\chi^2 = 16.9$. That is,

$$\chi^2 = df \times s^2 / \sigma^2 \quad \text{or} \quad 16.9 = 9s^2 / 16.35^2 = 0.033667 \ s^2.$$
We solve to find $s^2 = 501.98$ or $s = 22.40$. A sample standard deviation exceeding 22.4 would have less than 5% probability of occurring by chance alone.

**EXAMPLE: VARIABILITY IN PLATELET COUNT**

Is the variability in platelet count among a sample of wound patients treated with growth factor different from that for normal patients? In DB9, platelet count for $n = 20$ wound patients treated with growth factor has $m = 295,000$ and $s = 164,900$. Suppose, we found extensive data in the medical literature that gives the platelet count standard deviation as $\sigma = 60,000$ for normal patients. Is $s$ so much larger than $\sigma$ that it is improbable to have happened by chance? It turns out that the ratio of $s^2$ to $\sigma^2$, when multiplied by $df$, is a $\chi^2$ statistic with $n - 1$ $df$, or $\chi^2 = df \times s^2 / \sigma^2$ (see Chapter 13: Tests on variability and distributions, for further discussion of this topic).

Thus $\chi^2 = 19 \times 164,900^2 / 60,000^2 = 143.5$. From Table 6.3 for the row $df = 19$, we can see that at $\chi^2 = 30.14$, 5% of the area under the curve lies to the right, at $\chi^2 = 36.19$, 1% of the area under the curve lies to the right, and at $\chi^2 = 43.81$, 0.1% of the area under the curve lies to the right. Because 143.5 is much larger than 43.81,
the area under the curve to the right is very much smaller than 0.1%. The chance that a standard deviation as large as 164,900 arose from a sample of patients with normal platelet counts is much less than one in 1000; we conclude that the variability in the sample of patients who have had growth factor is larger than in normal patients.

**An effect of asymmetry in the chi-square distribution**

Table III (see *Tables of Probability Distributions*) provides the $\chi^2$ values that yield commonly used values of $\alpha$, that is, the probability that a randomly drawn value from the distribution lies in the right tail demarked by the tabulated $\chi^2$ value. Because the chi-square distribution is asymmetric, we cannot take an area in one tail and expect the area in the other tail to be the same. For areas under the left tail, another $\chi^2$ table is needed. Table III provides chi-square values for the more commonly used right tail, and Table IV (see *Tables of Probability Distributions*) provides values for the left tail. The mechanism of finding areas in a tail from the table is much the same as for the $t$: The desired area in the tail specifies the column, the $df$ specifies the row, and the critical $\chi^2$ value (the value demarking the tail area) lies at the row-column intersection.

**Exercise 6.3**

In DB7, the standard deviation of bone density of 18 patients with femoral neck fractures is $s = 24.01$. Take the standard deviation for the normal population as $\sigma = 16.12$. We want to know whether or not the variability for patients with a fracture is unusually larger than that for normals and therefore probably different. If we sampled repeatedly from the normal population, how often would we find an $s$ this large or larger? (Hint: By calculating $\chi^2$ as before, what percentage of the area under the curve is greater than that value of $\chi^2$?)

**6.5 THE F-DISTRIBUTION**

**The concept of F**

Section 4.8 reports that a ratio of two sample variances drawn from the same normal population has a distribution named $F$. Conceptually, $F$ may be thought of as the size of the top variance relative to the bottom variance. An $F$ of 2.5 indicates that the top variance is two and a half times the bottom variance. If the top is too much larger than the bottom, we believe it is unlikely that the two samples come from populations with the same variance. The probability distribution of $F$ defines “too much.”

**F and standard deviations**

Variances, the squares of the respective standard deviations, are used in the computations because they are easier to deal with mathematically. However, the concepts and
conclusions are the same for both variances and standard deviations. We may use variances in computations quite legitimately to make conclusions about the respective standard deviations.

**F and df**

Each variance (multiplied by a constant) is distributed chi-square, and the resulting $F$-distribution itself looks very much like $\chi^2$. Because both top and bottom variances have $df$, any particular $F$-distribution is specified by a pair of $df$. The top $df$ is always stated first. The member of the family of $F$-distributions specified for the pair 2, 6 $df$ is shown in Fig. 6.5, with the upper 5% tail area shaded. We note that 5.14 is the critical $F$ for 2, 6 $df$, that is, 5% of the $F$-distribution lies to the right of 5.14. Table 6.4, a segment of the $F$ table (Table V in *Tables of Probability Distributions*) shows from where this critical value comes.

**Using the $F$ table**

Table V in *Tables of Probability Distributions* (see also Table 6.4) provides the $F$-values corresponding to $\alpha = 0.05$, that is, a 5% probability that a randomly drawn value from the distribution exceeds the tabulated $F$-value. Because the $F$-distribution is asymmetric, we cannot take an $F$-value designating an area in one tail and expect the area in the other tail to be the same. However, the left tail is not used for the methods appearing in this book. Finding areas in the right tail is much the same as in $\chi^2$, except for having a pair of $df$ rather than a single $df$. The $df$ of the top variance specifies the
column, the \( df \) of the bottom variance specifies the row, and the distance greater than zero \((F)\) lies at the row-column intersection. Because \( df \) for both top and bottom must be used in tabulation, an \( F \) table can be given for only one \( \alpha \). Several pages of tables could be given, but because \( \alpha = 0.05 \) is commonly used in almost all medical research, only the table for 0.05 is given here. Calculation of \( F \) for other \( \alpha \)'s may be found in any good statistical computer package or in books of tables.

**EXAMPLE: VARIABILITY OF PROSTATE VOLUMES FOR POSITIVE VERSUS NEGATIVE BIOPSIES**

Let us calculate the variances for the prostate volumes found in Table DB1.1 for the groups of patients with positive and negative biopsy results. We want to know if the variability of volume for the seven patients with negative biopsy results is significantly larger than that for the three patients with positive biopsy results. Although variances are the statistics actually tested, think of the test as one of standard deviations. The sample sizes of three and seven give the \( df \) pair 2,6, leading to the 5.14 entry in the \( F \) table. Because of the 5.14 critical value, the test requires the variance for negative biopsy volumes to be more than five times the size of the variance for positive biopsy volumes to show evidence that it is larger with no more than a 5% chance of being wrong. On calculating the variances, we find \( s^2_1 = 326.89 \) (standard deviation = 18.08) and \( s^2_2 = 83.52 \) (standard deviation = 9.14). The calculated \( F = 326.89/83.52 = 3.91 \). Because this is less than the critical \( F \)-value of 5.14, we find that observing a ratio of 3.91 or greater between the two variances would occur more than 5% of the time even if the two population variances are truly equal.

**\( F \) and sample size**

The effect of sample size can be seen dramatically. The larger the sample size is, the stronger is our evidence. If we had obtained the same variances from double the sample sizes, that is, 6 and 14 patients, we would have had 5,13 as the \( df \) pair, leading to a
CHAPTER 6 Finding probabilities

critical $F$ (from Table V) of 3.03. Our 3.91 $F$ ratio would have well exceeded this value, and we would reach the reverse conclusion: we would have enough evidence to say there is less than a 5% chance that the disparity in variances occurred by chance alone, and a causal factor likely is present.

Exercise 6.4
In DB8, means of four glycosaminoglycan levels recorded for each of two types of assays were about 0.39 (type I) and 0.44 (type II), which are not very different. However, the standard deviations were 0.2421 (type I) and 0.1094 (type II). If the type II assay averages the same but is less variable, we would choose it as the preferable type of assay. The statistic $F$ is the ratio of the two variances. Is there less than a 5% chance of observing a $F$ statistic this large or larger if there were no difference between the population variances (or equivalently, standard deviations)?

6.6 THE BINOMIAL DISTRIBUTION

Binomial events defined

Often, we encounter situations in which only two outcomes are possible: ill or well, success or failure of a treatment, a microorganism does or does not cause a disease. Let us denote $\pi$ to be the probability of occurrence of one of the two outcomes (e.g., the first) on any random trial. (This symbol has nothing to do with the symbol for 3.1416... used in the geometry of circles.) If we have $n$ opportunities for the outcome to occur (e.g., $n$ patients), the binomial distribution will tell us how many occurrences of the outcome we would expect by chance alone.

Binomial table

Table VI (see Tables of Probability Distributions) gives the probability of $n_o$ occurrences of an event under scrutiny out of $n$ trials, given an occurrence rate of $\pi$. Table 6.5 shows the segment of Table VI needed for the example.

EXAMPLE: HAS THE SUCCESS OF LASER TRABECULOPLASTY IMPROVED?

As an example, the long-term success of laser trabeculoplasty as therapy for open-angle glaucoma was examined in a Norwegian study. At the end of 2 years, the failure rate was 1/3. Suppose, you perform trabeculoplasty on six patients. At the end of 2 years, you find only one failure. What is the probability that your improved failure rate of 1/6 is due to more than chance, given the Norwegian rate of 1/3? Table 6.5 (or Table VI) gives the probability of $n_o$ or more occurrences of the event under scrutiny out of $n$ trials, given an occurrence rate of $\pi$. The table gives only a “$\geq$” value; to find the probability that $n_o$ is 1 or 0, we must find $1 - (\text{the probability that } n_o \geq 2)$. 

The column is indicated by $\pi = 0.333$, lying between 0.30 and 0.35. The row is indicated by $n = 6$ and $n_o = 2$. The value for our $\pi$ is bracketed by 0.580 and 0.681, about 0.65. Finally, we take $1 - \pi$ the tabulated value, or 0.35, for our result. A more exact value, calculated with the aid of a computer, is 0.351, which is almost the same. With 0.35 probability that our result could have occurred by chance alone, our small sample gives us inadequate evidence of an improved success rate.

### The effect of sample size

Fig. 6.6 shows a binomial distribution for $\pi = 1/3$ and $n = 12$, double the sample size of the example. (The values for 11 and 12 occurrences are not 0 but are too small to appear on the graph.) We can see by combining the first two bars that the probability...
of \( n_o \leq 1 \) is near to 0.06, much closer to significance. A group of 60 patients with 10 failures rather than 6 patients with 1 failure would have given a probability of 0.003, which is strong evidence of improvement.

**The binomial for larger samples**

The binomial distribution is difficult to calculate when \( n \) is not quite small. For medium to large \( n \), we need an easier way to find the required probability. When \( \pi \), the occurrence rate is toward the center of the 0–1 interval, the normal distribution adequately approximates the binomial. The mean of the binomial sample is the observed occurrence rate \( p = n_o/n \), number of occurrences observed/sample size, and the variance is \( \pi(1 - \pi)/n \). Thus the mean difference divided by the standard deviation follows approximately the standard normal distribution, as

\[
z = \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}
\]

and probability questions may be addressed using Table I (see *Tables of Probability Distributions*). (The approximation can be improved slightly with an adjustment factor; see Section 9.6.) The similarity to a normal shape can be seen in Fig. 6.6. If \( \pi \) is close to 0 or to 1, the binomial becomes too skewed to approximate with the normal, but it can be approximated adequately by the Poisson distribution.

**Exercise 6.5**

*In a relatively large study on subfascial endoscopic perforating vein surgery (SEPS),\(^2\) 10% (= \( \pi \)) of ulcers had not healed by 6 months after surgery. In a particular clinic, \( n = 6 \) SEPS were performed and \( n_o = 2 \) patients had unhealed ulcers after 6 months. What is the probability that, if the clinic shared the theoretical \( \pi = 0.10 \), at least 2 out of 6 patients would have had unhealed ulcers?*

**6.7 THE POISSON DISTRIBUTION**

**Poisson events described**

The Poisson distribution arises from situations in which there is a large number of opportunities for the event under scrutiny to occur but a small chance that it will occur on any one trial. The number of cases of bubonic plague would follow Poisson: A large number of patients can be found with chills, fever, tender enlarged lymph nodes, and restless confusion, but the chance of the syndrome being plague is extremely small for any randomly chosen patient. This distribution is named for Siméon Denis Poisson, who published the theory in 1837. The classic use of Poisson
was in predicting the number of deaths of Prussian army officers from horse kicks from 1875 to 1894; there was a large number of kicks, but the chance of death from a randomly chosen kick was small.

**Poisson table**

Table VII (see *Tables of Probability Distributions*) provides the \( \alpha \) to test the hypothesis that \( n_o \) or more cases would occur by chance alone, given the occurrence rate \( \lambda ( = n\pi) \). Table 6.6 shows the segment of Table VII needed for the example.

---

**EXAMPLE: RATE OF ALZHEIMER’S DISEASE IN ALUMINUM CITY VERSUS SEATTLE**

As an example, a survey\(^3\) of medical records of 23,000 people older than 60 years from a Seattle HMO revealed 200 participants with indications of Alzheimer’s disease, a rate of 0.0087. A clinician in Aluminum City has 100 patients older than 60 years and reports 4 with Alzheimer’s disease. What would be the probability of observing this high rate of rate, or higher, if Aluminum City’s rate were the same as that of Seattle? Table 6.6 or Table VII provides the probability that \( n_o \) or more cases would occur by chance alone, given the occurrence rate \( \lambda ( = n\pi) \). The column in the table must be chosen using the value \( \lambda = n\pi = 100 \times 0.0087 = 0.87 \), which would lie between columns 0.8 and 0.9. The row is indicated by \( n_o = 4 \). The Poisson probability is between 0.009 and 0.014, about 0.012. The actual probability from a computer calculation is the same. This result indicates that the chance of observing four or more cases would be about 1 in 100 if Aluminum City’s rate were the same as that of Seattle, which provides adequate evidence to conclude that Aluminum City has a higher rate of Alzheimer’s disease than Seattle.
The appearance of the Poisson distribution

Fig. 6.7 shows the member of the Poisson family of distributions for $\lambda = 0.87$, the occurrence rate in the example. We might note that theoretically the Poisson distribution has no greatest possible number of occurrences as does the binomial distribution.

The Poisson for larger samples

Like the binomial, the Poisson distribution is difficult to calculate when $n$ is more than a few. For medium to large $n$, when $\pi$, the occurrence rate, is close to 0 or 1, the normal distribution adequately approximates the Poisson. Again, like the binomial, the mean of the Poisson sample is the observed occurrence rate $p = n_o/n$, the number of occurrences observed/sample size, but the variance is $\pi/n$. Thus the mean difference divided by the standard deviation follows, approximately, the standard normal distribution, as

$$z = \frac{p - \pi}{\sqrt{\pi/n}},$$

and probability questions again may be addressed using Table I. Section 9.7 provides a more detailed discussion of this topic.

Exercise 6.6

The probability of acquired immunodeficiency syndrome (AIDS) developing in a person during the first year of infection with human immunodeficiency virus (HIV) is about 0.01.
An investigator reviewed the records of 70 patients with HIV. During the year following their reported date of first possible exposure to another HIV-infected person, AIDS had developed in five. Does the investigator believe these patients’ reporting?

REFERENCES


6.7 The Poisson Distribution
7.1 HYPOTHESES IN INFERENCE

A clinical hypothesis

Suppose, we want to know if a new antibiotic reduces the time for a particular type of lesion to heal. We already know (or presume we know) the mean time $\mu$ to heal in the population, that is, the general public, without an antibiotic. We take a sample of lesion patients from the same population, randomizing the selection to ensure representativeness (see Section 1.9), and measure the time to heal when treated with the antibiotic. We compare our sample mean time to heal, say $m_a$ (a for antibiotic), with $\mu$. Our clinical hypothesis (not statistical hypothesis) is that the antibiotic helps, that is, the average time to heal when treated with the antibiotic is less than $\mu$. To be sure that our comparison gives us a dependable answer, we go through the formal sequence of steps that compose scientific inference.

Decision theory introduced

In general terms, the decision theory portion of the scientific method uses a mathematically expressed strategy, termed a decision function (or sometimes decision rule), to make a decision. This function includes explicit, quantified gains, and losses to reach a conclusion. Its goal is to optimize the outcome of the decision, that is, to jointly maximize gains and minimize losses. Gains might be factors such as faster healing, less pain, or greater patient satisfaction. Losses might be factors such as more side effects or greater costs—in time, effort, or inconvenience as well as money. Any number of possible decision functions exists, depending on the strategy selected, that is, on the gains and losses chosen for inclusion and their relative weightings. Is financial cost of treatment to be included? Is pain level more or less important than eventual overall satisfaction? In a given situation, quite different “optimum” decisions could be reached, depending on the decision function chosen. Those who wish to apply outcomes derived from an investigator’s use of decision theory should note that a personal or financial agenda may be involved in the choice of elements and weightings used in the decision function. To safeguard against
such agendas, a user should accept only decision functions with natures that have been clearly and explicitly documented.

**Operations research**

In most applications the decision problem is expressed as a strategy to select one of a number of options based on a criterion of minimum loss. The loss might include risk of making a wrong decision, loss to the patient (financial cost as well as cost in pain, decreased quality of life, or even death), and/or financial or time cost to the investigator—institution. In the industrial, business, and military fields, applied decision theory most often has come under the heading of operations research [or operational analysis (British)]. In medicine, some forms of applied decision theory using multiple sources of loss in the decision strategy appear under the heading of outcomes analysis.

**Decision-making by testing a statistical hypothesis**

Section 1.4 notes that medicine’s major use of statistical inference is in making conclusions about a population on the basis of a sample from that population. Most often, we form statistical hypotheses, usually in a form different from the clinical hypothesis, about the population. We use sample data to test these hypotheses. This procedure is a special case of two-option decision theory in which the decision strategy uses only one loss, that of the risk (probability) of making an erroneous decision. More specifically, the loss is measured as a known, controlled probability of each of two possible errors: rejecting the first hypothesis when it is true and failing to reject the first hypothesis when the second is true.

**How we get from observed data to a test**

From the data, we calculate a value, called a *test statistic* that will answer the statistical question implied by the hypothesis. For example, to answer the time-to-heal question, we may compare the statistic $m_a$ (mean time to heal using the antibiotic) with the parameter $\mu$ (mean time to heal for the untreated population). Such *test statistics* are calculated from data that follow probability distributions; therefore the statistics themselves will follow a probability distribution. Section 4.8 notes that a mean, the time-to-heal statistic here, is distributed (at least approximately) normal. Areas under the tails of such distributions provide probabilities, or risks, of error associated with “yes” or “no” answers to the question asked. Estimation of these error probabilities from distributional theory and the sample data constitutes the test.

**The null hypothesis**

The question to be answered must be asked in the form of a hypothesis. This *statistical hypothesis*, which differs from the clinical hypothesis, must be stated carefully to relate
to a statistic, especially because the statistic must have a known or derivable probability distribution. In most applications, we want to test for a difference, starting with the *null hypothesis*, symbolized $H_0$, that the population parameter giving rise to our sample is no different (hence “null”) from known information or assumed values. (The hypothesis to be tested must be based on the known or assumed parameter defining a distribution, in this case, of the established mean healing time. This is explained further later.) In testing for a *difference*, we hypothesize that the population mean time to heal with an antibiotic, $\mu_a$ (which we estimate by our sample mean, $m_a$), does not differ from the mean time to heal without the antibiotic, that is, $H_0$: $\mu_a = \mu$. In contrast to testing for a difference, we could test for *equivalence*. This chapter focuses primarily on the more familiar difference testing. Equivalence testing is addressed in Chapter 12, Equivalence testing.

**The alternative hypothesis**

After forming the null hypothesis, we form an *alternative hypothesis*, symbolized $H_1$, stating the nature of the discrepancy from the null hypothesis if such discrepancy should appear. Hypotheses in words may be long and subject to misunderstanding. For clarity, they usually are expressed in symbols, where the symbols are defined carefully.

**Forms the hypotheses can take**

To answer the time-to-heal question, we hypothesized that the population mean $\mu_a$ from which our sample is drawn does not differ from $\mu$. Our alternative hypothesis is formed logically as *not* $H_0$. If we believe the antibiotic could either shorten or lengthen the healing, we use a *two-sided* alternative hypothesis so that our hypotheses are

$$H_0 : \mu_a = \mu \quad \text{and} \quad H_1 : \mu_a \neq \mu.$$  \hfill (7.1)

The two-sided alternative is conservative and is the usually chosen form. In case we truly believe that the antibiotic *cannot lengthen* the healing, the alternative to no difference is shortened healing, and our alternative hypothesis would be $H_1: \mu_a < \mu$. The alternative here is known as a *one-sided* hypothesis. When a decision is made about using or not using a medical treatment, the sidedness may be chosen from the decision possibilities rather than the physical possibilities. If we will alter clinical treatment only for significance in the positive tail and not in the negative tail, a one-tailed test may be used.

**Why the null hypothesis is null**

It may seem a bit strange at first that our primary statistical hypothesis in testing for a difference says there is no difference, even when, according to our clinical hypothesis,
we believe there is one and might even prefer to see one. The reason lies in the ability to calculate errors in decision-making. When the hypothesis says that the distribution giving rise to our sample is no different from known information, we have available a known probability distribution and therefore can calculate the area under the distribution associated with the erroneous decision: A difference is concluded when in truth there is no difference. This area under the probability curve provides us with the risk for a false-positive result. The alternative hypothesis, on the other hand, says just that our known distribution is not the correct distribution, not what the alternative distribution is. Without sufficient information regarding the distribution associated with the alternative hypothesis, we cannot calculate the area under the distribution associated with the erroneous decision: No difference exists when there is one, that is, the risk for a false-negative result.

**A NUMERICAL EXAMPLE: HYPOTHESES ABOUT PROSTATE VOLUMES**

As a numerical example, consider the sample of the 10 prostate volumes in Table DB1.1. Suppose, we want to decide if the population mean $\mu_v$ from which the sample was drawn is the same as the mean of the population of 291 remaining volumes (presumed here to be the true population mean). Because the true mean giving rise to the sample volumes may be either larger or smaller than the population mean, the alternative is a two-sided hypothesis. Our hypotheses are

$$H_0 : \mu_v = \mu \text{ and } H_1 : \mu_v \neq \mu.$$  \hspace{1cm} (7.2)

$m$ estimates the unknown $\mu_v$ and, if $H_0$ is true, $m$ is distributed $N(\mu, \sigma_m^2)$. [$\mu$ is the mean, $\sigma$ the standard deviation of the 291 volumes, and $\sigma_m$ is the standard error of the mean (SEM) $\sigma/\sqrt{n}$ (see Section 4.9).] Standardization of $m$ provides a statistic $z$, which we know to be distributed $N(0,1)$, that is,

$$z = \frac{m - \mu}{\sigma_m} = \frac{m - \mu}{\sigma/\sqrt{n}}.$$  \hspace{1cm} (7.3)

We know that $m = 32.73$, $\mu = 36.60$, and $\sigma = 18.12$; $n = 10$ is the size of the sample we are testing. By substituting in Eq. (7.3), we find $z = -0.675$; that is, the sample mean is about two-thirds of a (population) standard deviation below the hypothesized population mean. Table 7.1, a small excerpt from Table I (see Tables of Probability Distributions), shows the two-tailed probabilities for the 0.60 and 0.70 standard deviations. The calculated $z$ lying between these two values tells us that the probability of finding a randomly drawn normal observation more than $0.675\sigma$ away from the mean is a little more than 0.484, or about 0.5. We conclude that the probability of observing a sample mean as or more indicative of the alternative hypothesis is 50% if the null
hypothesis, \( \mu_v = \mu_s \), is true. This is not a rare event, and hence there is insufficient evidence to conclude a difference. This result may be visualized on a standard normal distribution as in Fig. 7.1.

The critical value

A “cut point” between whether a statistic is statistically significant or nonsignificant is termed a critical value. In symbols the critical values of \( z \) leading to a two-tailed \( \alpha \) level test involving a normally distributed sample mean are \( z_{\alpha/2} \) and \( z_{1-\alpha/2} \). By symmetry, we need calculate only the right tail \( z_{1-\alpha/2} \) and use its negative for the left. The decision about the importance of the statistic \( z \) maybe stated as a formula. \( z \) does not lead us to conclude a statistical difference if

\[
- z_{1-\alpha/2} < \frac{m - \mu}{\sigma_m} < z_{1-\alpha/2},
\]

or, using Eq. (7.3),

\[
- z_{1-\alpha/2} < \frac{m - \mu}{\sigma_m} < z_{\alpha/2}.
\]

Figure 7.1 A standard normal distribution showing 2.5% tail areas adding to 5% risk of error from concluding that a random value did not arise from this distribution (risk of a false positive). The shaded areas are rejection regions. The 0.025 area under the right tail represents the probability that a value drawn randomly from this distribution will be 1.96σ or farther above the mean, σ taking the value 1 in this standard normal form. We can see that \( z = -0.675 \), about 2/3 of a standard deviation below 0, lies far from a rejection region.

Table 7.1 Excerpt from Table I, normal distribution.

<table>
<thead>
<tr>
<th>( z ) (no. standard deviations to right of mean)</th>
<th>Two-tailed ( \alpha ) (area in both tails)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.60</td>
<td>0.548</td>
</tr>
<tr>
<td>0.70</td>
<td>0.484</td>
</tr>
</tbody>
</table>
The most common types of statistics being tested and their associated probability distributions

The previous section demonstrated how a hypothesis having a general pattern of a sample mean divided by a known population SEM follows a standard normal distribution. In the same fashion a hypothesis about a mean when the standard deviation must be estimated (the population standard deviation is unknown) has the same general pattern, namely, a sample mean divided by a sample SEM, so it follows a t distribution. A hypothesis about a standard deviation uses a variance, which follows a chi-square distribution. A hypothesis comparing two standard deviations uses a ratio of two variances, which follows an F distribution. And finally, a hypothesis about a rate of occurrence follows a binomial or a Poisson distribution. Thus a large number of statistical questions can be tested using just these six well-documented probability distributions (introduced in Section 4.8).

Confidence intervals are closely related

Confidence intervals are closely related to hypothesis tests procedurally, in that null and alternative hypotheses could state that the interval does and does not, respectively, enclose the population mean. Indeed, a confidence interval is obtained by inverting the corresponding hypothesis tests and thus represents the set of null hypothesis values for which a hypothesis test would fail to reject the null given an observed set of data. However, the use of confidence intervals differs: A confidence interval is used for conveying the precision of our estimate of the population parameter, not to test. Confidence intervals are addressed in the next chapter.

Assumptions in hypothesis testing

Section 4.7 shows that various assumptions underlie hypothesis testing and that the result of a test is not correct when these assumptions are violated. Furthermore, the methods do not tell the user when a violation occurs or by how much the results are affected. Results appear regardless. It is up to the user to verify that the assumptions are satisfied. The assumptions required vary from test to test, and the assumptions required are noted along with each test methodology. The most common assumptions are that errors are normal and independent from one another and, for multiple samples, that variances are equal. In many cases we can find tests that do not require such assumptions to remain valid, at least in reasonably large samples.

The meaning of “error”

An “error” in an observation does not refer to an error in the sense of a mistake, but rather to the deviation of the individual observation from the typical. The term error, often and more appropriately called “residual,” is used for historical reasons; if you
find the term “error” in referring to observations confusing in your reading, read “deviation from typical” instead.

**The assumption of independence of errors**

When a set of observations is taken, we assume that knowledge of the error on one tells us nothing about the error on another, that is, that the errors are independent from one another. Suppose, we take temperature readings on a ward of patients with bacterial infections. We expect them to be in the 38°C–40°C range, averaging about 39°C. We assume that finding a 39.5°C reading for one patient (i.e., an error of 0.5°C greater than the average) tells us nothing about the deviation from average we will find for the patient in the next bed. How might such an assumption be violated? If the ward had been filled starting from the far end and working nearer as patients arrived, the patients at the far end might be improving and have lower temperatures. Knowledge of the temperature error of a particular patient might indeed give us a clue to the temperature error in the patient in the next bed. The assured independence of errors is a major reason for incorporating randomness in sampling. This assumption is made in almost all statistical tests.

**The assumption of normality of errors**

Second, many, but not all, tests also assume that these errors are drawn from a normal distribution. Two major exceptions are tests of means, where the sample mean is distributed approximately normally distributed in large samples (by the Central Limit Theorem), and rank-order (or nonparametric) tests.

**The assumption of equality of standard deviations**

The third most frequently made assumption occurs when the means of two or more samples are tested. It is often assumed that the standard deviations of the errors are the same. This assumption is stated more often as requiring equality of variances rather than of standard deviations. The term homoscedasticity sometimes encountered in a research article is just a fancy word for equality of variances. While the assumption of equal standard deviations is common when testing means there exist tests that do not require such an assumption. The tests are often preferred in practice because they remain valid whether the population standard deviations are the same or different.

**Exercise 7.1**

In DB4 a hypothesis to be investigated is that protease inhibitors reduce pulmonary admissions. Is this a clinical or a statistical hypothesis? What would be a statement of the other type of hypothesis?
Exercise 7.2
In DB14, we ask if the mean ages of patients with and without exercise-induced bronchoconstriction (EIB) are different. Write down the null and alternative hypotheses.

7.2 ERROR PROBABILITIES
Method presented through an example: Mean prostate-specific antigen between cancerous and healthy patients
We have a mean of prostate-specific antigen (PSA) readings from a cancer patient group and we want to compare it with the mean PSA of healthy patients to see if PSA can detect the presence of prostate cancer. The null hypothesis, $H_0$, states that population mean PSA is no different between groups.

Type I ($\alpha$) error
The null hypothesis may be rejected when it is true; we conclude that there is a difference in true mean PSA when there is no such difference. Such an error is denoted a type I error. In hypothesis testing, we bound the probability of a type I error (formally written as $P[\text{rejecting } H_0 | H_0 \text{ true}]$ and read: “probability of rejecting $H_0$ given $H_0$ is true”) at $\alpha$. The conclusion of a difference when there is none is similar to the false-positive result of a clinical test. Properly, $\alpha$ is chosen before data are gathered so that the choice of critical value cannot be influenced by study results.

Type II ($\beta$) error; power of a test
Alternatively, we may fail to reject the null hypothesis when the alternative hypothesis is true; we conclude that the true mean PSA is the same between groups when it is not. Such an error is denoted a type II error. Its probability of occurring by chance alone is denoted $\beta$, where $\beta$ depends upon the true difference between groups. Formally, $\beta = P[\text{Fail to reject } H_0 | H_0 \text{ false}]$. The failure to reject the hypothesis of no difference when there is one is similar to the false-negative result of a clinical test. $1 - \beta$ is the power of the test, which is referred to often in medical literature and used especially when assessing the sample size required in a clinical study (Chapter 21: Sample size estimation). Power is the probability of rejecting the null hypothesis when a difference truly exists. Again, power will depend upon the hypothesized true difference between groups.

$p$-Value
The true error probabilities, $\alpha$ and $\beta$, are characteristics of the population and are unknown. We must sample the population and estimate them. After the data have
been gathered, new information is available: the value of the decision statistic, for example, the (standardized) difference between control and experimental means. The probability of observing data as or more indicative of the alternative hypothesis when the null hypothesis is assumed true can now be estimated using sample data and is termed the \( p \)-value. In historic usage, if the \( p \)-value is smaller than \( \alpha \), we reject the null hypothesis; otherwise, we do not have enough evidence to reject it. However, a deeper look reveals that the interpretation is not so cut and dried. In reality, the \( p \)-value does not give a conclusion about the hypothesis. This is hard for most users to relate to a practical decision, as shown by the nearly ubiquitous adoption of a binary decision of an arbitrary 0.05, irrespective of practical benefit. The title of a paper by Betensky\(^1\) says it well: “The \( p \)-Value Requires Context, Not a Threshold.”

Let us look at it afresh. We have obtained a sampling of data from a population that bears upon a question about an effect. We have used these data to estimate the effect. We now want to know how much credence we can put into this effect. We performed a test on the data, conditioning upon the null hypothesis of no effect, and derived an estimate of the probability of observing the estimate that we did (or more extreme) if in fact there is no effect. This is the \( p \)-value. If the \( p \)-value is very small, there is little chance we would have observed the effect we did if the null hypothesis were true. If the \( p \)-value is very large, it would not be a rare event that we would have observed the effect we did if the null hypothesis were true. In short, the \( p \)-value is a measure of evidence, not a conclusion. Section 7.5 addresses this issue further.

**Relation among truth, decision, and errors**

Type of error depends on the relationship between decision and truth, as is depicted in Table 7.2, a form of truth table.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
<th>Fail to reject ( H_0 )</th>
<th>Reject ( H_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 ) true</td>
<td>Correct decision</td>
<td>Type I error</td>
<td>(Probability ( 1 - \alpha ))</td>
</tr>
<tr>
<td></td>
<td>True negative</td>
<td>False positive</td>
<td>(Probability ( \alpha ))</td>
</tr>
<tr>
<td>( H_0 ) false</td>
<td>Type II error</td>
<td>Correct decision</td>
<td>(Probability ( \beta ))</td>
</tr>
<tr>
<td></td>
<td>False negative</td>
<td>True positive</td>
<td>(Probability ( 1 - \beta ))</td>
</tr>
</tbody>
</table>
LOGICAL STEPS IN A STATISTICAL TEST

The logic of a statistical test of difference used historically is the following: (1) We choose a level of $\alpha$, the upper bound on the risk for a type I error, that we are willing to accept. (2) Because we know the distribution of the statistic we are using under the null hypothesis, for example, $z$ or $t$, we can use a probability table to find its critical value. (3) We take our sample and calculate the value of the statistic arising from it. (4) If our statistic falls on one side of the critical value, we do not have sufficient evidence to reject $H_0$; if on the other side, we do reject $H_0$.

The critical value and the errors illustrated

Illustrative null and alternative distributions and their $\alpha$ and $\beta$ values are shown in Fig. 7.2.

Choosing $\alpha$ but not $\beta$ is a common reality

We usually know or have reason to assume the nature of the null distribution, but we too seldom know or have enough information to assume that of the alternative distribution. Thus instead of minimizing both risks, as we would wish to do, we fix a small $\alpha$ and try to make $\beta$ small by taking as large a sample size as we can (usually subject to logistical or ethical constraints). In medical applications, choosing $\alpha = 0.05$ has become commonplace, but 0.05 is by no means required. Other risks, perhaps 0.01 or 0.10, may be chosen so long as they are stated and justified. And, of course, what we

Figure 7.2 Depiction of null and alternate distributions along with the error probabilities arising from selecting a particular critical value. To the right of the critical value is the rejection region; to the left, the insufficient-evidence-to-reject region.
calculate is the \( p \)-value, which is the estimated probability of observed data as or more indicative of the alternative hypothesis if the null hypothesis were in fact true.

**Providing a confidence interval**

Confidence intervals help to quantify the uncertainty in our estimates of population parameters and are also closely related to hypothesis testing. Simple algebra can change Eq. (7.4) into a confidence interval. Concluding a real effect, historically called significance, in Eq. (7.4) is equivalent to observing that the confidence interval does not cross zero. Thus the confidence interval does not add to the binary decision-making process. It does not tell us what the power is. However, it often provides useful insights into and understanding of the relationships among the parameter estimates involved.

**Accepting \( H_0 \) versus failing to reject \( H_0 \)**

Previously, it was common in the medical literature to speak of accepting \( H_0 \) if the statistic fell on the “null side” of the critical value (see Fig. 7.2). Users were often mystified to see that a larger sample led to contradiction of a conclusion already made. The issue is one of evidence. The conclusion is based on probabilities that arise from the evidence given by the data. The more accurate statement about \( H_0 \) is the following: The data provide inadequate evidence to infer that \( H_0 \) is false. With additional evidence, \( H_0 \) may be properly rejected. The interpretation is that our data-based evidence gives inadequate cause to say that it is untrue, and therefore we act in clinical decision-making as if it were true. The reader encountering the “accept \( H_0 \)” statement should recall this admonition.

**Testing for a difference versus equivalence**

Testing for a difference, for example, between mean times to resolve infection caused by two competing antibiotics, has a long history in medical statistics. In recent years the “other side of the coin” has been coming into practice, namely, testing for equivalence. If either antibiotic could be the better choice, the equivalence test is two sided. However, in many cases we have an established (gold) standard and ask if a new competitor is not appreciably worse than the standard, giving rise to a one-sided equivalence test. A one-sided equivalence test is often termed a *noninferiority* test. For example, we may have an established but expensive brand-name drug we know to be efficacious. A much cheaper generic drug appears on the market. Is it (roughly) equally efficacious? The concept is to pose a specific difference between the two effects, ideally the minimum difference that is clinically relevant, as the null hypothesis, \( H_0 \). If \( H_0 \) is rejected, we have evidence that the difference between the effects is at most clinically irrelevant, and we believe the generic drug is effective. The methodology of equivalence testing is
more fully addressed in Chapter 12, Equivalence testing; at this stage of learning, the reader need only be aware of this option.

Exercise 7.3
A clinical hypothesis arising from DB3 might be that the mean serum theophylline level is greater at the end of the antibiotic course than at baseline. (a) What probability distribution is associated with this hypothesis? (b) What assumptions about the data would be required to investigate this hypothesis? (c) State in words the type I and type II errors associated with this hypothesis. (d) How would the probability (risk) of these errors be designated? How would the power of the test be designated?

Exercise 7.4
A clinical hypothesis arising from DB9 might be: The standard deviation of platelet counts is 60,000 [which gives about 95% confidence (mean $\pm 2 \times 60,000$) coverage of the normal range of 240,000]; this hypothesis is equivalent to supposing that the variance is $60,000^2 = 3,600,000,000$. (a) What probability distribution is associated with this hypothesis? (b) What assumptions about the data would be required to investigate this hypothesis? (c) State in words the type I and type II errors associated with this hypothesis. (d) How would the probability (risk) of these errors be designated?

Exercise 7.5
A clinical hypothesis arising from DB5 might be: The variances (or standard deviations) of plasma silicone before and after implant are different. (a) What probability distribution is associated with this hypothesis? (b) What probability distribution is associated with this hypothesis? (b) What assumptions about the data would be required to investigate this hypothesis? (c) State in words the type I and type II errors associated with this hypothesis. (d) How would the probability (risk) of these errors be designated?

7.3 TWO POLICIES OF TESTING

A bit of history
During the early development of statistics, calculation was a major problem. It was done mostly by pen, sometimes assisted by bead frames (various forms of abaci) in many parts of the world (Asia, Middle East, Europe, the Americas). Later, hand-crank calculators and then electric ones were used. Probability tables were produced for selected values with great effort, the bulk being done in the 1930s by hundreds of women given work by the US Works Project Administration during the Great Depression. It was not practical for a statistician to calculate a $p$-value for each test, so the philosophy became to make the decision of acceptance or rejection of the
null hypothesis on the basis of whether the $p$-value was bigger or smaller than the chosen $\alpha$ (e.g., $p$-value $\geq 0.05$ vs $p$-value $< 0.05$) without evaluating the $p$-value itself. The investigator (and reader of published studies) then failed to reject $H_0$ if the $p$-value was not less than $\alpha$ (result not statistically significant) and reject it if the $p$-value was less (result statistically significant).

**Calculation by computer provides a new option**

With the advent of computers, calculation of even very involved probabilities became fast and accurate. It is now possible to calculate the exact $p$-value, for example, $p$-value $= 0.12$ or $0.02$. The user now has the option to make a decision and interpretation on the exact error risk arising from a test.

**Contrasting two approaches**

The later philosophy has not necessarily become dominant, including in medicine. The two philosophies have generated some dissension among users. Advocates of the older approach hold that sample distributions only approximate the probability distributions and that exactly calculated $p$-values are not accurate anyway; the best we can do is select a “significant” or “not significant” choice. Advocates of the newer approach—and these must include the renowned Sir Ronald Fisher in the 1930s—hold that the accuracy limitation is outweighed by the advantages of knowing the $p$-value. The action we take about the test result may be based on whether a highly doubted result suggests a most unlikely difference (perhaps $p$-value $= 0.80$) or is borderline and suggests further investigation (perhaps $p$-value $= 0.08$). Similarly, our action may be based on whether a negligibly doubted result is close to the decision of having happened by chance (perhaps $p$-value $= 0.04$) or leaves very little doubt in the reader’s mind (perhaps $p$-value $= 0.004$).

**Other factors must be considered**

The preceding comments are not meant to imply that a decision based on a test result depends solely on a $p$-value. The value of $\beta$, the risk of failing to conclude a difference when there is one, is germane. And certainly, the sample size and the estimated clinical difference must enter into the interpretation. Indeed, the estimated clinical difference is often, and should be, the most influential of values used in a test equation. The comments on the interpretation of $p$-values relative to one another hold little doubt for adequate sample sizes and realistic clinical differences.

**Exercise 7.6**

In DB14, healthy subjects showed a mean decrease of 2.15 ppb exhaled nitric oxide (eNO) from before exercise to 20 minutes after exercise. A level $\alpha = 0.05$ test of this mean against a
theoretical mean of 0 (paired t-test) yielded p-value = 0.085. Make an argument for the two interpretations of this p-value discussed in Section 7.3.

7.4 DISTINGUISHING BETWEEN STATISTICAL AND CLINICAL SIGNIFICANCE

The scientific community is often focused on whether results are “statistically significant,” by which they mean that a calculated p-value is lower than a prespecified decision point (usually taken to be 0.05 for a two-sided test). However, a valid criticism of the use of p-values for determining scientific importance is that they are dependent upon (1) the estimate of effect that one is testing, (2) the variance of the responses, and (3) sample size. Because of this, a p-value below the decision point can be obtained when the best estimate of an effect is clinically meaningless, and conversely, a p-value above that point can be obtained despite an observed clinically meaningful difference. As an example, consider the results of four hypothetical studies of the test of (potentially different) blood pressure medications as depicted in Table 7.3. In study A, we observe a reduction of 12.1 mmHg in systolic blood pressure (SBP) comparing treated patients to untreated patients. While these results would likely be clinically relevant, due to the sample size and variance, we obtain a p-value of 0.202 for testing whether the true difference in mean blood pressure between the groups is different from zero. In study B, we obtain the same p-value (still not statistically detectable), but we observe a clinically meaningless difference of 1.21 mmHg between the treatment groups. Despite having the same sample size, we obtain the same p-value because the variance of blood pressure in study B is less than that in study A. In study C, we again observe a reduction of 12.1 mmHg in SBP comparing treated patients to untreated patients, but this time the result would be called highly statistically significant by nearly any standard (a p-value computed as 0.00005 that we would report as < 0.001). The statistically detectable result is due to the increased sample size of study C relative to study A. Lastly, in study D, we have the same statistically detectable p-value despite observing a clinically meaningless difference in

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size per arm</th>
<th>SBP difference</th>
<th>Variance of SBP</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>12.1</td>
<td>225</td>
<td>.20215</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>1.21</td>
<td>2.25</td>
<td>.20215</td>
</tr>
<tr>
<td>C</td>
<td>50</td>
<td>12.1</td>
<td>225</td>
<td>.00005</td>
</tr>
<tr>
<td>D</td>
<td>50</td>
<td>1.21</td>
<td>2.25</td>
<td>.00005</td>
</tr>
</tbody>
</table>

SBP, Systolic blood pressure.
blood pressure. Again, this is driven by the increased sample size and smaller variance in study D. Thus a take-home message is that one cannot simply rely on statistical detectability for determining whether an effect is clinically meaningful. Further, when reporting statistical analyses, it is critical to not simply report a computed \( p \)-value but to also report the estimate of the parameter being tested as well as some indication of the precision of the estimate.

### 7.5 Controversies Regarding the Rigid Use and Abuse of \( p \)-Values

There has been a lot of criticism about using an arbitrary cut point for denoting what has been historically called significance versus nonsignificance. Much of the criticism rightly notes that the scientific method imposes caveats of uncertainty that should be considered. However, there is a difference between scientific inquiry, in which repeated studies tend to strengthen belief until a result is generally accepted, and applied decision, in which a choice must be made. For example, shall I switch to a new drug or continue using the old one? I must do one or the other. In this case a study using a somewhat arbitrary cut point that makes a binary decision may be better than no study.

Too frequently, investigators using a difference test often find a \( p \)-value of, perhaps, 0.08 and say that while there is “no statistical significance,” there is a “trend.” What they intend is that they did not satisfy the arbitrary requirement of \( p \)-value < 0.05, but they still believe there is practical evidence of a difference. We should not require that they “weasel word” their result, nor should we deny them further attention to a promising outcome.

One solution would be to scale evidence levels as sort of a Likert scale with nonstatistical descriptors that can be interpreted by anyone. Such a scale was hinted at by Goodman\(^2\) and proposed by Bland\(^3\). Doubtless the interpretations using such a scale would vary with the application, but a rule of thumb would be very helpful to the public. Bland uses the scheme shown in Table 7.4. We present this table more for concept than for general use.

### Table 7.4

An example of a \( p \)-value scale with interpretations meaningful to nonstatisticians.

<table>
<thead>
<tr>
<th>Range of ( p )-level</th>
<th>How strongly to believe in an effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt; 0.001)</td>
<td>Very strongly</td>
</tr>
<tr>
<td>(0.001 - &lt; 0.01)</td>
<td>Strongly</td>
</tr>
<tr>
<td>(0.01 - &lt; 0.05)</td>
<td>Moderately</td>
</tr>
<tr>
<td>(0.05 - &lt; 0.10)</td>
<td>Weakly</td>
</tr>
<tr>
<td>(\geq 0.10)</td>
<td>Not at all</td>
</tr>
</tbody>
</table>

The \( p \)-level is the probability of observing a result as or more indicative of the alternative hypothesis when the null hypothesis is true. (This table is given to stimulate discussion, not for general use.)
7.6 AVOIDING MULTIPLICITY BIAS

Another common complaint about the use of \( p \)-values in the scientific literature is that multiple hypothesis tests are performed in a single analysis, each resulting in a reported \( p \)-value. In this case, it is tempting to highlight which tests are statistically significant based upon a given threshold (again, it is common to take that threshold to be 0.05 for two-sided hypothesis tests). However, when this is done the probability of falsely concluding that a true association exists across all of the tests is typically far greater than 0.05. This is because for each test there is a 5% chance that one falsely concludes a true association exists when in fact no association exists. Of course, probability (and common sense) then tells us that the more tests we perform the more likely we are to yield a false-positive result. This is generally termed a *multiplicity bias* in that we bias the overall false-positive rate due to the multiple tests. How should one avoid this? The first solution is to carefully consider what the primary test (or tests) is and limit the reporting of \( p \)-values to only that (or those) comparison(s). This is careful and rigorous science and should be done. Additional investigations may be conducted and reported, but statistical significance should not be reported unless special procedures have been performed to ensure that the overall false-positive rate has been controlled for. Some approaches for this adjustment are considered in Chapter 11, Tests of location with continuous outcomes, and Chapter 27, Techniques to aid analysis.

7.7 ORGANIZING DATA FOR INFERENCE

The first step: Identify the type of data

After the raw data are obtained, they must be organized into a form amenable to analysis by the chosen statistical test. We often have data in patient charts, instrument printouts, or observation sheets and want to set them up so that we can conduct inferential logic. The first step is to identify the type of data to be tested. Section 2.4 shows that most data could be typed as categorical (we count the patients in a disease category, e.g.), rank order (we rank the patients in numerical order), or continuous (we record the patient’s position on a scale). Although there is no comprehensive rule for organizing data, an example of each type will help the user become familiar with the terminology and concepts of data types.

Categorical data

In the data of Table DB1.1, consider the association of a digital rectal examination (DRE) result with a biopsy result. We ask the following question: “Does DRE result give some indication of biopsy result, or are the two events independent?” Both results
are either positive or negative by nature. We can count the number of positive results for either, so they are nominal data. For statistical analysis, we almost always want our data quantified. Each of these results may be quantified as 0 (negative result) or 1 (positive result). Thus our variables in this question are the numerical values (0 or 1) of DRE and the numerical values (0 or 1) of biopsy. The possible combinations of results are 0,0 (0 DRE and 0 biopsy); 0,1; 1,0; and 1,1. We could set up a $2 \times 2$ table in the format of Table 7.5.

### Entering the data

After setting up the format, we go to the data (from Table DB1.1) and count the number of times each event occurs, entering the counts into the table. It is possible to count only some of the entries and obtain the rest by subtraction, but it is much safer to count the entry for every cell and use subtraction to check the arithmetic. The few extra seconds are a negligible loss compared with the risk of publishing erroneous data or conclusions. Entering the data produces Table 7.6. Our data table is complete, and we are ready to perform a categorical test.

### Rank data

Continuing with Table DB1.1, suppose, we ask whether average PSA density (PSAD) is different for positive versus negative biopsy. PSAD is a continuous-type measure;

| Table 7.5 Format for recording joint occurrences of two binary data sets. |
|------------------------|------------------|------------------|------------------|
|                        | DRE result       |                  |                  |
|                        | 0                | 1                | Totals           |
| Biopsy result          | 0                |                  |                  |
|                        | 1                |                  |                  |
|                        | Totals           |                  |                  |

*DRE, Digital rectal examination.*

| Table 7.6 Format of Table 7.5 with Table DB1.1 data entered. |
|------------------------|------------------|------------------|------------------|
|                        | DRE result       |                  |                  |
|                        | 0                | 1                | Totals           |
| Biopsy result          | 0                | 3                | 4                | 7                |
|                        | 1                | 2                | 1                | 3                |
|                        | Totals           | 5                | 5                | 10               |

*DRE, Digital rectal examination.*
that is, each value represents a point on a scale from 0 to a large number, which ordinarily would imply using a test of averages with continuous data. However, PSAD values come from a very right-skewed distribution. (The mean of the PSADs for the 301 patients is 0.27, which lies far from the center of the range 0−4.55.) The assumption of normality is violated; it is appropriate to use rank methods.

**Converting continuous data to rank data**

Sometimes ranks arise naturally, as in ranking patients’ severity during triage. In other cases, such as the one we are addressing now, we must convert continuous data to rank data. To rank the PSAD values, we write them down and assign rank 1 to the smallest, rank 2 to the next smallest, and so on. We associate the biopsy results and sort out the ranks for biopsy 0 and biopsy 1. These entries appear in Table 7.7. Our data table is complete, and our data entry is complete, and we are ready to perform a rank test.

**Continuous measurement data**

Suppose we ask whether average PSA is different for positive versus negative biopsy results. For each patient, we would record the PSA level and a 0 or 1 for biopsy result, ending with data as in columns 5 and 8 of Table DB1.1. PSA level is a continuous measurement, that is, each value represents a point on a scale from 0 to a large number. It was noted earlier that PSA is not too far from normal when patients with benign prostatic hypertrophy are excluded. Thus a continuous data type test of averages would be appropriate. We would need means and standard deviations. Their method of calculation was given in Chapter 5, Descriptive statistics. One way to display our data in a

<table>
<thead>
<tr>
<th>PSAD</th>
<th>PSAD ranks</th>
<th>Biopsy results</th>
<th>Ranks for biopsy = 0</th>
<th>Ranks for biopsy = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.24</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0.36</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>0.27</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>0.22</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>0.14</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>0.17</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>0.48</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
form convenient to use is shown in Table 7.8. Our data entry and setup are complete, and we are ready to perform a means test.

**What appears to be a data type may not act that way in analysis**

Suppose a list of cancer patients have their pathologic stage rated T1, T2, T3, or T4. Each stage is more severe than the previous one so that the ratings are ordered data. It appears that your analysis will be a rank-order method. But wait! When you examine the data, you find that all the patients are either T2 or T3. You could classify them into lower or higher categories. You actually analyze numbers of patients (counts) by category, using a categorical method. This phenomenon may also occur with continuous data, where actual data readings may force the type into rank-order or categorical form, thereby changing the method of analysis. The investigator must remain aware of and sensitive to such a data event.

**Exercise 7.7**

Of what data type is (a) the variable Respond versus Not Respond in DB6?, (b) the variable Nausea Score in DB2?, and (c) the variable Platelet Count in DB9?

**Exercise 7.8**

In DB10, rank the seconds to perform the triple hop for the operated leg, small to large.

**Exercise 7.9**

From DB14, set up tables as in Section 7.4: (a) EIB frequency by sex; (b) for the 6 EIB patients, 5-minute eNO differences by sex as ranks; (c) for the 6 EIB patients, 5-minute eNO differences by sex as continuous measurements.
Fundamentally, a study is asking a question

Clinical studies ask questions about variables that describe patient populations. A physician may ask about the PSA of a population of patients with prostate cancer. Most of these questions are about the characteristics of the probability distribution of those variables, primarily its mean, standard deviation, and shape. Of most interest in the prostate cancer population is the average PSA. Three stages in the evolution of scientific knowledge about the population in question are discussed in Section 1.2: description (the physician wants to know the average PSA of the cancerous population), explanation (the physician wants to know if and why it is different from the healthy population), and prediction (the physician wants to use PSA to predict which patients have cancer).

Description and prediction

When we know little about a distribution at question, the first step is to describe it. We take a representative sample from the population and use the statistical summarizing methods of Chapter 5, Descriptive statistics, to describe the sample. These sample descriptors estimate the characteristics of the population. We can even express our confidence in the accuracy of these sample summaries by the confidence methods to be seen in Chapter 8, Tolerance, prediction, and confidence intervals. The description step was addressed in Chapter 5. Prediction combines the result of the inference with a cause-explanatory model to predict results for cases not currently in evidence. This is a more sophisticated stage in the evolution of knowledge that is discussed further starting with Chapter 15, Linear regression and correlation. This section addresses statistical testing that leads to the inference from sample to population.

Testing

We often want to decide (1) whether our patient sample arose from an established population [does a sample of patients who had an infection have the same average white blood cell (WBC) count after treatment with antibiotics as the healthy population?] or (2) whether the populations from which two samples arose are the same or different (does a sample of patients treated for infection with antibiotics have the same average WBC count as a sample treated with a placebo?). In these examples the variable being used to contrast the differences is mean WBC. Let us subscript the means with $h$ for healthy patients, $a$ for patients treated with antibiotics, and $p$ for patients treated with a placebo. Recall that, $\mu$ represents a population mean. Then, (1) contrasts $\mu_a$ with $\mu_h$ and (2) contrasts $\mu_a$ with $\mu_p$. 
STEPS IN SETTING UP A TEST [WITH WHITE BLOOD CELL (WBC) COUNT AS AN EXAMPLE]

These contrasts are performed by statistical tests following the logic of inference (see Section 7.1) with measured risks of error (see Section 7.2). The step-by-step logic adhered to is as follows (assuming the simplest case of one question to be answered using one variable):

1. Write down the question you will ask of your data. (Does treating my infected patients with a particular antibiotic make them healthy again?)

2. Select the variable on which you can obtain data that you believe best to highlight the contrasts in the question. (I think WBC count is best to show the state of health.)

3. Select the descriptor of the distribution of the variable that will furnish the contrast, for example, mean and standard deviation. (Mean WBC count will provide the most telling contrast.)

4. Write down the null and alternative hypotheses indicated by the contrast.
   
   \[ H_0 : \mu_a = \mu_h; \quad H_1 : \mu_a \neq \mu_h. \]

5. Write down a detailed, comprehensive sentence describing the population(s) measured by the variable involved in that question. (The healthy population is the set of people who have no current or chronic infections affecting their WBC count. The treated population is the set of people who have the infection being treated and no other current or chronic infection affecting their WBC count.)

6. Write down a detailed, comprehensive sentence describing the sample(s) from which your data on the variable will be drawn. (My sample is selected randomly from patients presenting to my clinic who pass my exclusion screen (e.g., other infections).)

7. Ask yourself what biases might emerge from any distinctions between the makeup of the sample(s) and the population(s). Could this infection be worse for one age, sex, cultural origin, and so on, of one patient than another? Do the samples represent the populations with respect to the variable being recorded? (I have searched and found studies that show that mortality and recovery rates, and therefore probably WBC, are the same for different sexes and cultural groups. One might suspect that the elderly have begun to compromise their immune systems, so I will stratify my sample (see Section 1.8) to assure that it reflects the age distribution at large.)

8. Recycle steps 1–7 until you are satisfied that all steps are fully consistent with each other.

9. In terms of the variable descriptors and hypotheses being used, choose the most appropriate statistical test (see Chapter 2: Planning analysis: how to reach my scientific objective, for more examples) and select the \( \alpha \) level you will accept.

10. If your sample size is not preordained by availability and/or economics, ensure that you have an adequate sample size to answer the question (see also Chapter 21: Sample size estimation).

At this point, you are ready to obtain your data (which might take hours or years).
ADDITIONAL EXAMPLE: DURATION OF THE COMMON COLD

We would like to test (and lay to rest) the following assertion: “People who do not see a physician for a cold get well faster than people who do.”

1. Question being asked: Does seeing a physician for a cold retard the time to heal?
2. Variable to use: length of time (days) for symptoms (nasal congestion, etc.) to disappear.
3. Descriptor of variable that will provide contrast between patients who see a physician (group 1) and those who do not (group 2): mean number of days $\mu_1$ (estimated by $m_1$) and $\mu_2$ (estimated by $m_2$).
4. Hypotheses: $H_0: \mu_1 = \mu_2$; $H_1: \mu_1 > \mu_2$.
5. Populations: Populations are the sets of people in this nation who have cold symptoms but are otherwise healthy and see a physician for their condition (population 1) and those who do not (population 2).
6. Samples: For sample 1, 50 patients will be randomly chosen from those who present at the walk-in clinic of a general hospital with cold symptoms but evince no other signs of illness. Data for sample 2 are more difficult to obtain. A random sample of five pharmacies in the area is taken, and each is monitored for customers who have signs of a cold. Of these customers, a random sample of 10 who state they are not seeing a physician is taken from each pharmacy, with the customers agreeing to be included and to participate in follow-up by telephone.
7. Biases and steps to prevent them: There are several possible sources of bias, as there are in most medical studies. Two of the major questions of bias are the following: (1) Are patients and customers at and in the vicinity of our general hospital representative of those in general? and (2) Are customers who buy cold medicines at pharmacies representative of cold sufferers who do not see physicians? We can answer question (1) by analyzing the demographics statistically after our study is complete, whereas question (2) requires a leap of faith.
8. Recycle: These steps seem to be adequately consistent as they are.
9. Statistical test and choice of $\alpha$: Two-sample $t$-test with $\alpha = 0.05$.
10. Sample size: Chapter 21, Sample size estimation, explains that we need not only (1) $\alpha (= 0.05)$ but also (2) the power $(1 - \beta$, which we take to be power $= 0.80)$, (3) the difference between means that we believe to be clinically meaningful (which we choose as 2 days), and (4) $\sigma_1$ and $\sigma_2$. We estimate $\sigma_1$ to be 3 from a pilot survey of patients with cold symptoms who were followed up by telephone and assume $\sigma_2$ is the same. Section 21.4 explains that we need at least 36 in each sample; our planned 50 per group is a large enough sample.

Now we may begin to collect our data.
**Exercise 7.10**

Using the $2 \times 2$ table in DB2, follow the first nine steps of Section 7.8 in setting up a test to learn if the drug reduces nausea score.

**Exercise 7.11**

From DB13, we want to learn if our clinic results agree with those of the laboratory. Follow the first nine steps of Section 7.8 in setting up a test.

**REFERENCE**

8.1 OVERVIEW

The basis of tolerance, prediction, and confidence intervals

Hematocrit (Hct) values (measured in percent) for healthy patients are not all the same; they range over an interval. What is this interval? Because we see an occasional very high or low value in a healthy patient, we want an interval such that with a specified probability will include most of the healthy population; we name this a tolerance interval. This is an expression of relative frequency or likelihood. What is the width of this interval? We might say that the interval should include 95% of the healthy population, which is to say that a randomly chosen healthy patient has a 0.95 probability of being within the interval. This leads to the term tolerance interval, because our tolerance in specifying the certainty around this range leads up to a 5% \((1 - 0.95)\) error.

Now suppose that we have constructed a model for Hct values on the basis of patient characteristics (e.g., height, weight, sex, and age) for a sampled population. Such a model might be constructed using linear regression as presented in Chapter 16, Multiple linear and curvilinear regression and multifactor analysis of variance. Given this model, we might want to predict the Hct value for a new randomly sampled patient with a given set of characteristics. However, since our prediction is based upon estimates from sampled data, we also wish to quantify how good our prediction is. Thus we wish to construct an interval around our prediction that quantifies the uncertainty in our “guess.” This interval is termed a prediction interval. A prediction interval is a range of likely values for a new observation that has a specified set of characteristics used to aid in the prediction. The precision of a prediction interval is defined by the probability that the new observation will fall in the specified range. For example, if we produce a 95% prediction interval for the Hct value of a new patient we are treating, the actual Hct of the patient should lie within the interval with probability 0.95.

Another type of goal we might have is to characterize the range of plausible values for the true mean Hct value in a population of healthy individuals. Thus we desire a range of values that characterize the population mean Hct. This type of interval is termed a confidence interval, often denoted by CI. What do we mean by confidence?
Given that the true population mean Hct in a specified population is a fixed quantity, we wish to control the probability that our interval will contain the population mean. Thus if we produce a valid 95% CI for the population mean using a sample of data, we are ensuring that if we repeatedly sampled from the population and formulated such an interval, then 95% of these intervals would in fact cover the true population mean.

**Confusion between tolerance, prediction, and confidence intervals**

Throughout the scientific literature, there is often confusion between what researchers call tolerance, prediction, and confidence intervals. Each of these intervals is distinct in the scientific goal of the interval as well as the probability statements being made. Put simply, a tolerance interval is used for producing the *range of observations* likely to arise from a population. Here, the probability statement is on the range of values across the population. A prediction interval is used for producing a range of likely values for a *new observation* (possibly a patient) on the basis of characteristics of the patient. Here, the probability statement is on the likely values of a measurement conditional upon the characteristics of the patient and accounts for our uncertainty in the prediction. A CI is used for producing a range of plausible values for a *parameter*. Here, the probability statement is on the probability that the interval covers the parameter and accounts for our uncertainty in the sample statistic. The most commonly encountered intervals in medical research are CIs, and these are somewhat ironically the most misinterpreted. As such, we will primarily focus on CIs in this chapter.

**Error rate**

In the case of tolerance, prediction, and confidence, we let the symbol $\alpha$ represent the degree of certainty we desire in the interval. For a prediction interval, $\alpha$ represents the probability that a patient’s Hct will be outside the prediction interval; in this case, $\alpha = 0.05$. For a CI, $\alpha$ represents the probability that the parameter is not covered by the CI limits. (Note that the CI limits are random as they would vary from sample to sample). We usually choose $\alpha/2$ on each side of the interval. When a patient’s Hct does fall outside the prediction interval based upon healthy subjects, we think, “It is likely, but not certain, that the patient has arisen from an unhealthy population. We must compare the Hct with other indicators to derive a complete picture.”

**Where do these intervals come from?**

In Section 3.5, it was pointed out that probabilities of occurrences correspond to areas under portions of probability distributions. Thus, if we know the distribution of Hcts, we can find the Hct values outside of which 5% of the area (2.5% in each tail) will fall. This will give rise to a tolerance interval. If we have a prediction for a new value and we can estimate the distribution of the prediction, we can find the predicted Hct values outside of
which 5% of the area (2.5% in each tail) will fall. This will give rise to a prediction interval. If we have an estimate of a parameter and can estimate the sampling distribution of this estimate (e.g., we know from the central limit theorem that the sample mean is approximately normally distributed in large samples), then we can construct an interval that will cover the true parameter 95% of the time. The mechanism for deriving these intervals from probability distributions will be seen in the following sections.

Other uses for probabilities

The use of probabilities is by no means limited to risks of error in testing (as in the last chapter) or to CIs. Suppose we needed to know the chance of encountering a healthy patient with an Hct less than 30%. If we knew the Hct probability distribution, we could calculate this chance as the area under that distribution to the left of the horizontal axis value of 30%. Rarely must we calculate such probability values directly, because we may use computers and/or tables to find them. The method for finding probabilities from the distributions commonly met in statistics was seen in Chapter 6, Finding probabilities.

8.2 TOLERANCE INTERVALS FOR PATIENT MEASUREMENTS

Tolerance intervals are often used for characterizing typical responses from patients. Suppose, for example, we should find that healthy Hct values arise from a \( N(47,3.6^2) \) probability distribution. [Recall that \( N(47,3.6^2) = N(47,12.96) \) symbolizes a normal distribution with mean \( \mu = 47 \) and standard deviation \( \sigma = 3.6 \).] Table I shows that when \( 1 - \alpha \) (the area except for both tails) is 0.95, the area is enclosed by 1.96 standard deviations above and below the mean. This implies that the upper limit of the healthy interval lies at \( \mu + 1.96\sigma = 47 + 1.96 \times 3.6 = 54 \). Similarly, the lower limit lies at \( \mu - 1.96\sigma = 47 - 1.96 \times 3.6 = 40 \). Our healthy interval is 40%–54%. If we have a healthy patient, we would bet 95–5 (or 19–1) that the Hct will be in the interval, which is to say we are 95% confident that the interval will include a healthy patient.

We can restate as follows: The probability that a randomly drawn Hct from \( N(47,3.6^2) \) is contained in the interval \( 40,54 \) is 0.95.

A briefer and easier way to say the same thing would be \( P[40 < \text{Hct} < 54] = 0.95 \).

Accounting for uncertainty in distributional parameters

In the previous example, we assumed that Hct values arise from a \( N(47,3.6^2) \) probability distribution. However, in practice, we will generally have estimated the mean of 47 and standard deviation of 3.6 using a sample of data. These estimates will have uncertainty associated with them. In general, this uncertainty should also be accounted for when producing a tolerance interval.
EXAMPLE: A TOLERANCE INTERVAL FOR PATIENTS’ EXTENT OF TRACHEAL CARINA RESECTION

We find the extent of resection of the tracheal carina in DB12 to be approximately normal with \( m = 2.96 \) and \( s = 1.24 \). The sample size is large enough to treat \( s \) as \( \sigma \). In the normal table (Table I), we find that \( 1 - \alpha = 0.95 \) is associated with \( z = 1.96 \), or the tolerance interval on an observation extends \( \pm 1.96\sigma \) on either side of \( m \). Thus the end points of the interval would be \( 2.96 - 1.96 \times 1.24 = 0.53 \) and \( 2.96 + 1.96 \times 1.24 = 5.39 \). This tells us that a new patient presenting should have 95% probability to have a resection between about 0.5 and 5.5 cm if they are drawn from a similar population that was used to obtain our sample. In this particular instance, the tolerance interval on an individual patient is not very helpful clinically, since the surgically reasonable range is about 1–6 cm anyway. In other cases, such as choosing a healthy range for a laboratory test result, tolerance intervals on individuals can be quite useful.

Exercise 8.1

A probability distribution for variable \( X \) is expressed as \( N(2, 9) \). How would you transform it to \( N(0, 1) \), the standard normal distribution for variable \( Z \)?

Exercise 8.2

For the \( N(2, 9) \) distribution, what are the values on the x (horizontal) axis that enclose 95% of the area under the curve?

Exercise 8.3

In DB12 the age distribution of patients undergoing carinal resections is approximately normal with mean = 47.8 years and standard deviation = 14.8 years. The sample is large enough to take the standard deviation as if it were the population \( \sigma \). You have a 12-year-old patient whose tracheal carina requires resection. Does this patient fall within a 95% tolerance limit of age for individual patients, or is this patient improbably young?

Exercise 8.4

In DB13, let us take the 104 laboratory INR values’ standard deviation as the population \( \sigma \). We find \( m = 2.28 \) and \( \sigma = 0.63 \). Find the 95% tolerance interval for patients. Does this tolerance interval match the patient “in range” interval?

8.3 CONCEPT OF A CONFIDENCE INTERVAL FOR A PARAMETER

In the previous section, we used a distribution of patients to obtain a tolerance interval, in relation to which we could interpret an observation from a single patient. This observation may be thought of as a sample of size of 1 from a population with known
Although this is often useful in clinical practice, in research we are interested in the CI for a parameter, most often a mean or, less often, a standard deviation.

**Confidence interval for a parameter defined**

A random sample is a set of observations drawn from a population, in which the method of drawing is random. When we calculate a descriptive statistic from a random sample, we obtain an estimate of the equivalent parameter. For example, \( m \) and \( s \) are our best estimates of \( \mu \) and \( \sigma \), respectively. But what is the accuracy of these estimates? How much confidence do we have that the estimates are “on target”? We want an interval that will tell us the width of the target. A CI is an interval based on the probability distribution of our estimator, which expresses the confidence, or probability, that the interval contains the parameter being estimated.

Note that in the previous description, it is the interval that is random. Hence, when we refer to the probability that the interval contains the parameter, we are referring to the proportion of times that a CI computed in the same manner, over many samples of data, covers the true value of the parameter. We refer to this as the coverage probability of the CI.

**With what statistics can we use confidence intervals?**

The parameter for which a CI is most frequently computed is the population mean, but CIs are not restricted to the mean. We can put CIs on any parameter: a median (What is the 50th percentile of Hct values in the normal population?), a standard deviation (Is the effect of a new drug less variable and therefore more dependable than an older one?), or others. The distribution from which the interval values are chosen is specified by the statistic on which the interval is desired: normal for a mean, chi-square for a variance (with the square root taken of the results to apply to standard deviations), etc.

**A GENERAL STATEMENT OF CONFIDENCE**

A confidence interval (CI) need not be constrained to 0.95 and it may be derived from any distribution. A general statement for a CI with any desired interval may be stated as follows (with indentations designed to help identify subconcepts):

The probability

that the limits of the interval

as derived from the probability distribution of a sample statistic

contains a parameter

is given by

the area of the distribution over that interval.

The concept in Eq. (8.1) usually is written in the following format:

\[
P[\text{lower CI limit } < \text{parameter } < \text{upper CI limit}] = 1 - \alpha
\]  

(8.2)
which implies that the interval defined by the CI limits will enclose the parameter with probability $1 - \alpha$. Note here that it is the confidence limits that are random. That is, over repeated studies, each producing a CI, $100 \times (1 - \alpha)$% of them will cover the true value of the parameter of interest.

Caution: Confidence intervals and asymmetry

The reader of medical articles most often sees CIs as a statistic plus or minus ($\pm$) the confidence distance above or below. Sometimes a graph shows the statistic’s value as the height of a bar with a whisker above the bar showing the upper confidence bound, implying the lower bound to be an equal distance below the bar height. It should be noted that such expressions are not acceptable for statistics arising from asymmetric distributions. Confidence distances above and below the statistic are not equal for many statistics, including standard deviations (chi-square distribution), rates as a proportion quite different from 0.5 (binomial distribution), medians from skewed distributions, individual readings from skewed distributions, and some others. For these cases, separate upper and lower limits must be given, not plus or minus a single value.

8.4 CONFIDENCE INTERVAL FOR A POPULATION MEAN, KNOWN STANDARD DEVIATION

EXAMPLE POSED: UPPER CONFIDENCE LIMIT FOR MEAN TRACHEAL RESECTION

The extent of resection of the tracheal carina (DB12) has an approximately normal distribution with mean $m = 2.96$ and standard deviation $s = 1.24$. The sample size is large enough to treat $s$ as $\sigma$. What would be a 95% CI for the population mean?

As a variation on this question, suppose we are concerned only with larger resections that pose a threat because of their size; the smaller resections do not. Thus we seek a critical value cutting off $\alpha$ proportion in only the upper tail, which could be thought of as a one-tailed CI. Suppose we still want a 95% CI.

METHOD

Eq. (8.2) applied to means would be expressed as $P[\text{lower CI limit} < \mu < \text{upper CI limit}] = 1 - \alpha$. The limits on a population mean $\mu$ would be given by the sample mean $m$ plus or minus a number of standard errors of the mean (SEMs) as indicated by areas under the normal curve. The SEM is given by dividing the standard deviation of the individual observations by the square root of the number of observations, or $\text{SEM} = \frac{\sigma_m}{\sqrt{n}}$. The number (Continued)
(CONTINUED)
of SEMs is $z_{1-\alpha/2}$, the distance on the normal curve outside of which each tail $\alpha/2$ of the area lies. By substituting these elements in Eq. (8.2), we find

$$P[m - z_{1-\alpha/2}\sigma_m < \mu < m + z_{1-\alpha/2}\sigma_m] = 1 - \alpha.$$  (8.3)

In the probability statement given in expression (8.3) and throughout the remainder of this chapter, we are assuming the statistic (in this case, $m$) is random as opposed to be observed for a given study. Thus the value of the statistic would change over repeated sampling, leading to the previous probability statement.

**Method Applied to a 95% Confidence Interval**

Table 8.1 shows the most frequently used entries from Table I. If 95% confidence is wanted, we look in Table I or Table 8.1 for 0.950 under “two-tailed 1 $\alpha$.” To its left in the first column (i.e., under $z$), we find 1.960. Thus our CI limits are the sample mean $m \pm 1.96 \times \sigma_m$, the standard error of the mean (SEM; the standard deviation of $m$). In symbols,

$$P[m - 1.96\sigma_m < \mu < m + 1.96\sigma_m] = 0.95.$$  (8.4)

Table 8.1 The most used entries from Table I, the normal distribution.

<table>
<thead>
<tr>
<th>$z$</th>
<th>One-tailed $\alpha$</th>
<th>$1 - \alpha$</th>
<th>Two-tailed $\alpha$</th>
<th>$1 - \alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.645</td>
<td>.050</td>
<td>.950</td>
<td>.100</td>
<td>.900</td>
</tr>
<tr>
<td>1.960</td>
<td>.025</td>
<td>.975</td>
<td>.050</td>
<td>.950</td>
</tr>
<tr>
<td>2.326</td>
<td>.010</td>
<td>.990</td>
<td>.020</td>
<td>.980</td>
</tr>
<tr>
<td>2.576</td>
<td>.005</td>
<td>.995</td>
<td>.010</td>
<td>.990</td>
</tr>
</tbody>
</table>

*For distances ($z$) to the right of the mean, given are (a) one-tailed $\alpha$, area under the curve in the positive tail; (b) one-tailed $1 - \alpha$, area under all except the tail; (c) two-tailed $\alpha$, areas combined for both positive and negative tails; and (d) two-tailed $1 - \alpha$, the area under all except the two tails.

**Method Applied to Confidence Levels Other than 95%**

For any other level of confidence (e.g., 90% or 99%), for a sample mean with a known population standard deviation, we follow the same pattern, but looking in the 0.90 or 0.99 row in Table I.

**One-Tailed Confidence Interval**

If we are concerned with the mean falling on one side only, say the larger side, we would seek a CI limit cutting off all $\alpha$ proportion in only the upper tail. Rewriting Eq. (8.3), we get

$$P[\mu < m + z_{1-\alpha}\sigma_m] = 1 - \alpha.$$  (8.5)

For a $1 - \alpha = 0.95$ confidence interval, we look for 0.95 under the “one-tailed applications” column in Table 7.1 and find that $z_{1-\alpha}$ is 1.645. Substituting into Eq. (8.5), we obtain

$$P[\mu < m + 1.645\sigma_m] = 0.95$$  (8.6)
EXAMPLE COMPLETED: UPPER CONFIDENCE LIMIT ON MEAN TRACHEAL RESECTION

We want a 95% CI for the mean extent of carinal resection. From Table 8.1 (or Table I) under “two-tailed applications,” we find that $1 - \alpha = 0.95$ is associated with $z = 1.96$, or the CI for the mean extends $\pm 1.96\sigma_m$ on either side of $m$. $\sigma_m = \sigma / \sqrt{n} = 1.24 / \sqrt{134} = 0.107$. Thus the end points of the interval would be $2.96 - 1.96 \times 0.107 = 2.75$ and $2.96 + 1.96 \times 0.107 = 3.17$. By using the form in Eq. (8.4), the 95% CI is given by $(2.75, 3.17)$ and voiced as, “The probability that the interval 2.75 to 3.17 contains $\mu$ is 0.95.”

To find the one-tailed CI, we look under “one-tailed applications” in Table 8.1 (or Table I). For $1 - \alpha = 0.95$, we find $z_{1-\alpha}$ (number of standard deviations to right of mean) is 1.645, and we substitute these and the values for $m$ and $\sigma_m$ into Eq. (8.6) to obtain an upper bound of a 95% CI of $2.96 + 1.645 \times 0.107 = 3.136$. We are 95% confident that the mean resection will not exceed 3.14 cm.

ADDITIONAL EXAMPLE 1: CONFIDENCE INTERVAL FOR MEAN AGE OF TRACHEAL RESECTION

In the initial example of this section, we found a 95% CI for the mean extent of resection of the tracheal carina (DB12). Let us find one on patient age for the same sample of patients: $m = 47.84$ and we take $\sigma = 15.78$, so that $\sigma_m = 15.78 / \sqrt{134} = 1.36$. By substituting in Eq. (8.4), we find the 95% CI to be

$$(47.84 - 1.96 \times 1.36, 47.84 + 1.96 \times 1.36) = (45.17, 50.51)$$

We are 95% confident that the interval of 45.2–50.5 years will enclose the population mean age.

ADDITIONAL EXAMPLE 2: CONFIDENCE INTERVAL FOR MEAN TREATMENT TIME FOR ARM FRACTURES

An orthopedist is experimenting with the use of Nitronox as an anesthetic in the treatment of children’s arm fractures.1,2 He anticipates that it may provide an attractively short procedure. He treats $n = 50$ children and records the treatment time in minutes. He finds $m = 26.26$ and $\sigma = 7.13$. (The sample size is large enough to use the calculated standard deviation as $\sigma$.) He wants a 95% CI for mean treatment time. He will require $\sigma_m = \sigma / \sqrt{n} = 7.13 / \sqrt{50} = 1.008$. From Eq. (8.4) a 95% CI is given by

$$(m - 1.96 \times \sigma_m, m + 1.96 \times \sigma_m) = (26.26 - 1.96 \times 1.008, 26.26 + 1.96 \times 1.008) = (24.28, 28.24)$$

The orthopedist is 95% confident that the interval from 24 and 28 minutes contains the true mean time to treat.
Exercise 8.5
Find 90% and 99% CIs for the mean age of patients with carinal resection (DB12).

Exercise 8.6
The orthopedist in the Nitronox example at the end of Section 8.4 also is interested in the patients’ pain, which he has measured using the CHEOPS rating form. For his \( n = 50 \) patients, \( m = 9.16 \) and \( \sigma = 2.04 \). Find (a) 95%, (b) 99%, and (c) 90% CIs for mean pain rating.

Exercise 8.7
For the laboratory INR values of Exercise 8.4 (DB13), find a 95% CI for the mean.

8.5 CONFIDENCE INTERVAL FOR A POPULATION MEAN, ESTIMATED STANDARD DEVIATION

EXAMPLEPOSED:CONFIDENCIETRVALFORMEANPROSTATEVOLUME
Suppose we wanted to use the data from Table DB1.1 to put a 95% CI for the population mean prostate volume, but now we do not know the large sample standard deviation as in the last section; we must use the \( t \) distribution rather than the normal. We have found that the 10 prostate volumes in Table DB1.1 have \( m = 32.73 \text{ mL} \) and \( s = 15.92 \text{ mL} \). What is the 95% CI for \( \mu \)?

METHOD
More often than not in medical applications, we do not know \( \sigma \) and must estimate it by a small sample \( s \). Finding a confidence interval for \( \mu \) using \( m \) and \( s \) follows the same logic as using \( m \) and \( \sigma \), except that we use a \( t \) table rather than a normal table. Because the \( t \) distribution has a greater spread than the normal, the confidence interval will be slightly wider. If we want 95% confidence, we look under 0.95 for “two-tailed \( 1-\alpha \) (except both tails)” in Table II for the appropriate \( df \). Replacing the \( \sigma \) and \( z \) symbols in Eq. (8.3) with the equivalent \( s \) and \( t \) symbols, we obtain

\[
P[m - t_{1-\alpha/2}s_m < \mu < m + t_{1-\alpha/2}s_m] = 1 - \alpha. \tag{8.7}
\]

(We know from Section 4.9 that the sample standard deviation of \( m \), i.e., sample SEM denoted \( s_m \), is \( s/\sqrt{n} \).)

EXAMPLECOMPLETED:CONFIDENCIETRVALFORMEANPROSTATEVOLUME
We calculate \( s_m = s/\sqrt{n} = 15.92/3.16 = 5.04 \). To find the \( t \)-value for 95% confidence, we look under 0.95 for “two-tailed \( 1-\alpha \) (except both tails)” in Table II for 9 \( df \).
Table 8.2 shows a segment of Table II. We find $t_{1-\alpha/2} = 2.262$. Substituting for $m$ and $s_m$ in Eq. (8.7), we find the 95% CI to be

$$
(m - t_{1-\alpha/2}s_m, m + t_{1-\alpha/2}s_m) = (32.73 - 2.262 \times 5.04, 32.73 + 2.262 \times 5.04) = (21.33, 44.13)
$$

We are 95% confident that the interval (21.33, 44.13) encloses the true population mean.

**ADDITIONAL EXAMPLE 1: CONFIDENCE INTERVAL FOR MEAN RATINGS OF POSTPARTUM STRETCH MARKS**

A dermatologist is studying the efficacy of tretinoin in treating women’s ($n = 15$) postpartum abdominal stretch marks. Tretinoin was used on a randomly chosen side of the abdomen and a placebo on the other side. Neither the patient nor the investigator knew which side was medicated. The patient rated the improvement on each side on a 10-cm-long visual analog scale, and ratings were recorded as a reading between 0 and 10. The difference, treated-side rating minus untreated-side rating, indicating the excess of improvement with tretinoin versus the placebo, was calculated. $m = -0.33$ and $s = 2.46$. $s_m = s/\sqrt{n} = 2.46/\sqrt{15} = 0.64$. From Table II or 8.2, the 95% $t$-value for 14 df is 2.145. When these values are substituted in Eq. (8.7), the CI evolves as

$$
(m - t_{1-\alpha/2}s_m, m + t_{1-\alpha/2}s_m) = (-0.33 - 2.145 \times 0.64, -0.33 + 2.145 \times 0.64) = (-1.70, 1.04)
$$

The dermatologist is 95% confident that the mean stretch-mark improvement is between $-1.7$ and $+1.0$. The sample average is negative (untreated side better), the CI includes 0, and the upper confidence bound of 1 is not very important clinically; evidence is insufficient to conclude a benefit from tretinoin in this particular medical application.

**Table 8.2** A segment of the $t$ table, Table II.

<table>
<thead>
<tr>
<th>Two-tailed $\alpha$</th>
<th>.20</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
<th>.002</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-tailed $1 - \alpha$</td>
<td>.80</td>
<td>.90</td>
<td>.95</td>
<td>.98</td>
<td>.99</td>
<td>.998</td>
<td>.999</td>
</tr>
</tbody>
</table>

$df = 9$

| 14 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 15 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |

$df = 15$

| 14 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.373 | 4.073 |

$df = 15$

*Selected distances (t) to the right of the mean are given for various degrees of freedom (df), for two-tailed $\alpha$, areas combined for both positive and negative tails, and two-tailed $1 - \alpha$, area under all except the two tails.*
**ADDITIONAL EXAMPLE 2: CONFIDENCE INTERVAL FOR MEAN THEOPHYLLINE LEVELS**

In DB3, serum theophylline levels were measured in patients with emphysema before they were given azithromycin, at 5 days, and at 10 days. What are 95% CIs on mean serum theophylline levels at these three points in time? \( n = 16, \ m_0 = 10.80, \ s_0 = 3.77, \ m_5 = 9.81, \ s_5 = 4.61, \ m_{10} = 10.14, \) and \( s_{10} = 3.98. \) SEMs are 0.94, 1.15, and 1.00, respectively. From Table II or 8.2 the intersection of the column for two-tailed \( 1 - \alpha = 0.95 \) with the row for 15 \( df \) gives \( t_{1-\alpha/2} = 2.131. \) After substitution in Eq. (8.7), these values yield the following CIs:

- **Baseline:** \( (10.80 - 2.131 \times 0.94, 10.80 + 2.131 \times 0.94) = (8.80, 12.80) \)
- **5 Days:** \( (9.81 - 2.131 \times 1.15, 9.81 + 2.131 \times 1.15) = (7.36, 12.26) \)
- **10 Days:** \( (10.14 - 2.131 \times 1.00, 10.14 + 2.131 \times 1.00) = (8.01, 12.27) \)

**Exercise 8.8**

Among the indicators of patient condition following pyloromyotomy (correction of stenotic pylorus) in neonates is time (hours) to full feeding. A surgeon wants to place a 95% CI for mean time to full feeding. The surgeon has readings from \( n = 20 \) infants. Some of the data are 3.50, 4.52, 3.03, 14.53, ... The surgeon calculates \( m = 6.56 \) and \( s = 4.57. \) What is the CI?

**Exercise 8.9**

In DB5, plasma silicone level on \( n = 30 \) patients was measured before and again after silicone implantation. What is a 95% confidence level for the mean difference preop—postop?

**Exercise 8.10**

An emergency medicine physician samples the heart rate of \( n = 8 \) patients following a particular type of trauma. She finds the mean \( m = 67.75 \) beats/min and the standard deviation \( s = 9.04 \) beats/min. What is her 95% CI for the mean?

**Exercise 8.11**

In the kidney cooling data of DB15, the baseline (time 0) means are \( m_1 = 34.40 \) for saline infusion and \( m_2 = 36.12 \) for ice slush, each based on 12 readings. (Use 2 decimal places throughout for easier calculation.) The respective estimated standard deviations are \( s_1 = 3.25 \) and \( s_2 = 1.17. \) Find 95% CIs for the means for each treatment. Are they similar?

### 8.6 CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

**Proportions fall into two types**

Proportions fall in the interval (0,1). However, methods must be divided into two types: cases in which the proportion lies toward the center of the interval away from
the extremes 0 and 1, that is, the central proportion, which may be thought of as the rate of occurrence of a common event, and cases in which the proportion lies very close to 0 or 1, that is, the extreme proportion, which may be thought of as the rate of occurrence of a rare event. The reason is that they lead to different distributions (see later in the “Methods” section).

**EXAMPLE POSED, COMMON EVENT: CONFIDENCE INTERVAL FOR PROPORTION OF POSITIVE BIOPSIES**

Of our 301 urology biopsies from DB1, 95 had positive results, yielding a sample proportion \( p = 0.316 \). What is a plausible range for the positive rate? We want a CI for the theoretical proportion \( \pi \).

**EXAMPLE POSED, RARE EVENT: CONFIDENCE INTERVAL FOR PROPORTION OF CHILDREN WITH HIGH LEAD LEVELS**

Children with high lead levels are found in a certain hospital’s catchment. Of 2500 children sampled, 30 with high lead levels are found, yielding \( p = 0.012 \). How far may a rate deviate from this \( p \) before the hospital administration suspects an atypical situation? We want a CI for the theoretical proportion \( \pi \).

**METHOD**

The confidence interval we seek here is on the theoretical but unknown population proportion \( \pi \), which we estimate by the sample proportion \( p \).

**Common Event**

In this case, \( \pi \) is not close to 0 or 1 but is nearer to 0.5. It has been shown that if \( p \) is the sample proportion computed on the basis of random variables from a binomial distribution, the distribution of \( p \) is approximated by the normal distribution with sample mean \( \mu = p \) and standard deviation \( \sigma = \sqrt{p(1-p)/n} \). The mean \( p \) and standard deviation \( \sigma \) are substituted in the confidence interval pattern Eq. (8.3). Specifically, for 95% confidence, substitute in Eq. (8.4) to obtain

\[
P[p - 1.96 \times \sigma - 1/2n < \pi < p + 1.96 \times \sigma + 1/2n] = 0.95,
\]

where the \( 1/2n \) components are continuity corrections to improve the approximation.

**Rare Event**

In this case, \( \pi \) is very close to 0 or 1. The distribution of \( p \) can be approximated by the normal with standard deviation estimated as the smaller of \( \sigma = \sqrt{p/n} \) or \( \sqrt{(1-p)n} \). Again, \( p \) and \( \sigma \) are substituted in the confidence interval pattern of Eq. (8.3). Specifically, for 95% confidence, substitution in Eq. (8.4) yields

\[
P[p - 1.96 \times \sigma < \pi < p + 1.96 \times \sigma] = 0.95
\]

(8.9)
8.6 Confidence Interval for a Population Proportion

**EXAMPLE COMPLETED, COMMON EVENT: CONFIDENCE INTERVAL FOR PROPORTION OF POSITIVE BIOPSIES**

In this example, the mean \( p = 0.316 \). The standard deviation \( \sigma \) is given by

\[
\sigma = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.316 \times 0.684}{301}} = 0.0268.
\]

By using Eq. (8.8), construct a 95% CI for \( \pi \) as

\[
(p - 1.96 \times \sigma - 1/2n, p + 1.96 \times \sigma + 1/2n)
= (0.316 - 1.96 \times 0.0268 - 0.00166, 0.316 + 1.96 \times 0.0268 + 0.00166)
= (0.262, 0.370)
\]

The plausible values for the true proportion of positive prostate biopsies in the population of patients presenting with urological problems is in the range of 0.26—0.37.

**EXAMPLE COMPLETED, RARE EVENT: CONFIDENCE INTERVAL FOR PROPORTION OF CHILDREN WITH HIGH LEAD LEVELS**

Because \( p = 0.012 \), \( \pi \) is evidently close to 0, implying the Poisson distribution. \( \sigma = \sqrt{p/n} = \sqrt{0.012/2500} = 0.00219 \). The hospital administration wants to be sure that it does not have too many high-lead children in its catchment and therefore chooses a 99% CI. From Table 8.1, the 0.99 two-tailed \( 1-\alpha \) yields a corresponding \( z \) of 2.576. Replacing the 1.96 in Eq. (8.9) with 2.576, we obtain a 99% CI given by

\[
p - 2.576 \times \sigma, p + 2.576 \times \sigma
= (0.012 - 2.576 \times 0.00219, 0.012 + 2.576 \times 0.00219)
= (0.0064, 0.0176)
\]

By focusing on the right tail, the hospital administration may be 99.5% confident that no more than 1.8% of children in its catchment have high lead levels.

**ADDITIONAL EXAMPLE: PATIENT SATISFACTION WITH ANESTHESIA**

In a study anesthetizing patients undergoing oral surgery by a combination of propofol and alfentanil, 89.1% of 110 patients rated the anesthetic as highly satisfactory or excellent. What are 95% confidence limits on \( \pi \), the proportion of the population satisfied with the anesthetic? Because \( \pi \) is not near 0 or 1, the normal approximation to the binomial is appropriate. \( p = 0.891 \) and \( \sigma = \sqrt{0.891 \times 0.118/110} = 0.031 \) are substituted in Eq. (8.8) to yield the following CI:

\[
(0.891 - 1.96 \times 0.031 - 1/220, 0.891 + 1.96 \times 0.031 + 1/220) = (0.826, 0.956)
\]
Hence, the plausible values for the true proportion of patients that will rate the anesthetic as highly satisfactory or excellent range from 0.83 to 0.96.

**Rule of three: If number of rare events is 0**

A useful tool in the analysis of medical studies is the Rule of Three. Suppose there have been no events of a certain type, used most frequently for rare adverse events, in a sample of \( n \) patients. If \( n \geq 30 \), the approximation method for the binomial or Poisson distribution (either may be used) allows us to place confidence limits on the eventual number of such events. We are 95% confident that the true number of events is contained by an interval ranging from 0 to 1 in \( n/3 \). Thus if we had 60 patients with no events, we are 95% confident that the true (population) number of events is contained by an interval ranging from 0 to 1 in \( 60/3 \), or 5%.

**Exercise 8.12**

Based on a sample of 226 patients, a 5-year recurrence rate of basal cell carcinoma when lesions measured more than 20 mm was 0.261. Compute a 95% CI for the population proportion, \( \pi \).

**Performance of Wald-based confidence intervals**

The form of the CI given in (8.9) is referred to as a Wald-based CI. This is because the CI is computed by taking the estimated value, \( p \), then adding and subtracting the standard error of the estimate times the distribution of the standardized estimate. This type of CI is named after Abraham Wald, an American mathematician and statistician who developed much of the testing techniques that give rise to this form of CI.

For large values of \( n \) or values of \( \pi \) near 0.5, the Wald-based interval in (8.9) tends to perform well in practice, meaning that over repeated experimentation, if a \( (1 - \alpha) \% \) CI was formed for each experiment, then over many experiments \( (1 - \alpha) \% \) of the CIs would cover the true value of the unknown probability \( \pi \). That is to say that the **coverage probability** of these intervals is correct. However, multiple authors (see, e.g., Newcombe and Agresti and Coull) have noted that when \( n \) is small (say 50 or less) and \( \pi \) is near 0 or 1 (say 0.05, 0.1, 0.90, 0.95), then the coverage probability of the Wald-based CI can be much lower than the intended \( (1 - \alpha) \% \). As an example, when \( n = 20 \) and \( \pi = 0.10 \), simulation can show that the coverage probability of a 95% Wald-based CI is only around 85%.

**Accounting for the dependence of the variance on the mean**

While technical, the poor performance of the Wald-based confidence is primarily attributed to the fact that \( \sigma \) is determined by \( p \). The implication of this is that at the boundaries of the CI, the standard error of \( p \) is poorly estimated by \( \sigma \). To account for this, many authors (see Newcombe as one example) have suggested using an alternate
CI known as the score-based or Wilson CI for $\pi$. This confidence is slightly more complicated and takes the form

$$\left(\frac{2p + z_{1-(\alpha/2)}^2 - z_{1-(\alpha/2)}\sqrt{z_{1-(\alpha/2)}^2 + 4np(1-p)}}{2\left(n + z_{1-(\alpha/2)}^2\right)}\right) \pm \left(\frac{2p + z_{1-(\alpha/2)}^2 + z_{1-(\alpha/2)}\sqrt{z_{1-(\alpha/2)}^2 + 4np(1-p)}}{2\left(n + z_{1-(\alpha/2)}^2\right)}\right).$$

**EXAMPLE REVISITED: PATIENT SATISFACTION WITH ANESTHESIA**

We return to the study of anesthetizing patients undergoing oral surgery by a combination of propofol and alfentanil. In this case, 89.1% of 110 patients rated the anesthetic as highly satisfactory or excellent. While the sample size in this case is fairly large, we will consider the score-based CI relative to the previously computed Wald-based interval. Using the score-based CI and plugging in values of $p = 0.891$, $n = 110$, and critical value 1.96, we obtain a 95% CI for $\pi$, the proportion of the population satisfied with the anesthetic, to be $[0.819, 0.937]$.

### 8.7 CONFIDENCE INTERVAL FOR A POPULATION MEDIAN

**EXAMPLE POSED: CONFIDENCE INTERVAL ON MEDIAN VIRAL LOAD IN HIV PATIENTS**

Often the distribution of a sample is asymmetric. For example, consider a viral load ($VL$) assay on 165 HIV patients in DB18 (from the internet database set). The mean $VL$ is 10,100 with standard deviation 22,821. If we assume a normal distribution of $VL$ levels and use Eq. (8.7) to compute a CI for the mean, $t_{1-\alpha/2} = 1.974$ and the confidence bounds are $10,100 \pm 1.974 \times 1777$, yielding the interval $6592 - 13,608$. This result tells us some very wrong information about the distribution of viral load levels: only 6 $VL$ levels fall within the CI, with 134 (81%) below the lower bound and 25 (15%) above the upper bound. Of course, the CI is on the mean, not on the data per se, so we do not expect 95% of the data to lie in the interval. However, the extremity of irregularity serves to illustrate the issue. The distribution of $VL$ is asymmetric and quite irregular, with most of the levels small and a “bump” of large values at the upper end. The median $VL$ is 77. The extreme discrepancy between the median and mean is a strong indicator of asymmetry. To convey a reasonable description of this distribution, we need to use the median with a CI for that median.

**METHOD FOR CONFIDENCE INTERVAL FOR THE MEDIAN**

The derivation of a 95% confidence interval for the median (Mood and Graybill, pp. 407—409) is more mathematically involved than most readers of this book will want to recall.
The idea is to place all observations in order and use the binomial distribution to calculate the positions along the ordered values that will have 0.025 probability of the median being less than, and, separately, greater than, the values \( v_{\text{lower}} \) and \( v_{\text{upper}} \), respectively. These values, usually not equidistant from the median, will be the lower and upper bounds of the confidence interval.

Investigators who need confidence intervals on medians are advised to use statistical computer software to obtain them. Not all packages contain this capability; one that does is Stata.

**EXAMPLE COMPLETED: CONFIDENCE INTERVAL FOR MEDIAN VIRAL LOAD IN HIV PATIENTS**

It was noted that the sample median \( VL \) is 77 for the HIV patients in DB18. The 95% binomial confidence bounds are 0—223. In contrast with 4% of the data falling between the bounds on the mean under the normal assumption, 58% of the data fall between the bounds on the median using the rank-based method.

**ADDITIONAL EXAMPLE: CONFIDENCE INTERVAL FOR MEDIAN LOS FOR A GROUP A STREP OUTBREAK**

A sudden outbreak of Group A *Streptococcus pneumoniae* occurred at the Marine Corps Recruit Depot in San Diego in 2008, serious enough to have invoked special coverage by national television media. The data for this example may be found in DB20. The variable: length of stay in hospital (days) varied from 1 to 34. For 127 patients, about half stayed 1 or 2 days. The median and mode of the observed distribution are both 2. The sample mean and corresponding CI under the normal assumption are 3.3 and (2.6, 4.0). Thus the sample mean and even the lower bound of the corresponding CI are greater than the median. The asymmetric binomial CI for the median is (2, 3).

**8.8 CONFIDENCE INTERVAL FOR A POPULATION VARIANCE OR STANDARD DEVIATION**

**EXAMPLE POSED: CONFIDENCE INTERVAL FOR STANDARD DEVIATION OF PROSTATE VOLUMES**

Using the 10 prostate volumes from Table DB1.1, what is a 95% CI for the population standard deviation \( \sigma \)? We have calculated \( s = 15.92 \) mL that has \( n - 1 = 9 \) df. We know the probability distribution of \( s^2 \) but not of \( s \), so we shall find the interval on the population variance, \( \sigma^2 \), and take square roots. \( s^2 = 15.92^2 = 253.4464 \). How do we express a CI?
METHOD FOR CONFIDENCE ON $\sigma^2$ OR $\sigma$

We know (from Section 4.8) that sample variance $s^2$ drawn randomly from a normal population is distributed as $\chi^2 \times \sigma^2/df$. A confidence-type statement on $\sigma^2$ from this relationship, with the additional twist that the asymmetric distribution requires the chi-square values excluding $1 - \alpha/2$ in each tail to be found separately, is given by

$$P[s^2 \times df/\chi^2_R < \sigma^2 < s^2 \times df/\chi^2_L] = 1 - \alpha,$$

(8.10)

where $\chi^2_R$ is the critical value for the right tail found from Table III (see Tables of Probability Distributions) and $\chi^2_L$ is that for the left found from Table IV (see Tables of Probability Distributions). To find the confidence on $\sigma$ rather than $\sigma^2$, we just take square roots of the components within the brackets.

EXAMPLE COMPLETED: CONFIDENCE INTERVAL FOR STANDARD DEVIATION OF PROSTATE VOLUMES

Table 8.3 shows segments of Tables III and IV. Looking in Table III or Table 8.3 under right tail area = .025 for 9 df, we find 19.02. Similarly, in Table IV or Table 8.3, under left tail area = .025 for 9 df, we find 2.70. Substituting these values in Eq. (8.10), we find

$$(s^2 \times df/\chi^2_R, s^2 \times df/\chi^2_L)$$

$$= (253.45 \times 9/19.02, 253.45 \times 9/2.70)$$

$$= (119.92, 844.83)$$

Taking square roots of each limit, we obtain a 95% CI for the standard deviation of (10.95, 29.07).

We note from the first table of DB1 that the population standard deviation (excluding known BPH to make the distribution more symmetric) is 16.35 mL, falling well within the CI.

ADDITIONAL EXAMPLE 1: CONFIDENCE INTERVAL FOR PRECISION OF THERMOMETER

We are investigating the reliability of a certain brand of tympanic thermometer (temperature measured by a sensor inserted into the patient’s ear). The standard deviation will give us an indication of its precision. We want a 95% CI for this precision, that is, a range outside of which the standard deviation would lie no more than 5 times in 100 random readings. Sixteen readings (°F) were taken on a healthy patient at intervals of 1 minute. Data were 95.8, 97.4, 99.3, 97.1, . . . , yielding $s = 1.23$. $s^2 = 1.23^2 = 1.51$. We use Eq. (8.10). The chi-square value from Table III for 97.5% area except right is
\[ \chi^2_R = 27.49 \text{ for } 15 \, df \text{; and from Table IV for area except left is } \chi^2_L = 6.26. \]

We find the CI for the variance to be

\[ \left( s^2 \times \frac{df}{\chi^2_R}, \frac{s^2}{\chi^2_L} \times df \right) = \left( 1.51 \times 15/27.49, 1.51 \times 15/6.26 \right) = (0.8239, 3.6182) \]

Taking square roots of the previous limits, we obtain a CI for the standard deviation of \((0.9077, 1.9022)\). The plausible values for the true standard deviation of this thermometer lie in the interval \(0.9^\circ F - 1.9^\circ F\).

**ADDITIONAL EXAMPLE 2: CONFIDENCE INTERVAL FOR VARIABILITY OF PLASMA SILICON LEVEL**

Is plasma silicon level (DB5) less variable after implantation? Although this question properly should be tested by a formal hypothesis test as in Chapter 28, Methods you might meet, but not every day, we can make an informal comparison by contrasting the CIs on \(\sigma_{\text{pre}}\) and \(\sigma_{\text{post}}\). What are these CIs? \(s_{\text{pre}} = 0.098565, \ s_{\text{pre}}^2 = 0.009715, \ s_{\text{post}} = 0.076748, \text{ and } s_{\text{post}}^2 = 0.005890. \) \(df = 29. \) From Table 8.3 or Tables III and IV, the intersections of columns \(\alpha = 0.025\) and row \(df = 29\) yield \(\chi^2_R = 45.72\) and \(\chi^2_L = 16.05. \) Substitution in Eq. (8.10) yields a 95% CI for \(\sigma_{\text{pre}}^2\) of \((0.006162, 0.017554)\) and a 95% CI for \(\sigma_{\text{post}}^2\) of \((0.003736, 0.010642). \) By taking square roots of these limits, we find CIs for \(\sigma_{\text{pre}}\) and \(\sigma_{\text{post}}\) to be \((0.078, 0.132)\) and \((0.061, 0.103)\), respectively. The CIs are not very different from a clinical perspective.

**Exercise 8.13**

An infectious disease specialist samples the white blood cell count of \(n = 19\) patients who contracted the same nosocomial infection in an orthopedic ward. She finds that the mean count is not
unusual but suspects that the standard deviation may be larger than normal: \( s = 6000 \). What is her 95% CI for the standard deviation \( \sigma \) for the population of infected patients?

**Exercise 8.14**

The variability in platelet-derived growth factor (PDGF) has been estimated (DB9) as \( s = 19,039.7 \). How far from this may a sample standard deviation deviate before we suspect the influence of a new factor? Find a 95% CI for the population \( \sigma \).

**Exercise 8.15**

The surgeon referenced in Exercise 8.3 is concerned about the variability of time to full feeding in neonates. Even if the mean is satisfactory, if the variability is too large, the outlying neonates on the longer side of the scale would be at risk. Recall that \( n = 20 \), \( m = 6.56 \) hours, and \( s = 4.57 \) hours. Establish 95% CIs on (a) the variance and (b) the standard deviation.

**Exercise 8.16**

The ED physician of Exercise 8.4 is concerned not only about the average heart rate of the trauma patients but also about the variability. Even if the mean is satisfactory, if the standard deviation is too large, outlying patients would be at risk. Recall that \( n = 8 \), \( m = 67.75 \) beats/min, and \( s = 9.04 \) beats/min. Establish 95% CIs on (a) the variance and (b) the standard deviation.

**Exercise 8.17**

Using the sample sizes and standard deviations given in Exercise 8.11, find 95% CIs for the standard deviation of the baseline data for each treatment. Are they similar?

### 8.9 CONFIDENCE INTERVAL FOR A POPULATION CORRELATION COEFFICIENT

**EXAMPLE POSED: CONFIDENCE INTERVAL FOR CORRELATION BETWEEN AGE AND PLANTAR FLEXION**

An orthopedist is studying hardware removal in broken ankle repair. One of the variables is the maximum angle (degrees) of plantar flexion. The orthopedist wants to know whether age is related to plantar flexion. The sample correlation coefficient for the 19 patients is found to be 0.1945. What is a 95% CI for the population correlation coefficient?

**METHOD FOR A CONFIDENCE INTERVAL FOR A CORRELATION COEFFICIENT**

Confidence intervals on the population correlation coefficient \( \rho \), estimated by \( r ( = s_{xy}/s_{x}s_{y} ) \) for small sample sizes are usually too wide to be of much help. If we have a larger sample (Continued)
size, we can transform the correlation coefficient to have approximately the normal distribution. We recall that \( \ln \) denotes “natural logarithm” and \( e \) denotes the “natural number,” 2.71828 \ldots The old base-ten logarithms were used to facilitate certain difficult arithmetic in the days before computers and seldom are used anymore; \( \ln \) and \( e \) are found on every computer and most handheld calculating machines. The normally transformed correlation coefficient has sample mean

\[
m = \frac{1}{2} \ln \frac{1 + r}{1 - r}
\]

and the standard deviation estimated as

\[
\sigma = \frac{1}{\sqrt{n - 3}}
\]

Solving the transformation Eq. (8.11) for \( r \) yields a ratio of exponential expressions, which, together with Eq. (8.12), can be entered into the pattern of Eq. (8.13). This inequality can then be solved mathematically to obtain a \( 1 - \alpha \) confidence interval as

\[
p \left[ \frac{1 + r - (1 - r)e^{2z_{1-\alpha/2}/\sqrt{n-3}}}{1 + r + (1 - r)e^{2z_{1-\alpha/2}/\sqrt{n-3}}} < p < \frac{1 + r - (1 - r)e^{-2z_{1-\alpha/2}/\sqrt{n-3}}}{1 + r + (1 - r)e^{-2z_{1-\alpha/2}/\sqrt{n-3}}} \right] = 1 - \alpha.
\]

Eq. (8.13) holds for nonnegative correlation coefficients. If \( r \) is negative, symmetry properties allow a simple solution. Find the confidence limits as if the \( r \) were positive (i.e., use \( |r| \)), then change the signs on the resulting limits and exchange their positions. This mechanism is illustrated in the additional example.

**EXAMPLE COMPLETED: CONFIDENCE INTERVAL FOR CORRELATION BETWEEN AGE AND PLANTAR FLEXION**

The correlation coefficient between maximum angle of plantar flexion and age for 19 patients was 0.1945. Eq. (8.13) is not so intimidating when broken into pieces: \( 1 + r = 1.1945 \), \( 1 - r = 0.8055 \), \( n - 3 = 16 \), and \( \sqrt{16} = 4 \). For 95% confidence, \( z_{1-\alpha/2} = 1.96 \). The exponent of \( e \) will be \( \pm 2 \times 1.96/4 = \pm 0.98 \). By substituting these components into Eq. (8.13), the orthopedist finds a 95% CI given by

\[
\left( \frac{1.1945 - 0.8055 \times e^{0.98}}{1.1945 + 0.8055 \times e^{0.98}}, \frac{1.1945 - 0.8055 \times e^{-0.98}}{1.1945 + 0.8055 \times e^{-0.98}} \right) = (-0.2849, 0.5961).
\]

Because the population correlation coefficient may be anywhere from about \(-0.28\) to \(+0.60\) (the CI crosses zero), the orthopedist concludes that age has not been shown to be an influence.
ADDITIONAL EXAMPLE: CONFIDENCE INTERVAL FOR CORRELATION BETWEEN HCT AND ERYTHROPOIETIN

In a study on increasing red blood cell mass before surgery, the correlation between Hct and serum erythropoietin levels was of interest. An inverse correlation was anticipated for patients with normal renal function. Levels were measured in \( n = 126 \) patients and the correlation coefficient was calculated as \( r = -0.59 \). Let us find a 95% CI for the population correlation coefficient \( \rho \). Because \( r \) is negative, we first find the limits as if it were positive; we temporarily use \( r = +0.59 \). Then, \( 1 + r = 1.59 \), \( 1 - r = 0.41 \), \( n - 3 = 123 \), and \( \sqrt{123} = 11.0905 \). The exponent is \( \pm 0.3535 \). Substitution in Eq. (8.13) yields (0.4628, 0.6934). To convert the limits to the observed negative correlation, \( r = -0.59 \), we change 0.4628 to \(-0.4628\) and put it in the right (larger limit) position and then change 0.6934 to \(-0.6934\), putting it in the left (smaller limit) position in Eq. (8.13). The resulting CI for \( \rho \) when \( r = -0.59 \) from a sample of 126 becomes \((-0.6934, -0.4628)\).

Exercise 8.18

An orthopedist is experimenting with the use of Nitronox as an anesthetic in the treatment of arm fractures in children.\(^1,2\) He finds that it provides a shorter duration of surgery and less pain than other procedures he has used. In addition, he expects that the shorter the procedure, the less the perceived pain. Do the data bear out this expectation? More specifically, is the CI for the population correlation coefficient between the two variables entirely within the positive range? He treats \( n = 50 \) children, recording the treatment time in minutes and the score from a CHEOPS pain scale. He finds \( r = 0.267 \). Find the 95% CI for the population correlation coefficient.

REFERENCES

5. Missing sources. The sources of a few examples could not be found despite a strong effort to locate them. Such data that could not be referenced were slightly altered so as not to reflect on any investigator later appearing.
Tests on categorical data

9.1 CATEGORICAL DATA BASICS

Terms and symbols

Categorical or nominal data are data that fall into distinct categories rather than being measured as a point on a scale or ranked in order. The basic sample statistics are the count, obtained by counting the number of events per category, and the proportion or rate of events in a category, the count divided by the total number in the variable. If \( n \) denotes the total number of events and \( n_a \) the number in category \( a \), the proportion in \( a \) is \( p_a = n_a / n \). Percent, \( 100p_a \), is often used. \( \pi \) denotes the population proportion, which is known from theory or closely approximated by a very large sample.

EXAMPLE: RATE OF THROMBOSIS RECURRENCE

A vascular surgeon treated 29 patients endoscopically for thrombosis in the leg, after which thrombosis recurred in 6 patients. \( n = 29 \), \( n_a = 6 \), and \( p_a = 6/29 = 0.207 \). About 21% of the patients experienced recurrence of thrombosis.

Organizing data for categorical methods

If the relationship between two variables is of interest, such as whether they are independent or associated, the number of observations simultaneously falling into the categories of two variables is counted. For example, the vascular surgeon is interested in the relation between recurrence and ulcer condition. The first variable may be categorized as recurred or not recurred and the second as open ulcers or healed ulcers. This gives rise to the four categories of counts, which are shown together with example counts in Table 9.1. We can see that five is the number of recurrences, contingent on the ulcers being healed. Tables of this sort are called contingency tables, because the count in each cell is the number in that category of that variable contingent on also lying in a particular category of the other variable.

Contingency table symbols

Contingency tables often have more than four categories, that is, a \( 2 \times 2 \) table. For example, if the preulcerous conditions of heavy pigmentation were added to the ulcer
CHAPTER 9 Tests on categorical data

### Table 9.1 Contingency table of thrombosis recurrence and ulcer condition showing symbols used in formulas.

<table>
<thead>
<tr>
<th>Ulcers</th>
<th>Open</th>
<th>Healed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurred</td>
<td>Yes</td>
<td>$n_{11} = 1$</td>
<td>$n_{12} = 5$</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>$n_{21} = 7$</td>
<td>$n_{22} = 16$</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>$n_{-1} = 8$</td>
<td>$n_{-2} = 21$</td>
</tr>
</tbody>
</table>

In general, the number of rows is denoted by $r$ and the number of columns by $c$, yielding an $r \times c$ table. Because the values of $r$ and $c$ are unknown until a particular case is specified, the cell numbers and the marginal row and column totals must be symbolized in a way that can apply to tables of any size. The usual symbolism is shown in Table 9.1: $n$, for “number”, with two subscripts, the first indicating the row and the second indicating the column. Thus the number of cases observed in row 1, column 2 is $n_{12}$. The row and column totals are denoted by inserting a dot in place of the numbers we summed to get the total. Thus the total of the first row would be $n_1$, and the total of the second column would be $n_{-2}$. The grand total can be denoted just $n$ (or $n_{-}$.).

### Choosing categorical methods

Some descriptors of categorical data are listed in Table 2.1. A descriptor of the level of association between two categorical variables is the tetrachoric correlation coefficient, seen in Section 5.3. Methods to test the significance of association between two categorical variables exist for both counts and proportions. When all cells of the table of counts can be filled in, use the methods for counts. Tests of proportions are appropriate for cases in which some table totals are missing but for which proportions can still be found. A first-step guide to choosing a test was introduced in Table 2.2 in Section 2.10. The tests in this chapter address cells (1,1), (1,2), (2,1), (3,1), and (3,2) in Table 2.2.

### DEGREES OF FREEDOM FOR CONTINGENCY TESTS

A contingency test’s number of degrees of freedom ($df$) is given by the minimum number of cells that must be filled to be able to calculate the remainder of cell entries using the totals at the side and bottom (often termed margins). For example, only one cell of a $2 \times 2$ table with the sums at the side and bottom needs to be filled, and the others can be found by subtraction; it has 1 $df$. A $2 \times 3$ table has 2 $df$. In general, an $r \times c$ table has $df = (r - 1) \times (c - 1)$.
**Categories versus ranks**

Cases for which groups of counts or proportions can be reordered without affecting the interpretation are constrained to categorical methods. An example would be group 1: high hematocrit (Hct); group 2: high white blood cell count; and group 3: high platelet count. These groups could be written in order as 2, 1, 3 or 3, 2, 1 without affecting our understanding of the clinical implications. However, if sample groups fall into a natural order in which the logical implication would change by altering this order—for example, group 1: low Hct; group 2: normal Hct; and group 3: high Hct—rank methods are appropriate. Although these groups could be considered as categories and categorical methods could be used, rank methods give better results. Rank methods lose power when there is a large number of ties but are still more powerful than categorical methods, which are not very powerful. *When ranking is a natural option, rank methods of Chapter 11, Tests of location with continuous outcomes, should be considered.*

### 9.2 Tests on Categorical Data: 2 × 2 Tables

**What does a test tell us?**

Testing a contingency table will answer the question: are the row categories independent of the column categories? Independence implies that the rows’ variable has not been shown to be related to the columns’ variable; knowing the outcome of one of the variables gives us no information as to the associated outcome of the other variable. Dependence tells us that some association exists, but it does not tell us what the strength of this association is. Section 5.3 provides an estimate of the strength of association, and Section 10.2 and Chapter 17, Logistic regression for binary outcomes, provide estimates of the strength of association.

**Example: Is prediction (diagnosis) of prostate cancer by digital rectal examination better than chance?**

Table 9.2 shows contingency in which the 301 biopsy (BIOP) and digital rectal examination (DRE) pairings from DB1 have been tabulated. Does knowing a patient’s DRE give us any help in predicting whether or not he will have a positive BIOP, or would we do as well just to choose randomly? Let us conceive of a spinner in the center of a circle with circumference divided into 301 equal parts. We color a sector (slice of pie) of 95 parts red to match the 95 positive biopsies. To predict a patient’s BIOP result randomly, we spin. If the spinner stops on red, we choose positive, otherwise negative. If the DRE adds no information, the ratio of positive BIOP prediction to total for a positive DRE should be about the same as the 95/301 ratio our spinner would give. We would be inclined to say that BIOP result is independent of DRE outcome, but we need a formal test.
A test of independence

A formal test will tell us the probability that dependence is more likely to occur than by chance, information we cannot obtain from the descriptive statistics. Fisher’s exact test (FET) and the chi-square test of contingency are two ways to test the independence of categorical variables. FET is often required and usually preferred because it is exact, while the chi-square test is only an approximation. However, the derivation of the chi-square test is easier to understand, so its basis will be given in more detail to better assist the reader’s perception of the process.

9.3 THE CHI-SQUARE TEST OF CONTINGENCY

A more exact statement of the question being asked about digital rectal examinations

The question may be expressed statistically as follows: Are the numbers of DRE positive predictions and negative predictions, contingent upon actual outcome, distributed the same as they would be without the outcome information (as is the right-hand column)?

EXPECTED VALUES

This last question gives rise to what is termed an expected value. If the occurrence of predicted positive cases is unrelated to biopsy outcome, the number of correct positive predictions we would expect to occur (call it \( e_{11} \) to match \( n_{11} \)) in ratio to all positive predictions \( (e_{11}/n_{11}) \) is the same as the number of either prediction for positive biopsy in ratio to total cases \( (n_1/n) \), or \( e_{11}/n_{11} = n_1/n \). By multiplying both sides by \( n_{11} \), we find the expected number of correct positive predictions to be its row sum multiplied by its column sum divided, by the total sum. In general, for row \( i \) and column \( j \), the expected value of the \( ij \)th entry is given by

\[
e_{ij} = \frac{n_i n_j}{n}
\]  

9.2 Table 9.2 Contingency table of simultaneous biopsy (BIOP) and digital rectal examination (DRE) results from 301 patients.

<table>
<thead>
<tr>
<th></th>
<th>DRE 1</th>
<th>DRE 0</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOP 1</td>
<td>( n_{11} = 68 )</td>
<td>( n_{12} = 27 )</td>
<td>( n_1 = 95 )</td>
</tr>
<tr>
<td>BIOP 0</td>
<td>( n_{21} = 117 )</td>
<td>( n_{22} = 89 )</td>
<td>( n_2 = 206 )</td>
</tr>
<tr>
<td>Totals</td>
<td>( n_1 = 185 )</td>
<td>( n_2 = 116 )</td>
<td>( n = 301 )</td>
</tr>
</tbody>
</table>

\( a \) symbols are included to illustrate cell counts, marginal totals, and grand total.
For example, the expected value for the top left position in Table 9.2 is $95 \times 185/301 = 58.4$, in contrast to the observed value of 68. The expected values $e_{12}$, $e_{21}$, and $e_{22}$, respectively, are 36.6, 126.6, and 79.4.

**The basis of the chi-square test of contingency**

The chi-square test of contingency is based on the differences between the observed values and those that would be expected if the variables were independent. If these differences are small, there is little dependence between the variables; large differences indicate dependence. The actual chi-square statistic is the sum of squares of these differences in ratio to the expected value. A small chi-square statistic arises if the observed values are close to the values we would expect if the two variables were unrelated. A large chi-square statistic arises if the observed values are rather different from those we would expect from unrelated variables. If the chi-square statistic is large enough that it is unlikely to have occurred if the variables were truly independent, we conclude that it is significant and that the rows' variable is not totally independent of the columns variable. However, it does not follow that one can be well predicted by the other.

**EXAMPLE POSED: IS DIGITAL RECTAL EXAMINATION INDEPENDENT OF BIOPSY OUTCOME?**

This chi-square test should tell us if the DRE result is independent of the BIOP outcome. We would hope that it is not; the two should be closely related if we are to detect prostate cancer using the DRE. Let us detail the calculations in the test so we can answer this question.

**Calculation for the chi-square test of contingency**

The chi-square calculation is based on the difference between the observed cell count and the cell count that would be expected if the rows and columns were independent. The expected count for a cell is its row total times its column total divided by the grand total, as given in Eq. (9.1). After the four $e_{ij}$ are determined, the chi-square is found using the simple pattern: sum of $[(\text{observed} - \text{expected})^2/\text{expected}]$.

The symbolic form of the chi-square test is given as

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.$$  \hspace{1cm} \text{(9.2)}

One $\Sigma$ symbol tells us to add across the two rows and the other tells us to add over the two columns. The $p$-value can be found by calculation or from tables of chi-square probabilities. For a $2 \times 2$ contingency table the distribution of the chi-square statistic is approximated by a chi-square distribution with 1 df.

(Continued)
Assumptions Required for the Chi-Square Test of Contingency

The chi-square method is an approximation not dependable for small samples. In particular, it is valid only if the expected value (row sum × column sum/total sum) of every cell is at least 1 and a minimum count of 5 appears in every cell.

Yates’ correction

The reader should note that sometimes the chi-square statistic is calculated as the sum of \(|(\text{observed} - \text{expected}) - 0.5|^2/\text{expected}\), where the 0.5 term, called Yates’ correction, is subtracted to adjust for the counts being restricted to integers. Developed long before computers rendered FET computationally practical, it was used to provide a more conservative result for contingency tables with small cell counts. Currently, FET provides a better solution to dealing with small cell counts and is preferred. For larger cell counts, Yates’ correction alters the result negligibly and may be ignored. Thus the chi-square form Eq. (9.2) is used in this book.

Example continued: Digital Rectal Examination and Biopsy Outcome

Let us test the null hypothesis that DRE result and BIOP result are independent, that is, knowledge of the DRE result tells us nothing about the BIOP result. The expected values, \(e_{ij}\), were noted just following Eq. (9.1). The chi-square statistic is calculated \(\chi^2 = [(68 - 58.4)^2/58.4] + \cdots + [(89 - 79.4)^2/79.4] = 5.98\). Table III (right tail of the chi-square distribution; see Tables of Probability Distributions) will tell us if this chi-square value is significant. A fragment of Table III is shown as Table 9.3. The statistic 5.98 is between 5.02 and 6.63, associated with \(\alpha = 0.025\) and 0.01, respectively; therefore the \(p\)-value is about 0.015, a statistically significant result. (Exact calculation on a computer tells us \(p\)-value = 0.014.) We reject the null hypothesis of independence and conclude that the DRE result does indeed give us some information about the presence of prostate cancer. Note the wording: It gives us some information; it does not tell us how good this information is.

The effect of sample size on contingency tests

A significant chi-square result indicates that the types of categories are not independent, but it does not indicate how closely they are associated because the significance is influenced by both sample size and association. To illustrate this, Table 9.4 has each

<table>
<thead>
<tr>
<th>(\alpha) (area in right tail)</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>(df = 1)</td>
<td>2.71</td>
<td>3.84</td>
<td>5.02</td>
<td>6.63</td>
</tr>
</tbody>
</table>

*\(\chi^2\) values (distances to the right of 0 on a \(\chi^2\) curve) are tabulated for 1 \(df\) for four values of \(\alpha\).*
cell count in Table 9.2 replaced by one-fifth of its value, rounded to maintain integers. We have approximately the same pattern of proportions, which gives about the same level of association (37% correct positive prediction for the full sample and 38% correct positive prediction for the smaller one). However, now the chi-square has shrunk from 5.98 to 1.70. From Table 9.3, we can see that $\chi^2 = 2.71$ is associated with $\alpha = 0.10$, so that the $p$-value for $\chi^2 = 1.70$ will be greater than 0.10. (Calculating it directly yields $p$-value = 0.192). Reduction of the sample size while keeping the same general pattern of relationship between the variables has caused the $p$-value to grow from a significant 0.014 to a nonsignificant 0.192. Contingency test results must be interpreted with sample size dependency in mind. Statistical significance (not clinical significance) can always be found if the sample size is large enough.

What to do if we have percentages but not counts

Percent is just proportion with the decimal point moved two places. We can test a sample proportion against either a known population proportion or another sample proportion by methods to be seen in Sections 9.6 and 9.7.

Exercise 9.1

Have protease inhibitors reduced pulmonary complications from HIV? In DB4, test the contingency table using chi-square to decide whether or not access to protease inhibitors has reduced the rate of pulmonary admissions.

Exercise 9.2

Are tattoos using titanium ink harder to remove? In DB6, test the contingency table using chi-square to decide whether or not titanium-containing tattoo ink is harder to remove than other inks.

Exercise 9.3

Does ondansetron HCl reduce nausea in gall bladder patients? In Table 5.3, test the $2 \times 2$ contingency table using chi-square to decide whether or not nausea score is reduced by the drug.

Table 9.4 Contingency table of simultaneous biopsy (BIOP) and digital rectal examination (DRE) results with each cell showing one-fifth the count arising from the 301 patients.*

<table>
<thead>
<tr>
<th></th>
<th>DRE 1</th>
<th>DRE 0</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOP 1</td>
<td>$n_{11} = 14$</td>
<td>$n_{12} = 5$</td>
<td>$n_{1.} = 19$</td>
</tr>
<tr>
<td>BIOP 0</td>
<td>$n_{21} = 23$</td>
<td>$n_{22} = 18$</td>
<td>$n_{2.} = 41$</td>
</tr>
<tr>
<td>Totals</td>
<td>$n_{.1} = 37$</td>
<td>$n_{.2} = 23$</td>
<td>$n = 60$</td>
</tr>
</tbody>
</table>

*New cell counts are rounded to integers. The pattern of proportions is approximately maintained, but the $p$-value loses its significance due to the smaller sample size.
CHAPTER 9 Tests on categorical data

Exercise 9.4
Do clinic versus laboratory tests have different out-of-range rates? In DB13, combine below-range with above-range result categories (last two columns) to form an out-of-range category. Form an in-range versus out-of-range $2 \times 2$ contingency table and use $\chi^2$ to decide if clinic and laboratory result categories were discrepant.

9.4 FISHER’S EXACT TEST OF CONTINGENCY

Another test of independence between two categorical variables

In small sample sizes, $FET$, may be desired relative to the chi-square test. FET calculates the probability of independence directly and can be found in statistical packages on modern computers. (The test could well have been named the Fisher—Irwin test; Irwin and Fisher developed the theory independently and their work appeared in print almost simultaneously.) “Exact” implies that its results are calculated by methods from probability theory rather than approximated like results for the chi-square contingency test. The wide use of the chi-square test arose before there was adequate computing power for the exact test. In cases where there are cell counts less than 5 or there is an expected value less than 1, FET is often used. For larger cell counts and expected values, the two tests will give not very dissimilar results and either may be used. If there are many cells in the contingency table, perhaps 10 or 15 (depending on the speed of the computer), the exact method may still be too computation intensive and the chi-square approximation will have to be used. One issue with the FET is that it tends to be a conservative test, meaning that the true type I error rate of the test is usually less than the desired $\alpha$. This is a result of the exact nature of the test and that the resulting $p$-values from the test take on discrete values.

EXAMPLE CONTINUED: DIGITAL RECTAL EXAMINATION AND BIOPSY OUTCOMES

The dependence between the DRE and the BIOP result data appearing in Table 9.2 may be tested by FET. The chi-square statistic was calculated as $\chi^2 = 5.98$, which yields a significant $p$-value of 0.014, indicating dependence between DRE and BIOP. The FET $p$-value, calculated by a statistical software package, is a similar 0.016, yielding the same interpretation.

THE BASIS OF FISHER’S EXACT TEST (FET)

FET could be calculated by hand but is so difficult that it is almost always calculated by a statistical software package. Its $p$-value gives the probability of observing data as or more indicative of dependence between rows and columns if in reality no relationship exists.

(Continued)
FET uses a probability distribution known as the hypergeometric distribution for the observed counts, calculating the probability of all other $2 \times 2$ tables with the same marginal totals ($n$ counts). The $p$-value is the sum of probabilities of all such outcomes with counts less likely than the observed counts. The interpretation of the FET $p$-value is the same as the chi-square $p$-value. A significant FET or chi-square result indicates that the variables are not independent, but it does not indicate how closely they are associated.

**ADDITIONAL EXAMPLE: IMPROVEMENT WITH SURGICAL EXPERIENCE**

An ophthalmologist investigated his learning curve in performing radial keratotomies. He was able to perform a 1-month postoperative refraction on 78 of his first 100 eyes. He classified them as 20/20 or better versus worse than 20/20. The results appear in Table 9.5. The ophthalmologist notes that the percentage of eyes in the 20/20 or better group increased from 41% to 73% from earlier to later surgeries. The null hypothesis is that resultant visual acuity is independent of when in the sequence the surgery was done. FET yields $p$-value = 0.006. The ophthalmologist rejects the null hypothesis and concludes that his surgical skills have improved significantly.

The ophthalmologist could have approximated the FET result by a chi-square test. The expected value is calculated as $e_{ij} = n_i \times n_j / n$. $e_{11} = 44 \times 41 / 78 = 23.13$, $e_{12} = 20.87$, $e_{21} = 17.87$, and $e_{22} = 16.13$. Substitution in Eq. (9.2) yields

$$
\chi^2 = \sum_{i} \sum_{j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(17 - 23.13)^2}{23.13} + \ldots + \frac{(10 - 16.13)^2}{16.13} = 7.86.
$$

From Table III (see Tables of Probability Distributions) or Table 9.3, the critical value of $\chi^2$ for $\alpha = 0.05$ in one tail with 1 df is 3.84, which is much less than 7.86; again the null hypothesis is rejected. Calculation by a software package yields $p$-value = 0.005, which is similar to the $p$-value for FET.

**Table 9.5** Refraction of postoperative eyes by position in sequence of surgery.

<table>
<thead>
<tr>
<th></th>
<th>No. of eyes in first 50</th>
<th>No. of eyes in second 50</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>20/20 or better</td>
<td>17</td>
<td>27</td>
<td>44</td>
</tr>
<tr>
<td>Worse than 20/20</td>
<td>24</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>Totals</td>
<td>41</td>
<td>37</td>
<td>78</td>
</tr>
</tbody>
</table>
Exercise 9.5
Comparing methods of torn ankle ligament repair. Two methods of torn ankle ligament repair, modified Broström and Chrisman—Snook, were compared in an orthopedics department. The clinical success on 39 patients was rated as excellent or less than excellent. The contingency table is given in Table 9.6. Find the p-values using FET and the chi-square test. Are the two treatments significantly different?

9.5 TESTS ON $r \times c$ CONTINGENCY TABLES

Are the row categories independent of the column categories in tables having $r$ rows and $c$ columns ($r$ and/or $c > 2$)? A test yielding a nonsignificant $p$-value implies that there is insufficient evidence to establish that the rows' variable is related to the columns variable.

EXAMPLE POSED: IS THE USE OF SMOKELESS TOBACCO RELATED TO ETHNIC ORIGIN?
With increased restriction on smoking in ships and shore stations, many US Navy and Marine Corps personnel are changing to smokeless tobacco. Navy medicine is concerned about the effect of this pattern on health and seeks causes for this change. Cultural influences may be involved. Is ethnic background related to smokeless tobacco use? A count of the numbers in categories of use and ethnicity yielded Table 9.7. With two rows and three columns, $r = 2$ and $c = 3$. Smokeless tobacco use was ascertained from 2004 sailors and marines of 3 ethnic backgrounds from a wide variety of unit assignments. A total of 74.8% were of European background but represented 91.2% of users, 15.1% were of African background but represented 2.8% of users, and 10.1% were of Hispanic background but represented 6.0% of users. It was suspected that use is not independent of ethnic background, and the relationship was tested.

Method: Fisher’s exact test for $r \times c$ contingency tables
When the data are divided into $r$ row and $c$ column categories, an $r \times c$ contingency table is appropriate. FET uses a distribution named the hypergeometric for calculating the probability of occurrence of all other $r \times c$ tables with the same marginal totals.
The \( p \)-value is the sum of probabilities of all such outcomes with counts less likely than the observed counts.

**CHI-SQUARE APPROXIMATES THE RESULT FOR LARGER TABLES**

For large \( r \) and \( c \) (perhaps \( r + c > 7 \), depending on computer speed), calculating \( p \) for FET becomes lengthy, often too lengthy for even a modern computer. In this case, the chi-square contingency calculation of Eq. (9.3), similar to Eq. (9.2), is usually used. By denoting the observed count in row \( i \) and column \( j \) as \( n_{ij} \) and calculating the corresponding expected entry \( e_{ij} \) by row \( i \) total \( \times \) column \( j \) total/grand total, chi-square may be written as

\[
\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{9.3}
\]

with \( df = (r - 1)(c - 1) \).

The difference from Eq. (9.2) is that, instead of summing over 2 rows and 2 columns, we sum over \( r \) rows and \( c \) columns. Table III gives the chi-square \( p \)-value (looking it up as if it were \( \alpha \)). For example, a \( 3 \times 4 \) contingency table has \( (3 - 1) \cdot (4 - 1) = (2)(3) = 6 \) \( df \). Choose the largest \( \alpha \) having a table entry less than the calculated \( \chi^2 \); the \( p \)-value of the test, that is, the probability of observing results as or more indicative of dependence if independence were true, is less than that \( \alpha \). Alternatively, use a statistical package on a computer and calculate the \( p \)-value directly. (The distinction between using a tabulated critical value and a calculated \( p \)-value was discussed in Sections 7.2 and 7.3.)

**Large tables with violated chi-square assumptions**

If the chi-square assumption of cell entries at least 5 and expected values at least 1 is violated, but the table is too large for FET to be used, the table size can be reduced
by amalgamating cells with small entries or expectation. However, this procedure reduces both the \( df \) and the information contained in the analysis.

**EXAMPLE CONTINUED: SMOKELESS TOBACCO AND ETHNIC ORIGIN**

Here \( r + c \) is 5, small enough for practical use of the FET when the user has access to a computer statistical package. The (FET) \( p \)-value is less than 0.001. (Statistical software packages may display \( p \)-values like 0.000, implying that the first significant digit lies more than three digits to the right of the decimal point. However, values less than 0.001 are not accurate, so \( p \)-value, 0.001 should be reported.)

The chi-square contingency approximation formula appears in Eq. (9.3); the critical value may be found in Table III for \((r - 1) \times (c - 1) = 2 \ df\). The observed count in row \( i \) and column \( j \) is \( n_{ij} \) (e.g., \( n_{11} = 424 \)), and the corresponding expected entry \( e_{ij} \) is \( \frac{\text{row } i \text{ total} \times \text{column } j \text{ total}}{\text{grand total}} \) (e.g., \( e_{11} = \frac{465 \times 1499}{2004} = 347.82 \)). The chi-square component of the first cell is \( \frac{(n_{11} - e_{11})^2}{e_{11}} = \frac{(424 - 347.82)^2}{347.82} = 16.69 \). Following this pattern and summing over the six cells, the chi-square statistic becomes

\[
\chi^2 = 16.69 + 46.71 + 7.60 + 5.04 + 14.12 + 2.30 = 92.46.
\]

In Table 9.8 a fragment of chi-square Table III, the entry for \( \alpha = 0.001 \) and \( df = 2 \) is 13.80, which is much less than the calculated \( \chi^2 \). Our \( p \)-value is then reported as \( < 0.001 \). We reject the null hypothesis and conclude that smokeless tobacco use and ethnic background are related.

**ADDITIONAL EXAMPLE: CHOICE OF THERAPIST FOR PSYCHIATRY RESIDENTS**

In a study on the desirability of psychotherapy for psychiatry residents, a question asked was whether the choice of therapist was independent of whomever paid for the therapy. The responses compose Table 9.9. FET using a computer yielded \( p \)-value = 0.001. We conclude that choice of type of analyst is related to whomever pays for the analyst. To perform the chi-square test the expected values \( e_{ij} = \frac{n_i \times n_j}{n} \), which are \( e_{11} = 45 \times 38/155 = 11.032 \), \( e_{12} = 102 \times 38/155 = 25.006 \), \( \ldots \), \( e_{43} = 0.413 \) are calculated first. (Note that the last expected value is less than 1, which violates the assumptions required for the chi-square approximation.) \( df = (4 - 1) \times (3 - 1) = 6 \). Substituting the values in Eq. (9.3) yields \( \chi^2 = \frac{(15 - 11.032)^2}{11.032 + \ldots} = 24.45 \). The tail probability associated with 24.45 in the row for 6 \( df \) in

**Table 9.8 A fragment of Table III, the right tail of the chi-square distribution.**

<table>
<thead>
<tr>
<th>( \alpha ) (area in right tail)</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>( df = 2 )</td>
<td>4.61</td>
<td>5.99</td>
<td>7.38</td>
<td>9.21</td>
<td>10.60</td>
<td>13.80</td>
</tr>
</tbody>
</table>

*\( \chi^2 \) values (distances to the right of 0 on a \( \chi^2 \) curve) are tabulated for 2 \( df \) for seven values of \( \alpha \).
Table 9.9 Type of therapist chosen as contingent upon the chooser for psychotherapy given to psychiatry residents as part of their training.

<table>
<thead>
<tr>
<th>Type of Therapist chosen</th>
<th>Resident</th>
<th>Residency program</th>
<th>Insurance company</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychoanalyst</td>
<td>15</td>
<td>22</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>Psychiatrist</td>
<td>18</td>
<td>64</td>
<td>1</td>
<td>83</td>
</tr>
<tr>
<td>Psychologist</td>
<td>7</td>
<td>14</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>Social worker</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Totals</td>
<td>45</td>
<td>102</td>
<td>8</td>
<td>155</td>
</tr>
</tbody>
</table>

Table 9.10 Data on change in blood pressure (BP) in response to change in position.

<table>
<thead>
<tr>
<th>Rise in BP</th>
<th>No change in BP</th>
<th>Fall in BP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic BP</td>
<td>8</td>
<td>20</td>
<td>72</td>
</tr>
<tr>
<td>Diastolic BP</td>
<td>11</td>
<td>30</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>50</td>
<td>126</td>
</tr>
</tbody>
</table>

Table III is a little less than 0.0005. We record $p$-value < 0.001, which is similar to the FET $p$-value. The overwhelming majority of residents and residency programs selected psychiatrists and psychoanalysts. Psychologists were preferred by 63% of insurance companies, although the costs, interestingly, were similar.

**Exercise 9.6**

**Change in blood pressure (BP) resulting from change in position.** The influence of patient position during the measurement of BP has long been questioned. A 1962 article examined alterations in BP of 195 pregnant women resulting from a change from upright to supine positions. The data are given in Table 9.10. We ask the following: Is the pattern of change the same or different for systolic and diastolic BPs?

We note that rise, no change, and fall are in order and therefore a rank test would be appropriate. There are two unmatched samples as the independent variable and three ranked outcomes, leading to cell (1,3) in Table 2.2. The rank-sum test should be used, but the contingency table calculations may be done as an exercise.

**9.6 TESTS OF PROPORTION**

**What a test of proportion will do and why it is needed**

Suppose we are concerned with the efficacy of an antibiotic in treating pediatric otitis media accompanied by fever. We know that, say, 85% of patients get well within
CHAPTER 9 Tests on categorical data

7 days without treatment. We learn that, of 100 such patients, 5 (i.e., 95%) of those treated with the antibiotic got well within 7 days. Is this improvement statistically significant? Note that we cannot fill in a contingency table because we do not have the number of untreated patients or the total number of patients. However, we do have the theoretical proportion and so can answer the question using a test of proportions. Testing a proportion will answer the question: Is the probability of "success" giving rise to a sample proportion different from a theoretical or presumed population probability?

When a proportion is very small
If you are concerned with a rare event, perhaps a disease having prevalence 0.002, go to Section 9.7 following.

Basis of the test
Proportions for binary data (e.g., yes/no, male/female, survive/die) follow the binomial distribution, because outcomes from sampling must fall into one category or the other. Probabilities for the binomial distribution appear in Table VI (see Tables of Probability Distributions). Because the tabulation goes only to 0.50, we will say 15% of our otitis media patients remain ill past 7 days rather than 85% do not. We use the binomial distribution with theoretical proportion \( \pi = 0.15 \). The sample proportion of antibiotic-treated patients remaining ill past 7 days is \( p = 0.05 \). [The Greek symbol \( \pi \) is used for the binomial population proportion and the Roman \( p \) is used for the sample proportion. This \( p \) does not represent the common usage designating probability (e.g., \( p \)-value); the use of the same letter is coincidental. We must keep track of the usage by context.] We ask, “Could a rate as small as or smaller than our observed sample proportion \( p = 0.05 \) have occurred by chance alone if the true proportion were 0.15?” (The “smaller than” portion arises because we are really considering 1 minus the probability that the rate is more than 0.05.) The probability of observing a proportion \( p \leq 0.05 \), given \( \pi = 0.15 \), can be calculated or looked up in a table. The resulting probability is the \( p \)-value of the binomial test of proportion. In Table VI the binomial probabilities are given for \( \pi \) no larger than 0.50; for larger values of \( \pi \), just reverse the tails of the distribution as was done in this example.

A different method is used if \( n \) is large
For a large-sample \( n \), exact probabilities do not appear in the tables and they would be very time consuming to calculate. Fortunately, good approximations have been found for large samples. For the binomial distribution, when both \( n\pi \) and \( n(1-\pi) \) are greater than 5, the sample proportion is distributed approximately normal with mean equal to the theoretical proportion, that is, normal \( \mu = \pi \), and standard deviation also depending on \( \pi \). The null hypothesis is that the \( \pi \) for the distribution from which the
sample was drawn is the same as the theoretical \( \pi \). Knowing \( \mu \) and \( \sigma \), we may calculate a \( z \)-score as the difference between the sample and theoretical proportions divided by the standard deviation in much the same conceptual form described in Section 4.8. We then use normal tables to find \( p \)-values.

**Test of two proportions**

A generalization of the large sample test of a proportion exists to test for a significant difference between two population proportions. We sample \( n_1 \) and \( n_2 \) readings from two populations having proportions \( \pi_1 \) and \( \pi_2 \). The null hypothesis is \( H_0: \pi_1 = \pi_2 \). We estimate \( \pi_1 \) and \( \pi_2 \) by \( p_1 \) and \( p_2 \). We calculate a combined \( p_c \) pooled over the two samples and a standard deviation depending on \( p_c \). The difference between \( p_1 \) and \( p_2 \) divided by the standard deviation is again normal in form and we find \( p \)-values from the normal table.

For small samples, the sample sizes are usually known. From these and the sample proportion values, the actual counts can be found, which will permit the data to be posed in the form of a contingency table, as in the methods explained earlier in this chapter.

**Test of three or more proportions**

When three or more proportions are to be compared, small samples can be treated by posing the data in a contingency table larger than \( 2 \times 2 \), as in Section 9.5. For large samples, the associated distribution is the multinomial distribution, a generalization of the binomial. This problem is not met frequently enough to pursue here. However, in some cases in which it is met, a satisfactory answer can be found without the multinomial by combining categories and reducing the question to two proportions.

**METHODS OF TESTS OF PROPORTION**

**Test of a Sample’s True Proportion Against a Theoretical Proportion**

This method tests a sample’s true proportion against a theoretical or (previously known) population proportion, \( \pi \). We sample \( n \) observations randomly from a population with probability \( \pi_s \) of a “success.” (A “success” is the occurrence of the event we are testing, not necessarily what we would wish clinically.) We want to know if \( \pi_s \), the proportion for the population we sampled from, is the same as \( \pi \), the proportion for the theoretical population. Thus \( H_0 \) is \( \pi_s = \pi \). We obtain \( n_s \) successes, estimating \( \pi_s \) by the sample proportion \( p_s = n_s/n \).

**Case of Small \( n \)**

If \( n \) is less than 17, Table VI gives the \( p \)-value, the probability that, for expected chance of a success \( \pi \), we would find \( n_s \) successes or \( p_s \) proportion successes by chance alone. More extensive tabulation may be found in mathematical or statistical tables in books or on websites.

(Continued)
(CONTINUED)

Case of Large $n$

For larger $n$ a normal approximation to the binomial distribution is adequate, so binomial tables are not required. “Large” may be thought of in the context of the number necessary for the binomial to adequately approximate the normal, specifically $n\pi > 5$. The normal mean $\mu$ is $\pi$ and the standard deviation is

$$\sigma = \sqrt{\frac{\pi(1-\pi)}{n}}. \quad (9.4)$$

The difference between the sample and theoretical proportions is transformed to a normal variate $z$ by dividing by $\sigma$, or

$$z = \frac{p_s - \pi}{\sigma} = \frac{p_s - \pi}{\sqrt{\pi(1-\pi)/n}}. \quad (9.5)$$

The two-tailed $p$-value corresponding to the calculated $z$ is then found from the normal table (see Table I).

An adjustment factor to the approximation

Because a normal variable is continuous, and a binomial variable is discrete, the quantity $1/2 n$ may be subtracted from the absolute value of the numerator to increase the accuracy of the approximation. If $n$ is small, the exact form should be used and if $n$ is large, the adjustment adds little, so the adjustment is not used in this book. There may be cases of moderate-sized $n$ in which the adjustment is slightly beneficial.

CASE OF TWO PROPORTIONS

The smaller sample form can usually be treated as a contingency table. If $n_1\pi_1$, $n_1(1-\pi_1)$, $n_2\pi_2$, and $n_2(1-\pi_2)$ are all greater than 5, the larger sample form may be used. Here we test $H_0: \pi_1 = \pi_2$ by using $p_1 - p_2$. The standard deviation for the combined samples, $\sigma_c$, is a little more complicated than Eq. (9.4). First, we find a combined sample proportion estimate, $p_c$, as

$$p_c = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}, \quad (9.6)$$

and then the combined standard deviation as

$$\sigma_c = \sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \quad (9.7)$$

(Continued)
The normal $z$ statistic is given by

$$z = \frac{p_1 - p_2}{\sigma_c} = \frac{p_1 - p_2}{\sqrt{p_c(1-p_c)((1/n_1) + (1/n_2))}}.$$  \hspace{1cm} (9.8)

The critical value(s) to test $H_0$ can be found as area(s) under the normal curve. For example, a two-tailed test has critical values $\pm 1.96$, corresponding to 2.5% of the curve under each tail.

**Examples of tests of probabilities giving rise to a sample against theoretical proportions, small $n$**

**EXAMPLE: RATE OF POSITIVE BIOPSY RESULTS**

We want to compare the positive BIOP result rate from a sample size of 10 patients (sample proportion observed to be $p_i = 0.30$ (30%); see Table DB1.1) with a theoretical positive BIOP result rate defined as 0.25 (25%) of patients older than 50 years presenting to a urology clinic. We use the binomial distribution with theoretical proportion $\pi = 0.25$. To find the probability of drawing $n_s \geq 3$ out of $n = 10$ by chance alone, given $\pi = 0.25$, we go to Table VI. A fragment of Table VI appears as Table 9.11. The table entry is 0.474; the probability of observing such a result if the null hypothesis were true is nearly 0.5 (the $p$-value). This is clearly not a rare event. Thus we conclude that there is insufficient evidence to say that our proportion of positive BIOP results is larger than that found in the clinic.

**EXAMPLE: DOES RADIATION THERAPY IMPROVE THE PROPORTION SURVIVAL?**

The 2-year survival rate of an unusual form of malignant cancer has been $\pi = 0.20$. A radiation oncology department treated nine patients. Two years later, four patients had survived. Does the treatment improve survival? From Table 9.11, which is a segment of Table VI, the column for $\pi = 0.20$, the row for $n = 9$, and $n_o = 4$ yield a probability of 0.086 that four or more patients would survive if the 20% untreated

<p>| Table 9.11 A Fragment of Table VI.\textsuperscript{a} |
|-----------------|----|----|----|----|----|----|----|----|----|</p>
<table>
<thead>
<tr>
<th>$n$</th>
<th>$n_o$</th>
<th>.05</th>
<th>.10</th>
<th>.15</th>
<th>.20</th>
<th>.25</th>
<th>.30</th>
<th>.35</th>
<th>.40</th>
<th>.45</th>
<th>.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>4</td>
<td>.001</td>
<td>.008</td>
<td>.034</td>
<td>.086</td>
<td>.166</td>
<td>.270</td>
<td>.391</td>
<td>.517</td>
<td>.639</td>
<td>.746</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>.012</td>
<td>.070</td>
<td>.180</td>
<td>.322</td>
<td>.474</td>
<td>.617</td>
<td>.738</td>
<td>.833</td>
<td>.900</td>
<td>.945</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Values of cumulative binomial distribution, depending on $\pi$, $n$, and $n_o$ (number of occurrences observed).
survival rate were true. There is not enough evidence to strongly infer treatment effectiveness on the basis of such a small sample.

Examples of tests of probabilities giving rise to a sample against theoretical proportions, large \( n \)

**EXAMPLE: THE ISSUE OF PHYSICIAN CLEANLINESS 150 YEARS AGO**

In 1847, in Vienna, Ignaz Semmelweis experimented with the effect on patient mortality of physician hand cleansing between patients.\(^8\) Original data are not available (if they ever were), but a reconstruction from medical history accounts yields the following results: the long-term mortality rate for patients treated without the cleansing had been about 18%. Subsequently, for 500 patients treated with the cleansing, the mortality rate was 1.2%. The requirement \( n \pi > 5 \) (90 > 5) is satisfied. Because hand cleaning could not increase the mortality, a one-tailed test is appropriate. We substitute in the formula for a normal \( z \) value, obtaining

\[
z = \frac{p_\text{c} - \pi}{\sqrt{\pi(1 - \pi)/n}} = \frac{0.012 - 0.18}{\sqrt{0.18 \times 0.82/500}} = -9.78.
\]

By symmetry in the normal curve, we may use the upper tail, that is, \( z = 9.78 \). The area in the tail is far less than the smallest tabulated value in Table I for one-tailed \( \alpha \), which yields the \( p \)-value < 0.0002 (reported as \( p \)-value < 0.001) for \( z = 3.5 \). The cleansing significantly reduced the mortality rate. Nonetheless, as a matter of historical interest, the local medical society attacked Semmelweis’ findings viciously. He was reduced in rank and his practicing privileges were curtailed. He was broken and died in an asylum.

**EXAMPLE: ABSORBABLE MESH IN ABDOMINAL WALL CLOSURES**

The use of permanent prosthetic mesh has led to the formation of dense adhesions with later obstructions and fistula formations. We ask whether the population proportion \( \pi_a \) of these problems is less using absorbable mesh. \( H_0: \pi_a = \pi \) and \( H_1: \pi_a < \pi \). A number of studies in the literature report the proportion of dense adhesion formation from permanent mesh to be about 0.63, which is taken as \( \pi \). An experiment\(^4\) on 40 rats finds the proportion of dense adhesions from absorbable mesh to be \( p_a = 0.125 \) (5 rats), which estimates \( \pi_a \). The clinical difference is obvious, but we need to state the risk for error. From Eq. (9.4), \( \sigma = 0.0763 \); from Eq. (9.5), \( z = (|p_a - \pi|)/\sigma = (|0.125 - 0.63|)/0.0763 = 6.62 \). The area under the tail of a normal distribution more than 6.62 standard deviations away from the mean is smaller than any tabulated value, certainly less than 0.001, which is the smallest acceptable value to quote. (The discrepancy between the approximation and the true probability curve becomes more than the curve height in the far tails; therefore \( p \)-values smaller than 0.001 should not be reported, even though they can be calculated.)
Example of test of two population proportions

INTERNATIONAL NORMALIZED RATIO RANGE IN A COUMADIN CLINIC

In DB13, there are 104 international normalized ratio (INR) readings from each of the Coumadin Clinic and the hospital laboratory that are classed as below, in, or above range. If we combine the above and below ranges into an out-of-range category, we have in-range versus out-of-range. We find 33 out-of-range readings from the clinic for a rate of 31.7%, and 57 from the laboratory for a rate of 54.8%. Are the true rates from the clinic and the laboratory significantly different? We can test the two population rates using a test of two proportions. (That said, we should note that we are conducting a test of two independent rates and the readings in this database are paired. We will carry on for the purpose of example, but McNemar’s test in Section 9.8 is the proper form of analysis. These data are exemplified using McNemar’s test in that section.) $n_1 = n_2 = 104$, $p_1 = 0.317$, and $p_2 = 0.548$. We calculate $p_c = 0.4325$ from Eq. (9.6) and $z = 3.362$ from Eq. (9.8). From Table I, we see that a $z$ of 3.36 provides a two-sided area under the tails between 0.0010 and 0.0006; therefore we would report $p$-value $< 0.001$. The out-of-range rates are different for the clinic and the laboratory.

Exercise 9.7

Small n: are more boys than girls born in this family? A small-town obstetrician has been delivering babies from a particular family for three generations. He notes a high proportion of male babies and suspects he may be seeing a genetic phenomenon. Of 16 babies born during his experience, 11 are boys. Does he have evidence of a nonrandom causative factor or could it have occurred by chance?

Exercise 9.8

Large n: do boys and girls have limb fractures in equal numbers? Theoretically, the population proportion would be $\pi = 0.5$ if the proportion were sex independent. We recorded the sex of the next 50 children presenting to the orthopedic clinic with limb fractures and found 34 boys. What do we conclude?

9.7 TESTS OF RARE EVENTS (PROPORTIONS CLOSE TO ZERO)

What a test of a small proportion will do and why it is needed

Suppose we suspect a sudden increase in the occurrence rate of a rare disease, with prevalence known to be 0.002 under ordinary conditions. A test based upon our sample proportion will answer the question: Is probability giving rise to a sample proportion different from the theoretical or population proportion? The proportion 0.002 is too small to
 CHAPTER 9 Tests on categorical data

use the binomial methods discussed in Section 9.6. If \( \pi \) is near zero, and if the relationship \( n \gg n_\pi \gg \pi \) (recall that \( \gg \) denotes “much greater than”) may be assumed, the result can be well approximated using the Poisson distribution, for which probabilities appear in Table VII (see Tables of Probability Distributions). The tabulation of the Poisson probabilities uses a parameter \( \lambda = n \pi \) rather than \( \pi \) alone, which is used for binomial probabilities. Suppose our sample shows a prevalence of 0.006, which is three times the expected prevalence. The probability of observing a sample proportion \( p \geq 0.006 \) by chance alone, given \( \pi = 0.002 \), can be calculated or looked up in a table, as detailed in the methods and examples paragraphs that appear later in this section. The resulting probability is the \( p \)-value of the binomial test of proportion. (\( p \) denoting proportion and \( p \) denoting \( p \)-value use the same letter only coincidentally and must be differentiated by context.)

What to do if \( \pi \) is very close to 1 rather than 0

If \( \pi \) is near 1, we can use the complementary proportion, \( 1 - \pi \), because the distribution is symmetric. If survival from a disease is 0.998, we examine the prevalence of death, as 0.002.

What to do if no events are observed

Suppose an event is very improbable, say a side effect from a treatment. We take a sample of size \( n \) and observe no side effects: our sample proportion is 0. Can we make any conclusion? The rule of three (at the end of Section 8.6) allows us to place an upper confidence bound on the population proportion \( \pi \).

A different method is used for large samples

If \( n \) is large enough that the input value \( \lambda = n \pi \) does not appear in Table VII (i.e., \( \lambda > 9 \)), a large-sample approximation is used. It has been shown that, for the Poisson distribution, the sample proportion is distributed approximately normal with mean equal to the theoretical proportion \( \pi \), as for the binomial, but with a slightly different standard deviation \( \sigma \). By knowing \( \mu \) and \( \sigma \), we may calculate a \( z \) score and use a normal table to find \( p \)-values, as detailed in the methods and examples paragraphs that follow.

EXAMPLE POSED, SMALL \( \lambda \): ARE LARGE PROSTATES TOO FREQUENT?
The sample of \( n = 301 \) urologic patients contains 3 men with prostate volumes in excess of 90 mL. We might postulate that the distribution of volumes in the general population is normally distributed with mean 36.47 and standard deviation 18.04 mL. If so, would we be likely to find \( x = 3 \) volumes greater than 90 mL by chance alone, or should we suspect that our sample is not normal in form? In a normal curve with \( \mu = 36.47 \) and \( \sigma = 18.04 \) mL, 90 mL would lie 2.967 \( \sigma \) above the mean. From
Table I the area in the right tail would be about $0.002 \pi$. There is a large number of opportunities for a volume to be greater than 90 but only a small chance that anyone randomly chosen would be, suggesting the Poisson process. The Poisson constant is $\lambda = n \pi = 301 \times 0.002 = 0.602$.

**EXAMPLE POSED, LARGE $\lambda$: DOES A NEW DRUG CAUSE BIRTH DEFECTS?**

It has been established that $1.75\% (\pi = 0.0175)$ of babies born to women exposed during pregnancy to a certain drug have birth defects. It is believed that an adjuvant drug will reduce the rate of a birth defect. The two drugs are used together on 638 pregnant women, and 7 of their babies have birth defects. Does the adjuvant drug reduce the rate of defects? $\lambda = n \pi = 638 \times 0.0175 = 11.165$, which is larger than the value in Table VII, so the normal approximation is used.

**METHOD: TEST OF A VERY SMALL PROPORTION**

This method tests the true probability giving rise to a sample proportion against a theoretical (or previously known) population proportion, $\pi$, where $\pi$ is close to 0 (or to 1) and where the Poisson assumption $n \gg n \pi \gg \pi$ is satisfied. We sample $n$ observations randomly from a population with probability $\pi_s$ of a “success”. We want to know if $\pi_s$, the proportion for the population sampled, is the same as $\pi$. Thus $H_0$ is $\pi_s = \pi$. We obtain $n_s$ successes, estimating $\pi_s$ by the sample proportion $p_s = n_s/n$. The Poisson constant is $\lambda = n \pi$, which is also the number of occurrences expected on average. Calculate $\lambda = n \pi$ and verify the assumption.

**Case of Small $\lambda$**

Table VII gives the probability of $n_s$ or more successes if the null hypothesis is true which is also the $p$-value for the test that the sample observations have $\pi$ proportion success.

**Case of Large $\lambda$**

If $\lambda > 9$ (e.g., $n = 500$ and $\pi = 0.02$), a normal approximation to the Poisson distribution is adequate. The normal mean $\mu = \pi$ and the standard deviation $\sigma$ is

$$\sigma = \sqrt{\frac{\pi}{n}}$$

(Note the similarity to and the difference from Eq. (9.4).) Calculate usual standard normal form, here the difference between the observed and theoretical proportions divided by the standard deviation,

$$z = \frac{p_s - \pi}{\sqrt{\frac{\pi}{n}}} = (p_s - \pi) \sqrt{n/\pi}.$$  (9.10)

and test it using the normal Table I. (Either the expression in the middle or on the right may be used, as convenient.) The “continuity correction” $1/2n$ subtracted from the absolute value of the numerator may be beneficial in cases of moderate-sized samples, as noted in Section 9.6.
EXAMPLE COMPLETED, SMALL $\lambda$: ARE LARGE PROSTATES TOO FREQUENT?

The Poisson assumption $n \gg n\pi \gg \pi$ is satisfied: $301 \gg 0.6 \gg 0.002$. From Table 9.12 a fragment of Table VII, the Poisson probability of observing 3 or more volumes greater than 90 mL, given a Poisson constant of 0.6, is 0.023. With a $p$-value this small, we reject the hypothesis and conclude that our sample is not normal in form.

ADDITIONAL EXAMPLE, SMALL $\lambda$: SITE OF INJECTION AND INCIDENCE OF SARCOMA

A veterinarian postulates that cats contract vaccine-associated feline sarcoma when vaccinated in the back with a rate of 2 in 10,000 ($\pi = 0.0002$). She initiates a study in which 5000 cats are given their rabies and feline leukemia vaccinations in the rear leg. She finds $n_o = 3$ cases of the sarcoma. Is the incidence of sarcoma for leg vaccination different from that for the back? There are many opportunities for the sarcoma to occur, but the probability of it occurring in any randomly chosen case is near zero, indicating Poisson methods. If the incidences at the two sites are different, either one could be greater; therefore a two-tailed test is appropriate. In Table VII the table entry is the probability that $n_o$ or more sarcomas would have occurred by chance alone; therefore the critical values for $\alpha = 0.05$ will be 0.975 (i.e., $1 - 0.025$) for the case of fewer sarcomas in the new site and 0.025 for the case of more sarcomas in the new site. $\lambda = n\pi = 5000 \times 0.0002 = 1$. The assumption $n \gg n\pi \gg \pi$ is satisfied ($5000 \gg 1 \gg 0.0002$). In Table VII the probability is 0.080 between the two critical values. She concludes that the data did not provide sufficient evidence to conclude that leg vaccination gives a sarcoma rate different from back vaccination.

EXAMPLE COMPLETED, LARGE $\lambda$: DOES A NEW DRUG CAUSE BIRTH DEFECTS?

A large number of opportunities to occur (638) but a small chance of occurring on anyone opportunity (0.0175) indicates Poisson. Because the adjuvant drug would be used clinically if defects reduce but not if they stay the same or increase, a one-tailed test is appropriate. $\lambda = n\pi = 638 \times 0.0175 = 11.165$, which is larger than the value in Table VII, so the normal approximation is used. $ps = 7/638 = 0.0110$. Substitution in Eq. (9.10) yields $z = -1.2411$. By normal curve symmetry, we can look up 1.2411 in Table 9.13. By interpolation, we can see that the $p$-value is about 0.108.
ADDITIONAL EXAMPLE, LARGE $\lambda$: ANOREXIA IN THE MILITARY VERSUS THE PUBLIC

Anorexia among the general US population of women is known to be $\pi = 0.020$ (2%). We want to know whether this rate is also true for female military officers.\textsuperscript{10} Let us use $\pi_m$ to represent the true but unknown female military officer population proportion, which we will estimate by $p_m$. We wish to test the hypothesis $H_0$: $\pi_m = \pi$, and $H_1$: $\pi_m \neq \pi$. We take a sample of 539 female navy officers and find $p_m = 0.011$. Do we reject $H_0$? Because $\lambda = 0.02 \times 539 = 10.79$ is larger than the tabulated values, we use the normal approximation to the Poisson distribution. By using Eq. (9.10), we find

$$z = \frac{p_m - \pi}{\sqrt{\pi/n}} = \frac{0.011 - 0.020}{\sqrt{0.020/539}} = -1.48.$$ 

By symmetry of the normal curve the result is the same as for $z = 1.48$, which is between $z = 1.40$ and $z = 1.50$ in Table 9.13 (or Table I). The two-tailed $p$-values (found in the table as if they were $\alpha$’s) for these $z$ values are 0.162 and 0.135. Interpolation to $z = 1.48$ yields $p$-value = 0.157, a result not statistically significant. We do not reject $H_0$ and conclude that we do not have sufficient evidence to establish that the rate of anorexia in military officers is different from the rate of the general population.

**Exercise 9.9**

*Does blood-bank screening reduce hepatitis-infected blood?* Studies through 1992 found the rate of viral hepatitis infection caused by red blood cell transfusions in the United States to be 1 per 3000 units transfused, or $\pi = 0.000333$. Introduction of second- and third-generation hepatitis C screening tests is expected to reduce this rate.\textsuperscript{11} Test the significance of results from 15,000 transfused units reducing the rate to 1:5000.

**9.8 McNemar’s Test: Matched Pair Test of a 2 $\times$ 2 Table**

McNemar’s test assesses the association of categorical data in which a binary variable depends on two matched or paired independent categories. In DB13 the INR
readings of 104 diabetic patients were assessed by the Coumadin Clinic and the hospital laboratory. Is the rate of out-of-range readings different for the two assessors? We are not comparing a sample of clinic patients with a sample of laboratory patients; all patients have readings from both assessors, so the data are paired. This pairing was not the case for contingency tables and their methods. McNemar’s test fills this methodology gap. It tells us if one type of error is more prone than the other. In another view, it allows us to see if odds of correct to incorrect are the same for the two measures.

EXAMPLE POSED: OUT-OF-RANGE READINGS FOR CLINIC VERSUS LABORATORY
The INR readings of 104 diabetic patients were assessed by both the Coumadin Clinic and the hospital laboratory, and thus are paired. Is the out-of-range rate different?

ADDITIONAL EXAMPLE 1 POSED: IS SMOKING ASSOCIATED WITH LUNG CANCER?
We randomly choose the records of 10 patients with lung cancer and find 10 control patients without lung cancer with a match one-for-one in age, sex, health history, socioeconomic level, and air quality of living region. Then we record whether each patient smokes (yes) or not (no). Are the rates in cancer and non-cancer patients different?

METHOD: MCNEMAR’S TEST
McNemar’s test assesses the dependence of categorical data that are matched or paired. We want to determine whether a certain characteristic is associated with a disease (or other malady). We identify n patients having the disease and pair them one-for-one with control patients without that disease but having other possibly relevant characteristics the same. Then we record the presence or absence of the characteristic in each patient and summarize the result in a two-way table. The data are listed in the format of Table 9.14, where “yes” and “no” indicates the presence or absence, respectively, of the characteristic. A tally of the results can be recorded in a two-way table (Table 9.15), where a is the number of
Table 9.15 Recording format for tallying the data of Table 9.14.

<table>
<thead>
<tr>
<th>Control members of pair</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseased members of pair</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

pairs with a “yes—yes” sequence (as in patient 2), b is the number of pairs with a “yes—no” sequence (as in patient 1), and so on.

The difference in proportions of cases having and not having the characteristic under study is given by \((b - c)/n\) and its standard error by \(\sqrt{b + c}/n\). Chi-square with 1 df is calculated by Eq. (9.11) and the p-value is obtained from Table III.

\[
\chi^2_{1df} = \frac{(b - c - 1)^2}{b + c}
\]  

(9.11)

The p-value is the probability of observing data as or more indicative of the alternative hypothesis if the null hypothesis were in fact true.

**EXAMPLE COMPLETED: OUT-OF-RANGE READINGS FOR CLINIC VERSUS LABORATORY**

We tally the in-range and out-of-range INR readings for the clinic and the laboratory in the format of Table 9.15 to obtain Table 9.16. \(b = 27\), and \(c = 3\). By substituting in Eq. (9.11), we find \(\chi^2_{1df} = 23^2/30 = 17.633\), which is much larger than the \(\alpha = 0.05\) critical value of 3.84. The largest tabulated value for 1 df in Table III is 12.13, which is associated with a p-value = 0.0005, recorded as p-value < 0.001. We conclude that the out-of-range rates are different for the clinic versus the laboratory.

**ADDITIONAL EXAMPLE 1 COMPLETED: IS SMOKING ASSOCIATED WITH LUNG CANCER?**

We list the relevant data in the format of Table 9.14 as Table 9.17 and then tally the numbers of pairs in the “yes—yes”, “yes—no”, and so on categories in the two-way table (Table 9.18). The critical value of chi-square with 1 df corresponding to \(\alpha = 0.05\) is 3.84. For the lung cancer data, McNemar’s test statistic is

\[
\chi^2_{1df} = \frac{(5 - 1)^2}{5} = 3.2.
\]

From Table III a chi-square of 3.2 with 1 df lies between 0.05 and 0.10; for an \(\alpha\) of 0.05, we do not have sufficient evidence to strongly infer that smoking is related to
lung cancer given the precision afforded by our sample size. The actual \( p \)-value, calculated using a statistical software package, is 0.074.

**ADDITIONAL EXAMPLE 2: IS THERE A GENETIC PREDISPOSITION TO STOMACH CANCER?**

A company physician suspects a genetic predisposition to stomach ulcers and targets a specific gene. She chooses workers in the company who have ulcers. She matches each subject with a nonulcerous coworker in the same job (same work environment) with similar personal characteristics (sex, marital state, etc.). She then checks all subjects for the presence of the suspect gene, obtaining the data displayed in Table 9.19, and then tallies these data in the format of Table 9.15 to obtain Table 9.20. McNemar’s chi-square statistic is \( \chi^2_{1df} = (|b-c|-1) / (b + c) = (7-1)^2/9 = 4 \). From Table III the critical value of \( \chi^2 \) for \( \alpha = 0.05 \) with 1 \( df \) is 3.84. Because 4.0 \( > \) 3.84, the \( p \)-value is less than 0.05; therefore she concludes that evidence is present for a genetic predisposition based upon a level \( \alpha = 0.05 \) test. The actual \( p \)-value is 0.046.
Exercise 9.10

Are boys more accident prone than girls? You suspect that boys between the ages of 3 and 5 years are more accident prone than girls. You identify families with a boy and a girl and check their medical records for accidents during those ages. Your data are displayed in Table 9.21. Use McNemar’s test to decide if boys have significantly more accidents.
9.9 COCHRAN’S Q: MATCHED PAIR TEST OF A 2 × r TABLE

EXAMPLE POSED: CONSISTENCY OF CANCER RATING SEVERITY

In preparation for a radiology reliability study, a radiologist reads radiographs from 100 cancer patients and rates the severity of cancer on a 1—5 scale (DB35). If his rating is different from the known severity, it is scored 1; otherwise 0. After a month the radiographs are shuffled, and he rates the severity again. After a third month the process is repeated. Is his error rate consistent through time?

METHOD

Cochran’s Q test assesses the association of categorical data in which a binary variable depends on three or more matched or paired independent categories.

A table is set up with cases (e.g., patients) numbered in the first column and matched treatments across the top row. Each cell in the table is scored as 1 for survival, effective treatment, difference from standard, or such and 0 for the alternative event. \( r \) = the number of rows (e.g., number of patients) after deleting those that contain all 1’s or all 0’s, \( c \) = the number of columns to be compared (e.g., number of treatments), \( n \) = the total number of 1’s, \( r_i \) = number of 1’s in the \( i \)th row, and \( c_j \) = number of 1’s in the \( j \)th column.

The \( r_i \) are squared and added to form a sum of squares for \( r \), or \( SS_r \). The \( c_j \) are also squared and added, forming a sum of squares for \( c \), or \( SS_c \). Then the Q statistic is calculated as

\[
Q = \frac{c(c-1)SS_c - (c-1)n^2}{cn - SS_r}
\]

(9.12)

Q is distributed as \( \chi^2 \) with \( c - 1 \) df.

EXAMPLE COMPLETED: CONSISTENCY OF CANCER RATING SEVERITY

Of the 100 cases, 16 inconsistent data triplets occurred. A 1 indicates a reading that disagreed with the standard. The \( \chi^2 \) critical value for \( \alpha = 0.05 \) will be \( \chi^2 \) for \( c - 1 = 2 \) df, or 5.99. The data table appears as Table 9.22:

<table>
<thead>
<tr>
<th>Patient number</th>
<th>First rating</th>
<th>Second rating</th>
<th>Third rating</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Column totals</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>16 = n</td>
</tr>
</tbody>
</table>

Table 9.22 Cancer severity data.
\( c = 3, \quad r = 16, \quad n = 16, \quad SS_r = 16, \quad \text{and} \quad SS_c = 86. \quad Q = \frac{(3 \times 2 \times 86 - 2 \times 256)}{(3 \times 16 - 16)} = 0.125. \quad 0.125 \) is much less than the critical 5.99, so we have no evidence of a discrepancy among error rates of radiograph reading. Using computer software, a \( \chi^2 \) value of 0.125 with 2 df gives \( p \)-value = 0.939.

**Exercise 9.11**

**Does the time of eating affect total cholesterol?** The medical community has long believed that total cholesterol will be greater after a meal than after fasting. Is such an effect greater than would occur by chance? On different days, we measure the total cholesterol for three designs: after fasting, after a fatty meal, and after a low-fat meal. Using 200 mg/dL as the upper value of normal, we record 0 for normal and 1 for measures exceeding 200. (The first two variables are taken from DB37; the low-fat measures are a fictionalized addition for the exercise.) The data appear in Table 9.23.

### Table 9.23 Normal versus abnormal cholesterol data.

<table>
<thead>
<tr>
<th>Reading number</th>
<th>Fasting</th>
<th>Fatty meal</th>
<th>Low-fat meal</th>
<th>Row totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Column totals</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>16</td>
</tr>
</tbody>
</table>

9.10 Three or More Ranked Samples With Two Outcome Categories

This test is not used a great deal in medical applications, but, when data are appropriate for it, there is no other option. The need occurs when ranked independent variables give rise to a binary outcome.

**Example Posed: Mortality and Extent of Carinal Resection**

In DB12, let us categorize the extent of carinal resection as 1, 2, and 3 for 1–2.5, 3–3.6, and 4–6 cm, respectively. We ask if mortality is associated with extent of resection.
The ptrend test uses the fact that correlation indicates trend or slope in the relationship between two variables, say \( x \) and \( y \). It has been shown that \( n \times r_{xy}^2 \) follows the distribution of \( \chi^2 \) with 1 \( df \). All that is needed is to calculate and square the correlation coefficient \( r \), multiply it by the sample size, and look up the result in a \( \chi^2 \) table for 1 \( df \).

The test is sensitive to trend. In contrast, the FET detects just differences among the response rates without an increasing (or decreasing) order. An example of this contrast may be seen in the answer to Exercise 9.12.

Example completed: Mortality and extent of carinal resection
The independent variable is extent category with three groups that are ranked small, medium, or large. Survival is the dependent variable, 0 or 1. The outcomes are as shown in Table 9.24. We note that the mortality for a large extent is nearly 10-fold that for a small resection. The correlation coefficient between extent and survival is \( r = 0.334 \). \( 134 \times 0.334^2 = 14.95 \), considerably greater than the 3.84 value for \( \chi^2 \) with 1 \( df \) corresponding to for \( \alpha = 0.05 \). A computer calculation shows \( p \)-value < 0.001. There is strong evidence that increased mortality is associated with the extent of resection.

Exercise 9.12
An orthopedist wanted to know if additional detail about optional surgery influences a patient’s decision to undergo the procedure. 36 patients are randomized into three groups of 12. All groups are informed as to the benefits and risks of the surgery. The second group is additionally informed about realistic estimates of time required to resume normal activity and required exercises. The third group receives that plus views a video clip of the procedure. 11 of the first group continue with the surgery, 8 of the second, and 6 of the third. Is the pattern of diminishing numbers

<table>
<thead>
<tr>
<th>Extent of resection</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (survived)</td>
<td>66</td>
<td>24</td>
<td>27</td>
<td>117</td>
</tr>
<tr>
<td>97%</td>
<td>86%</td>
<td>71%</td>
<td>17%</td>
<td></td>
</tr>
<tr>
<td>1 (died)</td>
<td>2</td>
<td>4</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>3%</td>
<td>14%</td>
<td>29%</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>68</td>
<td>28</td>
<td>38</td>
<td>134</td>
</tr>
</tbody>
</table>

Table 9.24 Mortality by extent of carinal resection.
greater than would have occurred by chance? Calculate the ptrend statistic and make a conclusion based on its result. Test the hypothesis of independence using FET and explain the difference between the two results.

REFERENCES

4. Missing sources. The sources of a few examples could not be found despite a strong effort to locate them. Such data that could not be referenced were slightly altered so as not to reflect on any investigator later appearing.
10 Risks, odds, and receiver operating characteristic curves

10.1 ASSOCIATION MEASURES FOR CATEGORICAL DATA: RISKS AND ODDS

Method and example combined: Accuracy and errors in the diagnosis of appendicitis

Pain and tenderness in the lower right quadrant of the abdomen may or may not be appendicitis. Clinical diagnoses are often wrong. Of 250,000 cases treated yearly in the United States, 20% of appendicitis cases are missed, leading to ruptures and complications, and the appendix is unnecessarily removed in 15%—40% of cases. In a recent study at Massachusetts General Hospital, \(^1\) 100 cases of abdominal pain were diagnosed by computed tomography scans and followed for eventual verification of diagnosis. Fifty-two of 53 cases diagnosed as appendicitis were correct, and 46 of 47 cases diagnosed as other causes were correct. A case diagnosed as appendicitis when it is not is termed a false positive and one diagnosed as not appendicitis when it is, is termed a false negative. We can ask several different questions about the risks of being wrong and the odds of being right. The possible errors and their risks are discussed in this chapter.

Truth table format

A special case of the contingency table is the case in which one variable represents a prediction that a condition will occur and the other represents the outcome, that is, “truth.” Suppose, we are predicting the occurrence of a disease among a sample of patients by means of a clinical test (or, alternatively, from the fact of exposure or non-exposure to a disease). If we also know the outcomes of the test (or occurrence of the disease), we can count the possible relations between predictions and outcomes and array them as \(n_{11}\) through \(n_{22}\) in a \(2 \times 2\) truth table, as illustrated in Table 10.1. Four possible situations are named in the table: (1) true positive, the event of a predicted disease being present; (2) false negative, the event of predicting no disease when disease is present; (3) false positive, the event of a predicted disease being absent, and (4) true negative, the event of predicting no disease when disease is absent. These four concepts are...
Table 10.1 Truth table showing relationships and counts of the prediction of presence or absence of a malady (or exposure or nonexposure of a putative causal agent) as related to the truth (or occurrence) of that presence or absence.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Have disease (or exposed)</th>
<th>Do not have (or not exposed)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True positive</td>
<td>False negative</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Correct decision (probability $1 - \beta$) frequency $n_{11}$</td>
<td>Type II error (probability $\beta$) frequency $n_{12}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>False positive</td>
<td>True negative</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type I error (probability $\alpha$) frequency $n_{21}$</td>
<td>Correct decision (probability $1 - \alpha$) frequency $n_{22}$</td>
<td></td>
</tr>
<tr>
<td>Truth</td>
<td>Have disease</td>
<td>Do not have</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n_{1}$ (predict yes)</td>
<td>$n_{2}$ (predict no)</td>
<td>$n$ (or $n.$)</td>
</tr>
<tr>
<td>Totals</td>
<td>$n_{.1}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
used a great deal in several fields of clinical medicine and should be noted well. The $n$ values are the counts of these events' occurrences and will be used with their subscripts extensively in this chapter. Note that the first subscript refers to row number and the second to column number. A dot represents a sum over the row or column represented by its position. For example, $n_{.1}$ represents the sum over rows for the first column ($n_{.1} = n_{11} + n_{21}$).

**EXAMPLE DATA: EXPOSURE TO DUST AND OCCURRENCE OF COCCIDIOIDOMYCOSIS**

In Table 10.2 a table in the format of Table 10.1 shows some data on the relationship between exposure versus nonexposure to dust in a region believed to be heavily infused with *Coccidioides immitis* and the occurrence versus nonoccurrence of coccidioidomycosis in patients with fever, cough, and chest pains.

### Probability and odds compared

We know from Chapter 3, Probability and relative frequency, that the probability of an event is given by the number of ways the event in question can occur in ratio to the number of ways any event can occur. If $r$ out of $n$ patients have a disease, the chance that one of those patients chosen randomly has the disease is $r/n$. The *odds* that the randomly chosen patient has the disease are the number of ways the event can occur in ratio to the number of ways it does not occur, $r/(n - r)$. For example, the probability of a one spot on the roll of a die is $1/6$; the odds are $1 – 5$.

### Truth table statistics

A number of truth table statistics are shown in Table 10.3. Columns 1 and 2 give the names of the concepts and the formulas for computing them from the sample truth table, respectively. Columns 3 and 4 give population definitions and the relationships these concepts are estimating, respectively, that is, what they would be if all data in the population were available. These are the values that relate to error probabilities, as discussed in Chapter 3, Probability and relative frequency, and Chapter 4,
<table>
<thead>
<tr>
<th>Names of quantities</th>
<th>Sample estimate formulas</th>
<th>Definitions</th>
<th>Relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>False-positive rate</td>
<td>$n_{21}/n_2$</td>
<td>Probability of a false positive ($\alpha$)</td>
<td>$P(\text{predict Yes</td>
</tr>
<tr>
<td>False-negative rate</td>
<td>$n_{12}/n_1$</td>
<td>Probability of a false negative ($\beta$)</td>
<td>$P(\text{predict No</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>$n_{11}/n_1$</td>
<td>Probability of a true positive: power ($1 - \beta$)</td>
<td>$P(\text{predict Yes</td>
</tr>
<tr>
<td>Specificity</td>
<td>$n_{22}/n_2$</td>
<td>Probability of a true negative ($1 - \alpha$)</td>
<td>$P(\text{predict No</td>
</tr>
<tr>
<td>Accuracy</td>
<td>$(n_{11} + n_{22})/n$</td>
<td>Overall probability of a correct decision</td>
<td>$P(\text{predict No</td>
</tr>
<tr>
<td>PPV</td>
<td>$n_{11}/n_{11}$</td>
<td>Probability that a positive prediction is correct</td>
<td>$P(\text{Yes</td>
</tr>
<tr>
<td>NPV</td>
<td>$n_{22}/n_{22}$</td>
<td>Probability that a negative prediction is correct</td>
<td>$P(\text{No</td>
</tr>
<tr>
<td>RR</td>
<td>$n_{11}n_{2}/n_{12}n_{1}$</td>
<td>Probability of a disease when predicted in ratio to probability of the disease when not predicted</td>
<td>$P(\text{Yes</td>
</tr>
<tr>
<td>OR</td>
<td>$n_{11}n_{22}/n_{12}n_{21}$</td>
<td>Odds of a disease when predicted in ratio to odds of the disease when not predicted</td>
<td>Ratio (Yes/No</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>LR</td>
<td>$n_{11}n_{2}/n_{21}n_{1}$</td>
<td>Probability of correctly predicting a disease in ratio to probability of incorrectly predicting the disease</td>
<td>$P$ (predict Yes</td>
</tr>
<tr>
<td>AR</td>
<td>$(n_{11}/n_{1}) - (n_{12}/n_{2})$</td>
<td>Amount of incidence rate attributable to result of prediction or to exposure factor</td>
<td>$P$ (Yes</td>
</tr>
</tbody>
</table>

*Columns 3 and 4 show the population entities these values are estimating. The rightmost column shows disease occurrence rates from the data of Table 10.2. “Prediction” may be thought of as “exposure” in the epidemiological context.*
Distributions. As before, a vertical line, |, is read as “given.” Column 5 shows numerical estimates of the probabilities and odds in predicting coccidioidomycosis occurrences (Table 10.2). Interpretations of the statistics are discussed next.

**False-positive rate and false-negative rate**

The false-positive rate is the relative frequency of no disease when disease is predicted (exposure to dust occurred). The dust exposure estimates this chance as $35/397 = 0.09$. The false-negative rate is the relative frequency of disease occurring when no disease is predicted (exposure to dust did not occur), estimated as $39/59 = 0.66$.

**Sensitivity**

The sensitivity, $1 - \text{false-negative rate}$, is the sample estimate of the probability of disease occurring when it is predicted, that is, when exposure to dust occurs. This estimate of the probability of disease occurring when disease is predicted, the true-positive rate, is $20/59 = 0.34$. In the context of hypothesis testing, sensitivity is the sample estimate of the frequently met power.

**Specificity**

The specificity, $1 - \text{false-positive rate}$, is the sample estimate of the probability of correctly classifying the patient as free of disease, that is, ruling out disease when it is absent. This estimate of the probability of no disease occurring when the disease is not predicted, the true-negative rate, is $362/397 = 0.91$. In the context of hypothesis testing, specificity is the sample estimate of $1-\alpha$.

**Accuracy**

Sometimes, the interest is in the rate of overall accuracy, that is, the total percent correct, combining the true positives and true negatives. Accuracy is calculated as true-positive plus true-negative counts, divided by the total count. The accuracy of dust exposure as a predictor of the disease is $(20 + 362)/456 = 0.84$. However, this definition of accuracy assumes the errors of a false positive and a false negative have the same weight, that is, the same importance in the decision process, which they often do not. For example, a false negative in a test of carcinoma implies a missed cancer, whereas a false positive implies an unnecessary biopsy; clearly, the two errors are not equal in severity. Accuracy must be used judiciously and interpreted carefully.

**Positive and negative predictive values**

Predictive values indicate the relative frequency of a predictor being correct or an exposure leading to infection. The positive predictive value (PPV) of $20/55 = 0.36$ for dust exposure indicates that we will be correct just under 3 times out of 8 when we
predict that the patient has the disease, and the negative predictive value (NPV) of \( \frac{362}{401} = 0.90 \) for no dust exposure indicates that we will be correct 9 times out of 10 when we predict that a patient is free of disease. These measures are not used clinically as frequently as sensitivity, the rate at which we detect disease that is present, or specificity, the rate at which we rule out disease that is absent. However, in the epidemiological context, PPV and NPV are more useful than sensitivity or specificity.

**Relative risk**

The relative risk (RR) gives the rate of disease given exposure in ratio to the rate of disease given no exposure. It could also represent the ratio of incidence rates of disease under two different conditions or strategies (including treatment as one strategy and nontreatment as the other). In the dust example the RR gives the rate of coccidioidomycosis among exposed subjects in ratio to the rate among those not exposed; the 3.739 ratio tells us that the disease is between three and four times as likely in exposed subjects as in unexposed subjects.

RR sometimes appears under other names: risk ratio or relative rate.

**Odds ratio**

The odds ratio (OR) gives the odds of disease when predicted in ratio to the odds of disease when not predicted, indicating the usefulness of the prediction method. The coccidioidomycosis OR of \( \frac{20 \times 362}{35 \times 39} = 5.30 \) indicates that the odds of disease when exposed to dust is five times its odds when not exposed. Suppose we had recorded that 26 of the exposed patients who had been exposed to dust were exposed only once, with 5 contracting coccidioidomycosis, and the remaining 20 were exposed repeatedly, with 9 contracting the disease. Forming two separate 2 × 2 tables, we find that the OR for the single exposures would be 2.2, whereas that for the multiple exposures would be 7.6. It would be clear that repeated exposure carries a greater risk of disease. The OR gives us some indication of level of association between the two variables, but one drawback hinders interpretation: a negative OR lies in the region between 0 and 1, while a positive OR lies in the region between 1 and infinity. As a counter, the logarithm of the OR lies between −infinity and infinity, with no association falling at 0. The symmetry of log OR (LOR) allows easier interpretation. A test of association between the two variables, the LOR test, is addressed in Section 10.3.

**The relationship between relative risk and odds ratio**

The RR and OR seem rather similar in their definitions. The relationship between them is worth noting. If the disease or malfunction is rare, \( n_{11} \) and \( n_{12} \) are small so that their product \( n_{11} \times n_{12} \) is almost zero and drops out of \( RR = \frac{n_{11}(n_{12} + n_{22})}{n_{12}(n_{11} + n_{21})} = \frac{n_{11} \times n_{12} + n_{11} \times n_{22}}{(n_{11} \times n_{12} + n_{12} \times n_{21})} \approx \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} = OR. \)
In the case of a rare disease the OR approximates the RR. As a simple example, suppose a subject exposed to a factor has probability 0.01 of contracting an associated disease, whereas a subject not exposed has a probability of only 0.001. The RR is 10.00, and the OR is 10.09.

**Likelihood ratio**

The likelihood ratio (LR) gives the probability of correctly predicting disease in ratio to the probability of incorrectly predicting disease. The LR indicates how much a diagnostic test result will raise or lower the pretest probability of the suspected disease. An LR of 1 indicates that no diagnostic information is added by the test. An LR greater than 1 indicates that the test increased the assessment of the disease probability; if less than 1, it decreased. The LR of 3.845 in the dust example indicates that a patient with coccidioidomycosis is almost four times as likely to have been exposed as not exposed.

**Negative likelihood ratio**

Sometimes reference to a negative LR will be seen. Whereas the positive LR is the probability of a positive test in a patient with the malady in question in ratio to a positive test in a patient without that malady, the negative LR is the ratio of a negative test in a patient without the malady in ratio to a negative test in a patient with the malady. In the dust exposure example the negative LR \(\frac{n_{22}}{n_{12}} \times \frac{n_{1.}}{n_{2.}}\) is 1.379, indicating that a patient with no dust exposure is more than 1.37-times as likely to remain free of coccidioidomycosis as to contract it.

**Be wary of the term likelihood ratio**

LR, as used here, is a useful concept in describing and assessing diagnosis and treatment options, but the name is somewhat unfortunate. The term LR has long been used in statistical theory for concepts other than that adopted in the medical community as described here. The user should be careful to differentiate interpreting results and presenting results to others so that the meaning will not be misunderstood.

**Attributable risk**

Attributable risk (AR) is the portion of disease rate attributable to the exposure factor in the epidemiological context, the portion of correct diagnosis rate attributable to a positive predictive result (e.g., lab test) in the clinical context, or the portion of beneficial outcome rate attributable to a treatment. The AR is sometimes referred to as the risk difference. The AR of 0.266 in the dust example indicates that more than one-fourth of the disease occurrences were due to the exposure. AR is sometimes seen in other forms, and the reader must be wary. The form adopted here is perhaps its most
useful form. A related statistic sometimes erroneously called AR is an attributable fraction, the proportion that the occurrence would be reduced if the intervention (exposure, treatment, etc.) was removed. It can be calculated as $1 - 1/RR$.

**The relationship between relative risk and attributable risk**

RR is the ratio of rates of occurrence (among patients exposed, predicted, or treated to patients unexposed, not predicted, or untreated), while AR is the difference in those same rates of occurrence. RR could be thought of as the incidence rate of exposed patients as a multiple of the incidence rate of unexposed patients. AR could be thought of as the amount the rate increases due to the exposure.

**ADDITIONAL EXAMPLE: TRUTH TABLES AS RELATED TO CANCER SURVIVAL RATES**

The risks and odds related to various medical occurrences follow the same numerical patterns but must be interpreted according to their own definitions. Survival as related to treatment represents an interpretation rather different from the results of a lab test.

**PANCREATIC CANCER DATA**

Oncologists examined the effect of reoperation for pancreatic cancer on patient survival 1 year after surgery. Of 28 patients, 16 were resectable and 12 were not. All were followed for 1 year or until death, whichever came first. We know which patients survived for 1 year and which did not; we are interested in what resection portends about that survival. Results are given in Table 10.4.

**FALSE-POSITIVE AND FALSE-NEGATIVE RATES**

The false-positive rate is the chance of death given resection, or $n_{21}/n_{2} = 8/18 = 0.44$. The false-negative rate is the chance of survival given no resection, or $n_{12}/n_{1} = 2/10 = 0.20$. In a laboratory test to detect a disease, the values for false-positive and -negative rates, sensitivity, and specificity deal with the rates at which our predictions come true and have considerable meaning. In this example, they have less meaning, but the PPV and NPV relate more to what interests us.

**Table 10.4** Data on survival as related to treatment for pancreatic cancer.

<table>
<thead>
<tr>
<th></th>
<th>Resection</th>
<th>No resection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survive for 1 year</td>
<td>8 ($n_{11}$)</td>
<td>2 ($n_{12}$)</td>
</tr>
<tr>
<td>Death within 1 year</td>
<td>8 ($n_{21}$)</td>
<td>10 ($n_{22}$)</td>
</tr>
<tr>
<td></td>
<td>16 ($n_{.1}$)</td>
<td>12 ($n_{.2}$)</td>
</tr>
</tbody>
</table>
SENsitivitY, SPECifiCITy, AND ACCURaCY
The sensitivity is the chance of having had a resection, given survival, or \( n_{11}/n_1 = 8/10 = 0.80 \). The specificity is the chance of having had no resection given death, or \( n_{22}/n_2 = 10/18 = 0.56 \). The accuracy, that is, the overall chance of an appropriate resection, is \( (n_{11} + n_{22})/n = (8 + 10)/28 = 0.64 \).

PositivE AND NEGATivE PRDICTivE VALUES
The PPV is the chance of survival given a resection, or \( n_{11}/n_1 = 8/16 = 0.50 \). This useful information is the relative frequency with which resected patients survive. Similarly, NPV, the chance of death given no resection, or \( n_{22}/n_2 = 10/12 = 0.83 \), is the relative frequency with which nonresected patients die.

RELATivE RISK
The RR provides the rate of survival given resection in ratio to the rate of survival given no resection, which is essentially the improvement in survival rate due to resection. \( RR = n_{11}n_{2.}/n_{12}n_{.1} = 8 \times 12/2 \times 16 = 3 \) indicates that the resected patient’s chance of survival to 1 year is three times that of the nonresected patient.

ODDS RATIo
The OR gives the odds of resected patients' survival in ratio to the odds of nonresected patients' survival. \( OR = n_{11}n_{22}/n_{12}n_{21} = 8 \times 10/2 \times 8 = 5 \). The improvement in survival odds due to resection is fivefold.

LIKELIHOoD RATIo
The LR gives the ratio of successful resections to unsuccessful resections. \( LR = n_{11}n_{2.}/n_{21}n_{.1} = 1.8 \). The rate of successful resections is nearly double the rate of unsuccessful ones.

ATTRIBUTABLE RISK
AR gives the difference in rates of survival with and without resection. \( AR = n_{11}/n_1 – n_{12}/n_2 = 1/3 \) indicates that the estimated probability of survival at 1 year among resected patients is 0.33 higher than nonresected patients.

Exercise 10.1
Appendicitis diagnosis. Table 10.5 displays the Massachusetts General Hospital appendicitis data as a truth table. Calculate and interpret sensitivity, specificity, accuracy, PPV, NPV, RR, OR, LR, and AR.
10.2 Inference for the Risk Ratio: The Log Risk Ratio Test

**Introduction**

The chi-square and Fisher’s exact tests of contingency are so influenced by sample size that they are limited criteria of association. A small sample is unlikely to indicate a significant relationship between variables, however strong the association may be. However, a large enough sample will show evidence of association, no matter how weak the influence of that association is. Formulation of a test around an interpretable level of association is needed. Two such tests are the LOR test and the log RR (LRR) test. In this section, we consider the LRR test. The next section introduces the LOR test. In addition, AR may often be used as a clinical indicator of the relationship between two variables. The AR is estimated as the difference in sample proportions. As such, a test of the hypothesis that the AR is zero is equivalent to a two-sample test of proportions being equal. Thus the two-sample test of proportions that was discussed in Section 9.6 can be used to test whether the AR differs from zero in large samples.

**The log in the log relative risk test**

Is the RR large enough to indicate that the association between the two types of categories is significant in probability? This is the first question addressed in this section. If we have estimated the level of association by the RR, we can go directly to a test using that value. The RR has a difficult, asymmetric distribution (as does the OR). To put it into a form with a known and usable probability distribution, the natural logarithm of the RR, the LRR is used. LRR takes on all real values, unlike the RR that is strictly positive. LRR, divided by its standard error and then squared, is distributed as chi-square with 1 df. (The chi-square test of LRR is not the same test as the chi-square test of contingency; they just both use the chi-square probability distribution.)

**Table 10.5 Truth table for appendicitis data.**

<table>
<thead>
<tr>
<th>Appendicitis predicted</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendicitis Verified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>52 ($n_{11}$)</td>
<td>1 ($n_{12}$)</td>
</tr>
<tr>
<td>No</td>
<td>1 ($n_{21}$)</td>
<td>46 ($n_{22}$)</td>
</tr>
<tr>
<td></td>
<td>53 ($n_1$)</td>
<td>47 ($n_2$)</td>
</tr>
</tbody>
</table>
**CHAPTER 10 Risks, odds, and receiver operating characteristic curves**

Table 10.6 Contingency table of simultaneous biopsy and digital rectal examination (DRE) results from 301 patients.

<table>
<thead>
<tr>
<th></th>
<th>DRE 1</th>
<th>DRE 0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biopsy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$n_{11} = 68$</td>
<td>$n_{12} = 27$</td>
<td>$n_{1.} = 95$</td>
</tr>
<tr>
<td>0</td>
<td>$n_{21} = 117$</td>
<td>$n_{22} = 89$</td>
<td>$n_{2.} = 206$</td>
</tr>
<tr>
<td>Totals</td>
<td>$n_{.1} = 185$</td>
<td>$n_{.2} = 116$</td>
<td>$n = 301$</td>
</tr>
</tbody>
</table>

**EXAMPLE POSED: SIGNIFICANCE OF RELATIVE RISK IN PREDICTING A BIOPSY RESULT**

Using the method from Table 10.3 and the data from Table 10.6 giving digital rectal examination (DRE) and biopsy counts, the RR is $(68/185)/(27/116) = 1.58$. This RR is greater than the value of 1, indicating a potential association. Specifically, we estimate that the risk of a positive biopsy given a positive DRE is 58% higher, in a relative sense, when compared to a negative DRE. But is the true relative risk significantly greater than 1, or may this observation reasonably have been observed by chance if there were no association between DRE and biopsy results?

**METHOD FOR THE LOG RELATIVE RISK (LRR) TEST**

A test of the significance of $2 \times 2$ contingency requires the probability distribution of the measure. The LRR is one that can be put into a form having a chi-square distribution. (The log odds ratio might also be used and is discussed in the next section. Others exist, for example, the Mantel–Haenszel chi-square test but are more complicated, and no better for the simpler cases treated in this book.) Using the $n$ notation introduced in Section 10.1, and letting $p_1 = n_{11}/n_{.1}$ and $p_2 = n_{12}/n_{.2}$, the LRR is given by

$$LRR = \ln \left[ \frac{p_1}{p_2} \right]$$  \hspace{1cm} (10.1)

where "ln" denotes natural logarithm. The standard error of LRR is estimated by

$$se(LRR) = \sqrt{\frac{1 - p_1}{n_{.1}p_1} + \frac{1 - p_2}{n_{.2}p_2}}.$$  \hspace{1cm} (10.2)

To test the null hypothesis that the population LRR = $\theta$ (e.g., $H_0$: $\theta = 0$ implies, the category types are independent), the quantity

$$\chi^2 = \left( \frac{LRR - \theta}{se(LRR)} \right)^2$$  \hspace{1cm} (10.3)

(Continued)
EXAMPLE COMPLETED: SIGNIFICANCE OF RELATIVE RISK IN PREDICTING A BIOPSY RESULT

\[ LRR = \ln \left( \frac{p_1}{p_2} \right) = \ln \left( \frac{n_{11}/n_{1}}{n_{22}/n_{2}} \right) = \ln \left( \frac{68/185}{27/116} \right) = 0.4569 \]

Its standard error is

\[ se(LRR) = \sqrt{\frac{1 - p_1}{n_{1}p_1} + \frac{1 - p_2}{n_{2}p_2}} = \sqrt{\frac{1 - (68/185)}{68} + \frac{1 - (27/116)}{27}} = 0.1942. \]

The chi-square statistic to test against \( \theta = 0 \) is

\[ \chi^2 = \left( \frac{LRR - \theta}{se(LRR)} \right)^2 = \left( \frac{0.4569 - 0}{0.1942} \right)^2 = 5.535 \]

with one degree of freedom. This chi-square value falls between 0.025 and 0.01, as seen in Table III, so we may say that the \( p \)-value is significant. The actual \( p \)-value for chi-square with 1 \( df \) calculated from a software package is 0.019. We have adequate evidence from the sample to conclude that DRE result is associated with biopsy result.

**Exercise 10.2**

Is the quality of ligament repair significantly associated with the method of repair? Use the data in Exercise 9.5 on ligament repair (Table 9.6) to calculate LRR and \( se(LRR) \) and test the hypothesis that the LRR, \( \lambda \), is equal to 0, that is, that there is no difference between the Modified Bostrom and Chrisman—Snook methods of repair.

**10.3 INFERENCE FOR THE ODDS RATIO: THE LOG ODDS RATIO TEST**

**Introduction**

As previously noted, one could also base a test of association on whether the OR differs from 1. As with the RR, the OR has an asymmetric distribution. Thus it is again better to consider the natural logarithm of the OR, the \( LOR \). \( LOR \) is
symmetric about 0, unlike the OR. As with the LRR statistic, the square of LOR divided by its standard error is distributed as chi-square with 1 df.

**EXAMPLE POSED: SIGNIFICANCE OF ODDS RATIO IN PREDICTING A BIOPSY RESULT**

Again, consider the data from Table 10.6 giving DRE and biopsy counts, but now let us focus on the OR as the measure of association. In this case the OR is $\frac{68 \times 89}{117 \times 27} = 1.92$. As expected, the OR is in the same direction as the RR, but again we may ask: Is 1.9 significant in probability?

**METHOD FOR THE LOG ODDS RATIO (LOR) TEST**

As with the relative risk, the LOR is one that can be put into a form having a chi-square distribution. The LOR is given by

$$LOR = \ln \left[ \frac{(n_{11} + 0.5) \times (n_{22} + 0.5)}{(n_{12} + 0.5) \times (n_{21} + 0.5)} \right], \quad (10.4)$$

where again “ln” denotes natural logarithm. The 0.5 values are a continuity correction added to improve the approximation when the cell values are small. If all the cell values are bigger than, say, 10 or 15, all the 0.5 additions can be ignored. The standard error of LOR is

$$se(LOR) = \sqrt{\left( \frac{1}{n_{11} + 0.5} \right) + \left( \frac{1}{n_{12} + 0.5} \right) + \left( \frac{1}{n_{21} + 0.5} \right) + \left( \frac{1}{n_{22} + 0.5} \right)}. \quad (10.5)$$

To test the null hypothesis that population LOR = $\lambda$ (e.g., $H_0$: $\lambda = 0$ implies the category types are independent), the quantity

$$\chi^2 = \left( \frac{LOR - \lambda}{se(LOR)} \right)^2 \quad (10.6)$$

may be looked up in the chi-square table, Table III, for one degree of freedom. The resulting $p$-value is the probability of observing a LOR as big, or bigger, in magnitude than we did if the population LOR were truly 0 (indicating no association).

If the sample sizes ($n$-values) are large in every cell of the table, the 0.5 adjustments may be omitted.

**EXAMPLE COMPLETED: SIGNIFICANCE OF ODDS RATIO IN PREDICTING A BIOPSY RESULT**

$$LOR = \ln \left[ \frac{(68 + 0.5) \times (89 + 0.5)}{(117 + 0.5) \times (27 + 0.5)} \right] = \ln \left[ \frac{68.5 \times 89.5}{27.5 \times 117.5} \right] = 0.6404.$$
10.3 Inference for the Odds Ratio: The Log Odds Ratio Test

Its standard error is

\[
se(\text{LOR}) = \sqrt{\left(\frac{1}{n_{11} + 0.5}\right) + \left(\frac{1}{n_{12} + 0.5}\right) + \left(\frac{1}{n_{21} + 0.5}\right) + \left(\frac{1}{n_{22} + 0.5}\right)}
\]

\[
= \sqrt{\frac{1}{68.5} + \frac{1}{27.5} + \frac{1}{117.5} + \frac{1}{89.5}} = 0.2658.
\]

The chi-square statistic to test against \( \lambda = 0 \) is

\[
\chi^2 = \left(\frac{\text{LOR} - \lambda}{se(\text{LOR})}\right) = \left(\frac{0.6404 - 0}{0.2658}\right) = 5.8049
\]

with one degree of freedom. As with the test of the \( LRR \), the chi-square value for testing the \( LOR \) falls between 0.025 and 0.01, as seen in Table III, so we may say that the \( p \)-value is significant. The actual \( p \)-value for chi-square with 1 \( df \) calculated from a software package is 0.016. We have adequate evidence from the sample to conclude that DRE result is associated with biopsy result. For comparison, if we omit the 0.5 adjustments, \( \chi^2 = 5.9107 \) and \( p \)-value = 0.015, a negligible difference.

**ADDITIONAL EXAMPLE: RADIAL KERATOTOMY EXPERIENCE AND VISUAL ACUITY**

In the additional example on radial keratotomy surgery\(^5\) introduced in Section 9.4, the ophthalmologist found that postoperative visual acuity and position in his surgical sequence were dependent. The OR, calculated from the formula in Table 10.3, is 0.26, which is much less than 1, suggesting a negative relationship: the greater the number of surgeries, the fewer the eyes with poor refraction. Is the association as indicated by the OR statistically significant? The investigator substitutes data from Table 9.5 in (10.4), (10.5), and finally (10.6) to calculate

\[
\text{LOR} = \ln \left[ \frac{(n_{11} + 0.5) \times (n_{22} + 0.5)}{(n_{12} + 0.5) \times (n_{21} + 0.5)} \right] = \ln \left[ \frac{17.5 \times 10.5}{27.5 \times 24.5} \right] = -1.2993
\]

\[
se(\text{LOR}) = \sqrt{\left(\frac{1}{n_{11} + 0.5}\right) + \left(\frac{1}{n_{12} + 0.5}\right) + \left(\frac{1}{n_{21} + 0.5}\right) + \left(\frac{1}{n_{22} + 0.5}\right)}
\]

\[
= \sqrt{\frac{1}{17.5} + \frac{1}{27.5} + \frac{1}{24.5} + \frac{1}{10.5}} = 0.4791.
\]
From Table III the critical value of $\chi^2$ for $\alpha = 0.05$ for 1 df is 3.84. Since 7.35 is much larger than 3.84, he concludes that there is a statistically significant association between the two factors. The actual $p$-value = 0.007.

**CONSIDERATIONS WHEN CHOOSING BETWEEN THE RELATIVE RISK AND THE ODDS RATIO**

We have previously noted that when an event is rare (i.e., the probability of the event is small), the RR and the OR are similar. However, this relationship does not hold as the probability of the event increases. Specifically, the two start to diverge in meaningful ways as the probability of the event is greater than 0.10. Despite the clear mathematical difference between a probability and an odds, the scientific community has a tendency to misinterpret the OR as a ratio of probabilities. This is because it is more natural for us to think of the probability of an event rather than the probability of “success” divided by the probability of “failure.” Thus for interpretability, choosing to estimate and report the RR may be an attractive option. That being said, the OR has become ubiquitous in reported associations in medical applications. This is primarily for two reasons. First, the OR can be estimated in both prospective designs (where the outcome of interest is random and not fixed by design) and in retrospective designs. One example of a retrospective design is a case—control study design where one samples a fixed number of individuals that have experienced the event of interest (cases) and a fixed number of individuals that have not experienced the event (controls), then measuring the exposure or predictor of interest which is random (i.e., not fixed by design). This type of design is useful when the event of interest is rare, since random sampling from the population would likely yield very few events in a sample. However, in a retrospective design, one cannot estimate the probability of the event for each exposure group because the number of subjects with and without the event was fixed by design. Thus one cannot estimate the population RR in this case. However, the population OR can be estimated using retrospectively sampled data. This is due to a well-known result met in Section 3.1 known as Bayes’ Theorem (attributable to Reverend Thomas Bayes, an English statistician, philosopher, and Presbyterian minister).

The second reason the OR is often reported is due to its connection with logistic regression, which we introduce in Chapter 17, Logistic regression for binary outcomes. In this chapter, we have simply considered the association between two
variables. However, in observational studies, it is important to consider adjustment for confounding factors and effect modifiers as this will impact the estimated association between two variables. Logistic regression allows us to estimate the population OR after adjusting for other variables in the model.

Luckily, when it comes to testing whether the population RR or OR is different from 1, the two tests we have presented will generally agree (in fact, as our sample size grows larger and larger the two tests converge to one another). That being said, if one measure of association is reported, the corresponding test result should also be reported. That is, if the RR is reported as an estimate of association, the \( LRR \) test statistic and resulting \( p \)-value should be reported (as opposed to the \( LOR \) test statistic).

**CLINICAL RELEVANCE OF THE DEPENDENCE BETWEEN CATEGORIES IN A 2 \times 2 TABLE**

The clinical relevance of the dependence is another issue. The RR and OR do not answer this question well, since the actual rates of occurrence have been factored out. The clinical relevance may be indicated by the AR, the difference in rates of occurrence, and by the tetrachoric correlation coefficient \( r_t \) from Section 5.3.

The importance and interpretation of these statistics lie in clinical judgment rather than in a probability statement. For example, using the data from Table 10.6 on biopsy result versus DRE, we find the \( AR = 0.135 \). We get roughly a 1/7 greater chance of predicting an existing positive biopsy correctly from a positive DRE than from a negative DRE, not a startling improvement. For these data, \( r_t \) is a rather small 0.250. As another example, in the pancreatic cancer survival data of Table 10.4, \( AR = 0.333 \); resection increases survival by a third. \( r_t = 0.559 \), a notable correlation. As an example of the upper extreme, the appendicitis prediction data of Table 10.5 yields \( AR = 0.960 \); 96% of the detection of appendicitis is due to their prediction procedures. \( r_t = 0.996 \! \)

\( AR \) and \( r_t \) are not influenced by sample size. If we take any contingency table and increase the counts in all cells 10-fold, we find the same \( AR \) and \( r_t \). We must keep in mind that fact when we interpret these statistics. And while the AR retains rates of occurrence, these rates are factored out of \( r_t \). (Its formula uses \( \text{OR}^{\pi/4} \).) The interpretation of \( r_t \) is aided somewhat by using its standard error, the square root of the expression in Eq. (5.13).

**Exercise 10.3**

*Is the quality of ligament repair significantly associated with the method of repair?* Use the data in Exercise 9.5 on ligament repair\(^4\) (Table 9.6) to calculate LOR and se(LOR) and test the hypothesis that \( \lambda = 0 \), that is, that there is no difference between the Modified Bostrom and Chrisman—Snook methods of repair.
**Exercise 10.4**

*Coccidioidomycosis dependent on dust exposure.* For the data of Table 10.2, calculate and interpret the AR and $r_i$.

**Confidence interval on the odds ratio**

**METHOD**

To construct a confidence interval for the plausible values of the true population odds ratio, we use Eqs. (10.4) and (10.5) to find the two values $LOR \pm z_{1-\alpha/2} \times \text{se}(LOR)$ and then take the antilog of each to obtain a $100 \times (1 - \alpha)\%$ confidence interval given by:

$$
\left( e^{LOR - z_{1-\alpha/2} \times \text{se}(LOR)}, e^{LOR + z_{1-\alpha/2} \times \text{se}(LOR)} \right).
$$

(10.7)

An analogous confidence interval for the population relative risk can also be computed by replacing $LOR$ and $\text{se}(LOR)$ with $LRR$ from Eq. (10.1) and $\text{se}(LRR)$ from Eq. (10.2), respectively. As usual, use 1.96 for $z_{1-\alpha/2}$ to obtain 95% confidence.

**EXAMPLE: CONFIDENCE INTERVALS ON ODDS RATIO TO PREDICT BIOPSY RESULT FROM A DIGITAL RECTAL EXAMINATION**

In the preceding section, we found $LOR = 0.6404$ and $\text{se}(LOR) = 0.2658$ (retaining the 0.5 continuity correction, optional for samples of this size). A 95% confidence interval is then given by

$$
\left( e^{LOR - 1.96 \times \text{se}(LOR)}, e^{LOR + 1.96 \times \text{se}(LOR)} \right) = (1.1269, 3.1943).
$$

We infer that plausible values for the true OR are between 1.13 and 3.19.

**Exercise 10.5**

*Confidence on the OR of coccidioidomycosis.* From Table 10.2, we found $OR = 5.3040$. Use the table data to put a 95% CI on the true OR excluding the 0.5 continuity correction).

**10.4 RECEIVER OPERATING CHARACTERISTIC CURVES**

**WHAT A ROC CURVE IS**

A receiver operating characteristic (ROC) curve is a graph displaying the relationship between the true-positive (on the vertical axis) and the false-positive rates (on the horizontal axis). Brought into the medical field from engineering usage, the ROC curve is usually an
Abbreviated ROC curve. Fig. 10.1 shows the ROC curve for biopsy results predicted by prostate-specific antigen (PSA) from DB1. A ROC curve helps to choose the critical value at which a predictor best discriminates between choices, such as choosing the value of PSA that best predicts the presence or absence of a positive biopsy. The critical value is found as the value at which the curve’s deviation from the diagonal line from (0,0) (lower left) to (1,1) (upper right) is the greatest. However, there is an additional advantage in being able not just to maximize the correct response, but to weight it for relative loss associated with the two types of errors. For example, a missed cancer (false negative) is more costly to the patient than an unnecessary biopsy (false positive). If a false-negative error is assessed as four times as serious as a false-positive error, the corresponding point on the ROC curve can be located and the associated critical value determined.

**Calculating a Receiver Operating Characteristic Curve**

In Fig. 10.1, calculation started with the critical value of PSA at 0, where all patients were classed as noncancerous, and the error rates were evaluated. Then the critical value was taken as 0.1, then 0.2, etc., and the error rates for each critical value were calculated, and the points were plotted on the evolving ROC curve. Such a procedure is time-consuming and best done with computer software.
In our urological data a DRE is either positive or negative. There is no ROC curve for predicting biopsy result from DRE. (More exactly, its ROC curve would consist of a single point.) However, for prostate-specific antigen (PSA), we could choose any critical value. As one error type grows smaller, the other grows larger. The issue is choosing the trade-off of the two error values (or their complements, sensitivity, and specificity).

Choosing the best critical value

Most users will have access to software that lists the possible cut points with their corresponding sensitivities and specificities. However, it may be instructive to learn how such a list is formed, or to be able to form one in the absence of software. The strongest predictor of biopsy result will be indicated by where the perpendicular distance from (and above) the (45 degrees) line of equality is a maximum, which can be seen by inspection to be about at the point (0.36 and 0.75). We use these coordinates to select the number of negative biopsies predicted that led to that point, and from that number we find the critical value of PSA. If we calculated the ROC values ourselves, we just select it from the list of calculations. If the ROC curve was done on a computer, we have to “back calculate” it. We know that 95 patients have a positive biopsy, that is, \( n_1 = 95 \), and that 206 have a negative biopsy, that is, \( n_2 = 206 \). We use this information to find the \( n_{11} \ldots n_{22} \) elements of a new \( 2 \times 2 \) table. The user should draw such a table and fill it in step by step during the following calculations. Sensitivity \( \times n_1 = n_{11} \) or \( 0.75 \times 95 = 72 \). Because \( n_1 = n_{11} + n_{12}, \; n_{12} = 95 - 72 = 23 \). Similarly, \( n_{21} = (1 - \text{specificity}) \times n_2 = 0.36 \times 206 = 74.13 \), which we may round to 74. \( n_{22} = n_2 - n_{21} = 206 - 74 = 132 \). Finally, \( n_{12} + n_{22} = 23 + 132 = 155 \). Predicting the smallest 155 values of PSA to have a negative biopsy would lead to \( 1 - \text{specificity} = 0.36 \) and sensitivity \( = 0.73 \). We rank the PSA values in order smallest to largest and note that the prostate-specific antigen density (PSAD) value lying between the 155th and 156th PSA values is 5.95, the required PSA critical value. For clinical convenience, we may denote the cut point as \( PSA = 6 \).

Choosing the best weighted critical value

Suppose we were to require the rate of false-positive errors (the number of unnecessary biopsies) to be four times the rate of false-negative errors (the number of missed cancers), or \( 1 - \text{specificity} = 4 \times (1 - \text{sensitivity}) \). What critical value of PSA would provide this result? By inspection, \( 1 - \text{specificity} \) would be about 0.60 and \( 1 - \text{sensitivity} \) about 0.15. Following a procedure similar to that in the preceding paragraph, we find \( n_{21} = 124, \; n_{12} = 14, \; n_{22} = 82 \), and negative predictions number 96, which lead to critical \( PSA = 4.25 \).
Exercise 10.6

Can white blood cell (WBC) identify strep A? In a sudden outbreak of Streptococcus pneumonia A (DB20), can WBC separate strep A from other strains? A ROC curve shows sensitivity and specificity for strep A for all recordings of WBC from 128 cases of infection (Fig. 10.2). What would be a reasonable trade-off of sensitivity and specificity in seeking a critical value of WBC to predict strep A versus other strains? Find this WBC value.

10.5 COMPARING TWO RECEIVER OPERATING CHARACTERISTIC CURVES

Using the receiver operating characteristic curve to choose the better of two indicators (risk factors)

Another use of ROC curves is to compare two indicators, for example, PSA and PSAD, the ROC curves, of which are shown superposed in Fig. 10.3. A ROC curve that contains a larger area below it is a better predictor than one with a lesser area. However, if the curves cross, an indicator with a lesser area can be better in some regions of the indicator. In this example the area under the PSAD curve is greater than that under the PSA curve and the PSAD curve is uniformly higher than or equal to the PSA curve.
Even when one curve is uniformly greater than the other(s) (and, of course, has a larger AUC), the question remains: Is the difference statistically significant or could it have occurred by chance? A statistical test of the two (or more) ROC curves is needed. Such a test has been devised, but the calculation is more involved than most readers of this book will want to pursue. A number of software packages will perform this test. Two of the more prominent packages to do it are Stata and SAS.

The test uses the difference between AUCs divided by a function of the standard errors of the AUCs, which has been shown to be distributed as $\chi^2$ with $df = \text{number of curves less one}$.

**EXAMPLE:**

Testing the receiver operating characteristic curves for prostate-specific antigen versus prostate-specific antigen density The areas under the curves (AUCs) are shown in Fig. 10.3. Their standard errors are 0.032 for PSA and 0.031 for PSAD. The test produces $\chi^2 = 6.60$ with 1 df that we compare with the critical value for $\alpha = 0.05$, $\chi^2_{0.05} = 3.84$; the difference is significant. The $p$-value is given as 0.010.
Exercise 10.7

Is loudness or duration of snoring a better predictor of the overweight influence on sleep disturbance? In an overnight sample of 3776 patients experiencing sleep disturbance, snoring in excess of 50 dB was recorded for both average loudness and duration. The patients were classified using body mass index (BMI) ≤ 25 as normal and > 25 as overweight. We want to examine the influence of snoring measures on sleep disturbance, but first, we want to know if we must separate patients by BMI. Is the loudness or the duration of snoring more influenced by BMI? We create the following ROC curves for the two measures as distinguished by normal versus overweight classes of patients (Fig. 10.4).

The loudness $AUC = 0.691$ with $SE = 0.015$ and the duration $AUC = 0.491$ with $SE = 0.014$. $\chi^2 = 92.3$, yielding $p$-value $< 0.001$. Interpret these findings.

REFERENCES

CHAPTER 10 Risks, odds, and receiver operating characteristic curves


11. Tests of location with continuous outcomes

11.1 Basics of location testing

The organization of this chapter

We begin with tests comparing population means of distributions giving rise to samples with data characterized as continuous. These tests are examined in Sections 11.2–11.5. Then we continue with tests comparing distributions giving rise to samples with data characterized as ranked. These tests are examined in Sections 11.6–11.11. Finally, we look at tests of ranked distributions giving rise to samples whose data are also ranks. These tests are examined in Sections 11.12 and 11.13. The distinction among these classes of tests will become increasingly clear throughout the chapter. The distinction may be helped by the test classification table, Table 2.2 inside the book’s back cover.

The basic question being asked

We are interested in how events are located within a distribution. The most frequent question asked is what is typical of a distribution, or in other words what is the average. For continuous measures the sample average is the sample mean. After we estimate the average, often we are interested in making inference regarding the population mean, that is, to test it statistically. In this chapter, we will look extensively at tests of means and then at tests of rank. Rank tests do not test the median, but rather the order of ranks relative to a standard or to the ranks of another set of data.

The means question being asked

Is the population mean different from a hypothesized value or, alternatively, do two or more subpopulations have different means? The magnitude of difference between means does not tell the whole story. The difference depends on the scale. For example, the offset distance of a broken femur shows a larger magnitude if measured in centimeters rather than in inches. The distance must be standardized into units of data variability. We divide the distance between means by a measure of variability and achieve a statistic that does not depend on the scale (z if the population variability is known or closely estimated; t if it is estimated by small samples).
Furthermore, the observed difference in sample means could have occurred by chance even if the populations means are equal. We must assure ourselves whether this is likely or unlikely. Usually, this question is answered by testing the null hypothesis that the population means are equal and then rejecting or failing to reject this hypothesis. The concept of a hypothesis test was discussed in Sections 7.1 and 7.2. We could also ask if the population means are the same, rather than are different. This question falls under the heading of equivalence testing and is addressed in Chapter 12, Equivalence testing.

The assumption of normality and the concept of robustness

Tests of means were originally developed under the assumption that the sample was drawn from a normal distribution. Although usually not truly normal, a distribution that is roughly normal in shape is adequate for a valid test, even in small sample sizes. That is because the tests we will discuss are robust. A robust test is one that is affected little by deviations from the underlying assumptions. In large samples the sample mean is approximately normally distributed by the central limit theorem. Because of this, valid inference can still be made on means even if the underlying distribution of the data is nonnormal. If the data are highly nonnormal then rank-based methods can sometimes provide slight improvements in statistical power, but these are no longer tests of the population mean.

Other assumptions: Independently sampled data and equal variability

A normal shape to the frequency distributions is not the only assumption made in tests of means. Along with most other tests of hypotheses, means tests assume that the data being used are independent of each other and that the standard deviations are the same. Independence implies that knowledge of the value of one datum will provide no clue as to the value of the next to be drawn. An example of violating this assumption would be pooling repeated blood pressure measurements on several hypertensive patients. Knowledge of a pressure for one patient would give some information about what pressure would be expected from the next reading if it arose from the same patient but would give no information if the patients were different; thus the readings are not all independent. An example of different standard deviations might arise upon comparing white blood cells from a sample of healthy patients with those from a sample of patients having bacterial infections. The infected sample might be much more variable than the healthy sample. Techniques exist to adjust for unequal standard deviations, but not much can be done to salvage data with dependencies.
The effect of violated assumptions

The assumption of independent data is essential for all types of tests considered in this chapter; if it is badly violated, the results of any test will be spurious. If the assumption of equal variance is violated in a test of two means, the user should use an unequal-variance form of the means test. We present tests of means that assume equal variances and also provide tests that do not require this assumption. This is done for completeness. In general, however, one should usually default to tests that do not require the assumption of equal variances as they guard against invalid inference automatically if the equal variance assumption is violated and generally have few drawbacks. Further guidance will be given in Section 11.3.

We must specify the null and alternative hypotheses

The null hypothesis states that the mean of the population from which the sample is drawn is not different from a theorized mean or from the population mean of another sample. The alternative hypothesis may say that the two means are not equal or that one is greater than the other. The form of the alternate hypothesis was discussed in Section 7.1. We should specify the hypotheses before seeing the data so that our choice will not be influenced by the outcome.

The alternative hypothesis is used to choose a one- or two-tailed test

The alternative hypothesis will dictate a two-tailed or a one-tailed test. We often expect the result to lie toward one tail, but expectation is not enough. If we are sure the other tail is impossible, such as for physical or physiological reasons, we unquestionably use a one-tailed test. Surgery to severe adhesions and return motion to a joint frozen by long casting will allow only a positive increase in angle of motion; a negative angle physically is not possible. A one-tailed test is appropriate. There are cases in which an outcome in either tail is possible, but a one-tailed test is appropriate. When making a decision about a medical treatment, that is, whether or not we will alter treatment depending on the outcome of the test, the possibility requirement applies to the alteration in treatment, not the physical outcome. If we will alter treatment only for significance in the positive tail and it will in no way be altered for significance in the negative tail, a one-tailed test is appropriate.

11.2 SINGLE OR PAIRED MEANS: ONE-SAMPLE NORMAL (z) AND t TESTS

Single samples and paired samples are treated the same

Single and paired samples are treated by the same method, because we may create a single observation from a pair by subtraction, for example, the after measurement minus the before, or a patient’s response to drug 1 minus the response to drug 2.
EXAMPLES POSED, NORMAL TEST: EARLY VERSUS LATER MEMBERS OF OUR PROSTATE-SPECIFIC ANTIGEN SAMPLE

The sample mean of the first 10 patients of Table DB1.1 is observed to be \( m = 6.75 \). Due to potentially biased sampling, we ask if the mean prostate-specific antigen (PSA) of the population the first 10 patients are drawn from is different from \( \mu = 8.85 \)? (This is the mean for the remaining 291 patients, but we will treat it as a fixed hypothesize value for this example.)

EXAMPLES POSED, t TEST: DOES ASTHMA TRAINING REDUCE ACUTE CARE VISITS?

We include all pediatric asthma patients satisfying inclusion criteria presenting over a period of time, in this case 32, and record the number of acute care visits during a year. We then provide them a standardized course of asthma training and record the number of acute care visits for the following year. These “before-and-after” data allow us to analyze the change per patient: \( d \) (named for “difference” or “delta”) = number of visits before minus number after. A natural null hypothesis is \( H_0: \mu_d = 0 \), where \( \mu_d \) is the population mean of the within-patient difference. What is the alternative hypothesis? We would expect the training to reduce the number of visits, but we are not certain. Perhaps the training will increase the child’s awareness and fear, causing an increase in visits; unlikely, but we cannot rule it out. Therefore we cannot in good faith say that the error can lie in only one direction, so we must use a two-tailed test. \( H_1: \mu_d \neq 0 \).

METHOD: ONE-SAMPLE/PAIRED-SAMPLE z AND t TESTS

We want to test the hypothesis that the mean of the distribution from which we draw our sample, denoted \( \mu_0 \), is the same as the known (theoretical) mean \( \mu \); or \( H_0: \mu_0 = \mu \). (If we are dealing with differences, such as, before minus after, \( \mu_0 \) may be denoted \( \mu_d \).) The alternative hypothesis may be that \( \mu_0 \) is greater than or is less than \( \mu \), giving a one-tailed test, or either, giving a two-tailed test; respectively, \( H_1: \mu_0 > \mu \), \( H_1: \mu_0 < \mu \), or \( H_1: \mu_0 \neq \mu \). We must assume that the basic data are distributed roughly normal or that the sample size is large enough for the sample mean to be approximately normally distributed (say >20). Choose an appropriate \( \alpha \). The test is a normal \( z \) test or a \( t \) test in the form of a standardized mean. When the standard error of the mean \( \sigma_m \) is known theoretically or the sample is large enough that \( s_m \) is close to \( \sigma_m \) (say >50 or >100), use \( z \), calculated as in Eq. (11.1), and Table I:

\[
z = \frac{m - \mu}{\sigma_m} = \frac{m - \mu}{\sigma/\sqrt{n}}. \tag{11.1}
\]

When \( \sigma_m \) is unknown and, therefore, estimated by \( s_m \) and \( n \) is smaller than the guidance above, use \( t \), calculated as in Eq. (11.2), and Table II, with \( n - 1 \) df:

(Continued)
Follow these steps for either test:
1. Specify null and alternate hypotheses and choose $\alpha$.
2. Depending upon the sample size make a quick, informal frequency plot of the basic data to check for normal shape.
3. Look up the critical value in the appropriate table for the chosen $\alpha$.
4. Calculate the appropriate statistic from Eq. (11.1) or (11.2).
5. Make the decision to reject or not reject the null hypothesis.

**EXAMPLE COMPLETED, NORMAL TEST: AVERAGE EARLY VERSUS AVERAGE LATER PROSTATE-SPECIFIC ANTIGEN**

The distribution is shown for all 301 patients in Fig. 4.2. The null hypothesis is that the first 10 patients are drawn from a population having mean PSA $\mu = 8.85$. Before seeing data, we had no reason to anticipate whether the true mean of the first 10 patients should be larger or smaller than $\mu$, so we use a two-tailed test. The null hypothesis is tested by the normal statistic $z = (m - \mu)/\sigma_m$. For this example, we assume the true standard deviation of PSA measures in the population is $\sigma = 17.19$ (this is the standard deviation of the remaining 291 patients in the dataset and we will assume it known it is the true standard deviation in the first 10 patients for this example). Thus $\sigma_m = (\sigma/\sqrt{10}) = (17.19/\sqrt{10}) = 5.44$. From this, we have $z = (m - \mu)/\sigma_m = (6.75 - 8.85)/5.44 = -0.386$. Because we are concerned with how far $z$ is from 0 regardless of the direction, normal curve symmetry allows us to look up $+0.386$ in the table. Table 11.1 is a portion of Table I. In either table the two-tailed $p$-value (looked up as if it were $\alpha$) for this $z$ is a bit more than 0.690 (0.699 from exact calculation). The probability of observing a sample mean farther away from the hypothesized null than we did is nearly 70%; this is not a rare event and hence we

<table>
<thead>
<tr>
<th>$z$ (no. std. deviations to right of mean)</th>
<th>Two-tailed $\alpha$ (area in both tails)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>0.842</td>
</tr>
<tr>
<td>0.30</td>
<td>0.764</td>
</tr>
<tr>
<td>0.40</td>
<td>0.690</td>
</tr>
<tr>
<td>0.50</td>
<td>0.619</td>
</tr>
<tr>
<td>0.60</td>
<td>0.548</td>
</tr>
</tbody>
</table>

*aFor selected distances ($z$) to the right of the mean, given are two-tailed $\alpha$ values, the areas combined for both tails. Entries for the most commonly used areas are italicized.*
have insufficient evidence to conclude that the mean of the population the first 10 patients were drawn from are different from 8.85.

**EXAMPLE COMPLETED, t TEST: DOES ASTHMA TRAINING REDUCE ACUTE CARE VISITS?**

A plot of the data shows an approximately normal shape, satisfying that assumption. The 32 $d$ values were 1, 1, 2, 4, 0, 5, −3, 0, 4, 2, 8, 1, 1, 0, −1, 3, 6, 3, 1, 2, 0, −1, 0, 3, 2, 1, 3, −1, −1, 1, 1, 5. $m_d = 1.66$ and $s_d = 2.32$. $t = (m_d - \mu_d)/s_m = (1.66 - 0)/(2.32/\sqrt{32}) = 1.66/0.41 = 4.05$. In Table 11.2 or Table II, look under the two-tailed $\alpha$ column for $df = 30$ ($df$: degrees of freedom) to find the critical value of $t$; our $df = 31$ will yield a probability just less than the tabulated value, or about 2.04. Since 4.05 is greater than that, and in fact greater than the critical 3.64 for $\alpha = 0.001$, we can say that the probability of observing a within-subject mean difference of 1.66 or greater in magnitude is less than 0.001 if the true population within-subject mean difference were 0. This is a rare event and hence we reject the null hypothesis and conclude that the population within-subject mean difference differs from 0.

**ADDITIONAL EXAMPLE: IS A NEW DYSPEPSIA TREATMENT EFFECTIVE IN THE EMERGENCY DEPARTMENT?**

*Normal (z) TEST*

An emergency medicine physician wants to test the effectiveness of a “GI cocktail” (antacid plus viscous lidocaine) to treat emergency dyspeptic symptoms as measured on a 1−10 pain scale. $\ H_0$: $\mu = 0$. He suspects that the treatment will not worsen the symptoms, but he is not totally sure, so he uses $\ H_1$: $\mu \neq 0$, implying a two-sided test; he chooses $\alpha = 0.05$. He decides to accept as the population standard deviation that for scoring of a large number of patients without treatment, $\sigma = 1.73$. He samples $n = 15$ patients, measuring the difference in pain before minus after treatment. Data are 6, 7, 2, 5, 3, 0, 3, 4, 5, 6, 1, 1, 1, 8, 6. $m = 3.87$. He substitutes in Eq. (11.1) to find $z = (m - \mu)/\sigma_m = (3.87 - 0)/(1.73/\sqrt{15}) = 8.66$. Because the critical value from Table 11.1 or Table I is 1.96, which is much less than

<table>
<thead>
<tr>
<th>Two-tailed $\alpha$</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
<th>.002</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>$df = 11$</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
<td>4.025</td>
<td>4.437</td>
</tr>
<tr>
<td>14</td>
<td>1.761</td>
<td>2.145</td>
<td>2.624</td>
<td>2.977</td>
<td>3.787</td>
<td>4.140</td>
</tr>
<tr>
<td>26</td>
<td>1.706</td>
<td>2.056</td>
<td>2.479</td>
<td>2.779</td>
<td>3.435</td>
<td>3.707</td>
</tr>
<tr>
<td>30</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.385</td>
<td>3.646</td>
</tr>
<tr>
<td>40</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>3.307</td>
<td>3.551</td>
</tr>
<tr>
<td>60</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>3.232</td>
<td>3.460</td>
</tr>
<tr>
<td>100</td>
<td>1.660</td>
<td>1.984</td>
<td>2.364</td>
<td>2.626</td>
<td>3.174</td>
<td>3.390</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.090</td>
<td>3.291</td>
</tr>
</tbody>
</table>

*Selected distances ($t$) to the right of the mean are given for various degrees of freedom ($df$) and for two-tailed $\alpha$, areas combined for both positive and negative tails.*
8.66, he rejects H₀ and concludes that the treatment is effective. The actual p-value is 0 to more than 3 decimal places and so is stated as p-value < 0.001.

**t TEST**

Suppose the standard deviation came from a small sample instead of a large sample. If the physician had decided not to use σ = 1.73 because it arose from untreated patients, he would have estimated the standard deviation from the data as s = 2.50 and would have used the t test. From Table 11.2 or Table II the critical value for a two-tailed t with 14 df is 2.145. Substituting in Eq. (11.2), he finds \( t = \frac{(m - \mu)}{s_m} \) = 6.00. The conclusion is the same as that for the case of large test statistics.

**Exercise 11.1**

The original example that “student” used for his t. In W.S. Gossett’s classic 1908 paper³ introducing the t test, he used the following data. Two soporific drugs were used in turn on 10 insomniac patients and the number of hours of additional sleep each provided was recorded. Data were as in Table 11.3. We denote the unknown population’s mean of differences as δ, estimated from the sample by the mean of d. At the \( \alpha = 0.05 \) level of significance, test H₀: δ = 0 against H₁: δ ≠ 0.

### 11.3 TWO MEANS: TWO-SAMPLE NORMAL (z) AND t TESTS

The normal test and the t test are two forms of the two-sample means test

Two forms are addressed in this section: the case of known population standard deviations, or samples large enough that the sample standard deviations are not practically different from known, and the case of small-sample estimated standard deviations. The means test uses a standard normal distribution (z distribution) in the first case and a t distribution in the second.

#### Table 11.3

Gossett’s original data illustrating his t test.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Dextro</th>
<th>Laevo</th>
<th>Difference (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7</td>
<td>1.9</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>-1.6</td>
<td>0.8</td>
<td>2.4</td>
</tr>
<tr>
<td>3</td>
<td>-0.2</td>
<td>1.1</td>
<td>1.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>2.0</td>
<td>4.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Mean =</td>
<td>0.75</td>
<td>2.33</td>
<td>1.58</td>
</tr>
<tr>
<td>Standard deviation =</td>
<td>1.23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 11 Tests of location with continuous outcomes

(Review: The sample means are assumed normal. Standardizing places a standard deviation in the denominator. If the standard deviation is known, it behaves as a constant and the normal distribution remains. If the standard deviation is estimated from the sample, it follows a probability distribution and the ratio of the numerator’s normal distribution to this distribution turns out to be $t$ when the underlying data are normally distributed.)

**Assumptions required**

As discussed at the beginning of this chapter, the two-sample means test requires the assumptions that data sampled are independent from each other and that their frequency distributions are approximately normal, that is, roughly bell-shaped. The assumption of equal variances is addressed in the next paragraph.

**The effect of differing variances and/or sample sizes**

The particular form of the test to contrast two means depends on the relationship of the variances, which are pooled to obtain an overall estimate of variability. Unequal variances are pooled differently from equal variances. Table 11.4 provides classical advice for which type of test to use. We present this table, along with the tests of means that assume equal variances and also provide tests that do not require this assumption. This is done for completeness. In practice, however, we generally default to tests that do not require the assumption of equal variances as they guard against invalid inference automatically if the equal variance assumption is violated and generally have few drawbacks in terms of loss of power.

There is some controversy in the field of statistics about using unequal-variance normal and $t$ methods. Although they are approximately correct and acceptably usable, some statisticians would prefer to use the rank-sum test whenever variances are

| Table 11.4 Classical advice for selecting an appropriate two-sample means test. |
|-------------------------------|-----------------|-----------------|-----------------|-----------------|
| **Total sample size**          | **Subgroup sample size** | **Variances about equal** | **Variances somewhat different** | **Variances extremely different** |
| Large size or $\sigma$’s known | About equal         | Normal ($z$) test, equal variances | Normal ($z$) test, unequal variances | Rank-sum test     |
|                               | Very different   | $t$ Test, equal variances             | $t$ Test, unequal variances          | Rank-sum test     |
| Small size                     | About equal       | $t$ Test, equal variances             | $t$ Test, unequal variances          | Rank-sum test     |
|                               | Very different   | $t$ Test, equal variances             | $t$ Test, unequal variances          | Rank-sum test     |
unequal. There are some cons to this approach as well. Specifically, the rank-sum test is no longer a test of means (or medians) and hence lacks some interpretability. In addition, the test can suffer power loss relative to a test of means.

EXAMPLES POSED: ARE AGE OR PROSTATE VOLUME RISK FACTORS FOR PROSTATE CANCER? Examples for the tests of Table 11.4 are given in the following order: normal \((z)\) test, equal variances; \(t\) test, equal variances; normal \((z)\) test, unequal variances; and \(t\) test, unequal variances. Toward the end of this section, these results are compared with those using the rank-sum test, to be met in Section 11.8.

METHOD: TWO SAMPLE TESTS OF MEANS

Steps in method:
1. Make sure that the sample has been drawn such that it represents the population and that observations are independent from each other. This step is pure judgment based on the way the data have been collected. If these requirements are violated, statistics will not help very much.
2. For smaller sample sizes, make quick frequency plots of the two samples’ basic data to check normality and assess outlying observations that may be potential data errors.
3. Choose the \(z\) test or \(t\) test as appropriate. You have \(n\) data split between two samples of size \(n_1\) and \(n_2\). If \(\sigma\) (the population standard deviation under the null hypothesis) is known or if \(n\) is large (greater than 50 or 100), use \(z\). If \(\sigma\) is not known and \(n\) is small, use \(t\).
4. Specify null and alternative hypotheses. The null hypothesis usually will be \(H_0: \mu_1 = \mu_2\). Select the alternative as \(H_1: \mu_1 \neq \mu_2\) for a two-sided test, or \(H_1: \mu_1 < \mu_2\) or \(H_1: \mu_1 > \mu_2\) for a one-sided test.
5. Choose an appropriate \(\alpha\) and look up the associated critical value from Table I \((z)\) or Table II \((t)\) with \(n_1 + n_2 - 2\) degrees of freedom. For a two-sided test, use the “two-tailed \(\alpha\)” heading. For a one-sided test, use “one-tailed \(\alpha\).”
6. Calculate as appropriate the \(z\) or \(t\) statistic. The test is in the form of a standardized difference between means, that is, the difference \(m_1 - m_2\) divided by the standard error \(\sigma_d\) or \(s_d\).
   To use \(z\), calculate it by
   \[
   z = \frac{(m_1 - m_2)}{\sigma_d}, \tag{11.3}
   \]
   where
   \[
   \sigma_d = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{11.4}
   \]
   when variances are equal, or (Continued)
(CONTINUED)

\[ \sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \]  

(11.5)

when variances are unequal.

To use \( t \), calculate it by

\[ t = \frac{(m_1 - m_2)}{s_d}, \]  

(11.6)

where

\[ s_d = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \left[ \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \]  

(11.7)

when variances are equal. For the \( t \) test when variances are unequal (and \( ns \) are small, say up to 50 or 100), find \( s_d \) from Eq. (11.8) with \( df \) from Eq. (11.9). \( df \) is rounded to the integer next smallest below the rather peculiar expression of Eq. (11.9).

\[ s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]  

(11.8)

\[ \approx \text{approx}(df) = \frac{(s_1^2/n_1) + (s_2^2/n_2))^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \]  

(11.9)

7. Reject or fail to reject the null hypothesis depending on where the statistic lies relative to the critical value.

8. If a statistical software package is available, the actual \( p \)-value may be calculated to facilitate further interpretation of the decision (as in Section 7.2).

EXAMPLE, NORMAL (z) TEST, EQUAL VARIANCES: AGE AND PROSTATE CANCER

We ask if patients with prostate cancer are older on average than those without.

Steps in example, normal (z) test:

1. We judge that the data are independent and that the sample is adequately representative.

2. Fig. 11.1 shows a plot of the two frequency distributions. They appear adequately normal in shape with equivalent standard deviations.

3. Of \( n = 301 \), \( n_1 = 206 \) biopsies were negative and \( n_2 = 95 \) were positive. The samples are large enough to use the \( z \) test. From the data, we calculate \( m_1 = 66.59 \) years, \( s_1 = 8.21 \) years, \( m_2 = 67.14 \) years, and \( s_2 = 7.88 \) years.
4. $H_0: \mu_1 = \mu_2$. Since it is possible for positive biopsy patients to be either older or younger (although we might not expect younger), we choose a two-sided alternative, $H_1: \mu_1 \neq \mu_2$.

5. We choose $\alpha = 0.05$. Table 11.1 shows a fragment of Table I, the normal distribution. Under “Two-tailed $\alpha$”, for $\alpha = 0.05$, we find the critical $z = 1.96$.

6. We want to use the $z$ form. From the data set statistics in the first table of DB1 the standard deviation for the 301 ages is 8.10; we take it as $\sigma$. Then from Eq. (11.4),

$$\sigma_d = 8.10 \times \sqrt{\frac{1}{206} + \frac{1}{95}} = 1.00,$$

and from Eq. (11.3), $z = (m_1 - m_2)/\sigma_d = (66.59 - 67.14)/1 = -0.55$.

7. The statistic $z = -0.55$ is well within the null hypothesis acceptance bounds of $\pm 1.96$. We do not reject the null hypothesis.

8. $p$-value = 0.582. We have insufficient evidence to conclude that the two population means differ. We fail to reject the null hypothesis that the two means are equal.

**Could the $t$ test have been used instead of the $z$ test?**

What would have been the effect of using the $t$ test instead of the normal test of means in examining the age difference between positive and negative biopsy patients? Is the $t$ test appropriate? It is. The normal test assumes that the variances are obtained from the entire population, which often is an infinite number, so that the $z$ test really is an approximation to the true normal. Upon comparison of Tables I and II, we can see that the critical $t$ values for $\infty$ df are the same as the normal values, but our sample size is, of course, not infinite. Let us compare the critical values for a two-tailed $\alpha$ of
The normal critical value is 1.96. With samples of 206 negative biopsies and 95 positive, \( df = (206 - 1) + (95 - 1) = 299 \). We want to use Table II, \( t \) distribution, a fragment of which is shown as Table 11.2. Probabilities for 299 \( df \) will lie between the rows for 100 \( df \) and an immeasurably large \( df \) (symbolized by infinity, \( \infty \)). Under the column for two-tailed \( \alpha = 0.05 \) the \( t \) critical value will lie between 1.98 and 1.96, or about 1.97, quite similar to the 1.96 critical value for the normal. Let us illustrate the steps in performing the two-sample \( t \) test for these data.

**FOLLOWING THE STEPS FOR THE \( t \) TEST IN THE EXAMPLE OF AGE AND PROSTATE CANCER**

The first four steps and the choice of \( \alpha \) are identical to the \( z \) test for these data. Finding the critical value in the fifth step was addressed in the previous paragraph; our critical value is 1.97. Step six is actual calculation. The formula in Eq. (11.7) for the standard error of the mean in the case of small samples gives

\[
s_d = \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right) \left( \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right)} = \sqrt{\left( \frac{1}{206} + \frac{1}{95} \right) \left[ \frac{205 \times 8.21^2 + 94 \times 7.88^2}{206 + 95 - 2} \right]} = 1.01
\]

From Eq. (11.6):

\[
t = \frac{(m_1 - m_2)}{s_d} = \frac{(66.59 - 67.14)}{1.01} = -0.54.
\]

For the \( z \) test the \( z \) statistic was \(-0.55\), falling well within the \( H_0 \) acceptance region of \( \pm 1.96 \). Similarly, to follow step seven for the \( t \) test, the \( t \) statistic falls well within the critical value region of \( \pm 1.97 \). We note that, for samples of over 100, the difference between the methods is negligible. Indeed, upon calculating the exact \( p \)-value, we find it to be 0.582, which is identical with that for the \( z \) test.

**EXAMPLE, NORMAL (\( z \) TEST, UNEQUAL SAMPLE SIZES AND VARIANCES: IS PROSTATE VOLUME A RISK FACTOR FOR PROSTATE CANCER (CAP)?**

We omit the five patients with known benign prostate hypertrophy (BPH), because we want a population at risk for CaP without other prostate disease. We ask if the population mean volume for patients with negative biopsies (\( n_1 = 201 \)) is different from that for patients with positive biopsies (\( n_2 = 95 \)). The sample sizes are large enough to use methods for normal instead of \( t \). \( H_0: \mu_1 = \mu_2 \). As it is theoretically possible for either mean to be the larger, \( H_1: \mu_1 \neq \mu_2 \). Plots of the frequency distributions appear in Fig. 11.2 with normal curves fitted. The shape for negative biopsies (A) is a little skewed to the right, while that for positive biopsies seems adequately symmetric.
However, the central limit theorem assures us that the sample means will be approximately normally distributed. The descriptive data are as in Table 11.5.

The sample sizes are quite different and standard deviations are somewhat different. Using the methods of Section 13.3, the variances test significantly different with $p$-value $= 0.008$. The method for unequal variances is appropriate and should probably be the default to use in practice regardless.

From Eq. (11.5),

$$
\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{17.33^2}{201} + \frac{13.68^2}{95}} = 1.86,
$$

so that

$$
z = \frac{(m_1 - m_2)}{\sigma_d} = \frac{(36.85 - 32.51)}{1.86} = 2.33.
$$

From Table 11.1 or Table I, $z = 2.33$ corresponds to $p$-value $= 0.020$, which is statistically significant. We have sufficient evidence that the mean prostate volume for patients with negative biopsies is larger than that for those with positive biopsies.
EXAMPLE POSED FOR $t$ TEST, UNEQUAL SAMPLE SIZES AND VARIANCES:
IS PROSTATE VOLUME A RISK FACTOR FOR PROSTATE CANCER (CAP)?

Would the $t$ test of means have been appropriate in the example just above? Yes; for large samples, the normal and $t$ tests are just a little different. We have the same hypotheses and normality assumptions.

From Eq. (11.8),

$$s_d = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} = \sqrt{\frac{17.33^2}{201} + \frac{13.68^2}{95}} = 1.86,$$

so that

$$t = \frac{(m_1 - m_2)}{s_d} = \frac{(36.85 - 32.51)}{1.86} = 2.33,$$

which are the same standard error and standardized difference as for $z$. However, we must calculate $df$ in order to find the $p$-value. From Eq. (11.9),

$$df \approx \frac{(s_1^2/n_1+s_2^2/n_2)^2}{((s_1^2/n_1)^2/(n_1-1))+((s_2^2/n_2)^2/(n_2-1))}$$

$$= \frac{(17.33^2/201+13.68^2/95)^2}{((17.33^2/201)^2/200)+((13.68^2/95)^2/94)} = 228.8,$$

which rounds to $df=228$. From Table 11.2 or Table II, $t = 2.33$ for 228 $df$ corresponds to a $p$-value of 0.02, as with the $z$ test. A calculation on a computer to three decimal places yields $p$-value = 0.021, negligibly larger than that from the $z$ test.

Suppose we had used the rank-sum test to assess prostate volume

Since the positive biopsy distribution was somewhat skewed, some would suggest the use of a rank-based method, such as the rank-sum test from Section 11.8, which does not require assumptions on the underlying distribution. Upon doing so, the $p$-value turns out to be 0.097, which is not statistically significant, potentially changing our conclusion. Why might this discrepancy arise? Primarily, this is because the rank-sum test is not a test of means. It is closer to, though not exactly, a test of medians. In this example the means were 36.85 and 32.51 for a difference of 4.34, while the medians were 33.20 and 30.87 for a difference of 2.33, just over half as much. The observed difference in means is larger due to the skewed nature of the distributions and the impact of outlying observations on the means. Does this skewness imply that one should not consider inference for the population means? No. It simply implies that inference for the population means will not be the same as that for the population medians. Which should one focus on? That depends
on the scientific objective. In some cases, we want to be sensitive to outlying observations (e.g., when our outcome is a marker for safety we want to detect rare but large values). In this case a test of means is appropriate. In other situations we may wish to be robust to the influence of outlying observations. In this case a test of medians or use of rank-based methods may be preferred.

**ADDITIONAL EXAMPLES, z AND t TESTS, EQUAL VARIANCES: COMPARING THE EFFECTIVENESS OF TWO TREATMENTS**

An emergency medicine physician wants to compare the relative effectiveness of a “GI cocktail” (antacid plus viscous lidocaine) (treatment 1) versus IV ranitidine HCl (treatment 2) to treat emergency dyspeptic symptoms as measured on a 1–10 pain scale. He records data as pain before treatment minus pain 45 minutes after treatment for $n = 28$ patients, randomly assigned to the two treatments; 15 fall into the first group and 13 into the second. In the medical literature, he finds a large-sample standard deviation of pain difference from untreated patients, $\sigma = 1.73$, and uses it as the standard deviation for both treatments. $H_0$: $\mu_1 = \mu_2$. He takes $\alpha = 0.05$. The test is two-tailed, since either treatment could be the more effective one; $H_1$: $\mu_1 \neq \mu_2$. The critical values of $z$ from Table 12.1 or Table I are $\pm 1.96$. Data are as in Table 11.6:

$\bar{m}_1 = 3.87$ and $\bar{m}_2 = 4.08$. He substitutes in Eqs. (11.4) and (11.3) to find

$$\sigma_d = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.73 \sqrt{\frac{1}{15} + \frac{1}{13}} = 0.6556$$

and

$$z = \frac{\bar{m}_1 - \bar{m}_2}{\sigma_d} = \frac{3.87 - 4.08}{0.6556} = -0.32.$$  

$z$ is within $\pm 1.96$, so he fails to reject $H_0$; he has insufficient evidence to conclude that there is a difference between the treatments. He interpolates from Table I to find $p$-value $= 0.75$.

A critic validly comments that the population $\sigma$ of untreated patients may not be an appropriate estimate for treated patients. He reanalyzes his results, using standard deviations estimated by his data. He finds $s_1 = 2.50$ and $s_2 = 2.84$. He substitutes in Eq. (11.7) and then Eq. (11.6) to find

<table>
<thead>
<tr>
<th>Table 11.6 Pain relief for two treatments of dyspepsia.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
</tr>
<tr>
<td>Treatment 2</td>
</tr>
</tbody>
</table>
CHAPTER 11 Tests of location with continuous outcomes

Table 11.7 Pain relief data from two drugs.

<table>
<thead>
<tr>
<th>Treatment 1</th>
<th>2</th>
<th>0</th>
<th>3</th>
<th>3</th>
<th>0</th>
<th>7</th>
<th>1</th>
<th>4</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 2</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

\[ s_d = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \left[ \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right]} = \sqrt{\left(\frac{1}{15} + \frac{1}{13}\right) \left[ \frac{14 \times 2.5^2 + 12 \times 2.84^2}{15 + 13 - 2} \right]} = 1.0088 \]

and

\[ t = \frac{3.87 - 4.08}{1.0088} = -0.21. \]

From Table II the critical \( t \) value for 26 \( df \) is 2.056. \( t \) is well within \( \pm 2.056 \), so he fails to reject \( H_0 \); there is insufficient evidence to conclude that there is a difference between the treatments. From statistical software, \( p \)-value = 0.83.

**ADDITIONAL EXAMPLE, \( t \) TEST, UNEQUAL VARIANCES: PAIN RELIEF FROM TWO DRUGS**

In 1958 a study\(^2\) compared a new postoperative pain relief drug (treatment 1) to the established Demerol (treatment 2). Data, consisting of reduction in pain measured on a 1 – 12 rating scale, for 23 patients who completed the protocol were as in Table 11.7.

\( n_1 = 13, \ m_1 = 2.15, \ s_1 = 1.9513, \ s_1^2 = 3.8077, \ n_2 = 10, \ m_2 = 5.10, \ s_2 = 4.0125, \ s_2^2 = 16.1000. \) The variances seem quite different. (For a test of these variances, see Additional Example in Section 13.3.) We substitute in Eq. (11.8) and then Eq. (11.6) to find

\[ s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{3.8077}{13} + \frac{16.1000}{10}} = 1.3795 \]

\[ t = \frac{m_1 - m_2}{s_d} = \frac{2.15 - 5.10}{1.3795} = -2.14. \]

The degrees of freedom calculation is a bother in this case. From Eq. (11.9),

\[ \text{approx}(df) = \frac{\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)} \]

\[ = \frac{(3.0877/13)^2 + (16.1000/10)^2}{((3.0877/13)^2/12) + ((16.1000/10)^2/9)} = 11.66 \]
The next smallest integer gives \( df = 11 \). From Table II the critical value for two-tailed \( \alpha = 0.05 \) for 11 \( df \) is 2.201. Our calculated \( t \) is just shy of the critical value, so we do not quite pass the threshold for rejecting the null hypothesis. Statistical software gives \( p \)-value = 0.056. While not “statistically significant,” the probability of observing results as or more indicative of the alternative hypothesis if the null hypothesis were true is still fairly small. This should be emphasized along with observed magnitude of the difference when reporting the results of the test.

**Exercise 11.2**  
*(Equal variances). Tympanic temperatures for left and right ears.* We are investigating the reliability of a certain brand of tympanic thermometer (temperature measured by a sensor inserted into the patient’s ear). Sixteen readings (°F), 8 per ear, were taken on a healthy patient at intervals of 1 minute, alternating ears.\(^4\) Data are given in Exercise 11.3. As either ear may be higher, the alternative hypothesis is two sided. “L” denotes left, “R”, right. \( m_L = 96.41^\circ F \), \( s_L = 0.88^\circ F \), \( m_R = 97.86^\circ F \), and \( s_R = 1.12^\circ F \). At the \( \alpha = 0.05 \) level of significance, if we treat the samples within the patient as independent are the means of the left and right ear different?

**Exercise 11.3**  
*Survival of carinal resection patients.* In DB12 the sample is large enough to take the sample standard deviation as if it were a population standard deviation \( \sigma \). For convenience, carry the calculations to only two decimal places. Perform a normal (z) test to learn if there is evidence that (1) mean age and (2) extent of carinal resection is different for the patients who survived versus those who died.

**Exercise 11.4**  
*Theophylline level change for men versus women.* In DB3, use the difference between theophylline levels at baseline minus at 5 days and perform a t test to learn if there is evidence that the mean reduction in level due to the drug differs between men and women.

**Exercise 11.5**  
*Mean bone density for men versus women.* In DB7, perform a t test to learn if there is evidence that the mean bone density is different between men and women.

**Table 11.8** Human immunodeficiency virus levels for vaccinated versus unvaccinated patients.

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>…</th>
<th>24</th>
<th>25</th>
<th>…</th>
<th>44</th>
<th>45</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of virus Treatment</td>
<td>134</td>
<td>19,825</td>
<td>38,068</td>
<td>…</td>
<td>4315</td>
<td>8677</td>
<td>…</td>
<td>292</td>
<td>67,638</td>
<td>4811</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>…</th>
<th>24</th>
<th>25</th>
<th>…</th>
<th>44</th>
<th>45</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of virus Treatment</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>…</td>
<td>1</td>
<td>2</td>
<td>…</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Exercise 11.6
(Unequal variances). Testing the effectiveness of a vaccine to inhibit human immunodeficiency virus (HIV). The vaccine was tested by randomizing HIV patients into unvaccinated (treatment 1) and vaccinated (treatment 2) groups and compared the number of HIV per mL of blood. The data format was as Table 11.8.

\[ n_1 = 24, \; m_1 = 21,457.3, \; s_1 = 29,451.98, \; n_2 = 22, \; m_2 = 32,174.5, \; s_2 = 50,116.50. \]
The standard deviations appear quite different, so an unequal-variances t test was chosen. (An F test of the variances as in Section 13.3 yielded p-value < 0.001.) At the \( \alpha = 0.05 \) level, test \( H_0: \mu_1 = \mu_2 \) against \( H_1: \mu_1 \neq \mu_2. \) Is \( H_0 \) rejected?

11.4 THREE OR MORE MEANS: ONE-FACTOR ANALYSIS OF VARIANCE

Example posed: Prostate cancer’s risks related to age

PSA is often used to screen patients for risk of cancer. PSA < 4 often represents low risk, PSA between 4 and 10 is uncertain, and >10 is high risk. We will label the categories \( i = 1, 2, 3, \) respectively. Could age be associated with these PSA levels? We want to know if the population mean ages are different for these groups. \( H_0: \) There are no differences in the mean age across the groups. The alternate hypothesis is \( H_1: \) A difference in mean age exists somewhere among the groups.

Concept of one-way analysis of variance

We want to know whether the population means from \( k \) groups are the same or not. (For first-time reading, just replace \( k \) by 3 until the material is familiar.) Should we just make \( t \) tests for each possible pair? No. Each \( t \) test would increase the probability of committing a type I error. (For two independent level \( \alpha = 0.05 \) tests, we would have the chance of error on the first or the second or both, which is \( 1 - \) the chance of no error on any, or \( 1 - [0.95]^2 = 0.0975, \) nearly double.) For three groups the overall (or familywise) type I error rate is nearly 15% \((1 - [1 - \alpha]^3 = 14.3\%); \) the risk of error would exceed 26% for 4 groups and exceed 40% for 5 in the case of independent tests. We need to use a single test to detect any overall difference and, if one difference or more is embedded, find a way to detect which one(s) are significant without increasing the risk of error. This single overall test is an analysis of variance (acronymed ANOVA), because it analyzes or separates the variance components into that due to mean differences and that due to random influences. In order to distinguish this form of ANOVA from the analysis of two or more factors simultaneously, it is called a one-factor ANOVA, or often simply a one-way ANOVA.
One-way analysis of variance is like the \( t \) test generalized to three or more means

Our total sample has \( n \) observations, sample mean \( m \), and variance \( s^2 \). This sample is divided into \( k \) groups having \( n_1, n_2, \ldots, n_k \) observations with group sample means \( m_1, m_2, \ldots, m_k \). The method of one-way ANOVA may be thought of as a 3-or-more-mean extension of the two-mean \( t \) test. It looks very different, but underneath the mechanics are similar. If we took \( k \) to be 2, using ANOVA to test 2 means, we would find that ANOVA’s \( F \) statistic is just the square of \( t \) and the \( p \)-values come out the same. (The one-way ANOVA bears a relationship to the \( t \) test much as the Kruskal–Wallis bears to the rank-sum test, as seen later in this chapter.)

Assumptions required for analysis of variance

A legitimate ANOVA, as would be expected from a generalization of the two-sample \( t \) test, requires the same three assumptions: (1) The data are independent from each other; (2) the distribution of each group in the original data about its mean is normal; and (3) the variances (or standard deviations) about the means are the same for all groups. ANOVA is fairly robust against these assumptions, so we need not be stringent about them, but the data should not be extremely far off.

New terms in analysis of variance: Mean square and sum of squares

These are only new names for familiar concepts. The mean square (MS) is just the sample variance, and the sum of squares (SS) is the numerator in the sample variance calculation. To analyze the variance (in the classic sense of separating it into components), the total SS (SST) is separated into its component due to variability among the means and the remaining (residual or error) component. These component sums of squares are divided by their \( df \) to obtain variances (or MSs). The differences-among-means variance (MSM or mean square of means) divided by the error variance (MSE or mean square of error) yields the \( F \) statistic. This is because conceptually the differences-among-means component is a variance due to an identified cause and the error variance is a variance

<table>
<thead>
<tr>
<th>Source of variability</th>
<th>Sum of squares</th>
<th>Variances, or mean squares</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Designation</td>
<td>Formula</td>
</tr>
<tr>
<td>Mean</td>
<td>SSM</td>
<td>( \sum^k n_i (m_i - m)^2 )</td>
</tr>
<tr>
<td>Error</td>
<td>SSE</td>
<td>( SST - SSM )</td>
</tr>
<tr>
<td>Total</td>
<td>SST</td>
<td>( \sum^n (x_i - m)^2 )</td>
</tr>
</tbody>
</table>

\( MSE \), Mean square of error; \( MSM \), mean square of means; \( MST \), mean square of total; \( SSE \), sum of squares for error; \( SSM \), sum of squares for means.

The symbolism \( \sum^k \) implies “sum (from 1) up to \( k \)”; likewise, \( \sum^n \) implies “sum up to \( n \).”
assumed to be due to random influences alone. If the means have values close together, their differences-among-means variance is small; if the means are quite different one from another, their differences-among-means variance is large. If the differences-among-means variance is sufficiently larger than the error variance that it is unlikely to have happened by chance, we say a significant difference exists among the means. The relationships among the various components in the ANOVA are shown in Table 11.9.

The results of an ANOVA usually are displayed as in Table 11.10.

---

**METHOD AND INTERPRETATION OF ANALYSIS OF VARIANCE (ANOVA)**

Most statistical software packages will perform a one-way ANOVA upon command. The logic of a one-way ANOVA follows these steps:

1. **Pose** $H_0$: There are no differences among the population means, and $H_1$: There are one or more differences somewhere among the population means.
2. **Verify** approximate satisfaction of the assumptions: (1) Normal distributions and (2) equal variances in the original data of the $k$ groups.
3. **Choose** the error probability $\alpha$, such as 0.05, you are willing to accept and find the associated critical value of $F$, named $F_0$, in Table V for $k - 1$ (numerator) and $n - k$ (denominator) $df$.
4. **Calculate** $m$ and $s^2$ for the total sample as usual, recording the sum of squares for the total (SST) before dividing by $n - 1$.
5. **Calculate** the means $m_i$ for the $k$ groups.
6. **Calculate** the sum of squares for means (SSM) by squaring each difference $m_i - m$ and adding them.
7. **Calculate** the sum of squares for error (SSE) by SST - SSM.
8. **Calculate** $s_m^2$ (or $\text{MSM} = \text{SSM}/(k - 1)$ and $s_e^2$ (or $\text{MSE} = \text{SSE}/(n - k)$).
9. **Calculate** $F = \text{MSM}/\text{MSE} = s_m^2/s_e^2$.
10. **Compare** the calculated $F$ with $F_0$. If $F > F_0$, reject $H_0$; if $F < F_0$, do not reject $H_0$.

In actuality the error variance is due to random influences plus unidentified causal influences. Part of a good study design is to control the variability so that the influence of unidentified causes is small. The test is conservative in that a significant outcome implies (Continued)
that the numerator variance (MSM) is larger than random plus other causal variability, so that it certainly is larger than random variability alone. If the design is not controlled carefully, large unidentified causes may enter and an outcome that would have been significant when tested against random variability alone will not yield the precision needed to establish a difference in population means.

**Identifying the mean difference(s) that caused the significance: Multiple comparisons tests**

Suppose we reject $H_0$ and conclude that differences exist among the group means. Which pairs of population means are different and which not? Recall that we cannot just make $t$ tests of all possible pairings, because the overall type I error accumulates. Methods testing the possible pairings without increasing the overall type I error rate are called *multiple comparisons* tests (sometimes termed *post-hoc* comparisons). Several, named after their developers, have been devised, including Tukey’s HSD (honestly significant differences), Fisher’s LSD (least significant differences), Scheffé, Bonferroni, and Šidák that test all possible pairs. These are listed with comments comparing them in Table 11.11.

While the comments show some theoretical differences, in practice the tests do not give very different results. The Bonferroni method is the simplest to use, but is also the most primitive theoretically. It leads to excessive numbers of false-negative errors. Use it only when you have no better method available. Having no better method available, however, does occur with a number of frequently used methods, for example, some rank methods and multivariate methods.

Two other tests compare subsets of pairs, including Dunnett’s test, which pairs only the control mean with the treatment means, and Hsu’s MCB, which pairs the strongest result with the others. In addition, the same goal can be accomplished by what are called *multiple range* tests, including those by Duncan and Newman–Keuls. However, they are not as easy to use and will not be pursued further here.

<table>
<thead>
<tr>
<th>Method</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tukey HSD</td>
<td>Exact $p$-values if sample sizes are equal, conservative if they are different</td>
</tr>
<tr>
<td>Fisher LSD</td>
<td>Least conservative method; will find significance more often than others</td>
</tr>
<tr>
<td>Scheffé</td>
<td>Conservative method; find’s no significance if ANOVA not significant</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>Significance is $\alpha$ divided by number of pairs; overly conservative method</td>
</tr>
<tr>
<td>Šidák</td>
<td>A slight theoretical improvement on Bonferroni</td>
</tr>
</tbody>
</table>

ANOVA, Analysis of variance; HSD, honestly significant differences; LSD, least significant differences.

*The term “conservative” implies high specificity, that is, a small chance of showing a difference when there is none.*
MULTIPLE COMPARISON METHODS PERFORMED BY STATISTICAL SOFTWARE

The user will find some methods in one statistical software package, others in another package. Because of the similarity of results from the different methods, the user can use Table 11.11 to choose from among the options given. There will be little loss in using whatever is available.

When statistical software is used to make a multiple comparison test, the outcome is given as $p$-values adjusted so that each may be compared with our chosen overall $\alpha$. Different software packages display the results using various schemes. One common display type is presented here; the user who understands it can follow other schemes with little difficulty. The display forms a table in which the group names are shown as both columns and rows, providing a position for every possible pair. (Of course, positions in which row and column show the same group are omitted.) If we denote the pairwise adjusted $p$-value for the pair $i$ and $j$ as $p_{ij}$, it appears as in Table 11.12.

Table 11.12 Format of multiple comparisons display.

<table>
<thead>
<tr>
<th>Mean 1</th>
<th>Mean 2</th>
<th>$\ldots$</th>
<th>Mean $k-1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean 2</td>
<td>$p_{12}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean 3</td>
<td>$p_{13}$</td>
<td>$p_{23}$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>Mean $k$</td>
<td>$p_{1k}$</td>
<td>$p_{2k}$</td>
<td>$\ldots$</td>
</tr>
</tbody>
</table>

We then identify the pairs for which the displayed $p$-value less than $\alpha$.

Multiple Comparisons Carried Out by Hand

Sometimes a requirement for multiple comparisons arises when no statistical software is available or using methods for which the software does not include it. Without much effort, we can use the Bonferroni method to contrast the pairs. We make two-sample $t$ tests on each pair but choose the critical $t$ from an adjusted $\alpha$ rather than $\alpha = 0.05$. The number of pairs is $q = k(k - 1)/2$ and the Bonferroni adjusted $\alpha$ is $\alpha/q$. We find the $df$ for each pair’s $t$ test. From Table II, we obtain a critical $t$ value for that $df$, using $\alpha/q$ instead of $\alpha$ and interpolating as required. The calculated $t$’s in the test of the pair’s mean must exceed that critical $t$ value. The $t$ values obtained from the $q$ $t$ tests may be displayed in a format similar to the preceding $p$-value display, and each compared with its critical $t$ one by one.

Example completed: Age as related to CaP risks

**DESCRIPTIVE STATISTICS AND ASSUMPTIONS**

We have the ages of patients in three risk groups. The descriptive statistics are as in Table 11.13.
Are the assumptions of normality and equal variance (or standard deviation) in the original data approximately satisfied? We make a quick plot of the distributions and see that they are approximately normal. We examine the three standard deviations and decide that they are not strikingly different. (If we wished to be more formal, for example in preparing for journal publication, we could add evidence to our procedure by testing normality group by group using the methods of Section 13.6 and testing equality of variances using the methods of Section 13.4.)

**ANALYSIS OF VARIANCE CALCULATIONS**

We choose $\alpha = 0.05$. Degrees of freedom are $k - 1 = 3 - 1 = 2$ and $n - k = 301 - 3 = 298$. For numerator $df = 2$ and denominator $df = 298$, Table V gives a critical value of $F$ falling between 3.06 and 3.00, approximately 3.03. We calculate (carrying six significant digits) the additional statistics needed for a one-way ANOVA as follows:

- $m = 66.7641$
- $SST = 19,670.3$
- $s^2 = 65.5675$
- $SSM = \sum n_i (m_i - m)^2 = 89(66.1124 - 66.7641)^2 + \cdots = 449.196$
- $SSE = SST - SSM = 19,221.104$
- $s_m^2$ (or MSM) = $SSM/(k - 1) = 449.196/2 = 224.598$
- $s_e^2$ (or MSE) = $SSE/(n - k) = 19,221.104/298 = 64.5003$
- $F = s_m^2 / s_e^2$ (or MSM/MSE) = $224.598/64.5003 = 3.48$

An ANOVA result is usually displayed in a table, giving the outcomes of the calculations of interest. Table 11.14 in the format of Table 11.10 is such a table for this example. As $3.48 >$ critical 3.03, we reject the null hypothesis and conclude that a

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among means</td>
<td>449.196</td>
<td>2</td>
<td>224.598</td>
<td>3.48</td>
<td>0.032</td>
</tr>
<tr>
<td>Within groups (error)</td>
<td>19,221.104</td>
<td>298</td>
<td>64.5003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19,670.3</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
difference among population means exists. From a computer calculation the \( p \)-value is 0.032.

**WHICH AMONG THE POSSIBLE MEAN DIFFERENCES ACCOUNT(S) FOR THE SIGNIFICANCE?**

Does the significance arise from low to uncertain risk \((\mu_1 \text{ vs } \mu_2)\), low to high risk \((\mu_1 \text{ vs } \mu_3)\), uncertain to high risk \((\mu_2 \text{ vs } \mu_3)\), or some combination of these? For the example, Table 11.15 shows multiple comparisons calculated for all five methods listed in Table 11.11. It can be seen that the methods do not give very different results. In most cases, they will lead to the same conclusions. The mean ages between the low and the uncertain risk groups are far from significantly different for all methods, and the mean ages between the uncertain and high-risk groups are significant for all methods. The \( p \)-values for the low versus high-risk groups do not vary greatly, but straddle 0.05. Fisher’s LSD, the least conservative (higher sensitivity and lower specificity), and Tukey’s HSD, a “middle-of-the-road” method, show significance while the remaining, more conservative methods do not. We conclude that higher PSA values occur in older patients.

**SUPPOSE WE DO NOT HAVE A STATISTICAL SOFTWARE PACKAGE**

If we have no access to statistical software, we can use Bonferroni’s method to contrast the pairs. We make two-sample \( t \) tests on each pair but choose the critical \( t \) from an adjusted \( \alpha \) rather than \( \alpha = 0.05 \). As the number of possible pairings is \( q = 3 \), the

<table>
<thead>
<tr>
<th>Method</th>
<th>Low versus uncertain</th>
<th>Low versus high</th>
<th>Uncertain versus high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tukey HSD</td>
<td>.983</td>
<td>.043*</td>
<td>.035*</td>
</tr>
<tr>
<td>Fisher LSD</td>
<td>.860</td>
<td>.017*</td>
<td>.014*</td>
</tr>
<tr>
<td>Scheffé</td>
<td>.985</td>
<td>.058</td>
<td>.048*</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>1.000</td>
<td>.051</td>
<td>.041*</td>
</tr>
<tr>
<td>Šidák</td>
<td>.997</td>
<td>.050</td>
<td>.041*</td>
</tr>
</tbody>
</table>

*Significance: \( p \)-value < 0.05 is denoted by *.

<table>
<thead>
<tr>
<th>Method</th>
<th>Low versus uncertain</th>
<th>Low versus high</th>
<th>Uncertain versus high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tukey HSD</td>
<td>.983</td>
<td>.043*</td>
<td>.035*</td>
</tr>
<tr>
<td>Fisher LSD</td>
<td>.860</td>
<td>.017*</td>
<td>.014*</td>
</tr>
<tr>
<td>Scheffé</td>
<td>.985</td>
<td>.058</td>
<td>.048*</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>1.000</td>
<td>.051</td>
<td>.041*</td>
</tr>
<tr>
<td>Šidák</td>
<td>.997</td>
<td>.050</td>
<td>.041*</td>
</tr>
</tbody>
</table>

*HSI, Honestly significant differences; LSD, least significant differences.

Table 11.16 Values of \( t \) for pair-wise comparisons of the prostate-specific antigen data.

<table>
<thead>
<tr>
<th></th>
<th>Low risk</th>
<th>Uncertain risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertain risk</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>High risk</td>
<td>2.33</td>
<td>2.64</td>
</tr>
</tbody>
</table>
Bonferroni adjusted $\alpha/q = 0.05/3 = 0.016$. For the different pairings, $df$ varies from about 50 to about 150. In Table II, we see that the $t$-value for $df$ around 100 lying a third of the way from two-tailed $\alpha = 0.02$ to two-tailed $\alpha = 0.01$ is in the vicinity of 2.4. (A computer tells us that it is exactly 2.393.) Thus the calculated $t$’s in the tests of mean pairs must exceed 2.4 to maintain an overall type I error rate of 0.05. The $t$ values obtained from the three $t$ tests, using a format similar to the $p$-value display above, is shown in Table 11.16.

We see that the $t$ statistic for low to uncertain risk is far from significant, whereas the other two are a little below and a little above the critical value, respectively. The results agree with that shown for the Bonferroni method in Table 11.15.

**ADDITIONAL EXAMPLE: DOES STEROID DECREASE EDEMA FOLLOWING RHINOPLASTY? IF SO, WHAT LEVEL OF STEROID SHOULD BE USED?**

Following rhinoplasty, swelling may cause deformity during healing. Steroids may decrease the swelling, but the required level of steroid has not been known. $n = 50$ rhinoplasty patients were randomized into $k = 5$ groups of increasing steroid level having $n_1 = \cdots = n_5 = 10$. Swelling reduction was measured by MRIs before and after administration of the steroid. $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ was to be tested using

<table>
<thead>
<tr>
<th>Swelling reduction (mL)</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients 1–5</td>
<td>1.6</td>
<td>4.4</td>
<td>5.5</td>
<td>4.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Patients 6–10</td>
<td>2.3</td>
<td>5.8</td>
<td>6.4</td>
<td>8.3</td>
<td>6.2</td>
</tr>
<tr>
<td>Patients 11–15</td>
<td>2.3</td>
<td>6.4</td>
<td>6.9</td>
<td>5.8</td>
<td>3.1</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Means $m_i$</td>
<td>3.77</td>
<td>5.00</td>
<td>5.67</td>
<td>6.79</td>
<td>6.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS (sum of squares)</th>
<th>$df$</th>
<th>MS (mean square)</th>
<th>Calculated $F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SSM $= 56.57$</td>
<td>4</td>
<td>MSM $= 56.57/4 = 14.14$</td>
<td>MSM/MSE $= 4.26$</td>
<td>0.005</td>
</tr>
<tr>
<td>Error</td>
<td>SSE $= 149.57$</td>
<td>45</td>
<td>MSE $= 149.57/45 = 3.32$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SST $= 206.16$</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$MSE$, Mean square of error; $MSM$, mean square of means; $SSE$, sum of squares for error; $SSM$, sum of squares for means.
\( \alpha = 0.05 \). The first few data are given in Table 11.17. The total mean (all data pooled) was \( \bar{m} = 5.52 \) and total variance \( s^2 = 4.2074 \). The following calculations are needed:

\[
\begin{align*}
SST & = (n - 1)s^2 = 49 \times 4.2074 = 206.16, \\
SSM & = 10[(3.77 - 5.52)^2 + (5 - 5.52)^2 + \cdots + (6.35 - 5.52)^2] = 56.57, \\
SSE & = SST - SSM = 206.16 - 56.57 = 149.59.
\end{align*}
\]

Table 11.18 shows the ANOVA results. From Table V the critical \( F_{4,45,df} = 2.58 \). The calculated \( F = 4.26 \) is much larger, so \( H_0 \) is rejected; we conclude that steroid level is associated with swelling. From a statistical software package the calculated \( p \)-value = 0.005.

**IDENTIFYING THE STEROID DOSAGE**

Steroid level affects swelling, but which level should be selected for clinical use? A Bonferroni multiple comparisons procedure using a statistical software package yields significance levels \( (p-values) \) as in Table 11.19, adjusted to be interpreted according to the usual \( \alpha = 0.05 \), although the computer calculations were made so that overall type I error rate is bounded at 0.05. (The Bonferroni method was used so that the reader can reproduce the outcomes without statistical software. The \( p \)-values showing 1.000 are really slightly less, but round to 1.000.)

We see that level 1 is significantly worse than levels 4 or 5; we reject level 1 as an acceptable steroid level. The other levels are not significantly different. Any one of levels 2 through 5 may be chosen for clinical use. Examining the means, we see that level 4 gives the greatest reduction in swelling, so one might tend to choose level 4 pending additional information.

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>0.243</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>0.006</td>
<td>0.333</td>
<td>1.000</td>
</tr>
<tr>
<td>Level 5</td>
<td>0.028</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Table 11.19 Multiple comparisons for edema reduction example.**

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>279</td>
<td>378</td>
<td>282</td>
<td>381</td>
</tr>
<tr>
<td>338</td>
<td>275</td>
<td>335</td>
<td>346</td>
</tr>
<tr>
<td>334</td>
<td>412</td>
<td>335</td>
<td>340</td>
</tr>
<tr>
<td>198</td>
<td>265</td>
<td>282</td>
<td>471</td>
</tr>
<tr>
<td>303</td>
<td>286</td>
<td>250</td>
<td>318</td>
</tr>
<tr>
<td>Mean ( \bar{m} ) =</td>
<td>290.4</td>
<td>323.2</td>
<td>274.8</td>
</tr>
</tbody>
</table>

**Table 11.20 Data for parasite infestation in rats.**
Exercise 11.7

*Are experimental animals equally resistant to parasites?* In a study on the control of parasites, 8 rats were injected with 500 larvae each of the parasitic worm Nippostrongylus muris. Ten days later, they were sacrificed and the number of adult worms counted. The question arose, “Is there a batch-to-batch difference in resistance to parasite infestation by groups of rats received from the supplier?”

$k = 4$ batches of $n_i = 5$ rats each were tested. Data are shown in Table 11.20. Total mean $m = 314.9$ and total variance $s^2 = 4799.358$. At the $\alpha = 0.05$ level of significance, test $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

### 11.5 THREE OR MORE MEANS IN RANK ORDER: ANALYSIS OF VARIANCE TREND TEST

**Is there a pattern of change through the means?**

In one-way ANOVA, significance indicates that there is a difference in means of some sort. It does not tell us what sort of pattern the differences fall into. For a nominal independent variable, no ordinal pattern is postulated, so multiple comparisons are appropriate to compare each pairing of means.

However, if the independent variable is ordinal or rank-ordered (no drug, some drug, lots of drug), the change in pattern of means when moving across the horizontal axis could be an *increasing trend* (no reaction, weak reaction, strong reaction, respectively). Indeed, it could even be a *quadratic curve*, such as weak reaction, strong reaction, weak reaction again, as the stimulus moves from weak, to strong, to very strong—there may be a point of maximum effect after which the effect diminishes again. Let us denote the independent variable by $x$, the usual indicator of a horizontal variable, and the dependent variable by $y$, the usual indicator of a vertical variable. When $x$ is rank-ordered, it is useful to test for a pattern dependent on the ranking (amount of drug, time after administration, etc.). Statistical methods that test for a pattern of change depending on $x$ are designated polynomial contrast tests.

**The concept of testing for a linear pattern**

The concept is to use *weightings* to tease out straight line trends from the variability about the means and to test it for significance. Perhaps we find that mean level of blood oxygen $y$ goes down as body mass index (BMI) $x$ increases in obstructive sleep apnea patients. Such weightings are called orthogonal polynomials. For example, if we have four means, the weights are $-3, -1, 1, 3$, which add to 0. We find the SS for linearity using the weights multiplying the means and test it for significance. If significant, we have an increasing outcome with increasing $x$ (or decreasing outcome if the slope is negative). Then we can subtract (remove) the SS for linearity from the SS for means and test the remainder for significance of categorical differences.
Patterns other than linear

We are not limited to linear patterns. We can find orthogonal polynomials for almost any pattern we suspect. For example, if we suspect a quadratic pattern (blood flow from a wound increases until clotting, then decreases; the mortality rate from cancer for acquired immunodeficiency syndrome patients increases until an effective treatment—for example, HAART—is employed and then decreases), we can use weightings such as 1, \(-2\), 1 for three means or 1, \(-1\), \(-1\), 1 for four means. The reader may pursue ideas of statistical modeling, but polynomial contrast tests more complicated than linear will not be pursued further here.

Implementing the method

It is unlikely that the reader will want to attempt ANOVA trend tests by hand; the example will be limited to software analysis. The software for the ANOVA trend test is not easily found. It is accessible in a number of statistical software packages, but under various headings and not easily understandable. SPSS software contains an accessible and understandable option. Within one-way ANOVA, checking a tab labeled “contrasts” brings forth a command box. Checking “polynomial” and then “linear” will lead to the ANOVA separated into a linear trend and the residual effect of mean differences after the linear trend is removed. (A “quadratic” polynomial pattern can be chosen as well.)

**EXAMPLE: DOES HUMAN IMMUNODEFICIENCY VIRUS PATIENTS’ CD4 COUNT TREND DOWNWARD WITH YEARS OF DISEASE?**

Years of disease in 265 HIV patients were recorded as 1–6, 7–13, and 14 or greater. Group means of CD4 counts per year were 602, 586, and 446, respectively, and were subjected to a one-way ANOVA. The mean differences were separated into components due to a linear trend, and differences remaining after the linear trend were removed. The ANOVA table appears as Table 11.21.

An ordinary one-way ANOVA would consist of the means effect, within groups, and total line only, showing a significant \(p\)-value that implies mean differences. But

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>(F)</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means effect combined</td>
<td>1,256,472.2</td>
<td>2</td>
<td>628,236.1</td>
<td>6.912</td>
<td>.001</td>
</tr>
<tr>
<td>Linear component</td>
<td>1,012,985.4</td>
<td>1</td>
<td>1,012,985.4</td>
<td>11.114</td>
<td>.001</td>
</tr>
<tr>
<td>Residual component</td>
<td>253,486.8</td>
<td>1</td>
<td>243,486.8</td>
<td>2.679</td>
<td>.103</td>
</tr>
<tr>
<td>Within groups</td>
<td>23,815,033</td>
<td>262</td>
<td>90,897.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>25,071,505</td>
<td>264</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 11.21** Analysis of variance trend table for CD4 count in human immunodeficiency virus patients over time.
are these differences due to the decrease in year, to mean differences unrelated to
trend through time, or both? Performing the ANOVA trend test adds the other two
lines, with the 2 \( \text{df} \) separated into 1 \( \text{df} \) for each component. The linear trend is signifi-
cant, and the residual is not; CD4 count is dependent on patients’ years of HIV.

Lest the reader think this result is obvious and the test unnecessary, note that the
CD4 count nadir was also tested. Again, the linear trend was significant, but there was
a significant (\( p \)-value < 0.001) effect remaining in mean differences after the trend was
removed.

11.6 THE BASICS OF NONPARAMETRIC TESTS

The term nonparametric refers to the lack of assumptions about or dependence on
parameters of distributions. For example, a nonparametric test may not have to assume
that the data to be tested arose from a normal distribution. Nonparametric tests are
often called distribution-free tests though the two labels differ in meaning. The great
majority of such tests are based on rank-ordered data. They use the values of the ranks
rather than the original data values for statistical computation. Because only rank-based
nonparametric tests are included in this book, the remainder of this chapter focuses on
ranks.

What are ranks?

Ranked data are data entries put in order according to some criterion: smallest to larg-
est; worst to best; cheapest to costliest. Such rank-order data may arise from ranking
events directly, as a surgeon ranking in order of difficulty the five types of surgery per-
formed most often in his specialty. Alternatively, rank-order data may arise from put-
ting already recorded continuous-type quantities in order, as ordering the PSA of
Table DB1.1 from smallest to largest. These latter ranks would appear as in
Table 11.22.

Ranking categorized data that fall into a natural order

Note that if PSA were categorized into (1) PSA < 4, (2) PSA 4–10, and (3)
PSA > 10, the data could still be ranked and analyzed using rank-order methods,
although there would be so many ties that the analysis would be much less sensitive.

Table 11.22 Prostate-specific antigen (PSA) levels and ranks from Table DB.1.

<table>
<thead>
<tr>
<th>Patient number</th>
<th>PSA</th>
<th>PSA rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.6</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4.1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5.9</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>9.0</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>6.8</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>8.0</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>7.7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>4.4</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>6.1</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>7.9</td>
<td>8</td>
</tr>
</tbody>
</table>
CHAPTER 11 Tests of location with continuous outcomes

How do we use ranks?
To compare two groups that are in rank-order, we attach ranks to the data combined over the two groups and then add the rank values for each group separately, forming rank sums. If the group rankings are not much different, the rankings from the two groups will be interleaved and the rank sums will not be much different. If one group has most of its members preceding the other in rank, one rank sum will be large and the other small. Probabilities of rank sums have been tabulated, so the associated \( p \)-value can be looked up in the table and the decision about the group difference made.

When do we use ranks?
Continuous measurements contain more information than do ranks and ranks more information than counts. When the user can rank events but cannot measure them on a scale, rank-order statistical methods will give the best results. When the sampling distribution of continuous data is skewed or otherwise poorly behaved, the assumptions underlying continuous-based methods may be violated, giving rise to inefficient results or results that may be driven by a few outlying observations. Ranks do not violate these assumptions. When sample sizes are too small to verify the satisfaction of these assumptions, rank-order methods are sometimes considered.

Why do we use ranks?
Statistical methods using continuous measurements on variables depend on the probability distributions of those variables. We assume certain properties of those probability distributions, such as that we are sampling from a normal distribution. When we have small sample sizes, our sample frequency distributions are insufficient to tell us if the assumptions are justified. Rank-order methods do not require as stringent assumptions about the underlying distributions. Further, even when we have larger samples, we may have evidence that the assumptions are not satisfied. Recall that the distribution of prostate volumes (cf. Fig. 4.2) is skewed to the right. Using methods that assume a normal distribution would violate this assumption, because the sample distribution is not the same shape as the assumed distribution. While the central limit theorem would allow us to make valid inference about the population mean using the \( t \)-test, the test may not be as powerful as other options. Rank-order methods, not subject to the skew, could have greater power for testing distributions though these tests do not specifically test for differences in the mean. It is important to note, however, that rank-order methods do not yield inference for differences in means, so if the mean is the clinical focus then rank-order methods should not be used.
SUMMARY OF WHEN TO USE RANK-ORDER METHODS

Rank-order methods might be considered (1) when the primary data consist of ranks, (2) when samples are too small to form conclusions about the underlying probability distributions, or (3) when data indicate that necessary assumptions about the distributions are violated.

Ties in ranked data

If ranks are tied, the usual procedure is to average the tied ranks. For example, a sample of heart rates (HRs) in increasing order is 64, 67, 67, 71, 72, 76, 78, 89. We have the ranks 1–8 to assign. However, the second and third HRs are the same. We average 2 and 3 to assign 2.5 to each. The ranks are 1, 2.5, 2.5, 4, 5, 6, 7, 8. Some statisticians prefer to assign the potential ranks to tied values randomly to avoid ties, but this technique introduces a bit of false information. Many rank methods provide adjustments for ties. Ties reduce the quality of rank methods slightly, but, unless the number of ties is large, not enough to bother with adjusting for them.

11.7 SINGLE OR PAIRED SAMPLE DISTRIBUTION(S): THE SIGNED-RANK TEST

EXAMPLE 1 POSED, SINGLE SAMPLE: PROSTATE-SPECIFIC ANTIGENS VERSUS A POPULATION MEDIAN

Table DB1.1 gives PSA for the first 10 of the 301 urology patients. The median PSA of the next 291 is 5.7. We ask if values from the PSA distribution for the first 10 patients tend to be different from those of the remaining 291. We use a two-tailed test. We take the differences of the first 10 from the median of the remainder. If the medians are alike, the ranks of the deviations of the first 10 should vary approximately equally about 0.

EXAMPLE 2 POSED, PAIRED SAMPLE: DOES A DRUG CHANGE HEART RATE?

Anecdotal information suggests that HR changes in response to a certain ophthalmologic drug designed to reduce intraocular pressure. For a sample of eight patients, the HRs prior to beginning treatment, 24 hours following first dosage, and the change are as follows:

Before: 64, 67, 67, 71, 72, 76, 78, 89
After: 66, 58, 68, 65, 75, 67, 59, 74
Difference: −2, 8, −1, 6, −3, 9, 19, 15

Are the distributions giving rise to the before and after readings different?
**METHOD: THE SIGNED-RANK TEST**

Often referred to as the Wilcoxon signed-rank test, this method tests the hypothesis that the distribution of the differences is symmetric with a mean of 0. It may test

1. a set of observations deviating from a hypothesized common value or
2. pairs of observations on the same individuals, such as before-and-after data.

Steps to perform the test are as follows:

1. Calculate the differences of the observations as in (1) or (2) just above (0 differences are just omitted from the calculation, and the sample size reduced accordingly).
2. Rank the magnitudes, that is, the differences without signs, the smallest being rank 1. (For ties, use the average of the ranks that would have occurred without ties in the same positions.)
3. Reattach the signs to the ranks.
4. Add up the positive and negative ranks.
5. Denote by $T$ the unsigned value of the smaller sum.
6. Look up or compute the $p$-value for the test in Table VIII.

If computer capability is not available and $n > 12$, a large-sample approximation may be found in Section 11.9.

**One tail or two?**

If we have some sound nonstatistical reason why the result of the test must lie in only one tail, such as a physiological impossibility to occur in the other tail, we can double the $\alpha$ to obtain a one-tail error probability.

**Table 11.23** Prostate-specific antigen values’ deviations from median and ranks.

<table>
<thead>
<tr>
<th>Deviations</th>
<th>1.9</th>
<th>−1.6</th>
<th>0.2</th>
<th>3.3</th>
<th>1.1</th>
<th>2.3</th>
<th>2.0</th>
<th>−1.3</th>
<th>0.4</th>
<th>2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signless ranks</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Signed ranks</td>
<td>6</td>
<td>−5</td>
<td>1</td>
<td>10</td>
<td>3</td>
<td>9</td>
<td>7</td>
<td>−4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

**Table 11.24** A portion of Table VIII, signed-rank probabilities, two-tailed probabilities of $T$.^a

| Sample size $n$ |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $T$             | 4               | 5               | 6               | 7               | 8               | 9               | 10              | 11              | 12              |
| 3               | .625            | .313            | .156            | .078            | .039            | .020            | .010            | .005            | .003            |
| 4               | .875            | .438            | .219            | .109            | .055            | .027            | .014            | .007            | .004            |
| 6               | .801            | .438            | .219            | .109            | .055            | .027            | .014            | .007            | .004            |
| 9               | .844            | .469            | .250            | .129            | .065            | .032            | .016            |                 |                 |

^aFor a sample of size $n$ and value of $T$, the entry gives the $p$-value.
EXAMPLE 1 COMPLETED: PROSTATE-SPECIFIC ANTIGENS VERSUS A POPULATION MEDIAN

We obtain the deviations of our first 10 PSAs from the median of the remainder and rank them without regard to the sign (+ or −), then reattach these signs to the ranks. We add the positive ranks and add the negative ranks. If the first 10 observations are like the remainder, the positive rank sum will be similar to the negative rank sum. These deviations and their ranks are as in Table 11.23.

The sum of positive ranks is 46 and that of negative ranks is −9, which are quite different. Denote by $T$ (for “total”) the unsigned value of the smaller sum; in this case, $T = 9$. Table 11.24 shows a portion of Table VIII, signed-rank probabilities, from the back of the book. In Tables 11.24 and VIII the intersection of the row for $T = 9$ and the column for $n = 10$ yields $p$-value = 0.065.

EXAMPLE 2 COMPLETED: DOES A DRUG CHANGE HEART RATE?

The differences, before medication − after medication, are ranked by magnitude, that is, with their signs removed. The data, differences, and rank entries are in Table 11.25. The two unsigned rank sums are 6 for negative, 30 for positive; $T = 6$. In Table 11.24 the intersection of the $n = 8$ column and the $T = 6$ row yields $p$-value = 0.109, too large to conclude evidence of a change.
ADDITIONAL EXAMPLE: HARDWARE TO REPAIR ANKLE FUNCTIONALITY

An orthopedist installs the hardware in the broken ankles of nine patients. He scores the percent functionality of the joint. He asks “Is the average percent functionality less than 90% that of normal functionality?” His data in percentage are 75, 65, 100, 90, 35, 63, 78, 70, 80. A quick frequency plot of the data shows they are far from a normal distribution, so he uses a rank-based test. He subtracts 90% from each (to provide a base of 0) and ranks them, ignoring signs. Then he attaches the signs to the ranks to obtain the signed ranks. These results are as in Table 11.26.

The sum of positive signs will obviously be the smaller sum, namely, 3.5 = T. From Table 11.24 (or Table VIII) the p-value for \( n = 9 \) with \( T = 3.5 \) for a two-tailed test lies between 0.020 and 0.027, about 0.024. Because he chose a one-tailed test, the tabulated value may be halved, or \( p \)-value = 0.012, approximately, clearly significant. He concludes that the patients’ average functionality is significantly below 90%.

Exercise 11.8

Normality of thermometer readings on a healthy patient. We are investigating the reliability of a certain brand of tympanic thermometer (temperature measured by a sensor inserted into the patient’s ear). Eight readings (°F) were taken in the right ear of a healthy patient at 2-minute intervals. Data were 98.1, 95.8, 97.5, 97.2, 97.7, 99.3, 99.2, 98.1. When centered at 98.6°F, is there sufficient evidence to conclude that the resulting distribution giving rise to these data is not symmetric about 0?

11.8 TWO INDEPENDENT SAMPLE DISTRIBUTIONS: THE RANK-SUM TEST

What is being tested

Given two samples taken from two populations, the null hypothesis being tested is that the probability that a randomly chosen member from population 1 has a higher value than a randomly chosen member from population 2 is equal to 0.5. For practical purposes, many users think of it informally as testing whether or not the two distributions have the same median, but the user should keep in mind that it is not a test of medians.
STEPS IN CONDUCTING THE RANK-SUM TEST

Steps in method:

1. Satisfy yourself that the sample has been drawn such that it represents the population and such that observations are independent one from another. This step is pure judgment based on the way the data have been collected. If these requirements are violated, statistics will not help.

2. Specify $\alpha$ and hypotheses, which usually are null: probability that a randomly chosen member from population 1 has a higher value than a randomly chosen member from population 2 is equal to 0.5, and alternate: probability that a randomly chosen member from sample 1 has a higher value than a randomly chosen member from sample 2 is not equal to 0.5.

3. Name the sample sizes $n_1$ and $n_2$. (Assigning the name $n_1$ to the smaller reduces computation.) Combine the data, keeping track of the sample from which each datum arose.

4. Rank the data.

5. Add the ranks of the data from the smaller sample and name the sum $T$.

6. If $n_2 \leq 8$, calculate $U = n_1n_2 + n_1(n_1 + 1)/2 - T$. If $n_2 > 8$, calculate $\mu = n_1(n_1 + n_2 + 1)/2$, $\sigma^2 = n_1n_2(n_1 + n_2 + 1)/12$, and $z = (T - \mu)/\sigma$.

7. Look up $p$-value from the appropriate table: Table IX for $U$ or Table I for $z$.

8. Reject the null hypothesis if $p$-value $< \alpha$; fail to reject the null hypothesis if $p$-value $\geq \alpha$.

### Table 11.27 Prostate-specific antigen (PSA) versus biopsy outcomes.

<table>
<thead>
<tr>
<th>PSA</th>
<th>7.6</th>
<th>4.1</th>
<th>5.9</th>
<th>9.0</th>
<th>6.8</th>
<th>8.0</th>
<th>7.7</th>
<th>4.4</th>
<th>6.1</th>
<th>7.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranks</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Biopsy</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 11.28 Portion of Table IX, rank-sum $U$ two-tailed probabilities.\(^a\)

<table>
<thead>
<tr>
<th>$n_1$:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_2$:</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U$</td>
<td>7</td>
<td>.714</td>
<td>.352</td>
<td>.178</td>
<td>.094</td>
<td>.516</td>
<td>.230</td>
<td>.106</td>
<td>.052</td>
<td>.026</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>.904</td>
<td>.476</td>
<td>.246</td>
<td>.132</td>
<td>.666</td>
<td>.316</td>
<td>.148</td>
<td>.074</td>
<td>.038</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>.762</td>
<td>.428</td>
<td>.240</td>
<td>1.00</td>
<td>.528</td>
<td>.268</td>
<td>.138</td>
<td>.072</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>.914</td>
<td>.536</td>
<td>.310</td>
<td>.648</td>
<td>.344</td>
<td>.180</td>
<td>.098</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)For two samples of size $n_1$ and $n_2$ ($n_2 > n_1$) and the value of $U$, the entry gives the $p$-value.
Example: Is prostate-specific antigen different for positive versus negative biopsy?

Returning to Table DB1.1, we ask, are the PSA levels different for the \( n_1 = 3 \) positive biopsy results and the \( n_2 = 7 \) negative ones? (The symbol \( n_1 \) is assigned the smaller of the two \( n \)s.) The PSA levels, their ranks (small to large), and the biopsy results are as in Table 11.27.

Steps in example:

1. We ask if PSA is a risk factor for prostate cancer. Do our urological patients with cancer have PSA different from those without? We judge that the data are independent and that the sample is adequately representative.

2. \( \alpha = 0.05 \). Hypotheses are null: Probability that a randomly chosen member from the positive biopsy population has a higher value than a randomly chosen member from the negative biopsy population is equal to 0.5, and alternate: probability that a randomly chosen member from the positive biopsy population has a higher value than a randomly chosen member from the negative biopsy population is not equal to 0.5.

3. \( n_1 = 3; n_2 = 7 \).

4. Data arose from first and third rows in data display.

5. Data are ranked as in second row in data display.

6. The rank sum \( T \) for positive biopsies is \( 3 + 10 + 5 = 18 \).

7. \( n_2 = 7; U = 3(7) + 3(4)/2 \) \( - 18 = 9 \).

8. Table 11.28 gives the portion of Table IX including \( n_2 = 7, n_1 = 3 \). The \( p \)-value for \( U = 9 \) is 0.834.

9. \( p \)-Value is greater than \( \alpha \). There is not sufficient evidence to reject the null hypothesis.

Why \( U \) is calculated from \( T \)

The Mann–Whitney \( U \) is tabulated in this book rather than the rank-sum \( T \), also available in some books, because \( U \) requires a much smaller table size. The two lead to the same result.

Table 11.29 Hematocrit (Hct) for type of pylorus repair.

<table>
<thead>
<tr>
<th>Hct</th>
<th>Rank</th>
<th>Open (0) versus laparoscopy (1)</th>
<th>Hct</th>
<th>Rank</th>
<th>Open (0) versus laparoscopy (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.6</td>
<td>1</td>
<td>1</td>
<td>38.3</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>29.7</td>
<td>2</td>
<td>1</td>
<td>38.8</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>32.0</td>
<td>3.5</td>
<td>0</td>
<td>39.0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>32.0</td>
<td>3.5</td>
<td>1</td>
<td>42.0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>32.1</td>
<td>5</td>
<td>1</td>
<td>43.3</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>32.7</td>
<td>6</td>
<td>0</td>
<td>43.9</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>33.9</td>
<td>7</td>
<td>0</td>
<td>46.7</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>34.0</td>
<td>8</td>
<td>1</td>
<td>52.0</td>
<td>16</td>
<td>1</td>
</tr>
</tbody>
</table>
**A one-tailed test**

If we have some sound nonstatistical reason why the result of the test must lie in only one tail, such as a physiological limit preventing the other tail (a knee does not bend forward), we can obtain a one-tailed \( p \)-value by halving the tabulated \( p \)-value.

**Other names for this test**

This test may be referred to in the literature as the rank-sum test, Mann—Whitney \( U \) test, Wilcoxon rank-sum test, or Wilcoxon—Mann—Whitney test. Mann and Whitney published what was thought to be one test and Wilcoxon another. Eventually they were seen to be only different forms of the same test.

**ADDITIONAL EXAMPLE: HEMATOCRIT FOR LAPAROSCOPIC VERSUS OPEN PYLOROMYOTOMIES**

Among the indicators of patient condition following pyloromyotomy (correction of stenotic pylorus) in neonates is hematocrit (Hct) percent. A surgeon\(^{11}\) wants to compare the distribution of Hct among laparoscopic versus open pyloromyotomies using a sample of 16 randomly allocated observations. \( n_1 = n_2 = 8 \). A quick frequency plot shows the distributions to be far from normal; a rank method is appropriate. Data were as follows: Open: 46.7, 38.8, 32.7, 32, 42, 39, 33.9, 43.3. Laparoscopy: 29.7, 38.3, 32, 52, 43.9, 32.1, 34, 25.6. He puts the data in order and assigns ranks, obtaining data as in Table 11.29.

He finds the rank sums to be 77.5 for open pyloromyotomy and 58.5 for laparoscopy. Because of symmetry, either may be chosen; \( T = 77.5 \). He calculates \( U = n_1n_2 + n_1(n_1 + 1)/2 - T = 8 \times 8 + 8(9)/2 - 77.5 = 22.5 \). Table 11.30 is a segment of Table IX. From Table 11.30 or Table IX the \( p \)-value for \( n_1 = n_2 = 8 \) and \( U = 22.5 \) falls between 0.33 and 0.38, or about 0.335.

---

**Table 11.30** A segment of Table IX, rank-sum \( U \) probabilities.\(^a\)

<table>
<thead>
<tr>
<th>( n_1: )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U )</td>
<td>22</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( ^a \)Two-tailed probabilities for the distribution of \( U \), the rank-sum statistic. For two samples of size \( n_1 \) and \( n_2 \) \((n_2 > n_1)\) and the value of \( U \), the entry gives the \( p \)-value.

**Table 11.31** Temperature readings.

| Left ear: | 95.8 | 95.4 | 95.3 | 96.0 | 96.9 | 97.4 | 97.4 | 97.1 |
| Right ear: | 98.1 | 95.8 | 97.5 | 97.2 | 97.7 | 99.3 | 99.2 | 98.1 |
CHAPTER 11 Tests of location with continuous outcomes

Exercise 11.9

Does a certain tympanic thermometer measure the same in both ears? We are investigating the reliability of a certain brand of tympanic thermometer (temperature measured by a sensor inserted into the patient’s ear). Sixteen readings ($^\circ$F), eight per ear, were taken on a healthy patient at intervals of 1 minute, alternating ears. (We note that these are paired data, but they are also paired—matched—through time, because they all arise from the same patient.) Data were as in Table 11.31. Are the distributions of temperature readings for the two ears different?

11.9 LARGE SAMPLE-RANKED OUTCOMES

When do we need a large-sample approximation?

In most situations, computer capability and statistical software are available, in which case the test may be made exactly for any sample size and the associated probability will be displayed. Failing such capability, the ranking of large data sets is manpower intensive, and there are no probability tables, so that large sample approximations are needed.

Single or paired data

The case of single—or paired which become single by subtraction—data will be addressed, then two-sample data.

EXAMPLE POSED: BIAS IN EARLY SAMPLING OF PROSTATE BIOPSY PATIENTS

We ask if the PSA levels of the first 20 patients differ significantly from 8.96 (the average of the remaining 281). The data are $-0.04, 0.06, 0.96, 1.06, 1.26, 1.26, 1.36, 1.36, 2.16, 2.63, 2.86, 3.06, 3.26, 3.36, 3.66, 4.16, 4.36, 4.56, 4.86, 

and 7.66. When the sample is large, it exceeds the tabulated probabilities, but may be approximated by the normal distribution, with $\mu$ and $\sigma$ calculated by simple formulas.

METHOD: THE NORMAL APPROXIMATION TO THE SIGNED-RANK TEST

The normal approximation to the Wilcoxon signed-rank test formally tests the hypothesis that the distribution of differences is symmetric about 0. It may test (1) a set of observations deviating from a hypothesized common value or (2) pairs of observations on the same individuals, such as before-and-after data. The $p$-value will not be identical with the exact method, but only rarely will this difference change the outcome decision. The steps in performing the test by hand are as follows:

1. Calculate the differences between pairs or from a hypothesized central value.
2. Rank the magnitudes (i.e., the differences without signs).
3. Reattach the signs to the ranks.

(Continued)
4. Add up the positive and negative ranks.
5. Denote by $T$ the unsigned value of the smaller; $n$ is sample size (number of ranks).
6. Calculate $\mu = n(n + 1)/4$, $\sigma^2 = (2n + 1)\mu/6$, and then $z = (T - \mu)/\sigma$.
7. Obtain the $p$-value from Table I.

How large is a “large sample”?

The approximation is valid only if the sample is large enough, and the size of this “large enough” is not established. Like all large-sample approximations, the larger the sample, the better the agreement with exact tests. If reliable statistical software is available, it should be used to calculate the exact test. If such software is not available and the approximation is calculated by hand, a minimum sample size may be taken as 16, a number stated in the classic textbook by Snedecor and Cochran. Many statisticians accept an even smaller minimum sample size, some as small as 10.

**EXAMPLE COMPLETED: FIRST 20 PROSTATE-SPECIFIC ANTIGENS VERSUS REMAINDER**

We rank the differences by magnitude (i.e., regardless of sign) and then reattach the signs. We find the sums of positive and negative ranks. $T = 1$, the unsigned value of the smaller sum of ranks; $\mu = n(n + 1)/4 = 20(21)/4 = 105$; $\sigma^2 = (2n + 1)\mu/6 = 41(105)/6 = 717.5$; $\sigma = \sqrt{717.5} = 26.7862$; $z = (T - \mu)/\sigma = (1 - 105)/26.7862 = -3.8826$. This large a $z$ is off the scale in Table I, telling us that the $p$-value $\leq 0.001$. (Calculated exactly, a two-tailed $p$-value $= 0.0001$.)

**Exercise 11.10**

$n = 8$ temperature readings taken on a patient were 98.1, 95.8, 97.5, 97.2, 97.7, 99.3, 99.2, 98.1. When centered at 98.6°F, is there sufficient evidence to conclude that the resulting distribution giving rise to these data is not symmetric about 0? As an exercise, answer the question using the normal approximation to the signed rank test, even though the sample is smaller than appropriate.
Two large sample-independent ranked outcomes

Let us examine the method for a normal approximation to the rank-sum test for larger sample sizes.

**EXAMPLE POSED: PROSTATE-SPECIFIC ANTIGEN FOR PATIENTS WITH VERSUS WITHOUT BPH**

We obtain a sample of 12 patients with BPH. We want to compare PSA for these patients with that for the 10 patients in Table DB1.1. \( n_1 = 10 \) and \( n_2 = 12 \). The data appear in Table 11.32.
METHOD: THE RANK-SUM TEST FOR LARGE SAMPLES

Given two samples, the null hypothesis being tested is that the probability that a randomly chosen member from population 1 has a higher value than a randomly chosen member from population 2 is equal to 0.5.

1. Name the sizes of the two samples \( n_1 \) and \( n_2 \); \( n_1 \) is the smaller.
2. Combine the data, keeping track of the sample from which each datum arose.
3. Rank the data.
4. Add up the ranks of the data from each sample separately.
5. Denote as \( T \) the sum associated with \( n_1 \).
6. Calculate \( \mu = n_1(n_1 + n_2 + 1)/2, \sigma^2 = n_1n_2(n_1 + n_2 + 1)/12, \) and \( z = (T - \mu)/\sigma \).
7. Obtain the \( p \)-value from Table I for a two- or one-tailed test as appropriate.

EXAMPLE COMPLETED: PROSTATE-SPECIFIC ANTIGEN FOR PATIENTS WITH VERSUS WITHOUT BPH

We rank all 22 data in order as in Table 11.33, keeping track of the sample each came from:

\[
T = 105.5, \quad \mu = n_1(n_1 + n_2 + 1)/2 = 10(10 + 12 + 1)/2 = 115; \\
\sigma^2 = n_1n_2(n_1 + n_2 + 1)/12 = 10(12)(10 + 12 + 1)/12 = 230; \\
\sigma = \sqrt{230} = 15.1658; \\
z = (T - \mu)/\sigma = (105.5 - 115)/15.1658 = -0.626. \]

From Table I, \( z = 0.60 \) yields an area in both tails of 0.548, slightly larger than the \( p \)-value that would arise from a \( z \) of 0.626. (An exact calculation gives an area, that is, \( p \)-value, of 0.532.)

Exercise 11.11

In Exercise 11.9, \( n_1 = n_2 = 8 \) temperature readings were taken from a patient’s left and right ears, respectively. The question was, “Are the distribution centers for the two ears different (implying a two-tailed test)?” As an exercise, answer the question using the normal approximation to the rank-sum test, even though the sample size is too small for the approximation. Data were as in Table 11.34. Is the resulting \( p \)-value close to that for the exact method?
### Table 11.36 Prostate-specific antigen (PSA) ranks in three groups of patients.

<table>
<thead>
<tr>
<th>PSA</th>
<th>Rank for BPH patients</th>
<th>Rank for positive biopsy patients</th>
<th>Rank for negative biopsy patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1.4</td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1.6</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2.3</td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>3.1</td>
<td>5</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>3.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.4</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.6</td>
<td>13.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.6</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.9</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.7</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.4</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.4</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.8</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17.3</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals ($T$)</td>
<td>76.5</td>
<td>125.5</td>
<td>51</td>
</tr>
</tbody>
</table>

### 11.10 THREE OR MORE INDEPENDENT SAMPLE DISTRIBUTIONS: THE KRUSKAL–WALLIS TEST

**EXAMPLE POSED: IS PROSTATE-SPECIFIC ANTIGEN THE SAME AMONG BPH, CAP, AND NED PATIENTS?**

A urologist asks whether PSA level is different among three patient groups: benign prostatic hypertrophy (BPH), positive biopsy for prostate cancer (CaP), and negative biopsy and no evidence of disease (NED). The observed PSA levels in his sample are as in Table 11.35.

A quick data plot shows the distributions to be nonnormal. We want to compare the three groups with a rank-order-based test.
Methodologically, the Kruskal–Wallis test is just the rank-sum test extended to three or more samples. It is sometimes used when rank-order data arise naturally in three or more groups, or if the assumptions underlying the one-way analysis of variance test are not satisfied. The null hypothesis being tested is that the probability that a randomly chosen member from one population has a higher value than a randomly chosen member from another population is equal to 0.5. For practical purposes the user may think of it informally as testing whether the several distributions have the same median, though it is truly not a test of medians. The chi-square approximation is valid only if there are 5 or more members in each sample.

1. Name the number of samples \( k \) (3 or 4 or \ldots).
2. Name the sizes of the several samples \( n_1, n_2, \ldots, n_k \); \( n \) is the grand total.
3. Combine the data, keeping track of the sample from which each datum arose.
4. Rank the data.
5. Add the ranks from each sample separately, naming the sums \( T_1, T_2, \ldots, T_k \).
6. Calculate the Kruskal–Wallis \( H \) statistic, which is distributed as chi-square, by

\[
H = \frac{12}{n(n+1)} \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \ldots + \frac{T_k^2}{n_k} \right) - 3(n+1)
\]

(11.10)

7. Obtain the \( p \)-value from Table III (\( \chi^2 \) right tail) for \( k-1 \) degrees of freedom. (An adjustment for ties is available, but its effect is generally negligible.)

**Example Completed: Compare Prostate-Specific Antigen for Three Disease Groups**

The number of groups and numbers within groups are \( k = 3 \), \( n_1 = 6 \), \( n_2 = 8 \), \( n_3 = 8 \), and \( n = 22 \). We rank all 22 data in order, keeping track of the sample each came from, as shown in Table 11.36. The sums (totals) of ranks by group: \( T_1 = 76.5 \), \( T_2 = 125.5 \), and \( T_3 = 51 \). We calculate the Kruskal–Wallis \( H \) statistic:
Table 11.38  Performance and ranks for five types of surgical instruments.

<table>
<thead>
<tr>
<th>Instrument 1</th>
<th>Ranks 1</th>
<th>Instrument 2</th>
<th>Ranks 2</th>
<th>Instrument 3</th>
<th>Ranks 3</th>
<th>Instrument 4</th>
<th>Ranks 4</th>
<th>Instrument 5</th>
<th>Ranks 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.90</td>
<td>1</td>
<td>1.95</td>
<td>2.5</td>
<td>1.95</td>
<td>2.5</td>
<td>1.96</td>
<td>6</td>
<td>1.96</td>
<td>6</td>
</tr>
<tr>
<td>1.95</td>
<td>2.5</td>
<td>1.96</td>
<td>6</td>
<td>1.96, 1.96</td>
<td>6 × 2</td>
<td>1.96</td>
<td>6</td>
<td>1.96</td>
<td>6</td>
</tr>
<tr>
<td>1.97</td>
<td>9</td>
<td>1.98</td>
<td>12</td>
<td>1.98, 1.98</td>
<td>12 × 2</td>
<td>1.98, 1.98</td>
<td>12 × 2</td>
<td>1.98, 1.98</td>
<td>12 × 2</td>
</tr>
<tr>
<td>1.98</td>
<td>12</td>
<td>2.00</td>
<td>16.5</td>
<td>2.00</td>
<td>16.5</td>
<td>2.01</td>
<td>20.5</td>
<td>2.01</td>
<td>20.5</td>
</tr>
<tr>
<td>1.98, 1.98</td>
<td>12 × 2</td>
<td>2.01</td>
<td>20.5</td>
<td>2.01</td>
<td>20.5</td>
<td>2.03, 2.03</td>
<td>25 × 2</td>
<td>2.03, 2.03</td>
<td>25 × 2</td>
</tr>
<tr>
<td>2.00, 2.00</td>
<td>16.5 × 2</td>
<td>2.03</td>
<td>25</td>
<td>2.03</td>
<td>25</td>
<td>2.04</td>
<td>29.5</td>
<td>2.04</td>
<td>29.5</td>
</tr>
<tr>
<td>2.01</td>
<td>20.1</td>
<td>2.03</td>
<td>25</td>
<td>2.04</td>
<td>29.5</td>
<td>2.04</td>
<td>29.5</td>
<td>2.04</td>
<td>29.5</td>
</tr>
<tr>
<td>2.02</td>
<td>29.5</td>
<td>2.06</td>
<td>33.5</td>
<td>2.06</td>
<td>33.5</td>
<td>2.06</td>
<td>33.5</td>
<td>2.06</td>
<td>33.5</td>
</tr>
<tr>
<td>2.07</td>
<td>38</td>
<td>2.07</td>
<td>38</td>
<td>2.07</td>
<td>38</td>
<td>2.07</td>
<td>38</td>
<td>2.07</td>
<td>38</td>
</tr>
<tr>
<td>2.08</td>
<td>41</td>
<td>2.08</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.10</td>
<td>43</td>
<td>2.10</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.11</td>
<td>44.5</td>
<td>2.11</td>
<td>44.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.11</td>
<td>44.5</td>
<td>2.11</td>
<td>44.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T=</td>
<td>297</td>
<td>212</td>
<td>212.5</td>
<td>212.5</td>
<td></td>
<td>291</td>
<td></td>
<td>3.03</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 11.39 Sedimentation rate for three types of fluid removal.

<table>
<thead>
<tr>
<th></th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>94</td>
<td>30</td>
<td>108</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>25</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>26</td>
<td>25</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 11.40 Ranks for three skin sensitivity tests.

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

\[
H = \frac{12}{n(n+1)} \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} \right) - 3(n + 1) = \frac{12}{22(23)} \left( \frac{5852.25}{6} + \frac{15,750.25}{8} + \frac{2601}{8} \right) - 3(23) = 8.53
\]

Table 11.37 is a portion of Table III. In Table 11.37 or Table III the critical chi-square for \( k - 1 = 2 \) degrees of freedom for \( \alpha = 0.05 \) is 5.99. The \( p \)-value for \( H = 8.53 \) is calculated as 0.014. As \( H \) is greater than the critical \( \chi^2 \), and indeed the \( p \)-value is quite small, our urologist has sufficient evidence to conclude that the distributions of PSA levels are different among BPH patients, patients with biopsies positive for prostate cancer, and non-BPH patients with negative biopsies.

**ADDITIONAL EXAMPLE: PERFORMANCE OF FIVE DIFFERENT SURGICAL INSTRUMENTS**

As part of an instrument calibration, we wish to compare \( k = 5 \) disposable current-generating instruments used in surgery to stimulate (and thereby help locate) facial nerves.\(^{14}\) Among the variables recorded is current (milliamperes). \( n_1 = n_2 = n_3 = n_4 = n_5 = 10 \) readings are taken from each machine for a total of \( n = 50 \). A quick plot shows that the data are clearly not normal, so that a rank method must be used. The ordered current readings, separated by instrument from which each datum arose, and their corresponding ranks are as in Table 11.38.

We calculate the \( H \) statistic to obtain:
\[ H = \frac{12}{n(n+1)} \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \ldots + \frac{T_5^2}{n_5} \right) - 3(n+1) \]

\[ = \frac{12}{50(51)} \left( \frac{88,209}{10} + \frac{44,944}{10} + \frac{45,156.25}{10} + \frac{84,681}{10} + \frac{68,906.25}{10} \right) - 3(51) = 3.187 \]

From Table 11.37 or Table III the critical value of \( \chi^2 \) for \( \alpha = 0.05 \) with 4 df is 9.49, much larger than 3.187, so \( H_0 \) is not rejected; there is inadequate evidence to show the current to be different from instrument to instrument. Using a statistical software package, we find the \( p \)-value = 0.527.

Exercise 11.12

Are three methods of removing excess fluid in infected bursitis different? An orthopedist compares an inflow—outflow catheter (treatment 1), needle aspiration (treatment 2), and incision-and-drain surgery (treatment 3). Among his measures of effectiveness is posttreatment sedimentation rate (Sed). He uses each method on \( n_1 = n_2 = n_3 = 10 \) patients for a total of \( n = 30 \). Data are as in Table 11.39.

Draw rough frequency plots of Sed for groups of 10 to show that they are not normal and that a rank method is appropriate. Use the Kruskal—Wallis method to test the null hypothesis that Sed is the same for all treatments.

11.11 THREE OR MORE MATCHED SAMPLE DISTRIBUTIONS: THE FRIEDMAN TEST

EXAMPLE POSED: TWO SKIN TESTS FOR ALLERGEN SENSITIVITY

An allergist applies three prick skin tests (two competing stimuli and a control) to the inside forearms, randomizing the order proximal to distal, to each of eight patients. After 15 minutes, he ranks the wheal-and-erythema reactions for each patient by severity (most severe being 1). The results (ranks) were as in Table 11.40. Are the distributions of the outcomes different across the three tests?

METHOD: THE FRIEDMAN TEST

If three or more treatments are given to each patient of a sample, we have an extension of the paired-data concept. More exactly, it is called a randomized block design. In the example a dermatologist applied three skin patches to each of eight patients to test for an allergy. The three skin test results for each patient is called a “block.” (In contrast, three groups of patients with a different skin test being used on each would call for the Kruskal—Wallis test.) The hypothesis being tested is that the several treatments have the
same distributions. The chi-square approximation is valid only if there are five or more blocks, for example, patients, in the sample.

1. Name the number of treatments (3 or 4 or ...) \( k \) and blocks (e.g., patients) \( n \).
2. Rank the data within each block (e.g., rank the treatments for each patient).
3. Add the ranks for each treatment separately; name the sums \( T_1, T_2, \ldots, T_k \).
4. Calculate the Friedman \( F_r \) statistic, which is distributed as chi-square, by

\[
F_r = 12 \frac{(T_1^2 + T_2^2 + \ldots + T_k^2) - 3n(k + 1)}{nk(k+1)}.
\]  

5. Obtain the \( p \)-value from Table III (\( \chi^2 \) right tail) for \( k - 1 \) df.

EXAMPLE COMPLETED: TWO SKIN TESTS FOR ALLERGEN SENSITIVITY

The number of treatments is \( k = 3 \) and number of blocks is \( n = 8 \). The rank sums for the three treatments are 12, 15, and 21. We calculate the Friedman statistic:
In Table 11.37 or Table III the critical value of $\chi^2$ for $k - 1 = 3$ degrees of freedom for $\alpha = 0.05$ is 7.81. As $F_r$ is much greater than 7.81, so the null hypothesis of equal distributions over time was rejected. The $p$-value calculated from a statistical software package is 0.002.

### ADDITIONAL EXAMPLE: LEVEL OF GENTAMICIN TREATMENT OVER TIME

In treating an infected ear, gentamicin, suspended in a fibrin glue, can be inserted in the middle of the ear, and some makes its way into the system. The time after application that it resides in the system is unknown. $n = 8$ chinchillas, which have large ears anatomically similar to humans, were used. The serum gentamicin level (microgram/milliliter) was measured at 8, 24, 72 hours, and 7 days after the administration. If the declining levels at these times test significantly different, an approximate “fade-out” time can be inferred. A quick plot showed the data per time grouping were far from normal; a rank method must be used. As the data arose through time from the same animal, Friedman’s test was appropriate. The data, listed and then ranked for each animal, were as in Table 11.41.

The results were substituted in Eq. (11.11) to obtain

$$F_r = \frac{12}{nk(k + 1)} \left( T_1^2 + T_2^2 + \ldots + T_k^2 \right) - 3n(k + 1) = \frac{12}{8(3)(4)} \left( 144 + 225 + 441 \right) - 3(8)(4) = 5.25.$$
Do posttherapy CaP patients’ PSAs remain stable? \( n = 9 \) cancer-of-the-prostate (CaP) patients had been without clinical evidence of disease 10 years after a negative staging pelvic lymphadenectomy and definitive radiation therapy.\(^{18}\) PSA levels were then measured in three successive intervals about a year apart. Data were as in Table 11.42. The data distributions are not normal; a rank test is appropriate. Test the hypothesis that distribution of PSA is not changing through time.

### 11.12 Three or More Ranked Independent Samples With Ranked Outcomes: Cusick’s \textit{nptrend} Test

This test is not used a great deal in medical applications, but, when data are appropriate for it, there is no other option. It is a design in which three or more rank-ordered unmatched independent variables give rise to three or more categorical or rank-ordered outcomes.

**Example Posed: Location of Pediatric Snake Bite by Age**

In DB24, 202 pediatric snake bites in the San Diego region were tabulated according to location of bite [1 (hand/arm), 2 (face), or 3 (foot/leg)] by age group [1–3 years (toddler), 4–9 (child), 10–12 (tween), and 13–18 (teen)]. The results were as in Table 11.43, a \( 4 \times 3 \) contingency table, as number of occurrences with percent in parentheses. The rate of foot/leg bites is largest for toddlers (stepping on a snake) and grows smaller with age. By teenage, the rate of hand/arm (picking up a snake) is largest. We ask if this pattern is greater than would occur by chance if there were no association between bite location and age.

(The locations are ranked according to the following observations. A hand/arm bite arises from trying to pick up a snake, an intended, very attentive act. A face bite arises from tripping and falling with the face near a snake or from climbing past a snake-filled crevice in rocks; a rare bad luck event unrelated to attention. A foot/leg bite arises from stepping on/near a snake, a careless, inattentive act. Thus the three types of bite are ordered or ranked as attentive, neutral, or inattentive.)

In this case, both age groups and location groups are rank-ordered. The ranking in this method is bothersome, so that the computation is better done by computer software. Not all packages offer \textit{nptrend}; Stata is one that does.

**Method: The \textit{nptrend} Test**

The data table is arranged with the independent variable (input values) as rows and the dependent variable (outcome values) as columns. Let \( i \) denote the row numbers, \( i = 1, 2, \ldots \) (Continued)
(CONTINUED)

..., and let \( j \) denote the column numbers, \( j = 1, \ldots, k \). Denote table entries as \( n_{ij} \), the number of occurrences for row \( i \) and column \( j \). Denote the row sums as \( n_i \), the column sums as \( n_j \), and the total number as \( n \).

An unusual component of the calculation is the method of assigning ranks. The first \( n_1 \) values carry the first rank, so the assigned rank is the average number, a short cut calculation for which is \((1 + n_1)/2\). The 2nd column of \( n_2 \) values take on the succeeding \( n_2 \) ranks, with average \( n_1 + (1 + n_2)/2 \). The third column’s average rank is \( n_1 + n_2 + (1 + n_3)/2 \). And so forth.

We calculate the sum of ranks in the rows, \( R_i \), as the number in each cell times the mean rank for that column, added along the row. Finally, we calculate two more values, \( L \) and \( T \). \( L \) is the sum over \( i \) (input groups) of number indicator \( \times \) numbers in groups, that is \( L = \sum(i \times n_i) \). \( T \) is the sum over \( i \) (input groups) of number indicators \( \times \) rank sums, that is, \( T = \sum(i \times R_i) \). Under a null hypothesis of homogenous distributions, \( T \) is distributed as normal with mean

\[
m_T = \left( \frac{L}{2} \right) (n + 1)
\]

and standard error

\[
s_T = \sqrt{\frac{n + 1}{12} (n \sum n_i^2 - L^2)},
\]

so that

\[
z = \frac{(T - m_T)}{s_T}.
\]

An adjustment of \( s_T \) is available for ties, but its effect is small. In the event that this adjustment is wanted, it can be calculated as follows. The number of unique values of the outcome variable (the number of data columns) was defined as \( k \). The number of times the \( j \)th such value is repeated was defined as \( n_j \). An adjustment value \( a \) for ties is given by

\[
a = \frac{\sum_{j=1}^{k} n_j(n_j^2 - 1)}{n(n^2 - 1)},
\]

and \( s_T(\text{adj}) = \sqrt{1 - a \times s_T} \) is used as the divisor in Eq. (11.14).

EXAMPLE COMPLETED: LOCATION OF PEDIATRIC SNAKE BITE BY AGE

From the example’s table, age groups are ranked \( i = 1, 2, 3, 4 \); location groups are ranked \( j = 1, 2, 3 \) (\( k = 3 \)); the age group (independent variable) numbers \( n_i \) are 42, 66, 36, and 58; the outcome location group (dependent variable) numbers \( n_j \) are 92, 3, and 107; and the total number \( n = 202 \). The first outcome (hand/arm) carries the first \( n_1 \) ranks 1 through 92, giving a mean of \((1 + 92)/2 = 46.5 \). The second outcome
11.12 Three or More Ranked Independent Samples With Ranked Outcomes

Table 11.44 Dose response ranks in migraine treatment.

<table>
<thead>
<tr>
<th>Patient number</th>
<th>Low dose</th>
<th>Medium dose</th>
<th>High dose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Sums of ranks: $R_3 = 21$, $R_2 = 15$, $R_1 = 12$

(face) carries the next $n_2$ ranks 93, 94, 95, average $n_1 + (1 + n_2)/2 = 92 + 4/2 = 94$. The third outcome (foot/leg) carries an average rank $n_1 + n_2 + (1 + n_3)/2 = 92 + 3 + (1 + 107)/2 = 149$.

We calculate the sum of ranks in the rows, $R_i$, as the number in each cell times the mean rank for that column, added along the row. For toddlers, $R_1 = 16 \times 46.5 + 0 \times 94 + 26 \times 149 = 4618$. For child, $R_2 = 26 \times 46.5 + 2 \times 94 + 38 \times 149 = 7059$. Similarly, $R_3 = 3826.5$ and $R_4 = 4999.5$. $L = 1 \times 42 + 2 \times 66 + 3 \times 36 + 4 \times 58 = 514$. $T = 1 \times 4618 + 2 \times 7059 + 3 \times 3826.5 + 4 \times 4999.5 = 50,213.5$.

By substituting in Eqs. (11.12)–(11.14), we find $m_T = 52,171$, $s_T = 924.46$, and $z = -2.12$, yielding $p$-value = 0.034, significant. However, there exist a great many ties. Let us correct for ties. $k = 3$, $n_1 = 92$, $n_2 = 3$, $n_3 = 107$, and $n = 202$. By substituting in Eq. (11.15), we find $a = 0.24308$. $s_T(adj) = 804.29$, $z$ becomes $-2.4338$, and $p$-value = 0.015, significant. The risk of a false-positive is even lower than without the adjustment.

Comparison with a contingency test

Because the age grouping is rank-ordered rather than purely categorical, a Fisher’s exact test (FET) of the contingency table is not appropriate. Indeed, the FET’s $p$-value = 0.097, showing no significance. We see that the ranking of age groups provided additional sensitivity to the test, permitting the significance to be seen using nptrend when it was not perceptible using FET.

Exercise 11.14

Overnight recordings of snoring loudness by 3776 sleep disturbed patients were grouped by BMI into normal ($\leq 25$), overweight ($> 25$ but $\leq 30$), and obese ($> 30$) (DB27). The groups were designated 1, 2, and 3, containing 363, 1106, and 2307 patients, respectively. The data were ranked by average dB when snoring exceeded 50 dB. Calculations using Eqs. (11.13)–(11.15) yields $z = 18.89$. Interpret the result.
CHAPTER 11 Tests of location with continuous outcomes

11.13 THREE OR MORE RANKED MATCHED SAMPLES WITH RANKED OUTCOMES: PAGE’S L TEST

This test is not used a great deal in medical applications, but, when data are appropriate for it, there is no other option. It is a design in which three or more rank-ordered matched (paired) independent variables give rise to three or more categorical or rank-ordered outcomes.

EXAMPLE POSED: TREATING MIGRAINE-CAUSED NAUSEA
We assess levels of an antiemetic in treating migraine-caused nausea. We randomize three levels of dose for eight chronic migraine patients in treating different episodes. After the three episodes, each patient is asked to rank the relative effectiveness of the three treatments as 1 (the most effective treatment) to 3 (the least effective). Thus we have three matched groups of dosages with three ranked outcomes. We ask if effectiveness is associated with dose level. Our data are as shown in Table 11.44, with effectiveness rank under each dose level.

METHOD: PAGE’S L TEST
Let \( n \) denote the number of subjects (\( n = 8 \) patients in the example) and \( k \) denote the number of assessments (outcomes) for each subject (\( k = 3 \) in the example). Place subjects in columns and assessments in rows. If the outcomes for each patient are measures or ratings, replace them by their ranks, highest value ranked first, next highest ranked second, and so on. (In the example, we started with ranks, so this is not required.) Add the rank values per column to find the rank sums \( R_i \), \( i = 1, 2, \ldots, k \). The rank sums are given weights \( (w_i) \) as their ordinal numbers, that is, the weight for the first rank is \( w_1 = 1 \), for the second rank is \( w_2 = 2 \), and so on. Compute \( L = w_1R_1 + w_2R_2 + \cdots + w_kR_k \). \( \chi^2 \) with 1 df is given by

\[
\chi^2_{1 df} = \frac{[12L - 3nk(k + 1)]^2}{nk^2(k^2 - 1)(k + 1)}.
\] (11.16)

The method was developed for a two-sided test, so that the critical values must be found for 0.025 in the left tail from Table IV and 0.025 in the right tail from Table III. A \( \chi^2 \) value \(<0.001 \) or \( >5.02 \) indicates a resulting p-value \(<0.05 \).

---

Table 11.45 Ratings and ranks of cancer severity.

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Ratings 1</th>
<th>Ratings 2</th>
<th>Ratings 3</th>
<th>Ranks 1</th>
<th>Ranks 2</th>
<th>Ranks 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1.5</td>
<td>1.5</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Sums of ranks -- -- -- 16.5 22 21.5
Example completed: Migraine-caused nausea

\[ L = 1 \times R_1 + 2 \times R_2 + 3 \times R_3 = 12 + 30 + 63 = 105. \]

By substituting in Eq. (11.16), we find

\[ \chi^2 = \left[ \frac{12(105) - 3(8)(3)(4^2)}{8(3^2)(3^2 - 1)(3 + 1)} \right]^2 = 5.06. \]

5.06 > 5.02, so our result is significant at the (two-tailed) 0.05 level. A computer calculates 0.024 under the curve in the right tail, leading to \( p \)-value = 0.048.

Page’s \( L \) compared to some tests with similar appearing formats

To appreciate the subtlety of this application, let us contrast the design with some similar designs. If we had three different groups of patients in the example, one for each dose level, we would use the \textit{nptrend} test. If we had three different medications, we would use the Kruskal–Wallis test. If we had our group set up in the original matched fashion but used three different medications on the same patients, we would use the Friedman test.

Exercise 11.15

\textbf{Is experience associated with the rating of severity of cancer by radiologists reading radiographs?} Three radiologists – a PGY1 resident (1), a PGY4 resident (2), and a senior staff member (3) – read radiographs of 10 cancer patients and rate the severity of disease on a 1–5 scale. For each patient the ratings are ranked, rank 1 corresponding to most severe rating. The rank sums are given in Table 11.45. Does detection of severity increase with experience? The data are illustrated in Table 11.45. Identify \( n \) and \( k \). Calculate \( L \) and \( \chi^2 \). Use Table III to make a conclusion and state that conclusion.

11.14 Potential Drawbacks to Using Nonparametric Tests

It was previously noted that nonparametric tests are a popular choice when sample sizes are too small to assess assumptions or when the observed data do not adhere to the assumptions underlying the theory giving rise to parametric tests. In this case the power of nonparametric tests for being able to discriminate between distributions can be greater than that of parametric tests. As with anything in life, however, there are no free lunches and, in this case, there are potential drawbacks to using nonparametric tests. First, because the magnitude of measurements is ignored when focusing solely on the rank or sign of data, statistical power can be decreased relative to parametric tests like the \( t \) test. Perhaps more serious, from a scientific perspective the units of measurement of outcomes are generally meaningful and hence the magnitude of differences in locations of distributions between comparison groups is also of interest. As an example, if we wish to compare whether a new treatment is more effective in
reducing blood pressure than an existing treatment, we would generally seek to know not just if blood pressure tends to be lower on the new treatment but also by how much on average. A rank-based test would not test for differences in the mean, whereas the \( t \) test would specifically estimate and test for differences in mean blood pressure. This leads to a much more precise, and clinically interpretable result.

**Nontransitivity of the rank-sum test**

The rank-sum statistic that was introduced in Section 11.8 is commonly referred to and thought of as a test of medians between distributions. However, as we have noted, in general the rank-sum test does not test equality of medians between distributions. Instead, given two populations, the hypothesis being tested is whether or not the value for a randomly chosen member of the first population is probably smaller than one of the second population. Notationally, if \( X \) represents a randomly drawn observation from population 1 and \( Y \) represents a randomly drawn observation from population 2, then rank-sum tests the null hypothesis \( H_0: \Pr(X > Y) = 0.5 \). This is not a test of the medians and is in fact not a test of differences in any parameter characterizing the distributions of each population. This rather unintuitive null hypothesis can lead to circular conclusions from use of the rank-sum test. That is, the test does not necessarily yield a well-defined ordering of distributions. More plainly, if we wish to compare observations from three populations (e.g., three treatments), we can conclude that observations from population 1 tend to be larger than those from population 2, observations from population 2 tend to be larger than those from population 3, but that observations from population 3 tend to be larger than those from population 1. This phenomenon is known as *nontransitivity*, and because of this we refer to the rank-sum test as a *nontransitive test*. This is clearly problematic in terms of making scientific conclusions, and such circularity cannot occur if one focuses on testing differences in parameters such as the \( t \) test for differences in population means.

**EXAMPLE: NONTRANSITIVITY OF THE RANK-SUM TEST**

Consider a somewhat hypothetical setting where we are interested in treating the underlying cause of a symptom (e.g., nasal congestion) that may have arisen from one of three (unknown) different illnesses (say illnesses 1, 2, and 3). Further, suppose that if

<table>
<thead>
<tr>
<th>Illness</th>
<th>Proportion (%)</th>
<th>Untreated</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
left untreated the symptom would resolve on its own in 3 days (regardless of the underlying illness) and that we wish to compare the efficacy of two treatments, A and B, relative to leaving patients untreated. Suppose the true recovery times for patients are given in Table 11.46. Thus if a patient suffers from illness 1, the recovery time is 2 days if the patient receives treatment A and is instant (0 days) if the patient receives treatment B. Also listed in the table are the proportions of subjects suffering from illnesses 1, 2, and 3. Consider the resulting ordering that would arise from the rank-sum test. Here we would find that

1. treatment A is preferable to no treatment since 60% of people do better on A (A outperforms no treatment in those with illnesses 1 and 2),
2. treatment B is preferable to A, since 80% of people do better on treatment B (B outperforms A in those with illnesses 1 and 3), and
3. leaving the patients untreated is better than treatment B 60% of the time (untreated outperforms B in those with illnesses 2 and 3).

Thus the result we would find from the rank-sum test is nontransitive, resulting in a circular conclusion. However, if we were concerned with the mean time to recovering in the population we would prefer treatment B since the population mean recovery times would be B (2.4 days) < untreated (3 days) < A (3.2 days). There is a clear and well-defined ordering in this case.

Choosing the test to use

Where does that leave us when we have to choose between a rank and a parametric test? There is no formula for decision. If the distributions are well behaved, it does not make a lot of difference. But “well behaved” is difficult if even possible to define. It lies more in the art of statistics than the science. We can look at the behavior of data through a histogram, a box plot, the contrast between the mean and the median, and many other ways. We suggest the following logic. Begin by considering if there is a parameter of inherent scientific importance to compare between groups or if interest truly lies in just seeing if one group tends to have larger values than another. Many times, we wish to compare means between groups because they represent the center of mass of distributions and are interpretable. If more than half the values have the same value (often 0), the median tells us very little and we get a better picture of typical from the mean. In this case, parametric tests (preferably those with few assumptions like the two sample t test with unequal variances) that focus on the mean should be used. Next, consider if the distributions are likely to be well behaved. If so, the use of parametric methods is recommended. If a distribution has extreme outliers or other deformities that render the mean nonrepresentative of the “typical,” use of rank methods may be considered.
CHAPTER 11 Tests of location with continuous outcomes

REFERENCE


12.1 CONCEPTS AND TERMS

Concept
Suppose, you have a finger blood pressure sensor that you believe will measure blood pressure as well as a cuff. You want statistical evidence that the measure of diastolic blood pressure (DBP) from the finger device is not different from that of the cuff. You record the difference between the two measures on each of a number of subjects and test the hypothesis that the finger device is not inferior. How is inferior defined? In the light of measuring device variability and patient variability over time, you judge that a measurement in DBP of ± 5 mmHg has no clinical implication; the mean of DBP deviations of the finger device from the cuff device must be more than 5 mmHg to conclude that it is inferior. This sounds somewhat similar to the means of testing met in Chapter 11, Tests of location with continuous outcomes, but there are important differences.

The testing methods examined in previous chapters of this book are difference tests, in which we pose a statistical null hypothesis of no difference, say, of mean DBP deviations between the two devices from 0. We perform a test to see if there is sufficient evidence to reject the null hypothesis and assert that the mean pressures are different. In contrast, it is becoming ever more common to ask if a new treatment is as good, on average, as an established one, perhaps because it is less invasive or less costly. In this case the null hypothesis might state: The effectiveness of a new treatment is outside a magnitude judged to represent clinical similarity. The alternative hypothesis states: The mean difference is within the clinically relevant magnitude, and thus the treatments are deemed similar enough to be used interchangeably in clinical practice. Rejecting the null hypothesis implies evidence that the treatments are not appreciably different. The relatively recent development in statistical methodology of testing for no difference has been termed equivalence testing.

Equivalence tests are similarity tests
Equivalence tests might better be termed similarity tests, because they test whether or not means are approximately the same, that is, similar enough to be not clinically different, rather than exactly the same.
Equivalence versus noninferiority (or nonsuperiority)

We could ask if one treatment is not appreciably worse than another (noninferiority), leading to a one-sided test, or if neither treatment is appreciably better than the other (equivalence), leading to a two-sided test. The one-sided tests also may be referred to as nonsuperiority tests that address the question of one treatment not being appreciably better than another. The distinction lies in which mean is subtracted from the other. The two names refer to the same test. The name noninferiority is used primarily in this book.

Equivalence of means, proportions, and other parameters

This chapter is devoted to tests of one or two means or proportions. (Indeed, the conceptual paragraphs focus on means for simplicity.) Because most equivalence tests are conducted on means and proportions, these are the tests for which the concepts have been most thoroughly developed. Equivalence concepts apply in a number of other methods that involve testing—analysis of variance, regression, rank-order tests, and others. Methods for many of these cases have not been detailed in the literature. However, the logic is similar. In some cases the enterprising reader may be able to extend the ideas sufficiently to perform appropriate equivalence testing.

12.2 BASICS UNDERLYING EQUIVALENCE TESTING

The null hypothesis is the key

In both the difference and equivalence tests, we are assessing a difference between population means; the distinction lies in the null hypotheses. In a difference test, we hypothesize no difference. In an equivalence test, we hypothesize a difference of at least some clinically relevant amount, say, $\Delta$. In the DBP example of the first paragraph in this chapter, a difference had to be greater than 5 mmHg to be clinically relevant. In some studies a value for $\Delta$ will be obvious from clinical considerations; in other studies the size of $\Delta$ will not be clear at all and must be almost arbitrary. In the latter case an arbitrary 20% of baseline is often chosen. $\Delta$ is the value we test against. The test statistic consists of the sample difference, say, $d$, plus or minus the allowable number of standard errors, calculated using $\alpha$. In testing paired data, for example, the change from before to after treatment, $d$, might be taken as the mean of this change, $md$. In the finger blood pressure example, $d$ is the difference between the DBP by cuff and the DBP from the finger sensor.

Difference symbols

The word difference has been used so much that it might be confusing. Let us be clear in distinguishing the three difference terms in equivalence testing: (1) there exists a
true but unknown theoretical (or population) difference between two outcomes; we denote this difference as $\delta$. This would be the measured difference if we could measure the entire population. (2) From our data, we can calculate an estimate of $\delta$, the sample difference; we denoted this difference as $d$. This is the measured difference that we actually use. (3) The null hypothesis states that the inferiority (or inequality) of the one outcome compared with the other is at least a certain difference $\Delta$, our clinically relevant difference. This is the cut point differentiating similarity (equivalence) from dissimilarity.

**What the null hypothesis statement really means**

The null hypothesis should be stated as an equality rather than an inequality, because we attach a test of a specific probability parameter to it. In a noninferiority test, for example, we want to test the null hypothesis that the theoretical difference $\delta$ is at most equal to (i.e., equal to or less than) the clinically relevant difference $\Delta$. However, we state a null hypothesis of this sort: $H_0: \delta \leq \Delta$. Rejecting $H_0$ implies that $\delta$ is not greater than $\Delta$ and therefore $\delta \leq \Delta$.

**Assumptions**

We should note that the underlying assumptions made for difference tests—identically and independently distributed data for all tests, normal distributions for $t$ tests, and so forth—also apply to equivalence tests.

**One-sided versus two-sided tests**

A one-sided (noninferiority) test is rather straightforward. We postulate $\Delta$, look up a critical value to demark difference from no difference, and compare $d$ with $\Delta$ to conclude a relationship between $\delta$ and $\Delta$. A two-sided test (equivalence) is trickier. Whereas in difference testing, the null hypothesis (the hypothesis we test) is one-tailed and the alternative is two-tailed, in equivalence testing, the reverse is true: it is the null hypothesis that is two-tailed. We hypothesize that the theoretical difference $\delta$ either is $\Delta$ or more, or alternatively is $-\Delta$ or less. The logically consistent approach to this problem has evolved to be the use of two tests, one for each tail. This approach has been termed two one-sided tests, or TOST (sometimes Schuirmann’s TOST). TOST is further discussed in Section 12.5 after more groundwork has been laid.

**Using a confidence interval instead of a test statistic**

Other chapters have stated that confidence intervals use much the same calculations and provide much the same information as hypothesis tests, but the purpose is different; confidence intervals convey the precision of our estimate and tests provide a binary
decision regarding a specified hypothesis. That being said, many investigators find it easier to use confidence intervals than hypothesis tests in equivalence testing. We can place a confidence interval on \( \delta \) using \( d, \sigma_d \) (or \( s_d \)), and the appropriate statistic, such as \( z_{1-\alpha} \) (or \( t_{1-\alpha} \)). If this confidence interval fails to enclose \( \Delta \), we have evidence to reject \( H_0 \) and to conclude equivalence.

To illustrate the use of confidence intervals in decision making, Fig. 12.1 provides the evidence for different scientific decisions under five different scenarios. The figure depicts the difference in means. The vertical dotted lines represent clinically important improvements in the mean (here denoted with values on the left as might be the case for drugs designed to increase cognitive test response scores), no difference in the means, and clinically important harmful effects in the mean. For each of the five hypothetical trial results, we depict the observed difference, \( d \), and corresponding \( (1 - \alpha) \times 100\% \) confidence interval.

For a superiority trial the goal would be to show that the lower bound of the confidence interval rules out no difference. Thus a decision of superiority would result in setting A.

In the case of an equivalence study, the dashed lines for important benefit and harm would represent the equivalence range. To establish equivalence between new treatments, one would seek to rule out large differences (both positive and negative) between the treatments with high confidence. This would be the case if the lower bound of the confidence interval is above the important harm boundary and the upper bound of the confidence interval is below the important benefit boundary. Setting C illustrates the setting where equivalence would be established. In the case of a

![Figure 12.1 Confidence intervals in relation to decisions.](image)
noninferiority trial, the goal would be to rule out important harm. This value is
known as the noninferiority margin and is depicted by the vertical dashed line indicat-
ing clinically important harm. A decision for noninferiority would be supported if the
lower bound of the confidence interval is above the noninferiority margin. Settings
A–C would all meet this criterion and would result in a decision of noninferiority.
Settings D and E would fail to establish noninferiority. In fact, setting E illustrates
the case where harm with respect to the new experimental treatment is established,
since the upper bound of the confidence interval lies below no difference, thus ruling
out benefit.

12.3 CHOOSING A NONINFERIORITY OR EQUIVALENCE
MARGIN

In typical settings, where one wishes to establish superiority, the choice of the
null hypothesis is easy, since the null hypothesis represents no difference in
means. For noninferiority and equivalence trials, specification of the noninferior-
ity and equivalence margin is far from trivial. In the case of a noninferiority
design, the noninferiority margin represents how much worse would be deemed
acceptable for the new treatment relative to the active control if the new treat-
ment were adopted into practice. This requires one to specify the difference in
means that would represent important clinical harm. While this can obviously be
a subjective decision, empirical information regarding the efficacy of the active
control should be incorporated. For example, suppose that there is an existing
approved therapy (call it T1) to lower blood pressure and the therapy was
approved on the basis of the results of a clinical trial showing that when com-
pared to placebo the therapy reduced systolic blood pressure by 20 mmHg with a
95% confidence interval equals to 14 and 26 mmHg. If a new therapy (call it T2)
were then to be tested against the active control T1 using a noninferiority mar-
gin, then the largest noninferiority margin we should reasonably entertain would
be a difference of 20 mmHg. The reason for this is because if treatment T1
reduced systolic blood pressure more than 20 mmHg relative to treatment T2,
then based upon our historical evidence this would mean that treatment T2 may
likely be even worse than placebo (since T1 was 20 mmHg better than placebo).
Clearly, it would then be unethical to approve treatment T2 in this case and
hence we should not allow for a noninferiority margin that is greater than the
total effect of T1 relative to placebo based upon historical studies.

While we do not entertain noninferiority margins greater than the total effect
that has been observed for the active control, we are generally more conservative
than this. That is, we generally prefer to have a noninferiority margin less than this total effect. The reason for this is that we do not know the true total effect of the active control. Instead, we simply have an estimate of what it is based upon what we observed in past studies. In the case of the blood pressure example, we do not know that the true difference between treatment T1 and placebo is 20 mmHg. There is variability associated with this estimate of effect. The question is: How can we reasonably account for this variation to make sure that if noninferiority is established for treatment T2, treatment T2 will still be better than placebo? (We cannot make such a comparison because it would be unethical to randomize people to placebo, since we have already established that T1 works.) The answer is that we can use our estimates of variability from the prior study (or studies) of T1 versus placebo. One way of doing this is to use the confidence interval that was computed from the previous study. In the case of the blood pressure example, one reasonable approach would be to use the lower bound of the confidence interval from the prior study, 14 mmHg. In some sense, this value represents the “worst case” scenario for the total effect of treatment T1 versus placebo. Thus if we were no worse than this value when comparing T2 to T1, then we would be confident that T2 would be better than placebo.

The same principles apply when considering equivalence margins. The only difference is that in this case one would use both the upper and lower bounds of the confidence interval from the historical trial.

When data from more than one trial comparing the active control to placebo are available, these data are often combined via a metaanalysis (see Chapter 24: Metaanalyses) to obtain an “average” estimate of the total effect of the active control over all of the studies along with a confidence interval that incorporates variability of the estimate, both within and between studies. In this case the noninferiority margin should be no larger than this average effect and a reasonable conservative approach would be the bounds of the confidence intervals for this average effect.

### 12.4 METHODS FOR NONINFERIORITY TESTING

**EXAMPLE POSED: EFFECT OF SILICONE IMPLANTS ON PLASMA SILICON LEVELS**

DB5 gives data on the effect of silicone implants on plasma silicon levels. If postoperative plasma silicon level is not greater than the preoperative level, the implants had no deleterious effect on this measure, so we want a noninferiority test on the mean difference between preoperative and postoperative readings. We theorize that silicone implants cannot decrease plasma silicon levels, and therefore we have a one-sided test;
if postoperative plasma silicon levels should show a decrease in our sample, it would
be a random, not caused, effect. We judge that anything less than a 20% increase in
postoperative mean is not clinically relevant. The preoperative mean is 0.2257 μg/g
(dry weight). Thus 0.20 \times 0.2257 = 0.0451 μg/g increase is our Δ. The null hypo-
thesis is \( H_0: \delta = 0.0451 \). The alternative hypothesis is \( H_1: \delta < 0.0451 \) that equates in
medical practice to no clinically relevant difference.

**METHOD FOR NONINFERIORITY TESTING**

**Hypotheses**

We want a sample difference \( d \), estimating the population difference \( \delta \). For a single sample
mean \( m \) versus a theoretical mean \( \mu \), \( d = m - \mu \). For a change time-to-time, treatment-to-
treatment, and so on, \( d = \text{mean change} \). For a single-sample proportion \( p \) versus a theoreti-
cal proportion \( \pi \), \( d = p - \pi \). For two-sample means \( m_1 \) versus \( m_2 \), \( d = m_1 - m_2 \). For two-
sample proportions \( p_1 \) versus \( p_2 \), \( d = p_1 - p_2 \). (Changing the sign of \( d \) will toggle between
nonsuperiority and noninferiority.) To show evidence that the difference \( \delta \) is less than a clin-
ically important amount \( \Delta \), we want to reject a null hypothesis that claims \( \delta \) is \( \Delta \) or more.
However, we have to calculate a probability based on an equality; we cannot calculate
the probability of an unspecified “or more.” Thus the hypothesis is posed statistically as
\( H_0: \delta = \Delta \), implying that if we reject \( \delta \) being as large as \( \Delta \), we certainly reject it being larger.
The alternative hypothesis is \( H_1: \delta < \Delta \). We choose \( \alpha \), the risk for error if we reject the null-
hypothesis when it is true.

**Test Statistics**

We want to test if \( d \) is more than a critical value distance from \( \Delta \) in standardized
units. We standardize by dividing \( \Delta - d \) by the standard error of \( d \), \( s_d \). The test sta-

tistics are

\[
z = \frac{\Delta - d}{\sigma_d} \tag{12.1}
\]

or

\[
t = \frac{\Delta - d}{s_d}, \tag{12.2}
\]

according to standard deviation known or estimated, respectively. Formulas for the standard
error \( s_d \) and the names of the probability table to be used to find the critical value \( z_{1-\alpha} \) or
\( t_{1-\alpha} \) may be found in Table 12.1. For one-sample tests, rows 1 and 2 correspond to
Eq. (12.1) and row 3, to Eq. (12.2). For two-sample means or proportion tests, rows 4, 5, and
6 correspond to Eq. (12.1) and row 7, to Eq. (12.2). If \( z \geq z_{1-\alpha} \) or \( t \geq t_{1-\alpha} \), reject \( H_0 \). We have
evidence of noninferiority.
STEPS IN NONINFERIORITY TESTING

Following are the steps in noninferiority testing:

1. Pose a difference $\Delta$ such that greater than this difference is clinically important and less than this difference is clinically indifferent.

2. State null hypothesis $H_0$: $\delta = \Delta$; the theoretical difference is $\Delta$. (By implication the same conclusion will be reached if $\delta > \Delta$.) State an alternative hypothesis $H_1$: $\delta < \Delta$; the theoretical difference is less than clinically important.

3. Choose $\alpha$.

4. Look up the critical values $z_{1-\alpha}$ or $t_{1-\alpha}$, in the appropriate table for the chosen $\alpha$.

5. Calculate the appropriate statistic from Eq. (12.1) or (12.2) and Table 12.1.

6. If the test statistic is as large as the critical value, reject $H_0$ that implies accepting noninferiority.

Using a Confidence Interval
For known $\sigma$, $H_0$ is rejected if $(\Delta - d)/\sigma_d \geq z_{1-\alpha}$, Simple algebra can put the statement into the following form: $H_0$ is rejected if $\Delta \geq d + z_{1-\alpha}\sigma_d$, the interior of a one-sided confidence interval probability statement. For $s$ estimated from the sample, a similar statement occurs: $H_0$ is rejected if $\Delta \geq d + t_{1-\alpha}s_d$.

Table 12.1 Test cases and standard deviations with corresponding standard errors and probability tables from which to find critical values.$^{a,b}$

<table>
<thead>
<tr>
<th>Case</th>
<th>Standard deviation(s)</th>
<th>$sd$ (Standard error)</th>
<th>Probability table$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = m - \mu$</td>
<td>$\sigma$</td>
<td>$\sigma/\sqrt{n}$</td>
<td>$z$ Table (Table I)</td>
</tr>
<tr>
<td>$d = p - \pi$</td>
<td>Assumed$^d$</td>
<td>$\sqrt{\frac{n(1 - \pi)}{\pi}}$</td>
<td>$z$ Table (Table I)</td>
</tr>
<tr>
<td>$d = m_1 - m_2$</td>
<td>$\sigma_1, \sigma_2$</td>
<td>$\frac{\sigma}{\sqrt{n}}$</td>
<td>$t$ Table (Table II)</td>
</tr>
<tr>
<td>$d = m_1 - m_2$</td>
<td>$\sigma_1 = \sigma_2 = \sigma$</td>
<td>$\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$</td>
<td>$z$ Table (Table I)</td>
</tr>
<tr>
<td>$d = m_1 - m_2$</td>
<td>$\sigma_1 \neq \sigma_2$</td>
<td>$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$</td>
<td>$z$ Table (Table I)</td>
</tr>
<tr>
<td>$d = p_1 - p_2$</td>
<td>Assumed$^d$</td>
<td>$\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left((n_1p_1 + n_2p_2)/(n_1 + n_2)\right)} \left(1 - \frac{n_1p_1 + n_2p_2}{n_1 + n_2}\right)$</td>
<td>$z$ Table (Table I)</td>
</tr>
<tr>
<td>$d = m_1 - m_2$</td>
<td>$s_1, s_2$</td>
<td>$\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/\left(n_1 + n_2 - 2\right)\right]}$</td>
<td>$t$ Table (Table II)</td>
</tr>
</tbody>
</table>

$^a$Subscripts correspond to sample number.

$^b$Expressions for proportions assume moderate to large sample sizes.

$^c$See Tables of Probability Distributions.

$^d$Standard deviations for proportions are assumed from binomial approximation.
EXAMPLE COMPLETED: SILICONE IMPLANTS AND PLASMA SILICON LEVELS
We calculate postoperative minus preoperative differences from the data, plot frequencies, and see that the distribution is not far from normal, and then calculate the mean difference as $d = -0.0073$, $s = 0.1222$, and $n = 30$; therefore $s_d = 0.0223$. From Table II (see Tables of Probability Distributions), one-tailed $t_{0.95}$ for 29 df is 1.699. Using Eq. (12.2), we find that $(\Delta - d)/s_d = [0.0451 - (-0.0073)]/0.0223 = 2.3498$ that is larger than 1.699. We have evidence to reject $H_0$. The plasma silicon level after implantation has been shown to be clinically equivalent to the preoperative levels. Computer software shows the actual $p$-value $= 0.013$.

ADDITIONAL EXAMPLE: EXHALED NITRIC OXIDE FROM EXERCISE-INDUCED BRONCHOSPASM
DB14 gives data on the effect of exercise-induced bronchospasm (EIB) as measured by exhaled nitric oxide (eNO). eNO is not different between patients with and without EIB before exercise. Earlier, it was shown that the eNO difference became significant by 20 minutes of exercise. Do the data show that the eNO decrease at 5 minutes of exercise in $n_1 = 6$ patients with EIB has not yet become greater than that in $n_2 = 32$ patients without EIB? Let us follow the steps in noninferiority testing. (1) We deem a change of 20% from preexercise level, averaging 29.26 parts per billion to be clinically important. $\Delta = 0.20 \times 29.26 = 5.85$; (2) $H_0$: $\delta = 5.85$, for example, the true-but-unknown difference is at least 5.85. $H_1$: $\delta < 5.85$; (3) $\alpha = 0.05$; (4) the sample difference is the difference between the two means of eNO change, or $d = m_1 - m_2$, and the standard deviations are calculated from the sample, so we use the last row of Table 12.1. From the $t$ table (see Table II) for $n_1 + n_2 - 2 = 36$ df, $t_{0.95} = 1.689$ (by interpolation). (5) The mean and standard deviation are 0.70 and 4.64 for patients with EIB and 4.82 and 7.41 for patients without EIB, respectively.

Substitution in Table 12.1 entry yields,

$$s_d = \sqrt{\left(\frac{1}{6} + \frac{1}{32}\right) \left(\frac{5 \times 21.53 + 31 \times 54.91}{6 + 32 - 2}\right)} = 3.15.$$

The observed difference between means is $d = 0.70 - 4.82 = -4.12$. From Eq. (12.2) the test statistic $t = (\Delta - d)/s_d = [5.85 - (-4.12)]/3.15 = 3.17$. Because $3.17 > 1.689$, $H_0$ is rejected; we have evidence that the eNO difference between patients with and without EIB has not increased as of 5 minutes of exercise.
CHAPTER 12 Equivalence testing

Exercise 12.1
Data on recovery after orthopedic surgery are given in DB10. We ask if triple hop time on the operated leg is not larger than (noninferior to) that on the nonoperated leg. We judge that recovery, that is, $\Delta$, is defined as operated leg time no greater than 10% of nonoperated leg time. Do we have evidence of adequate recovery?

12.5 METHODS FOR EQUIVALENCE TESTING

This section addresses two-sided equivalence testing. In the finger BP sensor from the beginning of this chapter, we considered the mean difference of DBP readings between the two devices for inferiority testing, that is, the magnitude of the difference whether it was plus or minus. Suppose we must treat the patient differently if the error is on the high side than on the low side—we need a two-sided test, that is, a test of equivalence.

EXAMPLE POSED: CARDIAC INDEX BY BIOIMPEDANCE VERSUS THERMODILUTION

A method of measuring cardiac index (CI; cardiac output normalized for body surface area) is thermodilution (TD), in which a catheter is placed in the heart. A proposed noninvasive method is bioimpedance (BI), in which an instrument attached to the patient’s skin by electric leads indicates CI. For BI to be clinically useful, we need to provide evidence that it is equivalent to TD.1

METHOD FOR EQUIVALENCE TESTING

Hypotheses

A two-sided equivalence test asks the clinical question: Is $m_1$ similar to $m_2$? In contrast to the difference test, in which $H_0$ is one-sided and $H_1$ is two-sided, the equivalence test requires $H_0$ to be two-sided, whereas $H_1$ is one-sided. The question is posed statistically as three hypotheses: two null, subscripted “L” for left and “R” for right, and the alternative. They are $H_{0L}$: $\delta = -\Delta$, $H_{0R}$: $\delta = \Delta$, and $H_1$: $-\Delta < \delta < \Delta$. (Recall that the equal signs are used for exact calculation of a probability and are interpreted as “$\delta$ at most as small as $-\Delta$” and “$\delta$ at least as large as $\Delta$”, respectively.) $\Delta$ can occur on only one side, and we want $\alpha$ probability of error on that side, in contrast to a two-sided difference test in which $\alpha$ is split into two sides. However, we do not know in advance on which side of $d$ we will find $\Delta$, which gives rise to the concept of TOST, with error risk $\alpha$ for each. Two one-sided tests do not increase the complication of testing; we simply go through the motions twice, once for each side.

(Continued)
Test Statistics

The test statistics are given in Eqs. (12.3) and (12.4), with standard error and error probability table as in Table 12.1:

\[ z_L = \frac{d - (-\Delta)}{\sigma_d}, \quad z_R = \frac{\Delta - d}{\sigma_d} \]  \hspace{1cm} (12.3)

or

\[ t_L = \frac{d - (-\Delta)}{s_d}, \quad t_R = \frac{\Delta - d}{s_d}. \]  \hspace{1cm} (12.4)

For Eq. (12.3), if both \( z_L \geq z_{1-\alpha} \) and \( z_R \geq z_{1-\alpha} \), reject \( H_0 \). For Eq. (12.4), if both \( t_L \geq t_{1-\alpha} \) and \( t_R \geq t_{1-\alpha} \), reject \( H_0 \).

**STEPS IN EQUIVALENCE TESTING**

1. Pose a clinically important difference \( \Delta \).
2. State a pair of null hypotheses \( H_0 \): \( \delta = -\Delta \), the theoretical difference is \( -\Delta \), and \( H_0: \delta = \Delta \), the theoretical difference is \( \Delta \). (By implication the same conclusion will be reached if \( \delta < -\Delta \) or if \( \delta > \Delta \).) State an alternative hypothesis \( H_1: -\Delta < \delta < \Delta \), the theoretical difference is within a clinically unimportant interval.
3. Choose \( \alpha \).
4. Look up the critical values \( z_{1-\alpha} \) or \( t_{1-\alpha} \) in the appropriate table for the chosen \( \alpha \).
5. Calculate the appropriate left and right statistics from Eqs. (12.3) or (12.4) and Table 12.1.
6. If both test statistics exceed the critical value, reject \( H_0 \) that implies accepting equivalence.

**Using a Confidence Interval**

If we use the confidence interval approach, we have to form a \( 1 - 2\alpha \) level of confidence to have error risk \( \alpha \) in each tail. For known \( \sigma \), \( H_0 \) is rejected if \( \Delta \) falls outside the confidence interval \( (d - z_{1-\alpha}\sigma_d, d + z_{1-\alpha}\sigma_d) \). For \( s \) estimated from the sample, a similar statement occurs: \( H_0 \) is rejected if \( \Delta \) falls outside the confidence interval \( (d - t_{1-\alpha}s_d, d + t_{1-\alpha}s_d) \). Fig. 12.2 is a diagram of the \( \delta \)-axis showing illustrative locations of \( \Delta, d \), and the confidence interval for the case of the \( t \) distribution. Because the confidence interval does not include \( \Delta \), there is evidence to reject \( H_{0\delta:} \delta > \Delta \) and conclude equivalence.

![Figure 12.2](image-url)
In keeping with what is becoming a convention in equivalence testing in the absence of a $\Delta$ chosen for clinical importance, we will judge $BI$ to be equivalent if its mean is within 20% of the TD mean, which is known from repeated large samples to be $\mu_T = 2.75$ (L/min/m$^2$). Because $\mu_T$ is known, we will be testing a sample mean against a population mean. Because we want $BI$ to be neither greater nor less than TD, we need a two-sided test. We use TOST with $\alpha = 0.05$ in each tail. A 20% of 2.75 is 0.55, so $H_0$: $\delta = (\mu_B - \mu_T) = -0.55$; $H_{0R}$: $\delta = 0.55$; and $H_1$: $-0.55 < \Delta < 0.55$. We sample $n = 96$ patients and find $BI$ has mean $m_B = 2.68$ (L/min/m$^2$). $d = 2.75 - 2.68 = 0.07$ with standard deviation $s = 0.26$, so $s_d = s/\sqrt{n} = 0.0265$. From Table II $t_{0.95}$ for 95 df = 1.661. Substituting in Eq. (12.4), we find $t_L = [(0.07 - (-0.55))/0.0265] = 23.40$, which exceeds 1.661, and $t_R = (0.55 - 0.07)/0.0265 = 18.1132$, which also exceeds 1.661. We have evidence of equivalence, that is, that $BI$ readings are clinically similar to TD readings. The result could have been reached somewhat more easily using a confidence interval. Substituting in $(d - t_1 - \alpha \delta_d, d + t_1 - \alpha \delta_d)$ yields $(0.026, 0.114)$. $\Delta = 0.55$ is outside the confidence interval, so $H_0$ is rejected. There is evidence for equivalence.

**ADDITIONAL EXAMPLE: EFFECT OF PRIOR SURGERY ON RATE OF INTUBATION**

DB12 gives data on prior surgery and intubation for patients undergoing carinal resection. We ask if the rate of intubation is the same for 34 patients with prior surgery and 100 without. We will follow the steps in equivalence testing. (1) We deem that a change of 5% in the intubation rate because of prior surgery is clinically significant. Then, $\Delta = 0.05$; (2) $H_{0L}$: $\delta = -0.05$; $H_{0R}$: $\delta = 0.05$; and $H_1$: $-0.05 < \delta < 0.05$; (3) $\alpha = 0.05$; (4) row 6 of Table 12.1 indicates Table I (see Tables of Probability Distributions) to find the critical value. From Table I, in the “one-tailed $1 - \alpha$” column, $z_{1 - \alpha} = z_{0.95} = 1.645$, used twice, once for each $H_0$. (5) We calculate the intubation rate as 17.65% for patients with prior surgery and 13.00% for those without prior surgery. The difference in rates is $d = 0.0465$. From the formula in row 6 of Table 12.1, we find $\sigma_d$ as follows:

$$\sigma_d = \sqrt{\left(\frac{1}{34} + \frac{1}{100}\right)\left(\frac{34 \times 0.1765 + 100 \times 0.1300}{34 + 100}\right)\left(1 - \frac{34 \times 0.1765 + 100 \times 0.1300}{34 + 100}\right)} = 0.0693.$$

From Eq. (12.3), we find

$$z = \frac{d - (-\Delta)}{\sigma_d} = \frac{0.0465 + 0.05}{0.0693} = 1.39.$$
and

\[ z = \frac{\Delta - d}{\sigma_d} = \frac{0.05 - 0.0465}{0.0693} = 0.05. \]

Because neither 1.39 nor 0.05 exceeds 1.645, we have no evidence for equivalence. The data fail to demonstrate similarity.

**Exercise 12.2**

DB3 gives baseline serum theophylline levels for patients with emphysema. Perform an equivalence test to learn if the data are free from an initial sex bias, that is, if mean baseline level is equivalent for \( n_1 = 6 \) women and \( n_2 = 10 \) men. We consider a difference in means of at least 2 to indicate a bias. Sample calculations are \( m_1 = 12.67 \), \( s_1 = 3.46 \), \( m_2 = 9.68 \), and \( s_2 = 3.65 \).

### 12.6 JOINT DIFFERENCE AND EQUIVALENCE TESTING

Most often, an investigator knows whether to look for a difference or for a lack of difference. However, there are exceptions. Let us begin with an example.

**EXAMPLE POSED**

To treat a fractured ankle, suppose standard of care mandates pinning by device 1. An investigator compares device 1 with treatment by device 2, a new, cheaper, and more easily installed pinning device, in a randomized controlled trial with 30 patients treated by each device (data fabricated to simplify illustration). The measure of success used for comparison is the distance—measured in inches—covered in a triple hop on the injured leg 4 months after repair (the longer the hop, the better the healing). The investigator deems that a hop must exceed 6 in., that is, half the length of a foot, to have clinical meaning. The investigator has no idea whether the new device is better, worse, or no different than the current device so decides to carry out a joint difference/equivalence test. The hop distance distributions are approximately normal so that \( t \) testing is appropriate.

**Terminology**

In this section, we will treat the issue only for the case of difference or equivalence between two population means. Let us name by \( \Delta \) the minimum difference between the means of interest that is clinically meaningful, that is, the equivalence margin; by \( \delta \) the true but unknown difference between \( \mu_1 \) and \( \mu_2 \); by \( d \) the sample estimate of \( \delta \); by \( s_e \) the standard error of the variable being measured; and by \( t_c \) the critical value of the \( t \) distribution for the chosen \( \alpha \) (e.g., 0.05).
CHAPTER 12 Equivalence testing

History
The difference \( t \) test has been around since Gosset’s historic paper in 1908\(^3\) and equivalence since the 1980s. The concept of joint difference and equivalence testing was seen in Wald\(^4\) embedded in the method of sequential analysis, although it was not so named. This simultaneous use of both types of error is based on the approach of decision theory that many say should supplant NHST (see, e.g., Betensky\(^5\) and Matthews\(^6\)). Tryon\(^7\) and Tryon and Lewis\(^8\) proposed a joint difference and equivalence test for a two-sample \( t \) test, although their method relied on confidence intervals, yielding no indication of the strength of the difference or equivalence, and it used two confidence intervals on the respective means rather than one confidence interval on the difference between means. Riffenburgh\(^9\) provided a formal algorithm for a simultaneous test and also a method for simultaneously finding the estimates of the error risks.

METHOD
We can express the possibilities for joint difference and equivalence in terms of a confidence interval on \( \delta \), with its lower bound denoted by \( L = d - t_s \delta \) and its upper bound by \( U = d + t_s \delta \), leading to four possible outcomes:

1. \((L \leq -\Delta \text{ and } U \leq 0) \text{ or } (L \geq 0 \text{ and } U \geq \Delta)\) corresponding to evidence of difference and no evidence of equivalence. If this outcome is true, the estimate of error risk is found as twice the area under the \( t \) distribution curve to the right of: \(|d|/s_d\).

2. \((L \text{ lies between } -\Delta \text{ and } 0 \text{ and } U \text{ lies between } 0 \text{ and } \Delta)\) corresponding to evidence of equivalence and no evidence of difference. If this outcome is true, the conservative estimate of error risk is found as the area under the \( t \) distribution curve to the right of the lesser of the two values: \(|\Delta - d|/s_d \text{ and } |\Delta + d|/s_d\).

3. \((L \leq -\Delta \text{ and } U \geq 0) \text{ or } (L \geq 0 \text{ and } U \geq \Delta)\) corresponding to indeterminacy, that is, no evidence of equivalence or difference.

4. \((L \text{ and } U \text{ lie between } -\Delta \text{ and } 0) \text{ or } (L \text{ and } U \text{ lie between } 0 \text{ and } \Delta)\) corresponding to the degenerate case in which statistical variability is much smaller, that is, measurement precision is much larger than the practical considerations giving rise to \( \Delta \).

EXAMPLE COMPLETED
In the example the investigator took \( \Delta \) to be 6 in. Sample means and standard deviations were \( m_1 = 33, s_1 = 4; \) and \( m_2 = 35, s_2 = 6 \). The difference between means is \( d = 2 \), its \( df = 58 \), its standard deviation \( s_d = 5.0990 \), and its standard error \( s_d = 1.3166 \). For two-tailed \( \alpha = 0.05 \), \( t_c = 2.00 \). A confidence interval on \( \delta \) is \( L = -0.6354 \text{ to } U = 4.6354 \), leading to the second possible outcome \((-0.6354 \text{ lies between } -6 \text{ and } 0 \text{ and } 4.6354 \text{ lies between } 0 \text{ and } 6)\). We conclude evidence of equivalence and no evidence of difference. The lesser of the two error risks \(|\Delta - d|/s_d = 3.0381 \text{ and } |\Delta + d|/s_d = 6.0762\) leads to the area under the \( t \) curve to the right of 3.04 as 0.002. From a clinical decision
standpoint, we can have reasonable belief in our decision that device 2 is no different in benefit from device 1. Equivalence mandates not that we default to device 1, but rather that we choose the device on a basis unrelated to subject performance. In this example, we would change clinical practice by choosing device 2 based on ease of surgical installation and cost.

REFERENCES

13.1 BASICS OF TESTS ON VARIABILITY

Why should we be interested in testing variability?

The average does not tell the whole story. As a metaphor, consider two bowmen shooting at a target. Bowman A always hits the bullseye. Half of Bowman B’s arrows falls to the left of the bullseye and half to the right. Both bowmen have the same average, but Bowman A is the better shot. As an example in medicine, small amounts of two orally administered drugs reach a remote organ. The mean level is the same for both, but drug B is more variable (has larger standard deviation) than drug A. In some cases, too little of drug B gets through to be effective and in other cases a dangerously high level gets through. Thus the less variable drug is the better.

A test of variability serves two main purposes

(1) The need to detect differences in variability per se was just illustrated and (2) it tests the assumption of equal variances used in means testing.

How are two variances compared?

In dealing with variability, we usually use the variance, which is just the square of the standard deviation. The decisions made using it are the same, and the difficult mathematics associated with square roots is avoided in the derivation of probability functions. To compare the relative size of two variances, we take their ratio, adjusted for degrees of freedom (df). The tabulated critical value of the ratio for statistical significance also depends on the df.

The true variance of a population may be compared with a postulated value or another population’s variance

If we compare the true variance giving rise to a sample, estimated by \( s^2 \), against a postulated population variance \( \sigma^2 \) (the topic of Section 13.2), we find the significance level of \( \chi^2 = df \times s^2 / \sigma^2 \) in chi-square (Tables III or IV) for \( df = n - 1 \). If we compare two subpopulation variances, estimated by \( s_1^2 \) and \( s_2^2 \) (the topic of Section 13.3),
assigning subscript 1 to the larger, we find the significance level of \( F = s_1^2 / s_2^2 \) in Table V. This table involves df for both variances, \( df_1 = n_1 - 1 \) and \( df_2 = n_2 - 1 \).

**A nonsignificant test result must be interpreted carefully**

Tests of variance may be useful in assessing the validity of the equal-variance assumption required for normal (z) and t tests and the analysis of variance (ANOVA), but we must understand the limitation of this use. A significant difference between variances implies the conclusion that they are different, whereas like all difference tests of hypotheses, a nonsignificant result implies only that no difference has been demonstrated, not that one does not exist. As we note in Chapter 11, Tests of location with continuous outcomes, however, in practice it is advisable to use methods that make as few assumptions as necessary. When testing differences in population means, it is advisable to use tests that do not assume equal variances as this guards against violations of the equal-variance assumption with little impact on the power of the test.

### 13.2 TESTING VARIABILITY ON A SINGLE SAMPLE

**EXAMPLE POSED: PROSTATE-SPECIFIC ANTIGEN VARIABILITY**

Is the true prostate-specific antigen (PSA) variance of the first 10 urology patients (Table DB1.1) different from the variance of the remaining 291 (which for purposes of this example we assume to be fixed and known)? Using the subscript 0 to denote the variance of the population from which the first 10 patients were sampled, \( H_0: \sigma_0^2 = \sigma^2 \). We have no reason to anticipate whether \( \sigma_0^2 \) should be larger or smaller than \( \sigma^2 \), so we use \( H_1: \sigma_0^2 \neq \sigma^2 \) and, if \( \alpha \) is chosen as 0.05, we allow 0.025 for each tail. The reader might find it easier to think of the question in terms of standard deviations rather than variances. Let us consider the 291 readings as a population with \( \sigma = 17.19 \) and the first 10 as a sample with \( s = 1.61 \). Could the difference between \( \sigma \) and \( s \) have occurred by chance? We use variances for our test to avoid the mathematics of square roots. \( \sigma^2 = 295.50 \) and \( s^2 = 2.59 \).

**METHOD: TEST OF ONE SAMPLE VARIANCE**

We ask whether the variance \( \sigma_0^2 \) of a population from which we draw a sample is the same as a theoretical variance \( \sigma^2 \). \( \sigma_0^2 \) is estimated by the sample variance \( s^2 \). We assume the sample data are distributed normal. \( \sigma^2 \) may be known from some theory, or it may be estimated by a very large sample, because the sample variance converges on the population variance as the sample size grows large. The null hypothesis is that \( s^2 \) is drawn from the population having variance \( \sigma^2 \), or \( H_0: \sigma_0^2 = \sigma^2 \), implying that the ratio \( s^2 / \sigma^2 \) should be different from 1 only by a random influence. The statistic calculated is

(Continued)
This is a relatively simple test except for one aspect: the test statistic is distributed chi-square, which is not symmetric, so that a different table must be used for each tail. The alternate hypothesis $H_1$ specifies the probability table(s) to be used. For a chosen $\alpha$ the cases are as follows:

- **a.** $\sigma_0^2 > \sigma^2$: the chi-square critical value for $\alpha$ is found from Table III.
- **b.** $\sigma_0^2 < \sigma^2$: the chi-square critical value for $\alpha$ is found from Table IV.
- **c.** $\sigma_0^2 \neq \sigma^2$: split $\alpha$; the left $\chi^2$ critical value for $\alpha/2$ is found from Table IV and the right $\chi^2$ critical value for $\alpha/2$ is found from Table III.

The steps are as follows:
1. Verify that the sample frequency distribution is roughly normal.
2. Specify null and alternate hypotheses and choose $\alpha$.
3. Identify the theoretical $\sigma$.
4. Look up the critical value(s) for the chosen $\alpha$ in the appropriate $\chi^2$ table(s).
5. Calculate the statistic from Eq. (13.1): $\chi^2 = df \times s^2/\sigma^2$.
6. Make the decision whether or not to reject the null hypothesis.

**EXAMPLE COMPLETED: PROSTATE-SPECIFIC ANTIGEN VARIABILITY**

Let us follow the steps given in the Methods paragraph. We assume normality: we have seen that PSA data are roughly bell-shaped, although right-skewed. We believe the deviations are not so great as to invalidate this rather robust test. The remaining 291 PSAs form a sample large enough that its variance has converged closely to the population variance, so we may use the calculated variance as $\sigma^2$. Our test statistic is $s^2/\sigma^2$, which, when multiplied by $df$, is distributed $\chi^2$. The left-tail critical value must be found from Table IV and the right from Table III, because the chi-square distribution is not symmetric like $z$ or $t$. $df = 10 - 1 = 9$. Table 13.1 gives a fragment of Table IV. From Table 13.1 or Table IV, $\alpha = 0.025$ coupled with $df = 9$ yields $\chi^2 = 2.70$. If $df \times s^2/\sigma^2 < 2.70$, we reject $H_0$. Similarly, the same inputs into Table III (or Table 13.2) yield $\chi^2 = 19.02$; if $df \times s^2/\sigma^2 > 19.02$, we reject $H_0$. Now we can calculate our statistic. $\chi^2 = df \times s^2/\sigma^2 = 9 \times 2.59/295.50 = 0.0789$, which is far less than 2.70, the left critical value. We reject $H_0$ and conclude that the population giving rise to the first 10 patients has smaller variability than the population giving rise to the

**Table 13.1** A fragment of Table IV, chi-square distribution, left tail.

<table>
<thead>
<tr>
<th>$\alpha$ (Area in left tail)</th>
<th>0.0005</th>
<th>0.001</th>
<th>0.005</th>
<th>0.01</th>
<th>0.025</th>
<th>0.05</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$df = 9$</td>
<td>0.97</td>
<td>1.15</td>
<td>1.73</td>
<td>2.09</td>
<td>2.70</td>
<td>3.33</td>
<td>4.17</td>
</tr>
</tbody>
</table>

*Selected $\chi^2$ values (distances above zero) are given for 9 $df$ for various $\alpha$, the area under the curve in the left tail.*
remaining 291. Indeed, from Table 13.1, the $\chi^2$ for an 0.0005 $\alpha$ with 9 $df$ is 0.97.

As 0.0789 < 0.97, the resulting $p$-value is $< 0.001$.

**ADDITIONAL EXAMPLE: IS A TREATMENT FOR DYSPEPSIA IN THE EMERGENCY DEPARTMENT TOO VARIABLE?**

In the additional example of Section 11.2, an emergency medicine physician tested the effectiveness of a “GI cocktail” (antacid plus viscous lidocaine) to treat emergency dyspeptic symptoms as measured on a 1–10 pain scale. He concluded that the treatment was effective on average, but was it more variable? This question implies a one-tailed test. He will sample 15 patients, so $df = 14$. Table 13.2 provides a portion of Table III. From Table 13.2 or Table III with 14 $df$, the critical value of $\chi^2$ for $\alpha = 0.05$ is 23.69. The population standard deviation for the scoring difference upon being seen minus that after a specified number of minutes for a large number of patients without treatment was $\sigma = 1.73$. He treated $n = 15$ patients, measuring the difference in pain before minus after treatment. Data were 6, 7, 2, 5, 3, 0, 3, 4, 5, 6, 1, 1, 1, 8, 6. He calculated $m = 3.87$ and $s = 2.50$. He tested the population variance of his sample (observed variance $s^2 = 6.27$) against the assumed population variance of untreated patients (postulated here to be $\sigma^2 = 2.99$) at the $\alpha = 0.05$ level of significance. By substituting in Eq. (13.1), he found.

$$\chi^2 = \frac{df \times s^2}{\sigma^2} = \frac{14 \times 6.27}{2.99} = 29.36.$$ 

The calculated $\chi^2$ is greater than the critical value of 23.69. Indeed, it is slightly greater than the 29.14 value associated with $\alpha = 0.01$. He has evidence that the population giving rise to his sample is more variable than the comparison population. From a statistical software package, $p$-value = 0.009.

**Exercise 13.1**

*Is the variability of readings from a tympanic thermometer too large?* We are investigating the reliability of a certain brand of tympanic thermometer (temperature measured by a sensor inserted into the patient’s ear). Sixteen readings ($^\circ$F) were taken on a healthy patient at intervals
of 1 minute, left and right ears pooled. (Data are given in Exercise 11.9.) In our clinical judgment a reliable thermometer will be no more than 1°F off 95% of the time. This implies that 1°F is about 2σ, so that σ = 0.5°C. From the data, s = 1.230. At the α = 0.05 level of significance, is the variability of the tympanic thermometer reading on a patient unacceptably large?

13.3 TESTING VARIABILITY BETWEEN TWO SAMPLES

EXAMPLE POSED: DOES THE INITIAL UNREPRESENTATIVE VARIABILITY IN PROSTATE-SPECIFIC ANTIGEN EXTEND TO LATER PATIENTS?

In the example of Section 13.2, we concluded that the PSA variances (and therefore standard deviations) for the first 10 and the remaining 291 patients were quite different. We want to know if the first 10 per se are unrepresentative or if the variability increased gradually or at a later point. Let us test the variances of the populations given rise to the first 10 patients against each successive 10 and list the result. The normality-of-data assumption remains. By using “first” to denote the first sample of 10 and “later” to denote each later sample, \( H_0: \sigma^2_{\text{first}} = \sigma^2_{\text{later}} \) and \( H_1: \sigma^2_{\text{first}} < \sigma^2_{\text{later}} \).

METHOD: TEST OF TWO SAMPLE VARIANCES

Are two population variances, or standard deviations, the same? We assume that the data of both samples are distributed normal. \( H_0: \sigma^2_1 = \sigma^2_2 \). Since we test the larger variance over the smaller, Table V may be used for any \( H_1 \). Choose \( \alpha \). Sample 1 has sample variance \( s^2_1 \) calculated from \( n_1 \) observations and sample 2, \( s^2_2 \) from \( n_2 \), where the number 1 is assigned to the larger variance. The statistic is

\[
F = \frac{s^2_1}{s^2_2}
\]  

(13.2)

with \( n_1 - 1 \) and \( n_2 - 1 \) df. Look up the critical value of \( F \) in Table V or calculate the p-value with a statistical software package. The steps are as follows:

1. Verify that the data of each sample are roughly normal.
2. State the null and alternate hypotheses (\( H_0: \sigma^2_1 = \sigma^2_2 \) and \( H_1: \sigma^2_1 > \sigma^2_2 \)); choose \( \alpha \).
3. Look up the critical value for the chosen \( \alpha \) in Table V.
4. Calculate the statistic \( F = s^2_1/s^2_2 \) (where \( s^2_1 \) designates the larger variance).
5. Decide whether or not to reject the null hypothesis.

EXAMPLE COMPLETED: DOES THE INITIAL UNREPRESENTATIVE VARIABILITY IN PROSTATE-SPECIFIC ANTIGEN EXTEND TO LATER PATIENTS?

We choose \( \alpha = 0.05 \). Table 13.3 is a portion of Table V. From Table 13.3 or Table V, we find that the critical value of \( F \) for 9,9 df is 3.18. The \( F \) ratio must exceed 3.18 to become statistically significant.
CHAPTER 13 Tests on variability and distributions

Table 13.3 A portion of Table V, $F$ distribution.$^a$

<table>
<thead>
<tr>
<th>Numerator df</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denominator df</td>
<td>9</td>
<td>3.23</td>
<td>3.18</td>
</tr>
<tr>
<td>Denominator df</td>
<td>10</td>
<td>3.07</td>
<td>3.02</td>
</tr>
<tr>
<td>Denominator df</td>
<td>11</td>
<td>2.95</td>
<td>2.90</td>
</tr>
<tr>
<td>Denominator df</td>
<td>12</td>
<td>2.85</td>
<td>2.80</td>
</tr>
</tbody>
</table>

$^a$Selected distances ($F$) are given for $\alpha = 0.05$, the area under the curve in the positive tail. Numerator df appears in column headings, denominator df in row headings, and $F$ in the table body.

Table 13.4 $F$ ratios for subsamples of size 10.

<table>
<thead>
<tr>
<th>Sample numbers</th>
<th>$F$ ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>11–20</td>
<td>1.72</td>
</tr>
<tr>
<td>21–30</td>
<td>1.20</td>
</tr>
<tr>
<td>31–40</td>
<td>0.92</td>
</tr>
<tr>
<td>41–50</td>
<td>794.06</td>
</tr>
<tr>
<td>51–60</td>
<td>11.14</td>
</tr>
<tr>
<td>61–70</td>
<td>27.28</td>
</tr>
<tr>
<td>71–80</td>
<td>511.72</td>
</tr>
<tr>
<td>81–90</td>
<td>61.11</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

CALCULATIONS

The second 10 (numbers 11–20) yielded a standard deviation of 2.11. $F = (2.11)^2 / \sqrt{(1.61)^2} = 4.45 / 2.59 = 1.72$. As $1.72 < 3.18$, we conclude that the variances (and, of course, the standard deviations) are not different for the first and second sets of 10. A list of comparisons of the first 10 with successive samples of 10, compiled using the same numerical procedures, is as in Table 13.4.

INTERPRETATION

We note that the 41–50 set yielded a huge $F$. By examining the data, we find that PSA = 221 for patient 47! This turned out to be a useful way to discover outliers, those data so far different from the bulk of data that we suspect they arose from a unique population. However, even if we dropped the 221 reading, the $F$ would equal 8.88 for the remaining data, which is still significant. It appears that there are outliers and patterns of nonhomogeneity scattered throughout the data.

ADDITIONAL EXAMPLE: COMPARE THE VARIABILITY OF TWO PAIN-RELIEF DRUGS

In 1958, a study$^2$ compared a new postoperative pain-relief drug (treatment 1) to the established Demerol (treatment 2). Data consisted of reduction in pain measured on a 1–12 rating scale. For patients who completed the protocol, data were as in Table 13.5.
In order to test the means, we may wish to know whether the assumption of equal variances seems reasonable (though we could just as easily use the \( t \) test for unequal variances). We wish to test \( H_0: \sigma^2_1 = \sigma^2_2 \) against \( H_1: \sigma^2_1 \neq \sigma^2_2 \) at \( \alpha = 0.05 \). \( m_1 = 5.10, s_1 = 4.0125, s^2_1 = 16.1000, m_2 = 2.15, s_2 = 1.9513, s^2_2 = 3.8077 \). We substitute in Eq. (13.2) to find

\[
F = \frac{s^2_1}{s^2_2} = \frac{16.10}{3.81} = 4.23.
\]

From Table 13.3 or Table V, the critical value of \( F \) for 9,12 \( df \) is 2.80. As 4.23 is much larger, the variances (and, of course, the standard deviations) are significantly different. From a statistical software package, the \( p \)-value = 0.012. A test of means should use either the unequal-variance form of the \( t \) test from Section 11.3 (which should be the default as noted in Chapter 11: Tests of location with continuous outcomes) or possibly the rank-sum test from Section 11.8 depending on the scientific setting.

**Exercise 13.2**

*Is the variability of the tympanic thermometer readings the same in both ears?* We are investigating the reliability of a certain brand of tympanic thermometer (temperature measured by a sensor inserted into the patient’s ear). Sixteen readings (°F), eight per ear, were taken on a healthy patient at intervals of one minute, alternating ears.\(^2\) (Data are given in Exercise 11.9.)

We want to test the hypothesis that the true mean temperatures are the same in the left (L) and right (R) ears (Exercise 11.2), but we want to investigate if the equal-variance \( t \) test may be appropriate. \( m_L = 96.41\)°F, \( s_L = 0.88\)°F, \( m_R = 97.86\)°F, and \( s_R = 1.12\)°F. At the \( \alpha = 0.05 \) level of significance, are the population variances (or standard deviations) of the two ears different?

### 13.4 Testing Variability Among Three or More Samples

**Example Posed:** Can we assume equal variances in the test of classifying patients by risk of prostate cancer?

Of our sample of 301 patients, 19 are known to have cancer already and therefore are not in the risk population, so we must delete them to obtain an unbiased population, leaving 282. Three groups remain, identified by a clinical decision algorithm: those whose risk of prostate cancer is (1) low, (2) moderate, and (3) high. We note that mean PSA density (PSAD) seems quite different for the three groups: 0.07, 0.21, and 0.35. We want to test for differences in the population means using a one-way

| Treatment 1 | 2 | 6 | 4 | 12 | 5 | 8 | 4 | 0 | 10 | 0 |
| Treatment 2 | 2 | 0 | 3 | 3 | 0 | 0 | 7 | 1 | 4 | 2 | 2 | 1 | 3 |
ANOVA, but that requires the assumption of equal variances. Do we have evidence that we should not make that assumption?

**METHOD: BARTLETT’S TEST OF HOMOGENEITY OF VARIANCE**

We want to know whether three or more variances are different. Bartlett’s test compares the variances of \( k \) independent random samples, assumed to be distributed normal. It tests \( H_0: \sigma^2_1 = \sigma^2_2 = \cdots = \sigma^2_k \) against the alternative \( H_1: \text{not } H_0 \). Let \( i \) denotes sample number, \( i = 1, 2, \ldots, k \). There are \( n \) observations in total in the \( k \) samples with \( n_i \) observations \( x_{ij} \) in the \( i \)th sample. First find the variance \( s^2_i \) for each sample, just the way the usual sample variance is found (where \( j \) is the index to sum over):

\[
s^2_i = \frac{\sum x_{ij}^2 - n_i m_i^2}{n_i - 1}.
\]

(13.3)

Then pool the \( k \) sample variances to find the overall variance \( s^2 \) (now summing over \( i \)):

\[
s^2 = \frac{\sum (n_i - 1)s^2_i}{n - k}.
\]

(13.4)

Note that Eq. (13.4) is not the same as the overall variance of the \( n \) observations, as each is the variance about its own sample mean. The test statistic, Bartlett’s \( M \), is given by

\[
M = \frac{(n - k) \ln (s^2) - \sum (n_i - 1) \ln (s^2_i)}{1 + (1/(3(k - 1))))(\sum (1/n_i - 1) - (1/n - k))}.
\]

(13.5)

The \( \alpha \) for \( M \) arises from the right tail of a chi-square distribution with \( k - 1 \) df. A critical value for \( M \) may be found in Table III. An \( M \) less than this critical value implies inadequate evidence to reject \( H_0 \); otherwise, reject \( H_0 \).

The steps are as follows:

1. Note the null and alternate hypotheses (\( H_0: \sigma^2_1 = \sigma^2_2 = \cdots = \sigma^2_k \) and \( H_1: \text{not } H_0 \)).
2. Choose \( \alpha \).
3. Look up the critical value for the chosen \( \alpha \) in Table III for \( k - 1 \) df.
4. Calculate the variances \( s^2_i \) for each sample from Eq. (13.3) and the overall variance \( s^2 \) [note that \( s^2 \) is not exactly the same as the variance for the entire \( n \) observations combined; calculate it from Eq. (13.4) using the several \( s^2_i \)].
5. Calculate Bartlett’s \( M \) as in (13.5).
6. Make the decision to accept or reject the null hypothesis.

**EXAMPLE COMPLETED: CAN WE ASSUME EQUAL VARIANCES IN THE TEST OF CLASSIFYING PATIENTS BY RISK OF PROSTATE CANCER?**

Is the normal assumption satisfied? We make plots of the PSAD frequency distributions for the three groups with normal fits superposed, as shown in Fig. 13.1. We are not very happy with the one or two extremely high PSAD values in each group.
It would be advisable to go to rank methods and perform a Kruskal–Wallis test from Section 11.10. However, we will continue with Bartlett's test for the sake of illustration.

**CALCULATION**

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$. $H_1$: not $H_0$. We choose $\alpha = 0.05$. The 0.05 critical value from Table 13.2 or Table III for $2 \, df$ is 5.99; if $M > 5.99$, we reject $H_0$. We obtain the following values, carrying six significant digits in the calculations (so that we can take precise square roots). The $s_i^2$ are just the ordinary variances (standard deviations squared). Those variances and the associated $n$'s are as in Table 13.6.

![Figure 13.1 Plots of PSAD distributions of 282 patients for groups of low, moderate, and high risk of prostate cancer. PSDA, Prostate-specific antigen density.](image)

**Table 13.6** Sample sizes and variances for prostate-specific antigen density example.

<table>
<thead>
<tr>
<th>$n_i$</th>
<th>$s_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>0.00557546</td>
</tr>
<tr>
<td>164</td>
<td>0.0139188</td>
</tr>
<tr>
<td>29</td>
<td>0.0227769</td>
</tr>
</tbody>
</table>

It would be advisable to go to rank methods and perform a Kruskal–Wallis test from Section 11.10. However, we will continue with Bartlett’s test for the sake of illustration.
Substitution in Eq. (13.4) (summing over the index $i$) yields

$$s^2 = \frac{\sum (n_i - 1)s_i^2}{n - k}$$

$$= \frac{88 \times 0.00557546 + 163 \times 0.0139188 + 29 \times 0.0227769}{282 - 3} = 0.0122578$$

Finally, substitution in Eq. (13.5) yields

$$M = \frac{(n - k)\ln(s^2) - \sum (n_i - 1)\ln(s_i^2)}{1 + (1/(3(k - 1)))\left(\frac{\sum (1/n_i - 1) - (1/n - k)}{1/n - k}\right)}$$

$$= \frac{279\ln(0.0121762) - 88 \ln (0.00557546) - \cdots - 28 \ln (0.0227769)}{1 + (1/6)(1/88 + 1/163 + 1/28 - 1/279)}$$

$$= 29.1585$$

The $M$ of 29.1585 $> \alpha$ the critical 5.99, so $H_0$ is rejected. In fact, $M$ exceeds the critical value for the smallest chi-square in Table 15.21 for $\alpha = 0.0005$. We may state that $p$-value $< 0.001$. We reject the null hypothesis at level 0.05 and conclude a difference among the variances. A test of population means that does not account for unequal variances would not be appropriate.

**ADDITIONAL EXAMPLE: MEDICATION TO REDUCE EDEMA FOLLOWING RHINOPLASTY**

Following rhinoplasty, swelling may cause deformity during healing. Steroids may decrease the swelling, but the required level of steroid was not known. Is the equal-variance assumption satisfied for the one-way ANOVA of medication levels to reduce edema following rhinoplasty? A total of $n = 50$ rhinoplasty patients were randomized into $k = 5$ groups of increasing steroid level having $n_1 = \cdots = n_5 = 10$. Swelling reduction was measured by MRIs before and after administration of the steroid. The means, variances, and first few data are shown in Table 13.7.

The requirement was to carry out a one-way ANOVA to learn whether the means were all the same or whether there were differences among them (additional example

<table>
<thead>
<tr>
<th>Table 13.7 Data on edema following rhinoplasty.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swelling reduction (mL)</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>Patients 1–5</td>
</tr>
<tr>
<td>Patients 6–10</td>
</tr>
<tr>
<td>Patients 11–15</td>
</tr>
<tr>
<td>Means $m_i$</td>
</tr>
<tr>
<td>Variances $s_i^2$</td>
</tr>
</tbody>
</table>
from Section 11.4). The ANOVA requires the assumption that the variances of the various groups are equal, which may be examined by Bartlett’s $M$. We choose $\alpha = 0.05$. From Table 13.2 or Table III the critical value of $M$ (which follows the $\chi^2$ distribution) with $k - 1 = 4$ df is 9.49. By substituting in Eq. (13.4) and then Eq. (13.5), we find

$$s^2 = \frac{\sum (n_i - 1)s_i^2}{n - k} = \frac{9 \times 6.30 + 9 \times 3.80 + \cdots + 9 \times 2.25}{50 - 5} = 3.326$$

$$M = \frac{(n - k)\ln(s^2) - \sum (n_i - 1)\ln(s_i^2)}{1 + (1/(3(k - 1)))(\sum (1/n_i - 1) - (1/n - k))}$$

$$= \frac{45\ln(3.326) - 9\ln(6.30) - \cdots - 9\ln(2.25)}{1 + (1/12)((1/9) + \cdots + (1/9) - (1/45))} = 4.47$$

As $4.47 < 9.49$, we do not have enough evidence to reject $H_0$; we have insufficient evidence to conclude that the population variances are different. We will carry on with the one-way ANOVA. From a statistical software package, the calculated $p$-value = 0.341.

**Exercise 13.3**

*Is there a batch-to-batch variability of laboratory animals’ resistance to parasitic infestation?* In a study on the control of parasites, rats were injected with 500 larvae each of the parasitic worm Nippostrongylus muris. Ten days later, they were sacrificed and the number of adult worms counted. In Exercise 11.7, we concluded that there was no batch-to-batch difference in average resistance to parasite infestation by groups of rats received from the supplier. However, is there a batch-to-batch difference in the variability of resistance? We answer this question with Bartlett’s test on the homogeneity of variances. $k = 4$ batches of $n_i = 5$ rats each were tested. Data were given in Exercise 11.7. The four sample variances are 3248.30, 4495.70, 4620.70, and 3623.70. At the $\alpha = 0.05$ level of significance, test $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$.

### 13.5 Basics on Tests of Distributions

#### What do we usually ask about distributions?

Of the many questions that could be asked about distribution shape, two are most common: “Is the distribution normal?” and “Do two distributions have the same shape?”

#### Testing normality

Normality of underlying data is assumed in many statistical tests, including the normal test for means, the $t$ test, and ANOVA. While most of these procedures are quite
robust to the normality assumption (mostly due to the central limit theorem), a user who is not sure that the assumption of normality is justified may wish to test this assumption, although we must understand a limitation of this use. A significant outcome implies the conclusion that the distribution is not normal, whereas, like all difference tests of hypotheses, a nonsignificant outcome implies only that no deviation from normality has been demonstrated, not that it does not exist. An equivalence test would lead to a similar limitation; if we failed to reject the null hypothesis of difference, no equivalence has been demonstrated. At the present time, the usual procedure is to conduct the difference test and assume normality unless the null hypothesis of normality is rejected.

**Testing other distributions**

Note that the tests of normality are essentially various goodness-of-fit tests that may be used to test data fits to any known probability distribution. Other distributional forms will not be considered in this book, because the tests primarily are applied to the normal probability distribution. The methodology for other forms is similar.

**Testing equality**

The user may want to know if two distributions have the same form, in some cases to satisfy assumptions required for tests and in other cases to learn if the natural forces giving rise to these distributions are similar. The test does not address the issue of what the distribution shapes are, only whether they are the same.

### 13.6 TEST OF NORMALITY OF A DISTRIBUTION

**Types of normality tests and their comparison**

Keep in mind that the robustness of most tests, some more than others, allows the test to be used if the underlying data are *approximately* normal. Thus barring quite unusual circumstances, any good test of normality will be adequate. The three most frequently used tests are discussed here. The better tests are more complicated, as one might expect. Also as one might expect, the better tests are more recent. The chi-square goodness-of-fit test was first published in 1900, the Kolmogorov–Smirnov, usually abbreviated KS, test in 1933, and the Shapiro–Wilk test in 1965. The choice of the test to use depends on both the sample size and whether the user would rather err on the side of being too conservative or the opposite. The Shapiro–Wilk test tends to reject the null hypothesis more readily than one would wish, whereas the KS and chi-square tests are too conservative, retaining the null hypothesis too often. Table 13.8 may help the selection.
The role of computer software

Statistical software packages are usually used in practice but are not required. Attention here will be focused on what the KS and goodness-of-fit tests do and how they do it, not on commands required to conduct computer-based tests; the Shapiro—Wilk test is involved enough that it should be attempted only with a software package. Enough detail in the KS and chi-square goodness-of-fit methods is given to allow a user to apply the method even in the absence of software.

Using the Shapiro—Wilk test

Seek statistical computer software to use the Shapiro—Wilk test.

A limitation of the Shapiro—Wilk test is that some software packages allow only the sample parameters ($m$ and $s$, not $\mu$ and $\sigma$) to be used for the theoretical normal against which the data are being tested, whereas we would prefer to allow specification of the normal being tested by either theoretical or sample parameters.

Using the Kolmogorov—Smirnov test (one-sample form)

**EXAMPLE POSED: ARE THE AGES OF THE PATIENTS IN TABLE DB1.1 NORMAL?**

We might want to know if the sample of prostate biopsy patients follows a normal distribution. $H_0$: The distribution from which the sample was drawn is normal. $H_1$: The distribution is not normal.

**METHOD: ONE-SAMPLE KOLMOGOROV—SMIRNOV (KS) TEST**

There are two questions we might ask: (1) Is the sample normal in shape without any \textit{a priori} mean and standard deviation specified? and (2) did the sample arise from a specific normal distribution with a postulated mean and standard deviation? We ask question (1) to satisfy the normality assumption, that is $H_0$: The form of the distribution is normal, and we use the sample’s $m$ and $s$ as parameters. While we ask question (2) to verify that the distribution is normal and has mean $\mu$ and standard deviation $\sigma$, that is, $H_0$: The distribution from which the sample was drawn is not different from a specified normal. In the method that follows, use $m$ and $s$ or use $\mu$ and $\sigma$ according to the respective question. Calculate the

(Continued)
critical value from one of Eqs. (13.6a)–(13.6c) for the desired \( \alpha \). (These critical values are approximations but are accurate to within 0.001.)

\[
\begin{align*}
\frac{1.63}{\sqrt{n}} & - \frac{1}{3.5n} \quad \text{for } \alpha = 0.01 \\
\frac{1.36}{\sqrt{n}} & - \frac{1}{4.5n} \quad \text{for } \alpha = 0.05 \\
\frac{1.22}{\sqrt{n}} & - \frac{1}{5.5n} \quad \text{for } \alpha = 0.10
\end{align*}
\]

1. Arrange the \( n \) sample values in ascending order.
2. Let \( x \) denotes the sample value each time it changes. (If sample values are 1, 1, 2, 2, 3, the first \( x \) is 2, the second \( x \).) Write these \( x \)'s in a column in order.
3. Let \( k \) denotes the number of sample members less than \( x \). (When the sample values just above changed from 1 to 2 and \( x \) became 2, there were \( k = 2 \) values less than \( x \). At the next change, \( x \) takes on the value 3 and there are \( k = 5 \) values less than that \( x \).) Write these \( k \)'s next to their corresponding \( x \)'s.
4. Let \( F_n(x) \) denote \( k/n \) for each \( x \). This is the sample cumulative frequency sum. Write down each \( F_n(x) \) in the next column corresponding to the associated \( x \).
5. Now we need the expected cumulative sum against which to test the sample. For each \( x \), calculate \( z = (x - \mu)/\sigma \) for case (1) or \( z = (x - \mu)/\sigma \) for case (2).
6. For each \( z \), find an expected \( F_e(n) \) as the area under the normal distribution to the left of \( z \). This area can be found using a statistical software package or from a very complete table of normal probabilities. If neither of these sources is available, interpolate from Table I. Write down these \( F_e(x) \) next to the corresponding \( F_n(x) \).
7. Write down next to these the absolute value (value without any minus signs) of the difference between the \( F \)'s, namely \( |F_n(x) - F_e(x)| \).
8. The test statistic is the largest of these differences, say \( L \).
9. If \( L \) exceeds the critical value, reject \( H_0 \). Otherwise, do not reject \( H_0 \).

**EXAMPLE COMPLETED: ARE THE AGES OF THE 10 PATIENTS DISTRIBUTED NORMAL?**

\( n = 10 \). From (13.6b), choosing \( \alpha = 0.05 \), the critical value is found as

\[
\frac{1.36}{\sqrt{10}} - \frac{1}{4.5 \times 10} = 0.408.
\]

We set up Table 13.9 with headings representing the values that will be required in the calculation.
1. The first column consists of the ages in increasing order.
2. The second column (\(x\)) consists of the sample values at each change. The first change occurs as age goes from 54 to 61.
3. The third column (\(k\)) consists of the number of sample members less than \(x\). For the first \(x\), there is only one age less than 61, so \(k = 1\).
4. The fourth column consists of the values \(k/n\). For the first \(x\), \(k/n = 1/10 = 0.1\).
5. Column five consists of \(z\), the standardized values of \(x\). We note the mean \(m = 65.1\) years and the standard deviation \(s = 7.0\) years from Table DB1.1 and, for each \(x\), we calculate \((x - m)/s\).
6. The sixth column consists of normal distribution probabilities, the area under the normal curve to the left of \(z\). For the first \(x\), we want to find the area under a normal curve from the left tail up to the \(-0.586\) standard deviation position (just over half a standard deviation to the left of the mean). A statistical package gives it as 0.279. If we do not have such access, we can interpolate from Table I. Since Table I gives only the right half of the curve, we use the curve’s symmetry; the area to the left of \(-0.586\) will be the same as the area to the right of \(+0.586\). \(0.586\) lies 86% of the way from 0.50 to 0.60. We need the one-tailed area given by \(\alpha\) for 0.50 plus 0.86 of the difference between the \(\alpha\)’s for 0.50 and 0.60, or \(0.308 + 0.86 \times (0.274 - 0.308) = 0.27876\), which rounds to 0.279, the same result as from the statistical package.
7. The last column consists of the absolute difference (difference without regard to sign) between the entries in columns 4 and 6. For the first \(x\), \(|0.1 - 0.279| = |-0.179| = 0.179\).
8. The largest value in the rightmost column is \(L = 0.179\).
9. Finally, we compare \(L\) with the critical value. Because \((L = 0.179 \leq 0.408\) (\(=\) critical value), we do not reject \(H_0\); the distribution from which the sample was drawn has not been shown to be different from normal. The \(p\)-value was computed to be 0.472 using a statistical software package.
**ADDITIONAL EXAMPLE: A TEST OF NORMALITY ON A POTENTIAL HUMAN IMMUNOVIRUS VACCINE**

In a study on a potential vaccine for human immunovirus (HIV), the number of HIV per milliliter blood (denoted \( h \)) was measured on a baseline sample of \( n = 45 \) patients.\(^5\) It was required that the data be approximately normal in order to conduct other statistical procedures. A frequency plot is clearly far from a normal distribution, which is superposed on the plot (Fig. 13.2A). Natural logarithms (ln) were taken in the hopes of transforming the data to normal. A plot of ln \( h \) with a superposed normal (based on the mean and standard deviation of the data) appears somewhat better (Fig. 13.2B) but is deviant enough from bell-shaped to require a test of normality.

**STEPPING THROUGH THE TEST PROCEDURE**

\( \alpha \) is chosen as 0.05. The critical value of the test is found from Eq. (13.6b): substituting \( n = 45 \) yields 0.1978. Data arranged in ascending order appear in Table 13.10 in the format of Table 13.9. \( m = 9.04 \) and \( s = 1.80 \). As there are no ties in ln \( h \), \( x \) will be the same as the sample values. \( k \) is one less than the count number of each datum. As before, \( F_n(x) = k/n \), \( z = (x - m)/s \), and \( F_n(x) \) is the area under the normal curve to the left of the \( z \) associated with that \( x \). The last column is the absolute value of the difference between the two \( F \)s. The largest value in the last column is \( L = 0.1030 \). As \( L < 0.1978 \), we fail to reject the hypothesis of normality.

**Exercise 13.4**

**KS.** Hematocrit’s (Hct’s) need to be distributed normally to find a confidence interval. In the additional example of Section 11.8, we examined postoperative Hct on pyloromyotomized neonates\(^6\) and concluded that Hct was not different for laparoscopic versus open surgeries; thus we pool the data. The \( n = 16 \) Hct readings in ascending order are 25.6, 29.7, 32.0, 32.0, 32.1, 32.7, 33.9, 34.0, 38.3, 38.8, 39.0, 42.0, 43.3, 43.9, 46.7, 52.0. Test for normality using the Kolmogorov–Smirnov test of normality.
13.6 Test of Normality of a Distribution

Table 13.10 Human immunovirus per milliliter of blood data in the format of Table 13.9.

| \( \ln h \) and \( x \) | \( k \) | \( F_n(x) \) | \( z \) | \( F_e(x) \) | \( |F_n - F_e| \) | \( \ln h \) and \( x \) | \( k \) | \( F_n(x) \) | \( z \) | \( F_e(x) \) | \( |F_n - F_e| \) |
|-----------------|-----|-------------|-----|-------------|--------------|-----------------|-----|-------------|-----|-------------|--------------|
| 5.2575          | 1   | 0.0222      | -2.0437 | 0.0205      | 0.0017       | 9.5623          | 23  | 0.5111      | 0.2902 | 0.6141      | 0.1030       |
| 5.3613          | 2   | 0.0444      | -1.9058 | 0.0283      | 0.0161       | 9.5988          | 25  | 0.5556      | 0.3104 | 0.6219      | 0.0663       |
| 5.6095          | 3   | 0.0667      | -1.7938 | 0.0364      | 0.0302       | 9.7200          | 26  | 0.5778      | 0.3778 | 0.6472      | 0.0694       |
| 5.8377          | 4   | 0.0889      | -1.7790 | 0.0376      | 0.0513       | 9.7600          | 27  | 0.6000      | 0.4000 | 0.6554      | 0.0554       |
| 6.2206          | 5   | 0.1111      | -1.5663 | 0.0586      | 0.0525       | 9.7712          | 28  | 0.6222      | 0.4062 | 0.6777      | 0.0355       |
| 6.4599          | 6   | 0.1333      | -1.4334 | 0.0759      | 0.0575       | 10.1036         | 29  | 0.6444      | 0.5909 | 0.7227      | 0.0783       |
| 7.4719          | 7   | 0.1556      | -0.8711 | 0.1918      | 0.0363       | 10.1956         | 30  | 0.6667      | 0.6420 | 0.7396      | 0.0729       |
| 7.7841          | 8   | 0.1778      | -0.6977 | 0.2427      | 0.0649       | 10.1963         | 31  | 0.6889      | 0.6424 | 0.7397      | 0.0508       |
| 7.8466          | 9   | 0.2000      | -0.6301 | 0.2537      | 0.0537       | 10.4942         | 32  | 0.7111      | 0.8079 | 0.7904      | 0.0793       |
| 7.9124          | 10  | 0.2222      | -0.6264 | 0.2655      | 0.0433       | 10.6105         | 33  | 0.7333      | 0.8725 | 0.8085      | 0.0752       |
| 8.0830          | 11  | 0.2444      | -0.5316 | 0.2975      | 0.0530       | 10.6144         | 34  | 0.7556      | 0.8747 | 0.8091      | 0.0536       |
| 8.1026          | 12  | 0.2667      | -0.5208 | 0.3013      | 0.0346       | 10.6794         | 35  | 0.7889      | 0.9108 | 0.8188      | 0.0410       |
| 8.1158          | 13  | 0.2889      | -0.5134 | 0.3038      | 0.0149       | 10.8751         | 36  | 0.8000      | 1.0195 | 0.8460      | 0.0460       |
| 8.4036          | 14  | 0.3111      | -0.3536 | 0.3618      | 0.0507       | 10.9183         | 37  | 0.8222      | 1.0435 | 0.8516      | 0.0294       |
| 8.4602          | 15  | 0.3333      | -0.3221 | 0.3736      | 0.0404       | 10.9501         | 38  | 0.8444      | 1.0612 | 0.8557      | 0.0113       |
| 8.5834          | 16  | 0.3556      | -0.2537 | 0.3999      | 0.0443       | 11.2159         | 39  | 0.8667      | 1.2089 | 0.8866      | 0.0200       |
| 8.6778          | 17  | 0.3778      | -0.2012 | 0.4203      | 0.0425       | 11.2270         | 40  | 0.8889      | 1.2150 | 0.8878      | 0.0011       |
| 8.8796          | 18  | 0.4000      | -0.0891 | 0.4645      | 0.0645       | 11.2511         | 41  | 0.9111      | 1.2284 | 0.8904      | 0.0208       |
| 8.9383          | 19  | 0.4222      | -0.0565 | 0.4775      | 0.0552       | 11.3553         | 42  | 0.9333      | 1.2863 | 0.9008      | 0.0325       |
| 8.9408          | 20  | 0.4444      | -0.0555 | 0.4780      | 0.0336       | 11.3634         | 43  | 0.9556      | 1.2907 | 0.9016      | 0.0540       |
| 9.1607          | 21  | 0.4667      | 0.6701  | 0.5267      | 0.0601       | 11.4586         | 44  | 0.9778      | 1.3437 | 0.9105      | 0.0673       |
| 9.4340          | 22  | 0.4889      | 0.2189  | 0.5866      | 0.0977       |               |     |             |        |             |              |

Using the chi-square goodness-of-fit test (large sample test of normality)

Characteristics of the Chi-Square Goodness-of-Fit Test

The time-honored Pearson’s goodness-of-fit test is a relatively easy concept but has limitations, which include that it is an approximate test and that the calculated value of \( \chi^2 \) depends on the user’s choice of interval widths and starting points. The test is similar to a chi-square contingency table test in that it tests the difference between expected and observed values, using, however, areas under the normal curve for expected values.

Example Posed: Is the Distribution of Ages of the 301 Prostate Patients Normal?

We ask if the data are just normal, not drawn from a specific normal with theoretically given \( \mu \) and \( \sigma \), so we use \( m = 66.76 \) and \( s = 8.10 \) from DB1. With this large sample, we can use the chi-square goodness-of-fit test.
METHOD: CHI-SQUARE GOODNESS-OF-FIT TEST OF NORMALITY

The hypotheses are $H_0$: The distribution from which the sample was drawn is normal or, alternatively, is normal with parameters $\mu$ and $\sigma$, and (two-tailed) $H_1$: The distribution is different. Choose $\alpha$. Look up the critical $\chi^2$ in Table III.

1. We define the data intervals, say $k$ in number, as we would were we to form a histogram of the data. We form a blank table in the format of Table 13.11.

2. To use a normal probability table, we standardize the ends of the intervals by subtracting the mean and dividing by the standard deviation. The “expected” normal is usually specified by the sample $m$ and $s$, although it could be specified by a theoretical $\mu$ and $\sigma$.

3. To relate areas under the normal curve to the intervals, we find the area to the end of an interval from a table of normal probabilities, such as Table I, and subtract the area to the end of the preceding interval.

4. To find the frequencies expected from a normal fit (name them $e_i$), we multiply the normal probabilities for each interval by the total number of data $n$.

5. We tally the number of data falling into each interval and enter the tally numbers in the table. Name these numbers $n_i$.

6. Calculate a $\chi^2$ value [the pattern is similar to Eq. (9.2) in pattern] using Eq. (13.7):

$$\chi^2 = \sum \frac{(n_i - e_i)^2}{e_i} = \sum \frac{n_i^2}{e_i} - n,$$

(13.7)

where the first form is easier to understand conceptually and the second is easier to compute.

7. If calculated $\chi^2$ is greater than critical $\chi^2$, reject $H_0$; otherwise, conclude that there is insufficient evidence to determine that the data did not arise from a normal distribution.

EXAMPLE COMPLETED: IS THE DISTRIBUTION OF AGES NORMAL?

We choose $\alpha = 0.05$. The critical value of chi-square uses 9 $df$ (number of intervals – 1). From Table 13.12 (or Table III), chi-square (9 $df$) for $\alpha = 0.05$ is 16.92.

1. We define 10 intervals as $<50$, 50 up to but not including 55, . . . , 85 up to but not including 90, and $\geq 90$. We form a blank table in the format of Table 13.11 and complete it in the following steps to become Table 13.13.

Table 13.11 Format for table of values required to compute the chi-square goodness-of-fit statistic.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Standard normal $z$ to end of interval</th>
<th>Probability</th>
<th>Expected frequencies ($e_i$)</th>
<th>Observed Frequencies ($n_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. To use a normal probability table, we need to standardize the ends of the intervals by subtracting the mean and dividing by the standard deviation. We standardize the end of the first interval by \((50 - 66.76)/8.1 = -2.0691\). The second is \((55 - 66.76)/8.1 = -1.4519\).

3. We need to relate areas under the normal curve to the intervals, which we do by finding the area to the end of an interval and subtracting the area to the end of the preceding interval. Table I includes only positive \(z\)-values, so we use the normal symmetry: we change the sign of \(z\) and use \(\alpha\) instead of \(1 - \alpha\). For the first interval, \(z = -2.0691\) interpolates as \(0.023 + 0.691 \times (0.018 - 0.023) = 0.0195\). There is no prior interval yielding a probability to subtract, so 0.0195 is the probability for the first interval. For the second, \(z = -1.4519\) gives \(0.081 + 0.519 \times (0.067 - 0.081) = 0.0737\). Subtracting 0.0195 for the prior area results in a second-interval probability of 0.0542.

4. To obtain the expected frequencies \((e_i)\), we multiply the probability associated with an interval by \(n\). \(e_1 = 0.0195 \times 301 = 5.87\), etc.

5. To obtain the observed frequencies \((n_i)\), we tally the data for each data interval and enter the frequencies \((3, 17, 32, \ldots)\) in the table.

6. The chi-square statistic is obtained as
CHAPTER 13 Tests on variability and distributions

Table 13.14 Human immunovirus per milliliter blood data in the format of Table 13.11.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Standard normal z to end of interval</th>
<th>Probability</th>
<th>Expected frequencies (e_i)</th>
<th>Observed frequencies (n_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5—&lt; 6</td>
<td>-1.6864</td>
<td>0.0459</td>
<td>2.0655</td>
<td>5</td>
</tr>
<tr>
<td>6—&lt; 7</td>
<td>-1.1318</td>
<td>0.0830</td>
<td>3.7350</td>
<td>2</td>
</tr>
<tr>
<td>7—&lt; 8</td>
<td>-0.5772</td>
<td>0.1530</td>
<td>6.8850</td>
<td>4</td>
</tr>
<tr>
<td>8—&lt; 9</td>
<td>-0.0226</td>
<td>0.2091</td>
<td>9.4095</td>
<td>10</td>
</tr>
<tr>
<td>9—&lt; 10</td>
<td>0.5320</td>
<td>0.2117</td>
<td>9.5265</td>
<td>8</td>
</tr>
<tr>
<td>10—&lt; 11</td>
<td>1.0866</td>
<td>0.1588</td>
<td>7.1460</td>
<td>10</td>
</tr>
<tr>
<td>11—&lt; 12</td>
<td>1.6412</td>
<td>0.0882</td>
<td>3.9690</td>
<td>6</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum \frac{n_i^2}{e_i} - n = \frac{3^2}{5.87} + \frac{17^2}{16.31} + \ldots - 301 = 7.895 \]

7. The calculated chi-square of 7.895 is less than the critical chi-square of 16.92, so the result is taken as not significant and we do not reject \( H_0 \). The \( p \)-value was computed on a statistical software package to be 0.555.

**ADDITIONAL EXAMPLE: NORMALITY OF HUMAN IMMUNOVIRUS DATA**

Let us test the normality of the ln HIV data\(^5\) introduced earlier using the chi-square goodness-of-fit test, even though the sample size is just borderline smaller than we would choose. We used 7 intervals, as can be seen in Fig. 13.2B, leading to 6 \( df \). We use \( \alpha = 0.05 \). From Table 13.12 (or Table III) the critical value of \( \chi^2_{6, df} = 12.59 \). Recall that \( m = 9.0408 \) and \( s = 1.8031 \). Results in the format of Table 13.11 appear in Table 13.14. By substituting in Eq. (13.7), we find

\[ \chi^2 = \sum \frac{n_i^2}{e_i} - n = \frac{5^2}{2.0655} + \frac{2^2}{3.7350} + \ldots + \frac{6^2}{3.9690} - 45 = 10.91. \]

Since 10.91 < 12.59, we conclude that deviation from normality is not demonstrated. (The actual \( p \)-value = 0.091.)

**Exercise 13.5**

**\( \chi^2 \) Goodness-of-fit test.** In the example of Section 11.5, we assumed that the patient ages in the three groups were distributed normal. In particular, the distribution of \( n = 164 \) PSAs in the equivocal 4–10 range was uncertain. Test this distribution for normality using the \( \chi^2 \) goodness-of-fit test at the \( \alpha = 0.05 \) level with the frequencies tallied for nine intervals from Table 13.15. (The notation “45—< 50” represents the interval including 45 and up to but not including 50, etc.). \( m = 66.2988 \) and \( s = 7.8429 \).
The two-sample Kolmogorov–Smirnov test

The two-sample KS test will compare two data sets to decide whether they were sampled from population distributions of the same shape. (The one-sample form to test normality can be seen in Section 13.6.)

EXAMPLE POSED: ARE TWO PROSTATE-SPECIFIC ANTIGEN SAMPLES FROM DB1 DISTRIBUTED THE SAME?

We want to know if the distribution from which the first 10 PSA values were drawn is the same as that for the next 16 PSA values. We are suspicious that the first 10 are not representative of the remainder and may be subject to a sampling bias. The question of distribution shape accompanied by the tests for means and standard deviations will answer the question.

METHOD: THE TWO-SAMPLE KOLMOGOROV–SMIRNOV TEST

The hypotheses are $H_0$: The two samples arose from the same population distributions, and the two-tailed $H_1$: The samples arose from different population distributions. Choose $\alpha$. The sample sizes are $n_1$ and $n_2$, where $n_1$ is the larger. Calculate the critical value for $\alpha = 0.01, 0.05$, and $0.10$ from Eqs. (13.8a), (13.8b), or (13.8c), respectively. (These critical values are approximations. They are adequate for $n_2 > 10$. For smaller $n$'s, they may be used if the calculated statistic is much greater or lesser than the critical value; if borderline, the user should arrange for calculation using a statistical software package.)

$$1.63 \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \text{ for } \alpha = 0.01 \quad (13.8a)$$

$$1.36 \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \text{ for } \alpha = 0.05 \quad (13.8b)$$

$$1.22 \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \text{ for } \alpha = 0.10 \quad (13.8c)$$

The test consists of calculating the cumulative data frequencies for the two samples and finding the probability of the greatest difference between these cumulative frequencies.

1. We form a blank table in the format of Table 13.16. Combine the two data sets, keeping track of which datum belongs to which sample, and enter in ascending order.

(Continued)
2. Going down the data list, each time a datum belonging to sample 1 is different from the datum above it, record an entry for $k_1$, the number of data in sample 1 preceding it. Repeat the process for sample 2 data, recording entries for $k_2$.

3. For every $k_1$, calculate $F_1 = k_1/n_1$ and enter it under the $F_1$ column. Wherever a blank appears (corresponding to sample 2 data), write down the $F_1$ from the line above. Repeat the process for sample 2 data, recording entries for $F_2$ and writing down the $F_2$ value from the preceding line to fill in blanks. The $F$'s are the cumulative sums for the two samples.

4. Calculate and record $|F_1 - F_2|$, the absolute difference (difference without any minus signs) between the two cumulative sums, for every datum.

5. The test statistic is the largest of these differences; call it $L$.

6. If $L >$ critical value, reject $H_0$. Otherwise, do not reject $H_0$.

The two-sample Kolmogorov—Smirnov on large samples
If the sample sizes are large, the number of differences to be computed may be reduced to a manageable number by collapsing the data into class intervals as one would do in making a histogram. The method, however, then acquires the faults of the goodness-of-fit test: it becomes an approximation and becomes dependent on the choice of interval position and spacing.

**EXAMPLE COMPLETED: ARE TWO PROSTATE-SPECIFIC ANTIGEN SAMPLES FROM DB1 DISTRIBUTED THE SAME?**
We choose $\alpha = 0.05$. The larger sample is designated number 1. Then $n_1 = 16$ PSA readings, which are 5.3, 6.6, 7.6, 4.8, 5.7, 7.7, 4.6, 5.6, 8.9, 1.3, 8.5, 4.0, 5.8, 9.9, 7.0, and 6.9, and $n_2 = 10$, with the PSA values from Table DB1.1. We use Eq. (13.8b) to find the critical value:

$$1.36 \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.36 \sqrt{\frac{26}{160}} = 0.5482.$$  

| Ordered data | $k_1$ | $k_2$ | $F_1$ | $F_2$ | $|F_1 - F_2|$ |
|--------------|-------|-------|-------|-------|------------|
|              | :     | :     | :     | :     | :          |
|              | :     | :     | :     | :     | :          |

Table 13.16 Format for table of data and calculations required for the two-sample Kolmogorov—Smirnov test of equality of distributions.
1. Using the format (headings) as in Table 13.16, record in Table 13.17 the combined data sets in ascending order, keeping track of which datum belongs to which sample.

2. Going down the data list, each time a datum belonging to sample 1 is different from the datum above it, record an entry for $k_1$, the number of data in sample 1 preceding it. 1.3 has no data preceding it, so we enter $k_1 = 0$. 4.0 has one datum preceding it; we skip to 4.6, which has two sample 1 data preceding it. And so forth until PSA value 9.9, for which $k_1 = 15$. Repeat the process for sample two data, recording entries for $k_2$.

3. For every $k_1$, calculate $F_1 = k_1/n_1$ and enter it under the $F_1$ column. Wherever a blank appears (corresponding to sample 2 data), write down the $F_1$ from the line above. $F_1 = 0/16 = 0$ for the first PSA value. $F_1 = 1/16 = 0.0625$ for the second. The third PSA value is from sample 2 and no $k_1$ was recorded, so we repeat...
0.0625. We follow the same process for sample 2 data, recording entries for \( F_2 \) and writing down the \( F_2 \) value from the preceding line to fill in blanks. The \( F \)'s are the cumulative sums for the two samples.

4. Calculate and record \(|F_1 - F_2|\), the absolute difference (difference without any minus signs) between the two cumulative sums for every datum. For example, for the fourth PSA value, 4.4, \(|F_1 - F_2| = |0.0625 - 0.1| = |-0.0375| = 0.0375\).

5. The test statistic \( L \) is the largest of these differences. We can see that it is 0.3375.

6. The \( L \) of 0.3375 is less than 0.5482, the critical value, so we do not reject \( H_0 \).

(A statistical software package gives \( p \)-value = 0.397.)

**ADDITIONAL EXAMPLE: A POTENTIAL VACCINE FOR HUMAN IMMUNOVIRUS**

The study included a control sample of \( n_1 = 24 \) patients (placebo saline injection) and an experimental vaccine injection sample of \( n_2 = 20 \) patients, randomly allocated into the two groups.\(^5\) The knowledge of whether the experimental and control groups follow the same distribution aided in the development of the physiological theory. The number of HIV per milliliter of blood (denoted \( h \)) was measured before treatment and at a fixed period of time after treatment. Its natural logarithm (ln \( h \)) was calculated in order to reduce the extreme skewness of the frequency distribution. Then the difference \( d(\ln h) = \ln h \) before \( -\ln h \) after was taken. A frequency plot of the two samples appears in Fig. 13.3. The distributions are not obviously different. We use the KS test of equality of distributions with a two-sided \( \alpha = 0.05 \).

The data and analysis operations appear in Table 13.18 in the format of Table 13.16. The values of \( d(\ln h) \) are put in order. The number of data in the respective sample preceding the current observation is listed for control \((k_1)\) and vaccine \((k_2)\). \( F_1 = k_1/24, \ F_2 = k_2/20, \) and \(|F_1 - F_2|\) are entered. Inspection of the last column shows that the largest value is \( L = 0.4000 \), indicated by an asterisk. The critical value for the test is found from Eq. (13.8b):

![Figure 13.3 Plots of distributions of control (A) and vaccine-treated (B) differences in log number HIV per milliliter of blood before minus after treatment. HIV, Human immunovirus.](image-url)
Since 0.4000 < 0.4118, we do not reject $H_0$ and proceed on the premise that the distributions are not different.

**Exercise 13.6**

Are the distributions of functionality the same for two types of ankle repair? An orthopedist installs hardware in the repair of broken ankles in $n = 19$ patients. He randomly selects two groups: leave the hardware in place after adequate healing ($n_1 = 10$) and remove it ($n_2 = 9$). He judges the post healing percent functionality of the ankle joint. In assessing the relative success of the two groups, he observes that their distributions are far from normal in shape. However, are the distributions both the same, whatever they may be? His data are as in Table 13.19.

Perform a KS test of equality of distributions using $\alpha = 0.05$. 

13.7 Test of Equality of Two Distributions
CHAPTER 13 Tests on variability and distributions

REFERENCES

14.1 WHAT ARE ASSOCIATION AND AGREEMENT?

Association

Association between two (or more) variables is just lack of independence: occurrence in one variable provides some information about occurrence in the other. For example, measurements of a patient’s systolic blood pressure (SBP) at bedtime and upon awakening are associated. Knowing the value of SBP at bedtime does not tell you the value of SBP in the morning, but clearly, the two are not independent. A number of measures of association have been addressed earlier in this book. For example, correlation coefficients were met in Chapter 5, Descriptive statistics, and the odds ratio was met in Chapter 10, Risks, odds, and receiver operating characteristic curves.

It has been said before but bears repeating: association does not imply causation. Causation may be present, but the fact of association does not prove it. An association between the dosage of an nonsteroidal antiinflammatory drug (NSAID) and a reduction in marking of a pain analog scale is not evidence that the NSAID reduced the pain; the pain could have waned in the absence of treatment.

Agreement

Agreement between variables is the special case of association in which the same variable’s measurement is recorded on the same issue. For perfect agreement, all paired readings would be identical. For example, SBP measured on the same patient at the same time of day on two successive days could be considered agreement, whereas SBP and diastolic BP (DBP) at the same time could not (the paired readings could not be identical). Some examples of data susceptible to agreement measures might be cancer stage on a sample of patients assessed by two pathologists; measuring the same clinical reading on a sample of patients by two competing instruments (perhaps wrist vs upper arm BP cuffs); selecting a diagnosis on a sample of patients by clinical indicators versus by laboratory tests; or determining if BP in the presence of a white-coated physician agrees with that in the presence of a nurse in scrubs.
Using association measures to assess agreement

A measure of association can act as a measure of agreement in some circumstances while not in others. Whether or not association indicates agreement depends on the measures used and the interpretation of the outcome. A correlation coefficient between readings on patient temperatures and C-reactive protein levels would indicate association but not agreement. The correlation coefficient between readings on patient temperatures measured orally and rectally would indicate a pattern of agreement. Would the correlation coefficient between a patient’s oral temperatures taken mid-morning and midafternoon over time indicate agreement? That depends on the interpretation. If we were asking if the morning and afternoon temperature followed the same pattern, the correlation would give us information about agreement. If we were monitoring to detect an abnormality, it would not. (There are better ways to monitor; the example is illustrative.)

Reliability as agreement

Reliability has many aspects, most of which are not related to association or agreement. The use of the word “reliability” should be distinguished between the concept of being reliable and the measurement of reliability. Reliability measurement in the sense of repeatability is a form of agreement. Repeatability shows reliability in an instrument if it gives approximately the same readings over and over under the same conditions. In contrast, reliability may be shown by a measure of association when we reach the same conclusion while using different variables that purport to provide the same information.

This chapter addresses descriptors of association and agreement

Measures of association and agreement are descriptive statistics. There exist tests to determine significance for several of these measures, discussed elsewhere in this book. This chapter will focus on descriptive measures; methods of testing will not be repeated here.

14.2 CONTINGENCY AS ASSOCIATION

Contingency tables

Contingency tables in the context of detecting the presence of significant association, that is, tests of contingency, were addressed in Chapter 9, Tests on categorical data. Descriptive statistics, addressed in Chapter 10, Risks, odds, and receiver operating characteristic curves, were provided as rates for individual contingency table cells and as odds ratios for pairs of cells within a category. However, here we are interested primarily in descriptive measures for the full set of cells composing a contingency table as a whole. For example, in the coccidioidomycosis data shown in Table 10.2, the rate
of the disease in patients exposed to the suspicious dust is \( \frac{20}{55} = 0.36 \) and the odds ratio of the disease for patients exposed to the dust over those not exposed is 5.3. However, these statistics do not give a measure of overall association between coccidioidomycosis and the dust.

**2 × 2 Tables**

If the contingency table of interest is \( 2 \times 2 \), the phi coefficient \( \Phi \), suggested by Karl Pearson in the early 1900s, is given by

\[
\Phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_1n_2n_1n_2}}
\]

(14.1)

\( \Phi \) lies between \(-1\) and 1 as does Pearson’s correlation coefficient and is interpreted much the same way. A value of 0 implies no association, a value of 1, perfect association, and a value of \(-1\), perfect inverse association.

**EXAMPLE: ASSOCIATION BETWEEN SUSPICIOUS DUST AND COCCIDIOIDOMYCOSIS**

As an example, we substitute the coccidioidomycosis and suspicious dust data in Eq. (14.1). We find \( \Phi = 0.26 \). There is a perceptible association.

**2 × c Tables**

It can be shown that

\[
\Phi = \sqrt{\frac{\chi^2}{n}}
\]

(14.2)

which generalizes to a \( 2 \times c \) contingency table. Thus to calculate \( \Phi \), find Pearson’s \( \chi^2 \), divide by \( n \) (recall also denoted by \( n \ldots \)), and take the square root. However, in this case, \( \Phi \) may range only from 0 to 1.

**EXAMPLE: USE OF SMOKELESS TOBACCO BY ETHNIC GROUP**

As an example, consider the use versus nonuse of smokeless tobacco among three ethnic groups, European, African, and Hispanic. Data appeared in the \( 2 \times 3 \) Table 9.7. \( \chi^2 \) was calculated as 92.46 for the \( n = 2004 \) subjects. Upon substituting in Eq. 14.2, we find that \( \Phi = 0.21 \), a noticeable but not impressive level of association. However, the test in Section 9.5 showed \( p \)-value \(< 0.001 \). This apparent discrepancy illustrates the
CHAPTER 14 Measuring association and agreement

extensive influence of sample size on the $\chi^2$ test of contingency; almost any level of association can lead to test significance if the sample size is large enough. An investigator should use measures of association level in interpreting the contingency test results.

Exercise 14.1

For the $2 \times 2$ table in Table 5.3 [constructed from database (DB) 2] illustrating the frequency of patients with no nausea versus nausea against drug versus placebo, calculate and interpret $\Phi$ using Eqs. (14.1) and (14.2).

Larger ($r \times c$) tables with nominal categories

Harald Cramér published in 1946 his statistic $V$ that generalizes Eq. (14.2) to greater numbers of categories, provided these categories are not rankable, that is, cannot be placed in a logical order. This would be true of ethnic groups but not of cancer stages. If we denote by $k$ the lesser of $r$ and $c$, then Eq. (14.2) becomes simply

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}.$$  \hspace{1cm} (14.3)

As with $\Phi$, $V$ varies from 0 (no association) to 1 (perfect association).

EXAMPLE: CHOOSING A THERAPIST FOR PSYCHIATRY RESIDENTS

As an example, let us recall Table 9.9, showing data about who chose the type of therapist for $n = 155$ psychiatry residents. We have a $4 \times 3$ contingency table, so $k = 3$. $\chi^2$ was calculated as 24.45, so $V = \sqrt{[24.45/(155 \times 2)]} = 0.28$. There is a meaningful association between who was selecting the therapist and the type of therapist chosen.

Exercise 14.2

DB18 lists the presence or absence of HSV2 in human immunodeficiency virus patients by ethnic origin. If we denote European—Americans by $E$, African—Americans by $A$, and Hispanic—Americans by $H$, the data are shown in Table 14.1.

Calculate $V$.

Table 14.1 The presence or absence of HSV2 in human immunodeficiency virus patients by ethnic origin.

<table>
<thead>
<tr>
<th></th>
<th>$E$</th>
<th>$A$</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No HSV2</td>
<td>59</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>HSV2</td>
<td>46</td>
<td>34</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>105</td>
<td>52</td>
<td>28</td>
</tr>
</tbody>
</table>
Larger \((r \times c)\) tables with rankable categories

Suppose, we have an \(r \times c\) table, but the classes in one of the dimensions have a natural order. To use \(\Phi\) would be to lose the information contained in the ranking. An appropriate method is best explained by example.

A statistic called gamma \((\gamma)\) has been developed\(^2\) to describe the association of two ordered measures. \(\gamma\) is calculated as

\[
\gamma = \frac{C - D}{C + D}, \tag{14.4}
\]

the difference between concordant number of pairs \((C)\) and discordant number of pairs \((D)\) divided by the number of all pairs. \(\gamma\) can range from \(-1\) (perfect discordance) through \(0\) (no concordance) to \(1\) (perfect concordance). \(C\) and \(D\) are defined in the following example.

**EXAMPLE: CLINIC VERSUS LABORATORY VALUES OF INTERNATIONAL NORMALIZED RATIO IN DIABETIC PATIENTS**

Let us consider the International Normalized Ratio (INR) below range \((-1)\), in range \((0)\), and above range \((1)\) frequencies for our clinic compared with our laboratory given in DB12, in which the range status of the two data sources can be tabulated as Table 14.2.

If a cell entry increases in both sources, it is called concordant; if a cell entry decreases in one while increasing in the other, it is called discordant. For example, from the \((0,0)\) position in the center, moving down and to the right represents an increase in both sources \([\text{to } (1,1)]\), that is a concordance. There are \(44 \times 12\) pairings in which this event occurs. Moving down (clinic increases) but to the left (laboratory decreases) represents a discordance. There are \(44 \times 0\) pairings in which this event occurs. Thus the number of concordant pairs \((C)\) is given as the number in every cell times the sum of all cell entries lower and right: \(18(44 + 7 + 2 + 12) + 1(7 + 12) + 20(2 + 12) + 44(12) = 1997\). The number of discordant pairs \((D)\) is given as the number

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Laboratory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(-1)</td>
</tr>
<tr>
<td>(-1)</td>
<td>18</td>
</tr>
<tr>
<td>(0)</td>
<td>20</td>
</tr>
<tr>
<td>(1)</td>
<td>0</td>
</tr>
</tbody>
</table>
in every cell times the sum of all cell entries lower and left: $1(20 + 0) + 0(20 + 44 + 0 + 2) + 44(0) + 7(0 + 2) = 34$. For the clinic versus laboratory INR ranges, $\gamma = 0.97$, quite high; they agree very well.

**Exercise 14.3**

*Let us consider the alterations in BP of pregnant women from changing position from upright to supine, given in Table 9.10. Is SBP or DBP more closely associated with a drop in BP from lying down? (Moving down in the table represents going from SBP to DBP.)*

**Exercise 14.4**

*For the $2 \times 5$ table in DB2 for nausea scores of 1–5 against drug versus placebo, calculate and interpret $\gamma$.*

### 14.3 Correlation as Association

**Correlation: Continuous or rank-order data**

Correlation coefficients, met in Chapter 5, Descriptive statistics, and treated more thoroughly in Chapter 15, Linear regression and correlation, are statistics used to describe the similarity of patterns between two variables, say $x$ and $y$.

When $x$ and $y$ are continuous and the $x$-to-$y$ relationship is a straight line, *Pearson’s correlation* statistic for a sample, denoted $r$, may be defined as the covariance divided by the product of standard deviations:

$$r = \frac{s_{xy}}{s_x s_y}$$

(14.5)

when $x$ and $y$ are ranks, the equivalent *rank correlation coefficient* is

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

(14.6)

without ties, where $d_i$ denotes the difference between members of the $i$th ranked pair and $n$, the sample size.

Some renditions use $\rho$ to denote $r_s$; others use $\rho$ to denote the tetrachoric correlation coefficient $r_t$. In this book the Greek letter $\rho$ has been used to denote the theoretical or population correlation coefficients leaving Roman letters to denote sample coefficients. The subscript $s$ in $r_s$ honors originator Charles Spearman.\(^3\)
There exists another form of rank correlation coefficient, Kendall’s $\tau$. $r$, and $\tau$ use different scales and therefore yield slightly different values of the coefficient. While Kendall’s coefficient has a more meaningful interpretation, Spearman’s is easier to calculate and therefore is used more often.

*Kendall’s $\tau$* arises from a count of concordances and discordances, somewhat similar to the $\gamma$ statistic. Let us start with two columns of $n$ paired rankings. Place the first in increasing order so that, in the result, $x_r, y_r$ is 1, 2, 3, ... . Attach alphabetic labels A, B, C, ... to $y_r$ ordered from top to bottom. Eq. (3.8) tells us that we have $n(n - 1)/2$ possible pairs of letters. Make a list of such pairs, AB, AC, ..., BC, BD, ... . Record a 1 or $-1$ next to each pair such that if the ranks, that is, the $y_r$ values, show an increase from the first member of the pair to the second, record 1, and if they show a decrease, record $-1$. The number of 1s is the number of concordances $C$, and the number of $-1$s is the number of discordances $D$.

The correlation coefficient $\tau$ is the difference between $C$ concordances and $D$ discordances divided by the number of pairs, or

$$\tau = \frac{2(C - D)}{n(n - 1)}$$  \hspace{1cm} (14.7)

While $\tau$ is more difficult to calculate than $r$, we now have a simple and very useful interpretation available: the odds ratio of concordance to discordance (likelihood of positive association to negative association) is $C/D$.

**EXAMPLE: HOP DISTANCES BETWEEN INJURED AND HEALTHY LEGS**

As an example, let us consider correlation coefficients between the $n = 8$ distances covered in hops on a leg that had undergone hamstring or quadriceps surgery versus the healthy leg in DB10. Upon ranking the operated leg column and adding ranks, we obtain Table 14.3.

**Table 14.3** Distance (cm) hopped by eight patients on postoperated leg (op) and healthy (nonoperated) leg (non), ranked by op, with associated ranks for each column.

<table>
<thead>
<tr>
<th>Op dist</th>
<th>Nonop dist</th>
<th>Op ranks</th>
<th>Nonop ranks</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>319</td>
<td>424</td>
<td>1</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>360</td>
<td>450</td>
<td>2</td>
<td>2</td>
<td>B</td>
</tr>
<tr>
<td>385</td>
<td>481</td>
<td>3</td>
<td>3</td>
<td>C</td>
</tr>
<tr>
<td>436</td>
<td>504</td>
<td>4</td>
<td>4</td>
<td>D</td>
</tr>
<tr>
<td>489</td>
<td>527</td>
<td>5</td>
<td>5</td>
<td>E</td>
</tr>
<tr>
<td>523</td>
<td>568</td>
<td>6</td>
<td>7</td>
<td>F</td>
</tr>
<tr>
<td>541</td>
<td>553</td>
<td>7</td>
<td>6</td>
<td>G</td>
</tr>
<tr>
<td>569</td>
<td>606</td>
<td>8</td>
<td>8</td>
<td>H</td>
</tr>
</tbody>
</table>
Calculations for Pearson’s correlation coefficient \( r \), using Eq. (14.5), are \( s_{\text{op}} = 91.74 \), \( s_{\text{non}} = 61.48 \), \( s_{\text{op,non}} = 5539.75 \), \( r = 0.982 \). The distances are very highly correlated. Eq. (14.6) for Spearman’s rank correlation coefficient yields \( r_s = 1 - 6 \times 2/8 (8^2 - 1) = 0.976 \). The rank coefficient is similarly high. To find Kendall’s \( \tau \), we count the number of concordances (\( C \)) and discordances (\( D \)). Going from A to B, the Nonop ranks go from 1 to 2; as this is increasing, we count the first concordance. From A to C, the ranks go from 1 to 3, also increasing; we add a second concordance. We continue in this fashion through A to H and then start on B to C, etc. Actually, we can see from inspection that only 1 pairing—F to G—will be decreasing, so \( D = 1 \). Subtracting 1 from the sample size \( n = 8(8 - 1)/2 = 28 \) gives \( C \) as 27. From Eq. (14.7), \( \tau = (27 - 1)/28 = 0.929 \). This result is a less than \( r_s \), but still quite high. In addition to the value of \( \tau \), we can obtain the odds ratio as \( \text{OR} = C/D = 27 \). Odds are estimated to be 27 to 1 that an increase in hop distance on the operated leg will be paired with an increase in hop distance on the nonoperated leg.

We note that the high level of correlational association lends itself to easy interpretation. Two comments are in order.

First, we have seen a high correlation between leg performances. Does this mean that the knowledge of one performance will tell us approximately what the other is? Not at all. We note that the mean distances are 453 cm on the operated leg and 514 cm on the other, quite discrepant. The correlation coefficient tells us that the patterns of behavior are similar, not the actual values.

Second, if the rank coefficients had been of the order of 0.50 (implying \( C \) is about 21 and \( D \) about 7), the user might ask if this is high or low relative to desirability; interpretation would not have been as straightforward. In this case, \( \text{OR} = 21/7 = 3 \). We could say that an agreement between rank performances on an operated leg and a nonoperated leg would be 3 to 1 in ratio to a disagreement. This helps us considerably in assessing the importance of the 0.50 correlation coefficient.

**Exercise 14.5**

*From DB10, use the time to perform the hops rather than the distance and calculate \( r, r_s, \tau, \) and the OR.*

### 14.4 CONTINGENCY AS AGREEMENT

**Agreement among contingency categories**

The phi coefficient and Cramér’s \( V \) were addressed in Section 14.2 as measures of association among (nonrankable) categories in a contingency table, \( \Phi \) for a \( 2 \times 2 \) or \( 2 \times c \) table and \( V \) for a larger table. The statistic \( \gamma \) was presented for contingency tables with rankable entries. Inasmuch, as agreement is a form of association, theoretically all three of these methods could be used to measure agreement, so long as the
data meet the agreement requirement of measuring the same phenomenon to answer the same question.

Another option to measure agreement between categories is the tetrachoric correlation coefficient \( r_t \) [see Eq. (5.12)]. However, \( r_t \) assumes that \( x \) and \( y \) are continuous (bivariate normal) variables separated into binary categories by cut points and is not so often appropriate for measuring categorical agreement.

**EXAMPLE: CANCER SEVERITY RATED BY THREE RADIOLOGISTS**

The pairing issue can best be clarified through an example. Let us consider DB35 that shows cancer severity as 0—5 rated independently from radiographs on 100 patients by three radiologists, G, C, and S. The readings are paired (matched) so that the severity on each patient is given for G, C, and S. If we set up a contingency table of rating frequency (vertical dimension) by rater (horizontal dimension), we lose the information of pairing. To maintain pairing, we can set up a table of two raters, say, G (vertical dimension) by C (horizontal dimension), with the joint rating entered in the table. We may then estimate the agreement by \( \gamma \) (as 0.979). However, we cannot set up a two-way contingency table that maintains pairing for all three raters. In order to measure agreement on paired data, we should use Cohen’s kappa, introduced in Section 14.6.

**14.5 CORRELATION AS AGREEMENT**

The correlation coefficients \( r, r_s, \) and \( \tau \) addressed in Section 14.3 can measure agreement as a form of association so long as the paired variables in the coefficient measure the same phenomenon to answer the same question. For example, if a pair of variables record SBP and DBP from an upper arm cuff device, the correlation coefficient indicates association, not agreement. If the pair of variables record SBP from an upper arm cuff and a wrist cuff, the correlation coefficient indicates the level of agreement.

The user should recognize clearly that the level of agreement shown by a correlation coefficient is an agreement only on the pattern of data behavior and needs supplemental information to indicate complete agreement. Body temperature measured at the same time and location in degrees Celsius and degrees Fahrenheit will show almost perfect correlation, but the readings are not interchangeable. A correlation coefficient of 1 tells us that when one variable changes, the other changes in the same direction in the same proportion, even though they may be offset by a constant (they will have different means) or one is a multiple of the other (they will have different standard deviations). If the data are ratings or rankings in which the rank correlation measures \( r_s \) or \( \tau \) would be used, Cohen’s kappa (Section 14.6) may be used and does not have the same drawbacks. For continuous data using Pearson’s \( r \), tests of means and variances,
and perhaps tests of distributions, are needed to supplement the correlation measures in order to verify the agreement.

**EXAMPLE: ESTIMATION OF PATIENT WEIGHT IN THE EMERGENCY DEPARTMENT**

A patient’s weight influences laboratory test interpretation and drug prescription. At times, a patient in the emergency department (ED) is comatose or irrational. Does the next of kin’s reporting of the weight agree with the patient’s adequately for clinical use? DB26 reports (inter alia) weight estimates of 37 patients and their next of kin. By using Eqs. (14.5)—(14.7), we find \( r = 0.910, r_s = 0.952, \) and \( \tau = 0.851 \). These correlations are high and indicate that the patterns of weight estimates are similar for patients and their next of kin. May we conclude acceptance for clinical use? Not yet. Let us test at least the means and variances. A paired \( t \)-test between the patient and the next of kin yields \( p \)-value = 0.010. We also note that the patients’ mean weight as measured by the hospital is 77.8, which rounds to the same kilogram amount as the patients’ mean of estimates. The within-pair differences show that kin tends to underestimate weight. From the paired \( t \)-test, we have evidence that the next of kin significantly underestimates the weight. A variance ratio test yields \( p \)-value = 0.284, not significant. We may conclude that a next of kin estimate of a patient’s weight does not agree adequately with the patient’s estimate for clinical use, but that an added constant may improve the agreement.

**Exercise 14.6**

DB13 gives paired INR readings for 104 patients paired from the hospital laboratory and the Coumadin Clinic. For these data, \( r = 0.900 \), showing a strong pattern of agreement. However, the mean of differences is 0.125. A paired \( t \)-test with the null hypothesis stating that the mean = 0 yields \( p \)-value < 0.001, standard deviations are 0.536 and 0.628, and a variance ratio test yields \( p \)-value = 0.111. Interpret these results.

**14.6 AGREEMENT AMONG RATINGS: KAPPA**

Recall that ranks are assigned to choices relative to each other, while ratings are chosen for each choice irrespective of ratings for other choices. Ratings may all be the same, whereas ranks should be all different except for the occasional tie. In real-life clinical decision-making—a forced decision (diagnosis, treatment choice) leads to rankings, whereas patient judgment (patient satisfaction, judgment of nausea, quality-of-life categorization) more often leads to ratings. Cohen’s kappa measures interrater agreement in excess of the agreement that is expected to occur by chance. Kappa = 0 implies no agreement better than chance, and kappa = 1 implies perfect agreement. Different versions allow for two raters and multiple items; multiple raters and two items; and multiple raters and multiple items.
The mathematical bases for them are rather different, but their development arose from the same initial form, so they retain the name.

Cohen’s kappa is a form of intra-class correlation (ICC). Further comments on ICs may be seen in Section 14.9.

**METHOD FOR KAPPA**

2 *Raters.* Consider a case of 2 raters rating 10 tissue slides as normal (0) or pathological (1) with the results shown in Table 14.4. We note that the layout is similar to that for McNemar’s test (Section 9.8).

**Table 14.4** Results of 2 raters rating 10 tissue slides as normal (0) or pathological (1).

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Using the dotted *n* notation introduced in Section 9.1, the rate of agreement is given by 

\[ p_o = (n_{11} + n_{22})/n = (4 + 3)/10 = 0.7, \]

while the expected agreement by chance is

\[ p_e = (n_1 \cdot n_{.1} + n_2 \cdot n_{.2})/n^2 = (6 \times 6 + 4 \times 5)/10^2 = 0.5. \]

Kappa is the amount that observed agreement exceeds chance agreement as a proportion of the maximum possible amount, or

\[ \kappa = \frac{p_o - p_e}{1 - p_e}. \]  

(14.8)

For this example, \( \kappa = 0.4. \)

*More than 2 raters.* Kappa methods following the same philosophy have been developed for cases of more than 2 raters but are rather involved mathematically. The user is advised to employ statistical software for these cases.

**EXAMPLE: AGREEMENT AMONG RADIOLOGISTS RATING CANCER SEVERITY**

Let us consider DB35 that shows cancer severity as 0—5, rated independently from radiographs on 100 patients by three radiologists, G, C, and S. To start more simply, let us consider the first two radiologists, G and C, and reduce the number of items to 2. We replace the lower severity stages 0, 1, and 2 with *Low* and the higher stages 3, 4, and 5 with *High*. We can make a 2 \( \times \) 2 table, **Table 14.5**.

\[ p_o = (45 + 48)/100 = 0.93, \]

\[ p_e = (51 \times 46 + 49 \times 54)/100^2 = 0.50. \]

\[ \kappa = (0.93 - 0.50)/(1 - 0.50) = 0.86. \]

The agreement on cancer severity rating from radiographs by G and C agree rather well.
Increasing the number of items to 6, Table 14.6 shows the original data for radiologists G and C.

$$p_o = (32 + 3 + \cdots + 9)/100 = 0.82. \quad p_e = (40 \times 32 + 4 \times 9 + \cdots + 9 \times 12)/100^2 = 0.2519. \quad \kappa = (0.82 - 0.2519)/(1 - 0.2519) = 0.759. $$

With the added information, agreement is somewhat lower but still shows an acceptable level.

Let us add the data for radiologist S. By using software for the three-way matching over the 100 patients, we find $$\kappa = 0.761.$$  

**Exercise 14.7**

Let us continue the clinic versus laboratory INR of Exercise 14.6. DB13 also presents paired in-range (0) versus below-range (−1) and above-range (1) readings.

We are interested in determining whether the clinic and laboratory agree on the reading, not on quality control (keeping the INR within range). The calculation of kappa shows observed agreement rate to be 71% but the expected agreement rate to be 40%, resulting in $$\kappa = 0.52. $$ What are the interpretation and clinical implication of this result?

### 14.7 AGREEMENT AMONG MULTIPLE RANKERS

Consider a case of ranking made simultaneously among $k$ categories by $n$ judges. How is agreement measured? Rank correlation will measure the pattern of agreement...
among multiple rankers over a pair of categories or between two rankers over multiple categories. It will not address agreement overall or of multiple rankers about a single category. Kappa analyzes multiple subjects assigned to nominal or ordered categories rather than ranking possible choices about a single subject and will not measure agreement about a single decision choice across a set of rankers. Two methods exist for measuring the agreement among multiple rankers about a single subject: Kendall’s $W$ and Riffenburgh’s $A$.

**EXAMPLE POSED: EVALUATION OF A PROPOSED RESEARCH PROTOCOL BY AN IRB**

At an IRB meeting composed of $n = 7$ voting members, a proposed research protocol is considered. There are $k = 4$ possible decisions: (1) pass as is, (2) pass with minor revisions, (3) table for major revisions, or (4) reject. The Board must select one. Each Board member is asked to rank the decision choices Pass, Revise, Table, or Reject. Perhaps a member’s first choice would be Revise and her second choice Table. If the protocol cannot be revised or tabled, she would select Pass, reserving Reject for the last choice. After a particular proposal was presented and discussed, members ranked their choices 1–4. The original data appear as Table 14.7.

What can be said about agreement among IRB members? Kendall’s $W$ provides a measure of the overall agreement. Riffenburgh’s $A$ provides measures of agreement for each decision choice, that is, the decision to pass the protocol, the decision to revise the protocol, and so forth.

<table>
<thead>
<tr>
<th>IRB member</th>
<th>Pass</th>
<th>Revise</th>
<th>Table</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$T = \sum\text{ranks}$</td>
<td>15</td>
<td>8</td>
<td>19</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 14.7 Rankings of four possible decisions about a medical research protocol proposal by seven IRB members.
KENDALL’S COEFFICIENT OF CONCORDANCE W

Kendall’s $W$, which he termed a coefficient of concordance, provides a single value representing how well the judges agree on what to do with the question at hand as a whole.

For $k$ decision choices ($i = 1, \ldots, k$) to be made by $n$ judges ($j = 1, \ldots, n$), with sum of ranks for choice $i$ denoted $T_i$, Kendall showed the mean rank sum to be $n(k+1)/2$. He defined the sum of squares of deviations of rank sums from the mean rank sum as

$$S = \sum_{i=1}^{k} \left[ T_i - \frac{n(k+1)}{2} \right]^2. \quad (14.9)$$

He showed the maximum possible value of $S$ to be

$$\text{max}(S) = \frac{1}{12} n^2 (k^3 - k). \quad (14.10)$$

The measure of agreement among judges will be the observed $S$ in ratio to the maximum $S$, or

$$W = \frac{12S}{n^2(k^3 - k)}. \quad (14.11)$$

$W$ may range from 0, implying complete disagreement, to 1, implying perfect agreement. Because $W$ arises from a sum of squares, it follows a form of $\chi^2$ distribution for testing and developing confidence intervals.

EXAMPLE CONCLUDED FOR W

Kendall’s coefficient of concordance $W$ provides a single value representing how well the Board as a whole agrees on what to do with the protocol. $k = 4$ and $n = 7$. The mean rank sum is $n(k+1)/2 = 17.5$. $S = \sum (T - 17.5)^2 = 2.5^2 + 9.5^2 + 1.5^2 + 10.5^2 = 209$. The maximum $S$ value is $n^2(k^3 - k)/12 = 245$. Then $W = 209/245 = 0.85$. The overall agreement is rather high.

Riffenburgh’s A

Riffenburgh’s $A$ measures agreement among all rankers (judges) on each choice category. Thus the judges could see how well they all agreed on the first decision choice, how well they all agreed on the second decision choice, etc., providing guidance for a group decision. Like Kappa, $A$ measures agreement remaining after chance agreement is removed. $A = 0$ implies agreement is not better than chance, and $A = 1$ implies perfect agreement.

We are concerned with $k$ decision choices ranked by each of $n$ clinicians. We define $n_1, n_2, \ldots, n_k$ frequencies of ranks, such that $n_i$ is the number of rankers choosing the $i$th rank. The sequence $n, 0, \ldots, 0$ would represent perfect agreement on the first decision choice. The sequence $n/k, n/k, \ldots, n/k$ would represent random
14.7 Agreement Among Multiple Rankers

assignment, that is, no agreement. The measure of agreement $A$ will be defined as a weighted sum of rank frequencies. The nature of the weights can be defined in a number of ways, but the recommended form uses median rank as the weight, where we define the median ($m_{d}$) for each possible decision choice as the rank position at which half the rankers chose a rank equal or below and half chose equal or above.

We start with an $n \times k$ table listing rankers as the left column and choices as the top row, with rank chosen by each ranker as table entries. A second table is constructed from data in the first table, having a row for each of the $k$ choices with $k$ cells in each row, one for each possible ranking. For each choice the number of rankers selecting each rank for that choice is recorded.

Let us think of agreement as the complement of disagreement ($\text{agreement} = 1 - \text{disagreement}$) and of disagreement as the sum of choices in cells other than the median cell, weighted by the rank difference from the median. The observed disagreement about a potential choice $d_{\text{obs}}$ is given by

$$d_{\text{obs}} = \sum_{i=1}^{k} (n_{i} \times |m_{d} - i|),$$  

(14.12)

where $|m_{d} - i|$ is the weight for the $i$th rank. The expected disagreement $d_{\text{exp}}$ under random assignment (the vector $[n/k, n/k, \ldots, n/k]$), that is, given no agreement, is given by

$$d_{\text{exp}} = \frac{n}{k} \sum_{i=1}^{k} \frac{|k + 1}{2} - i.$$

(14.13)

The level of agreement $A$ is shown in the reference article to be given by

$$A = 1 - \frac{d_{\text{obs}}}{d_{\text{exp}}}.$$

(14.14)

Most of the time, this coefficient will lie between 0 and 1, indicating how much greater than chance the agreement is. If it is negative, it indicates that disagreement is greater than expected.

Tables of probabilities that such a value of $A$ would occur by chance are given in the reference article.

EXAMPLE CONCLUDED FOR A

Riffenburgh’s $A$ will give an agreement level for each possible choice. The first step is to reformat the information of Table 14.7 into the format allowing the desired calculations, namely, Table 14.8.
For example, as the first choice, one member ranked Pass, six ranked Revise, and no one ranked Table or Reject, leading to the entries in the column under first Rank. The median is calculated from rows. In the first row, under Pass, the sequence of ranks is 1, 2, 2, 2, 2, 3, 3. The median of these is 2. The calculation of \( d_{\text{obs}} \) for Pass would be \( 1 \times |2 - 1| + 4 \times |2 - 2| + 2 \times |2 - 3| + 0 \times |2 - 4| = 3 \). The calculation of \( d_{\text{exp}} \) would be \( 7(|2.5 - 1| + |2.5 - 2| + |2.5 - 3| + |2.5 - 4|)/4 = 7 \). Then \( A_{\text{Pass}} = 1 - 3/7 = 0.57 \). The calculated values of \( A \) for other choices are listed in the rightmost column in Table 14.8.

We can see that everyone agrees on Reject being the last choice; no one believes the protocol should be rejected. The next strongest agreement is Revise, with \( A = 0.86 \). It appears that Revise is the most desirable recommendation for the IRB to make.

**Exercise 14.8**

**Results of a tumor board’s deliberations.** Tumor boards are conferences used to provide multidisciplinary input into clinical decision-making for cancer patients. A tumor board composed of eight members considered the following case. A 68-year-old man had a gradual onset of dysphagia associated with a right neck mass. Workup revealed a lesion of the right tongue base, staged T3N3 (AJCC Staging Manual, 2002), with preepiglottic space invasion. Further workup revealed multiple 8 mm bilateral lung nodules on a spiral computer tomography (CT) performed the prior week. Positron emission tomography revealed only the robust disease above the clavicles. Past medical history indicated a significant pneumonia 5 years previously that cleared subsequently by CT, but that CT was not of spiral quality. The patient has been on dialysis for renal failure for 3 years. His performance status is 80, and he has lost 25 lb from the baseline of 205 within the last 4 months. The tumor board may consider the following five treatment choices: (1) surgery followed by radiation therapy (SurgXRT), (2) hyperfractionated XRT alone (HypXRT), (3) chemoradiation (ChemRad), (4) induction chemotherapy followed by radiation or surgery depending on response (ChemThen), or (5) palliative therapy (Palliat). Each Board member ranks the possible treatments as judged most appropriate. A choice-by-rank table, Table 14.9, is generated, similar in format to Table 14.8. The last two columns provide the median ranks and the associated agreement coefficients. Calculate \( W \) and the five values of \( A \).
14.8 Reliability

Internal consistency as reliability

Cronbach’s $\alpha$ is often taken as a measure of reliability. More exactly, it measures internal consistency. Technically, $\alpha$ is the square of the expected correlation with a data set of similar format having perfect reliability. As such, it lies between 0 and 1. Many investigators will ask what internal consistency is and what $\alpha$ really measures.

**Example Posed: Reliability of Four Types of Thermometer**

Suppose we have a number $n$ of paired readings on $k$ items. Let us look at an example. DB31 includes readings of five temperature measurements made 10 minutes apart on each of 48 patients by four types of ED thermometers: esophageal, tympanic, temporal arterial scanner (TAS) on forehead, and TAS on forehead plus ear. Looking at the first reading for each thermometer type, we have four columns (items) of 48 readings made simultaneously on each patient. Are the thermometers providing the same information on the same phenomenon (bodily temperature) with the same reliability?

If we measured the Pearson correlation coefficient $r$ for every pairing of items, we would have $k(k - 1)/2$ pairs [from Eq. (3.8)]. If we averaged these $rs$ to produce, say, $m_r$, we would have some answer to our questions. If the $rs$ are all high, they must agree in pattern and are measuring much the same thing. However, we need to standardize this $m_r$ in such a way that it lies between 0 and 1 for interpretation. The standardized Cronbach’s $\alpha$ is given by

$$\alpha = \frac{km_r}{1 + (k - 1)m_r}.$$  \hspace{1cm} (14.15)

It can be seen at a glance that $\alpha = 1$ if $m_r = 1$ and $\alpha = 0$ if $m_r = 0$. There are different forms of $\alpha$, but this is the simplest and most intuitively meaningful.

<table>
<thead>
<tr>
<th>Potential therapy</th>
<th>Rank</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>SurgXRT</td>
<td></td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>HypXRT</td>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ChemRad</td>
<td></td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ChemThen</td>
<td></td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Palliat</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>
**CHAPTER 14 Measuring association and agreement**

**EXAMPLE COMPLETED: RELIABILITY OF FOUR TYPES OF THERMOMETER**

In the thermometer example posed, we have four columns of 48 paired temperatures from the four instruments. \( k = 4 \) and \( n = 48 \). Calculating \( r \) for all possible pairs of thermometer types (say a, b, c, and d) leads to six values of \( r \) (from pairs ab, ac, ad, bc, bd, cd). These six values are 0.4001, 0.3996, 0.3656, 0.2642, 0.1934, and 0.6009. \( m_r = 0.3706 \). By substituting in Eq. (14.15), we find \( \alpha = 0.70 \). The different thermometers agree adequately but not outstandingly.

**Exercise 14.9**

Is the esophageal thermometer reliable, that is, does it give internally consistent temperatures? The thermometer is read repeatedly five times on each patient. \( k = 5 \) and \( n = 48 \). \( 5(5 - 1)/2 = 10 \) possible pairs. The correlation coefficients for these pairs are 0.9069, 0.8676, 0.7769, 0.5406, 0.9797, 0.9217, 0.6768, 0.9470, 0.6857, and 0.7064. Calculate Cronbach’s \( \alpha \) and conclude whether or not the esophageal thermometer is reliable.

**Correlation as reliability**

If we think of reliability as repetition of the same result, reliability is an agreement in the sense that a second result agrees with the first. In this case, we simply have Cronbach’s \( \alpha \) with \( k = 2 \) so that \( \alpha = 2r/(1 + r) \). For example, a correlation coefficient between two trials on the same instrument of 0.94 will yield \( \alpha = 0.97 \). However, we must remain mindful of the comments in Section 14.5; a high \( \alpha \) indicates that the two trials show internal consistency, not identity. As mentioned, simultaneous readings of body temperature by Celsius and Fahrenheit thermometers will give a high \( r \) and a high \( \alpha \), but the readings will be quite different and will not be interchangeable.

**14.9 INTRACLASS CORRELATION**

The Pearson correlation is a measure of association between ordered pairs of continuous measurements from two groups, such as pulse oximeter readings before versus after exercise or height versus weight of a sample of patients. If we cannot assign a reading to one group (height) or the other (weight), as would be the case in studies on twins or using two DBP readings drawn randomly from a patient’s chart, or if we have three or more matched readings (three such BP readings), then we cannot calculate the Pearson correlation and must use what is known as intraclass correlation (ICC).

Basically, ICC follows the logic of one-way analysis of variance, separating a total mean square into mean squares between (or among) matched sets (MSB) and mean square within them (MSW). While the Pearson correlation depends on the variability of each group, ICC depends on the variability from pooling matched members.
The more similar matched members are in value compared to each other relative to comparison with corresponding members of other groups, the smaller the $MSW$ will be relative to the $MSB$. A classic form of ICC is given by

$$ICC = \frac{MSB - MSW}{MSB + MSW}.$$  \hspace{1cm} (14.16)

If $MSB = MSW$, there is no more association with matched members than between a member of one subject and that of another, that is, there is no correlation, and $ICC = 0$. When matched members are identical, that is, there is perfect correlation, $MSW = 0$ and $ICC = 1$. These outcomes and their interpretation agree with those of the Pearson correlation.

There are several types of ICCs. When dealing with ICC, the user must choose the appropriate one, specifying and justifying this choice. The kappa statistic ($\kappa$), addressed in Section 14.6, is currently the best known and most frequently used form of ICC. Fleiss$^7$ showed that $\kappa$ is superior in several respects to a number of other forms of ICC coefficients. While the various forms of ICC are rather complicated for the level of this book relative to their frequency of use, $\kappa$ has already been introduced, so is the only form detailed in this book. An investigator who is faced with the need to use an ICC other than $\kappa$ is advised to consult a qualified statistician.

REFERENCES

5. Kendall, M. G. Rank Correlation Methods, London: Charles Griffin, 1948
Linear regression and correlation

15.1 INTRODUCTION

We introduced scatterplots in Section 5.6, a depiction of paired readings on two different variables plotted by placing the value of one variable on a horizontal, or \( x \), axis and its pair on a vertical, or \( y \), axis. Suppose that the \( y \)-variable usually increases as the \( x \)-variable increases, as would occur with people’s height and weight or white blood cell (WBC) count with count of bacterial infection. Fig. 15.1 shows some WBC count readings as depending on the paired infection level count.

We would like to be able to represent the joint increase, that is, the correlation between \( x \) and \( y \), graphically and to estimate the association between \( x \) and \( y \). One way would be to fit a line to the data such that the distances between each datum and the line (or more usual the squares of the distance) were as small as possible.

We need a mathematical expression for such a line. It can be uniquely represented by two pieces of information. If we think of the straight-line segment as a piece of straight wire, we can anchor it at a point about which it can rotate, say the position on the graph given by the mean of \( x \)-values paired with the mean of \( y \)-values. Then we can specify its slope, or gradient, leading away from that point, and the wire is fixed.

LINES FITTED TO DATA

We want an equation for a set of points in a scattergram modeled as a straight-line fit. By using the symbol \( b_1 \) for the estimate of the theoretical, unknown slope \( \beta_1 \), we can represent the line passing through the mean point \((m_x, m_y)\) in the \( xy \)-plane by the equation

\[
y - m_y = b_1(x - m_x).
\]

Fig. 15.2 shows the fit from Eq. (15.1) superposed on the data in Fig. 15.1. We want to know how to find such a line and what to do with it. Such a line is called a regression line specified by minimizing the sum of squares of vertical distances (called least squares) from the set of points to the line.
A frequent use of a straight-line fit is to predict a dependent variable on the $y$-axis using an independent variable on the $x$-axis. In the leukocyte example, WBC can be predicted by the infection culture count. (The example is used for illustration; clinical use of this prediction is unlikely.) To find the predicted $y$ value from a value on the $x$-axis, we would move directly up to the line of fit, and from there directly left to...

Figure 15.1 WBC data as dependent on culture count of initial infectious agent at a fixed period of time after exposure. WBC, White blood cell.

Figure 15.2 WBC data as dependent on culture count with least squares (regression) fit superposed. Crosshairs show the point composed of means, about $(41, 16)$. Note that the line rises about 8 U over the 50-U horizontal increase (10–60) implying a slope of about $8/50 = 0.16$. WBC, White blood cell.
Exercise 15.1
It is believed that respiration rate in infants decreases with age. To examine that belief, respiration rate was recorded for 232 infants from 0 to 24 months. The scatter diagram of respiration rate against age is shown as Fig. 15.3. Identify the dependent and independent variables. Is there a clear pattern of data behavior to model? Could a straight-line regression model be used?

15.2 REGRESSION CONCEPTS AND ASSUMPTIONS

What is regression?
Regression is a statistical tool that describes and assesses the relationship between two or more variables, at least one of which is an independent variable, that is, a possible causal factor, that predicts the outcome, and at least one of which is a dependent variable, the predicted outcome factor. The description is based on an assumed model of the relationship. This model may describe, or fit, the relationship well or poorly; the method tells us how well the model fits. Model selection is examined more thoroughly in Chapter 16, Multiple linear and curvilinear regression and multifactor analysis of variance. The model may be a straight line (simple regression, Section 15.3), a curved line (curvilinear regression, Section 16.5), or a dependent variable predicted by two or more independent variables (multiple linear regression, Section 16.2).
(The case of two or more dependent variables, that is, simultaneous outcomes, called multivariate regression, is not addressed in this book.) A multitude of regression methods are discussed throughout the text from this chapter on. Chapters 15, 16, and 20 focus on regression methods where the outcome is taken as a continuous variable (Chapter 20: Analysis of repeated continuous measurements over time, considers correlation between observations). Chapter 17, Logistic regression for binary outcomes, considers regression methods for a binary outcome (e.g., a yes or no indicator of whether a participant is diseased or not). Chapter 18, Poisson regression for count outcomes, addresses counts as outcomes, perhaps the number of seizures experienced by an patient over a given period of time. Chapter 19, Analysis of censored time-to-event data, considers regression methods for time-to-event data where the time of an event may not be observed on all patients (e.g., we may wish to examine the time to death among patients treated with an experimental agent and compare this to those treated with a placebo, but not all patients will have died while on study). Section 15.6 lists the various types of regression.

**The term regression**

The term “regression” has a historical origin unnecessary to remember. It may well be thought of as just a name for the process of estimating the association among multiple variables. We say that we “regress the outcome on a set of variables,” meaning that we are attempting to estimate the association between the set of predictors and the outcome variable.

In case the reader is interested, the name arose from a genetic observation. Evidence that certain characteristics of outstanding people were genetic anomalies rather than evolutionary changes was given by fitting the characteristics of the children of outstanding people to those of their parents and grandparents. As the children’s characteristics could be better predicted by their grandparents’ characteristics than their parents’, the children were said to have “regressed” to the more normal grandparent state.

In particular, a group of people were selected for being unusually tall. Most turned out to be anomalies who still carried the genetic character of people of ordinary height. Therefore their children tended to have more ordinary heights that were more closely related to their grandparents than their parents. This is an example of selection bias in sampling.

**Confirmatory versus exploratory uses of regression**

Regression is most useful when the model is dictated by an a priori question being asked regarding the association between the outcome and a set of variables, that takes into account an understanding of the underlying physical or physiological factors.
hypothesized to cause the relationship. In this case, regression allows us to estimate the association(s) of interest, appropriately quantify our uncertainty in these associations, predict a value of the outcome that will arise from a particular clinical reading on the independent variables and to quantify the quality of this prediction. However, in the absence of an a priori model, regression can be used as an exploratory tool suggesting which factors are related and/or the nature of such relationship. It can serve the sequence of steps in acquiring scientific knowledge: detecting the existence of a relationship between variables, describing this relationship, testing the quality of this description, and, finally, predicting outcome values based on independent variables. However, as we have discussed multiple times in the text, such data-driven analyses lead to increased risk of finding spurious associations by chance and hence limiting replicability. We should, therefore, interpret and present such exploratory findings accordingly.

**Five classic assumptions underlying linear regression**

1. There is the usual assumption in statistical methods that the errors in data values (i.e., the deviations from average) are independent one from another and that they have mean 0. Four other assumptions remain that are important to note, as they often are not recognized or assessed in day-to-day clinical research.

2. Obviously, regression depends on the appropriateness of the model used in the fit. Fig. 15.4 shows two fits on a data set. In the linear regression model, we are modeling the mean of the response as function of the predictors. The first-degree model,
$E[y] = \beta_0 + \beta_1 x$, where $E[y]$ denotes the mean of the response, provides a horizontal regression line ($\beta_1 = 0$), indicating no relationship between $x$ and $y$. A parabolic (second-degree) model, $E[y] = \beta_0 + \beta_1 x + \beta_2 x^2$, shows an almost perfect fit. Note that in many cases, the appropriateness of the model depends upon the scientific question being asked. For example, if our goal is to estimate the first-order relationship between $y$ and $x$, then a first-degree association is appropriate. As acclaimed statistician George Box once said, “all models are wrong, but some are useful.”

3. The independent ($x$) readings are measured as exactly known values (measured without error). This arises because the least squares method minimizes the sum of squared errors vertically, that is, on the $y$-axis but not on the $x$-axis. This assumption would be reasonable if we were relating yearly number of illnesses to patient age, as age is recorded rather exactly. However, if we were relating yearly number of illnesses to patient self-reported obesity [body mass index (BMI) calculated from patient-reported height and weight], this assumption would be violated, since this measure of obesity is rather inaccurate. What do we do in the face of a violation? We cannot eliminate the inaccuracies, and refusal to calculate the regression would deny us useful information; usually, we proceed and make careful note that the quality of, and therefore confidence in, the regression is weakened. However, in some cases it is possible to use a solution called Deming regression, addressed in Section 15.5. Fig. 15.5 shows a regression line fitted to data with a vertical error shown. (The correlation aspect will be addressed in Section 15.8.)

4. The variance of $y$ is the same for all values of $x$. As we move from left to right along the regression line, the standard deviation of $y$ about this line is constant.

Figure 15.5 A regression or correlation line (data chosen so line will be either) fitted to a set of data. The deviations, or “errors,” between a point and the line as assumed in regression and in correlation are labeled. Note that the regression error is vertical, indicating no $x$ variability, whereas the correlation error allows variability in both $x$ and $y$. 
5. The distribution of \( y \) is approximately normal for all values of \( x \). As we move from left to right along the regression line, distribution of data remains normal for each \( x \) with mean at the regression line. Fig. 15.6 illustrates assumptions (4) and (5) simultaneously. Linear regression is reasonably robust to this last assumption; one need not worry too much unless the violations are severe or the computation of prediction intervals (discussed later in this chapter) is of primary scientific concern.

### 15.3 SIMPLE REGRESSION

#### Concept of a line fitted to a set of points on a graph

It seems intuitively obvious that a fit should use all the data and satisfy the most important mathematical criteria of a good fit. The simplest and most commonly used fitting technique of this sort is named **least squares**. The name comes from minimizing the squared vertical distances from the data points to the proposed line. A primitive mechanical way of thinking about it would be to imagine the points as tacks, each holding a rubber band. Our strong wire segment from our earlier discussion representing the line segment is threaded through all the rubber bands. The tension on each rubber band is the square of the distance stretched. The wire is shifted about until it reaches the position at which the total tension (sum of tensions of each rubber band) is a minimum.

**EXAMPLE POSED: DAY 10 THEOPHYLLINE LEVEL PREDICTED FROM BASELINE**

DB3 contains data on levels of serum theophylline (a vasodilator to treat asthma, emphysema, etc.) just prior to administering an antibiotic (baseline) and 10 days later. The relationship between the two levels can be modeled as a straight-line fit, shown in Fig. 15.7. We can calculate the inputs to the equation for this regression line as \( m_x = 10.7988 \), \( m_y = 10.1438 \), \( s_{xy} = 11.8253 \), and \( s_x^2 = 14.1787 \). What is the equation for the regression line?
CHAPTER 15 Linear regression and correlation

**METHOD FOR SIMPLE REGRESSION**

**The Regression Equation**

A useful form of a regression line, using the slope $b_1$ of the line and the mean point $(m_x, m_y)$ in the $x$, $y$-plane, was given by Eq. (15.1) as $y - m_y = b_1(x - m_x)$. [From here on we omit the residual corresponding to the fitted value for ease of exposition.] Recall that theoretical or population coefficients in models are denoted by $\beta$'s, while sample-based estimates of the $\beta$'s are denoted by $b$'s. The estimate of $\beta_1$, $b_1$, is the covariance of $x$ and $y$ divided by the variance of $x$, or $s_{xy}/s^2_x$, so that Eq. (15.1) may be written as

$$y - m_y = b_1(x - m_x) = \frac{s_{xy}}{s^2_x}(x - m_x). \quad (15.2)$$

(The formulas for the $m$'s and $s$'s may be found in Sections 5.1 and 5.2.) This is the best fit by least squares (and also other mathematical criteria) expressing the relationship between $x$ and $y$.

**Interpreting the Association Between $y$ and $x$**

An easier form of the estimated regression in (15.2) is given by the slope-intercept form, $y = b_0 + b_1x$, in which $b_0 = m_y - b_1m_x$. This form provides interpretable parameter estimates. Specifically, $b_0$ is the intercept of the model and is interpretable as the estimated mean of $y$ when $x = 0$. $b_1$ is the slope, or first-order trend, associated with $x$ and is interpretable as the estimated difference in the mean of $y$ comparing two subpopulations differing in $x$ by 1 unit.

(Continued)
Predicting \( y \) From \( x \)

The most likely value of \( y \) as given by the regression model, that is, its prediction, for a chosen value of \( x \) is just the value of \( y \), say \( y|x \) (read “\( y \) given \( x \)”), arising from substituting the \( x \) value in Eq. (15.1) and solving for \( y \). If many predicted values are to be calculated, the slope-intercept form of the regression is easiest to consider:

\[
y = b_0 + b_1x. \tag{15.3}
\]

It should be noted that a prediction is valid only over the range of existing data. (In an example of rat survival, to be seen in Section 16.6, extending the days after malarial infection far enough will appear to bring dead rats back to life.)

Let us return to the infection example addressed in Section 15.1. For the data plotted in Fig. 15.1, \( m_x = 40.96 \), \( m_y = 16.09 \), \( s_x^2 = 163.07 \) (\( s_x = 12.77 \)), and \( s_{xy} = 25.86 \). Substituting these quantities in Eq. (15.2) and simplifying to Eq. (15.3), we obtain the fit:

\[
y = 0.1585x + 9.5978. \tag{15.4}
\]

From the above, we estimate that the difference in mean WBC count is \( 0.1585 \times 10^9 \) comparing two populations differing in culture count by \( 1 \times 10^9 \). Let us find the predicted \( y \) value (WBC \( \times 10^9 \)) from an \( x \) value (culture count \( \times 10^9 \)) of 30. We can find it approximately by moving up from 30 on the \( x \)-axis to the fitted line and then moving left to the \( y \)-axis, where we find \( y \) is a little more than 14. Upon substituting in Eq. (15.4), we obtain a more exact \( y \) value as \( 0.1585 \times 30 + 9.5978 = 14.35 \).

**EXAMPLE COMPLETED: DAY 10 THEOPHYLLINE LEVEL PREDICTED FROM BASELINE**

Substitution in Eq. (15.2) yields the line \( y = 10.1438 = 0.8340(x - 10.7988) \). Solution for \( y \) yields a simpler form for prediction: \( y = 1.1376 + 0.8340x \). Suppose a patient has a baseline serum theophylline level of 10 and we wish to predict the postantibiotic level, \( y|x (x = 10) \). (Recall the vertical line in the formula can be said “given.”) We substitute \( x = 10 \) to obtain \( y|_{10} = 9.4776 \), or about 9.5.

**ADDITIONAL EXAMPLE 1: PREDICTING TEMPERATURE OF LUNG-CONGESTED INFANTS**

Lung congestion often occurs in illness among infants but is not easily verified without radiography. Are there indicators that could predict whether or not lung opacity will appear on an X-ray? A study of 234 infants included age (months), respiration rate (breaths/min), heart rate (beats/min), temperature (degrees Fahrenheit), pulse...
oximetry (percent), clinical appearance of illness on physical exam, and lungs sounding congested on physical exam. Lung X-rays were taken to be “truth” and recorded as clear or opaque. However, the use of a binary outcome (yes or no, + or −, 0 or 1, etc.) as a dependent variable requires logistic regression and will be addressed in Chapter 17, Logistic regression for binary outcomes. Until then, relationships between continuous measurements less important than the X-ray result will be used as illustrations.

**PREDICTING TEMPERATURE BY AGE**

The following question arises: Is temperature associated solely with illness or is it influenced by age? Let us fit a regression line to temperature as predicted by age. If the line is horizontal, knowledge of the age gives us no ability to predict temperature. However, if the line has a marked slope (testing the slope statistically is treated in Section 15.4), age provides some ability to predict temperature and would seem to be one factor influencing lung congestion. Regression line inputs from age (taken as $x$) and temperature ($y$) were calculated as $m_x = 10.9402$, $m_y = 101.5385$, $s_{xy} = 5.6727$, and $s_x^2 = 45.5329$. $b_1 = s_{xy}/s_x^2 = 0.1246$. The slope–means form of the regression line, from Eq. (15.1), is $y = 101.5385 + 0.1246(x - 10.9402)$. $b_0 = 100.1754$, leading to the slope–intercept form as $y = 100.1752 + 0.1246x$. If an infant is 20 months of age, the most likely prediction of temperature would be $y|20 = 100.1752 + 0.1246 \times 20 = 102.6672$, or about 102.7°F. The data and the regression line are shown in Fig. 15.8.

![Figure 15.8 Temperature depending on age for 234 infants.](image)
**ADDITIONAL EXAMPLE 2: DOES MENTAL PATIENT HOSPITAL STAY RELATE TO INTELLIGENCE QUOTIENT?**

At a mental hospital, a psychologist’s clinical experience suggests that brighter patients seem to stay in the hospital longer. If true, such a finding might result from brighter patients being harder to treat. The psychologist collects intelligence quotient (IQ) measurements ($x$) and days in the hospital ($y$) for the next 50 patients treated. Calculation of means, variances, and the covariance yields $m_x = 100$, $m_y = 10.5$, $s_x = 10$, $s_y = 11.2$, and $s_{xy} = 39.43$. The slope $b_1$ is given by $s_{xy}/s_x^2 = 39.43/100 = 0.3943$, leading to the regression line $y - 10.5 = 0.3943 (x - 100)$ in the form of Eq. (15.1). The data from the 50 patients with the regression line superposed are shown in Fig. 15.9. What would be the best prediction of days in the hospital for a patient with a 90 IQ? With 110 IQ? Substitution of 90 for $x$ in the equation for the regression line yields about 6.6 days to be the expected stay for a patient with 90 IQ. Similarly, about 14.4 days is the best prediction of hospital stay for a patient with 110 IQ.

**Exercise 15.2**

Descriptive statistics for the data in Exercise 15.1 (where subscripts $a$ and $r$ represent age and respiration rate, respectively) are $m_a = 10.9$, $m_r = 39.0$, $s_a = 6.7$, $s_r = 11.8$, $s_{ar} = -15.42$. Calculate the equation for the straight line regression of $r$ on $a$. Lay a thin paper over the scattergram of respiration rate by age and sketch in this line.

**Exercise 15.3**

As part of a study on circulatory responses to intubation during laryngoscopy, the following question arose: Could changes in systolic blood pressure (SBP, $y$-axis) due to laryngoscopy be predicted by skin vasomotor reflex amplitude (SVmR, $x$-axis)? Readings (these readings from a

![Figure 15.9](image-url) Regression of days in the hospital on IQ for 50 mental patients. IQ, Intelligence quotient.
Exercise 15.4
In DB14, we ask if the change in eNO due to exercise is related to the initial (preexercise) eNO, or if they are independent. The scatter diagram is shown in Fig. 15.10. Descriptive statistics (where subscripts 0 and 20 represent preexercise eNO (0 minutes) and eNO at 20 minutes, respectively) are approximately $m_0 = 29.3$, $m_{20} = 28.1$, $s_0 = 25.6$, $s_{20} = 25.3$, $s_{0,20} = 621.5$. Calculate the equation for the straight-line regression prediction of $eNO_{20}$ by $eNO_0$. Lay a thin paper over the scattergram and sketch this line.

15.4 ASSESSING REGRESSION: TESTS AND CONFIDENCE INTERVALS

Tools of assessment
Before pursuing assessment, let us note some tools to use in assessment: the coefficient of determination $R^2$ and standard errors.
\( R^2 \), the coefficient of determination

Returning to the Additional Example 2 of Section 15.3, how much of the variation in hospital stay does IQ account for? A good indication of this is given by a statistic named the coefficient of determination, designated \( R^2 \). \( R^2 \) could be defined as the proportion of possible perfect prediction represented by the regression model. When only one predictor is used, as in this chapter, it is nothing more than the square of the correlation coefficient; for curvilinear and multiple regression, it is not so straightforward. In this case, \( r = 0.35 \), so \( R^2 = 0.12 \), implying that IQ counts for only about 12% of the variation in hospital days. Our psychologist would be inclined to conclude that IQ is a real but rather minor influence on length of hospital stay. Would he be justified? Let us consider the satisfaction of three of the assumptions upon which this regression analysis was based. Is the assumption that the \( x \)-values are measured without error satisfied? The patient’s true IQ is a fixed but unknown value. The measured IQ is an effort to estimate the true IQ, but its accuracy certainly is recognized as less than perfect. In addition to accuracy missing the mark, measured IQ varies from time to time in a patient and varies from type to type of IQ test. No, the assumption of exact \( x \)-values is not well satisfied. Is the assumption of normal shape about the line for each \( x \) satisfied? There are too few data to be very sure. From the appearance in Fig. 15.9, a case might be made toward the center, but data in the tails appear to be rather skewed. Satisfaction of this assumption is questionable. Finally, is the assumption of equal variability for each \( x \)-value satisfied? A glance at Fig. 15.9 will show that the data spread out more and more as we move to the right along the regression line. This assumption seems violated. So where does violation of the assumptions underlying the method leave the investigator? It is possible to adjust for some of the violations by using more sophisticated statistics, but the sample size per IQ value is still too small. All he can say is that he has obtained an estimate of the association between IQ and hospital days based upon some informal pilot results and attempt to quantify the uncertainty in this estimate. These results suggest a small influence by IQ on the length of hospital stay, on the basis of which he will carry out a more carefully designed study with a much larger sample and seek the assistance of a biostatistician at both planning and analysis time.

Standard errors

The standard error of the mean is often symbolized SEM. However, in regression, we shall have to find the standard errors of other statistics and use them in formulas, so we need a more succinct and flexible notation. Because a standard error of a statistic is just its standard deviation, it will be convenient to use the standard deviation symbol \( s \) with the respective statistic indicated as a subscript. Thus the estimated SEM becomes \( s_{m} \). We will need four more standard errors: the estimated standard error of the residual (residuals are the observations’ deviations from the regression line), usually denoted \( s_{e} \) in statistics (\( e \) for “error”); the estimated standard error of the estimate of the regression
slope $b_1$, $s_b$; the estimated standard error of the estimate of the mean values of $y$ for each $x$, $s_{m|x}$; and the estimated standard error of the estimate of the individual predictions of $y$ for each $x$, $s_{y|x}$. As the residual standard error $s_e$ is used in the others, it is given first.

Starting with the sum of squares of deviations from the regression line, we define

$$s_e^2 = \frac{\sum (y - b_0 - b_1 x)^2}{(n-2)}$$

where $n-2$ will be degrees of freedom (df), because the sample size $n$ lost a df to each of the two $b$’s. Some algebra will yield the following forms. Use whichever is easier computationally.

$$s_e = \sqrt{\frac{n-1}{n-2} \left( s_y^2 - b_1^2 s_x^2 \right)} = s_y \sqrt{\frac{n-1}{n-2} (1 - R^2)}$$

(15.5)

**EXAMPLE POSED: TESTING THE THEOPHYLLINE PREDICTION**

In the example of Section 15.3 using DB3, serum theophylline at day 10 was predicted from baseline level by the regression line $y = 1.1376 + 0.8340 x$ ($b_1 = 0.8340$). The calculated components we will need are $n = 16$, $m_x = 10.8$, $s_{xy} = 11.8253$, $s_x = 3.7655$, and $s_y = 3.9816$. Is the association between day 10 level and baseline significantly different from 0? How strong is this association? What is the confidence in our estimate of the association between day 10 level and baseline? Of individual predicted level?

**Methods of assessment**

Five questions might be asked of the regression and will be considered in turn. (1) Does the regression show that $y$ is significantly associated with $x$? (2) Does the model account for a significant portion of the observed variation in $y$? (3) What will be confidence limits on the mean of $y$ for a population of patients with covariate value $x$? (4) What will be prediction limits on the best predicted $y^*$ for a newly sampled patient with covariate value $x^*$

(1) **IS $X$ SIGNIFICANTLY ASSOCIATED WITH $Y$?**

If the regression line is horizontal, the predicted $y$ is the same for every $x$; $x$ has no linear association with $y$. If the line is sloped, each $x$ yields a different $y$; there is an association. Is this slope significantly different from horizontal, or could an apparent slope be due only to chance? In a test of this question, $H_0$: $\beta_1 = 0$ versus $H_1$: $\beta_1 \neq 0$ (or $<0$ or $>0$ if one tail is
impossible or clinically irrelevant). As in a hypothesis test of a mean, we ask if the ratio of a statistic to its standard error is larger than a critical \( t \) value. We find the standard error of the estimated slope, \( b_1 \), as

\[
s_b = \frac{s_e}{\sqrt{(n-1)s_x}}. \tag{15.6}
\]

To test \( H_0 \), \( t \) with \( n-2 \) df is simply

\[
t_{(n-2)df} = \frac{b_1}{s_b}, \tag{15.7}
\]

and we reject \( H_0 \) if \( b_1/s_b \) is farther out in the tails of the \( t \) distribution than the critical \( t \).

Testing Against a Theoretical Slope

Suppose the slope is to be compared not to zero slope but to a theoretical slope, say \( \beta^0_1 \). Situations in which such a \( \beta^0_1 \) might arise are, for example, when a relationship has been posed on the basis of established physiology or when comparing with previously published results. The standard error \( s_b \) remains the same, as does the \( n-2 \) df. The hypotheses are \( H_0: \beta_1 = \beta^0_1 \) versus \( H_1: \beta_1 \neq \beta^0_1 \). The \( t \) statistic becomes

\[
t_{(n-2)df} = \frac{b_1 - \beta^0_1}{s_b}, \tag{15.8}
\]

which is tested and interpreted the same as for zero slope.

Confidence Interval on the Slope

The general pattern of confidence intervals was given in Section 8.3. To follow that pattern here, we need the estimate of the statistic (sample slope \( b_1 \)), the critical value of the probability distribution (\( t_{n-2,\alpha/2} \)), and the standard error for that statistic (\( s_b \)), all of which have been given earlier. A \( 100 \times (1 - \alpha)\% \) confidence interval for the population slope \( \beta_1 \) is given by

\[
(b_1 - t_{1-\alpha/2} s_b, b_1 + t_{1-\alpha/2} s_b), \tag{15.9}
\]

where the critical \( t \) has \( n-2 \) df. A tight confidence interval indicates a precisely estimated relationship.

(2) DOES THE MODEL ACCOUNT FOR A SIGNIFICANT PORTION OF THE VARIABILITY IN Y?

The coefficient of determination indicates the portion of observed variability in the dependent variable that is accounted for by the independent variable. This is generally designated
(CONTINUED)

$R^2$. $r^2$ is sometimes used in the case of a single $x$ predicting $y$ by a straight line, but we shall use $R^2$ in all cases to be consistent. Note well that $R^2$ does not evaluate how much of the variation in $y$ is accounted for by $x$ but rather how much of the variation in $y$ is explained by the model. Sometimes a straight line is a very poor predictor, whereas a curved line such as a parabola is a good one, as was illustrated in Fig. 15.4. In this case, $y = b_0 + b_1x$ yields a very small $R^2$ (0.002 in the figure), but $y = b_0 + b_1x + b_2x^2$ yields a large one (0.970 in the figure). We know that an $r$, and therefore $R^2$, of 0 indicates no relationship and that an $r$ of $\pm 1$, and therefore an $R^2$ of 1, indicates perfect predictability. At what value of $R^2$ does the relationship become greater than chance? It turns out that this value is the same as the value at which $\beta_1$ becomes significantly greater than 0, as tested by Eq. (15.7). Algebra will permit this form to be rewritten as a test of $R^2$ (for straight line single $x$ regression only!). The null hypothesis $H_0: \rho^2 = 0$ is tested by

$$t_{(n-2)df} = \sqrt{\frac{(n-2)R^2}{1-R^2}}. \quad (15.10)$$

Because Eq. (15.10) tells us nothing new over Eq. (15.7), $R^2$ usually is used in interpreting rather than testing the $x$, $y$ relationship.

(3) WHAT IS A CONFIDENCE INTERVAL FOR THE MEAN OF $Y$ FOR A GIVEN VALUE OF $X$?

Because there is only one regression line, there is only one prediction of $y$ from a given $x$. The mean $y$ for a given $x$ is the same as the most likely $y$ for a given $x$. However, the confidence interval for a mean $y$ is not the same as that for the prediction of an individual patient’s $y$, so these intervals (termed prediction intervals) are given separately in this and the next paragraph. The confidence interval for the mean of $y$ for a given value of $x$ follows the same pattern as before: the estimated value on the $y$-axis $\pm$ a critical $t$ multiplied by the standard error. The estimate $b_1$ and the critical $t$ are the same as used in Eq. (15.9); only the standard error differs. We symbolize the sample mean value of $y$ for a given $x$ as $m|x$, which estimates the population mean value $\mu$ for a given $x$, $\mu|x$. Its standard error is

$$s_{m|x} = s_x \sqrt{\frac{1}{n} + \frac{(x-m_x)^2}{(n-1)s_x^2}}. \quad (15.11)$$

A $100 \times (1 - \alpha)\%$ confidence interval for $\mu|x$, where the critical $t$’s have $n-2$ df, is given by

$$(m|x - t_{1-\alpha/2} s_{m|x}, m|x + t_{1-\alpha/2} s_{m|x}). \quad (15.12)$$
WHAT IS A PREDICTION INTERVAL FOR Y FOR A NEW PATIENT WITH COVARIATE VALUE X?

The mean response often is of interest in research. In clinical practice, the predicted value for an individual patient often is of interest. The prediction will be the same; the mean is the most likely value and our best guess. However, the standard error is slightly different, leading to an altered interval which is known as a prediction interval. Suppose we wish to predict the response for a newly sampled patient with covariate value \( x \):

We write the predicted response as \( \hat{y} \mid x \). The standard error of \( \hat{y} \mid x \) is

\[
s_{\hat{y} \mid x} = s_{\hat{y}} \sqrt{1 + \frac{1}{n} + \frac{(x^* - m_x)^2}{(n-1)s_x^2}}
\]

A 100 \((1 - \alpha)\%\) confidence interval for \( \hat{y} \mid x^* \), where the critical \( t \)'s have \( n - 2 \) df, is given by

\[
(\hat{y} \mid x^* - t_{1-\alpha/2} s_{\hat{y} \mid x}, \hat{y} \mid x^* + t_{1-\alpha/2} s_{\hat{y} \mid x}^*).
\]

EXAMPLE COMPLETED: TESTING THE THEOPHYLLINE PREDICTION

(1) Is \( x \) Significantly Associated With \( y \)?

From Eq. (15.5), \( s_{\hat{y}} = 2.5335 \). To test \( H_0: \beta_1 = 0 \), we need \( s_b \). Substitution in Eq. (15.6) yields \( s_b = 0.1737 \). We use Table II to find a 0.05 two-tailed critical value of \( t \) for \( n - 2 = 14 \) df as 2.145. From Eq. (15.7), the calculated \( t = 4.80 \), which is much larger than critical. We conclude that the slope of the line is significantly different from 0 and that there is an association between \( y \) and \( x \). To put a confidence interval on the slope, we substitute \( b_1, s_b \), and critical \( t \) in Eq. (15.9) to find the confidence interval:

\[
(0.8340 - 2.145 \times 0.1737, 0.8340 + 2.145 \times 0.1737) = (0.46, 1.21).
\]

We can say with 95% confidence that the plausible values for the first-order trend between baseline and 10-day levels range from 0.46 to 1.21. Note that this interval excludes 0, which is what we expect given the result of our test of \( H_0: \beta_1 = 0 \).

(2) Does the Model Account for a Significant Portion of the Variability in \( y \)?

Baseline theophylline level is a significant predictor, but is it a major predictor of day 10 level? Calculation as in Eq. (5.10) yields \( r = 0.7887 \), so \( R^2 = 0.6220 \). We can say that about 62% of the variation in day 10 serum theophylline level is explained by the straight-line model; 38% remains for all other factors combined plus randomness. Although baseline level does not predict day-10 level exactly, we can conclude that baseline level accounts for a fairly high proportion of the total variance in day-10 levels.
What Is a Confidence Interval for the Mean of $y$ for a Given Value of $x$?

Suppose we have a group of patients with mean baseline serum level equal to 10 mg/dL. The predicted $m|x$ is, of course, the value of the regression obtained by substitution of that $x$, in this case $m|(x = 10) = 9.477$. We want a confidence interval on the mean day 10 level for a population with baseline serum level equal to 10 mg/dL. Use of Eq. (15.11) yields the standard error of this prediction statistic as $s_{m|x} = 0.6738$. We found the 0.05 two-tailed critical value of $t$ for 14 df to be 2.145. By substituting in Eq. (15.12), we find the confidence interval on mean day 10 level for a baseline level of 10 mg/dL to be

$$(9.477 - 2.145 \times 0.6738, 9.477 + 2.145 \times 0.6738) = (8.032, 10.922).$$

What Is a Prediction Interval for $y$ for a New Patient With Covariate Value $x$?

Suppose we have an individual patient that enters the office with baseline serum level equal to 10 mg/dL. The predicted $y^*|(x^* = 10)$ is still 9.477. We want a prediction interval on this individual’s day 10 level. Use of Eq. (15.13) yields the standard error of this prediction statistic as $s_{y^*|x^*} = 2.622$. The critical value of $t$ remains 2.145. By substituting in Eq. (15.14), we find the prediction interval on the patient’s day 10 level for a baseline level of 10 mg/dL to be


We note that the prediction interval on the individual’s predicted 10-day level is much wider than that the confidence interval on the mean. This is because the prediction interval not only accounts for our uncertainty in the estimated mean but also the individual variance of 10-day serum levels in the population the patient is sampled from.

**ADDITIONAL EXAMPLE: INFANT LUNG CONGESTION CONTINUED**

In the lung congestion study introduced in the Additional Example 1 of Section 15.3, we asked if temperature could be predicted by infant age. The regression line was calculated to be $y = 100.1752 + 0.1246x$ ($b_1 = 0.1246$). The regression prediction for a 20-month infant was 102.6674°F. The interim statistics we need are $n = 234$, $m_x = 10.9402$, $s_{xy} = 5.6727$, $s_x = 6.7478$, and $s_y = 2.2557$. From Eq. (15.5), $s_e = 2.0977$.

(1) Is $x$ Significantly Associated With $y$?

To test $H_0: \beta_1 = 0$, we need $s_b$. Substitution in Eq. (15.6) yields $s_b = 0.0204$. We interpolate in Table II to find a 0.05 two-tailed critical value of $t$ for $n - 2 = 232$ df as about 1.97. From Eq. (15.7), the calculated $t = 6.1078$, which is much larger than the critical $t$. Indeed, because $t$ is larger than the interpolated Table II $t = 3.33$ for two-sided $\alpha = 0.001$, we find that $p$-value $< 0.001$. We conclude that the slope of the line and therefore association between $y$ and $x$ is
significant. This tells us that age is significantly associated with temperature. To put a confidence interval on the slope, we substitute $b_1$, $s_b$, and critical $t$ in Eq. (15.9) to find the 95% confidence interval:

$$(0.1246 - 1.97 \times 0.0204, \ 0.1246 + 1.97 \times 0.0204) = (0.08, \ 0.16).$$

(2) Does the Model Account for a Significant Portion of the Variability in $y$?

The use of Eq. (5.10) yielded $r = 0.3727$, so $R^2 = 0.1389$. We can say that about 14% of the variation in temperature is accounted for by the straight-line model including age, the remaining 86% arising from other causal factors and randomness. This tells us that age accounts for a fairly minor portion of the variation in temperature.

(3) What Is a Confidence Interval for the Mean of $y$ for a Given Value of $x$?

Let us consider a group of 20-month-old infants. The predicted $m|\hat{x}$ is, of course, the value of the regression obtained by substitution in that $\hat{x}$, in this case 102.6672. We want a confidence interval on the mean temperature of a population that is 20 months of age. Use of Eq. (15.11) yields the standard error of this prediction statistic as $s_{m|\hat{x}} = 0.2299$. We found the 0.05 two-tailed critical value of $t$ for 232 $df$ to be 1.97. By substituting in Eq. (15.12), we find the 95% confidence interval on mean temperature of 20-month-old infants to be

$$(102.6674 - 1.97 \times 0.2299, \ 102.6674 + 1.97 \times 0.2299) = (102.2, \ 103.1).$$

(4) What Is a Prediction Interval for $y$ for a New Patient with Covariate Value $\hat{x}$?

Suppose we have an individual infant at age 20 months. The predicted temperature for the infant is still 102.6672. We want a prediction interval on this value. The use of Eq. (15.13) yields the standard error of this prediction statistic as $s_{y|\hat{x}} = 2.1103$. The critical value of $t$ remains 1.97. By substituting in Eq. (15.14), we find the 95% prediction interval on the 20-month infant to be

$$(102.6674 - 1.97 \times 2.1104, \ 102.6674 + 1.97 \times 2.1104) = (98.5, \ 106.8).$$

While the confidence interval on the mean might be useful in research, the prediction interval on temperature as predicted by age for an individual infant highlights how imprecise our prediction is, as it contains almost the entire range of possible readings. If we want to predict temperature for clinical use, we will have to incorporate additional information, including more important causal factors in our model, via multiple regression.

**Exercise 15.5**

Calculate the coefficient of determination from the descriptive data in Exercise 15.2. How is this coefficient of determination interpreted?
CHAPTER 15 Linear regression and correlation

Exercise 15.6
Calculate the coefficient of determination for the data of Exercise 15.4. How is this coefficient of determination interpreted?

Exercise 15.7
Exercise 15.3 poses the question: Could change in SBP (y) caused by laryngoscopy be predicted by SVmR (x)? Required interim statistics are n = 26, m_x = 0.1771, s_{xy} = 2.2746, s_x = 0.1553, and s_y = 20.0757. Find s_e and s_b. Look up the 95% critical t. Test H_0: β_1 = 0. Is the slope significantly greater than 0? Find and test R^2. Find a 95% confidence interval for β_1. Is SVmR a major predictor? What is the proportion of total variation in the response that remains for other predictors and randomness? For SVmR = 0.3, calculate s_{m|x} and s_{y^*|x^*} and find a 95% confidence interval for μ|x and 95% prediction interval for y^*|x^*.

15.5 DEMING REGRESSION

The usual regression assumption of exact measurements, that is, without error, on the x-axis was presented in Section 15.2. It was stated that, if the x-axis error is too large to pretend it to be absent, there exists Deming regression to account for it. In this case, the error may be visualized as the error labeled “correlation error” illustrated in Fig. 15.5 with both an x and a y component. The method depends on knowing the ratio λ of theoretical variances of individual x and y measurements,

\[ \lambda = \frac{\text{var (a y reading)}}{\text{var (an x reading)}}. \]

(15.15)

Therein lies the reason that this form of regression is seldom used: these variances usually are unknown.

History
One form of adjusting for the presence of an x component in the error was developed by Adcock\(^5\) and then Kummell\(^6\) in the late 1870s, but was little used. Koopmans\(^7\) in the 1930s found the method useful in economics. It was introduced into statistics more generally by Deming\(^8\) in the early 1940s (before he brought effective quality control to Japan after World War II) and, therefore, bears his name.
HOW DEMING REGRESSION ACCOUNTS FOR THE X-ERROR

The sample parameter estimates $m_x$, $m_y$, $s_x^2$, $s_y^2$, and $s_{xy}$ remain as used earlier. We use the slope-intercept form of the regression line equation $y = b_0 + b_1x$, where $b_0$ is the position at which the line crosses the $y$-axis (the value of $y$ when $x = 0$), but now $b_1$ is calculated by

$$b_1 = \frac{s_y^2 - \lambda s_x^2 + \sqrt{(s_y^2 - \lambda s_x^2)^2 + 4\lambda s_{xy}^2}}{2s_{xy}}.$$  \hspace{1cm} (15.16)

The difficulty from not knowing $\lambda$ should be apparent from Eq. (15.1). If the $x$ error is small, we usually lose little by ignoring it and using simple regression. If we think the error in the $x$ dimension cannot be ignored and is of a magnitude similar to that in the $y$ dimension, we could assume $\lambda$ to be 1. If we believe the $x$ error to be larger than the $y$ error, we must find some usable estimate of $\lambda$. Since the ratio is at issue rather than individual measures of variability, a case could be made for using a ratio of variability measures other than variability of errors on single readings for $x$ and $y$. However, if we try to use our sample estimates of variance to estimate $\lambda$, $b_1$ simplifies to $s_y/s_x$. One solution to measure $\lambda$ would be to use the ratio of squares of average deviations. [The average deviation is a historical measure of variability used before the standard deviation was shown to be more useful in mathematical development of statistical methods; it is the average (mean) of absolute differences between observations and their means, or the form $\left(\sum |x_i - m_x| \right)/n$.] This ratio will be of the same order of magnitude as $\lambda$ and may be better than no adjustment for an $x$ that has a large error.

ADDITIONAL EXAMPLES 1 AND 2 OF SECTION 21.3 CONTINUED

We note that, in Additional Example 1, $x$ is the infants’ age in months. Such a recording is almost always accurate and without random error. There is no need for Deming regression. We leave the result as analyzed in Section 15.3.

In Additional Example 2, however, IQ is generally recognized to have considerable error in individual readings. IQ tests are usually standardized to have a standard deviation of 10–15 IQ units. The distribution of days in hospital for patient might have a standard deviation of 5–7 days as an experienced guess. Using such a solution to the unknown $\lambda$ depends largely on personal impressions rather than data and will vary with the user; it is, therefore, not a scientific approach to the problem. However, this example both illustrates the difficulty with Deming regression and provides a numerical example of the method, so will be continued. The ratio of 5–7 as a standard
deviation on $y$ to 10–15 on $x$ is about half and squaring it leads to a $\lambda$ of approximately one-fourth. Substitution in Eq. (15.16) yields $b_1 = 0.25$ compared to the unadjusted slope calculated in Section 15.3 as $b_1 = 0.39$. The line in Fig. 15.9 would be flatter as a result of the large error in IQ readings.

15.6 TYPES OF REGRESSION

Classifying types of regression models

To this point in this chapter, treatment of regression has been restricted to prediction of continuous variables as outcomes. However, investigators often require other types of outcome. Several classes of regression models available to address the various types of outcome are enumerated in Table 15.1. Note that these classes refer only to the dependent variable, not the predictors. Each type may have first-degree, higher degree, or nonlinear predictor variables. Each type may have a single or multiple predictor variables, and the predictors may be continuous or discrete in nature.

Examples of the several types

An example of an outcome variable in clinical research associated with each regression type follows. *Ordinary regression*: Prostate-specific antigen levels are predicted by BMI. The outcome is a continuous variable. *Ordered regression*: Insomniacs are asked to rank the effectiveness of four sleeping aids. The outcome is a rank-order variable. *Logistic regression* (Chapter 17: Logistic regression for binary outcomes): Breast cancer in post-radiation therapy (RT) patients may have recurred or not, predicted by RT dose. The outcome is a binary variable. *Multinomial regression*: 1 year after treatment for actinic keratoses, patients may be classed as cured, recurred keratoses, or progressed to squamous-cell carcinoma. The outcome is a three-category variable. *Poisson regression* (Chapter 18: Poisson regression for count outcomes): Diet plus supplements, antibiotic use, and exercise levels are used to predict the number of infectious illnesses over a 3-year period. The outcome variable is composed of the number of occurrences (counts) as the outcome. *Cox proportional hazards regression* (Chapter 19: Analysis of censored

<table>
<thead>
<tr>
<th>Nature of outcome (dependent) variable</th>
<th>Name of regression type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>(Ordinary) Regression</td>
</tr>
<tr>
<td>Ranks</td>
<td>Ordered regression</td>
</tr>
<tr>
<td>Categorical: two categories</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Categorical: several categories</td>
<td>Multinomial regression</td>
</tr>
<tr>
<td>Counts (number of occurrences)</td>
<td>Poisson regression</td>
</tr>
<tr>
<td>Survival (independent variable: time)</td>
<td>Cox (proportional hazards)</td>
</tr>
</tbody>
</table>
time-to-event data): Survival depending on time as predicted by certain modes of treatment. The outcome is composed of the observed time for each patient and an indicator of whether the event of interest (e.g., death) was actually observed.

**Regression models in statistical software**

All even modestly capable software packages can analyze ordinary regression, including multiple regression. Curvilinear and simpler nonlinear regression models are included, because the analysis is the same as multiple regression, except that higher degree or nonlinear terms are substituted for additional variables. For example, a second-degree model with one predictor is just the first-degree model with two predictors, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, with the squared term replacing $x_2$ to form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$. The other forms listed in Table 15.1 appear in various software packages. The more complete packages, for example, NCSS, SAS, R, SPSS, Stata, Statistica, Systat, and others, provide the capability for all or most types. Often identification of the required data format and just what commands to use is not clearly specified.

**Where regression types are found in this book**

The goal of this book is to present the most frequently needed forms of analysis. This chapter addresses simple (ordinary) straight-line regression (Sections 15.2–15.4), Chapter 16, Multiple linear and curvilinear regression and multifactor analysis of variance, addresses curved and multiple regression, Chapter 17, Logistic regression for binary outcomes, addresses logistic regression, Chapter 18, Poisson regression for count outcomes, addresses Poisson regression, and Chapter 19, Analysis of censored time-to-event data, addresses Cox proportional hazards regression. Other types of regression, such as general liner models, are addressed in Chapter 20, Analysis of repeated continuous measurements over time. Statistical sophistication is increasing every year. In another decade, it is likely that the medical investigator frequently will require analyses not presented in this edition.

**15.7 CORRELATION CONCEPTS AND ASSUMPTIONS**

**What is correlation?**

“Correlation” as a popular term implies simply a relationship among events. In statistics, it refers to a quantitative expression of the interrelationship, or association, namely a coefficient of correlation. The population coefficient, which may be hypothesized but is generally unknown, is denoted by $\rho$ and its sample estimate by $r$. In the context of statistics, the term correlation usually refers to a correlation coefficient and will be used that way here. The (Pearson) correlation coefficient for ordered pairs of continuous variables was met in Section 5.3, where its calculation was given as the covariance of $x$ and $y$ divided by both standard deviations. The division has the effect of standardizing the coefficient;
it always represents points scattered about a 45 degrees line. In correlation, the level of association is measured by how tightly or loosely the \((x,y)\) observations cluster about the line, not by the slope of the line as in regression. Because this coefficient is standardized by dividing by the standard deviations, it lies in the range \(-1\) to \(+1\), with 0 representing no relationship at all and \(\pm 1\) representing perfect predictability. A positive coefficient indicates that both variables tend to increase or decrease together, whereas with a negative coefficient, one tends to increase as the other decreases. We would agree that a correlation coefficient of 0.10 implies little if any relationship between the two variables and that one of 0.90 indicates a strong relationship. Fig. 15.11 illustrates relationships between the coefficient and the scatter pattern of data.

In the case of unordered pairs (there are two groups but no way to assign a member of the pair to one particular group), or in case of three or more readings in each matched set, there exist *intraclass correlation coefficients*, or ICCs. More complicated than a Pearson coefficient, ICCs were met in Section 14.9.
Assumptions underlying correlation

Let us list assumptions about continuous-variable, or Pearson, correlation and compare them with the five regression assumptions from Section 15.2.

1. Correlation and regression require the same assumption: The errors in data values are independent one from another.

2. Correlation always requires the assumption of a straight-line relationship. A large correlation coefficient implies that there is a large linear component of relationship, but not that other components do not exist. In contrast, a zero correlation coefficient only implies that there is not a linear component; there may be curved relationships, as was illustrated in Fig. 15.4.

3. The assumption of exact readings on one axis is not required of correlation; both x and y may be measured with random variability, as was illustrated in Fig. 15.5.

4. and.

5. These assumptions take on a different form, because x and y vary jointly, so what is assumed for y relative to x is also assumed for x relative to y. x and y are assumed to follow a bivariate normal distribution that might be visualized as a hill in three dimensions, where x and y are the width and length and the height is the probability (could be read relative frequency) of any joint value of x and y. The peak of the hill lies over the point specified by the two means and the height of the hill diminishes in a normal shape in any direction radiating out from the means point. If the bivariate normal assumption is badly violated, it is possible to calculate a correlation coefficient using rank methods.

Rank correlation

Sometimes the data to be analyzed consist of ranks rather than continuous measurements so that a measure of correlation based on ranks is required, as was met in Section 5.3. At other times, the assumption of a bivariate normal distribution is violated, requiring that the ranks of the data replace the continuous measurements and a rank correlation is required. The coefficient that emerges, often referred to as Spearman’s correlation coefficient, may be interpreted in the same way.

Names for correlation

The correlation coefficient between two continuous variables, often called Pearson’s correlation, was originated by Francis Galton. British statistician Karl Pearson (who credits Galton, incidentally), along with Francis Edgeworth and others, did a great deal of the work in developing this form of correlation coefficient. Another name for this coefficient sometimes seen is product-moment correlation. Moments are mathematical entities related to the descriptors we use, such as the mean (first moment) and the variance (second moment about the first moment). Moments derived using a product of x and y are called
CHAPTER 15 Linear regression and correlation

product moments. The covariance used in calculating the correlation coefficient is a form of a product moment. These names are not really necessary except in the context of distinguishing correlation coefficients based on continuous variables from those based on ranks or categories. The rank correlation coefficient was first written about by C. E. Spearman, who also simplified the formula. The subscript $s$ (for Spearman) is attached to the population $\rho$ or sample $r$ to signify this form. (In the early part of the 20th century, before the convention of using Greek letters for population symbols and Roman for sample became standard, $\rho$ sometimes was used for the Spearman coefficient; if you encounter it, consider it a historical leftover.) A measure of correlation for categorical variables, the tetrachoric correlation coefficient, was addressed at the end of Section 5.3, and it might be well for the reader to review it here.

15.8 CORRELATION COEFFICIENTS

EXAMPLE POSED: THEOPHYLLINE LEVEL EXAMPLE CONTINUED

In the example of Section 15.3, DB3 data were used to predict serum theophylline level 10 days after antibiotic treatment from the baseline level. The covariance $s_{xy}$ was given as $11.8253$ and the regression line as $y = 1.1376 + 0.8340x$. Additional statistics are the standard deviations $s_x = 3.7655$ and $s_y = 3.9816$. What is the correlation coefficient $r$? What is the rank correlation coefficient $r_s$? Are they similar?

Methods for correlation coefficients

WHEN TO USE CORRELATION AS OPPOSED TO REGRESSION

Correlation is used when the question of interest is how closely $x$ and $y$ are associated (in a straight-line relationship). They are not thought of as dependent (outcome) and independent (causal) variables. There is no intent to predict one from the other as in regression, where $x$ is a factor (partially) causing $y$, or at least related to such causal factors. Cause and effect is not implied by correlation.

CALCULATING CORRELATION

The calculation of the correlation coefficient was given in Section 5.3 as

$$r = \frac{s_{xy}}{s_x s_y} = b_1 \frac{s_x}{s_y}. \quad (15.17)$$

The first expression, the covariance divided by the standard deviations, usually is used. However, if a regression has already been calculated, the second expression, the regression slope multiplied by the ratio of standard deviations, may be more convenient.
CALCULATING THE RANK CORRELATION

For each of $n$ pairs of $x$ and $y$ ranks in the sample, find the difference $x$-rank $- y$-rank; these are denoted $d_i$, the difference in $x$ and $y$ ranks for the $i$th patient. Square and add together these $d_i$. Spearman’s formula (simplified from the formula for the continuous coefficient) is

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$  \hspace{1cm} (15.18)

EXAMPLE COMPLETED: THEOPHYLLINE LEVEL EXAMPLE

By using the relationship of Eq. (15.17), $r = 11.8253 / (3.7655 \times 3.9816) = 0.7887$. Note that the second part of Eq. (15.1) can also be used, yielding $r = 0.8340 \times 3.7655 / 3.9816 = 0.7887$. This correlation coefficient of about 0.79 is rather high, indicating a close straight-line association. Were the assumptions underlying $r$ justified? The frequency distribution of baseline is not far enough from normal to worry us much, but the day 10 readings are. It is preferable to use $r_s$. Table 15.2 shows the DB3 data with ranks and $d_i^2$. $n = 16$ and $\sum d_i^2 = 192$. Substituting in Eq. (15.18), we find $r_s = 0.7176$, about 0.72, not far different from $r = 0.79$.

ADDITIONAL EXAMPLE 1: INFANT LUNG CONGESTION CONTINUED

In Additional Example 1 of Section 15.3 on infant lung opacity, $^2$ we posed the question: Is temperature associated with age? The covariance is $s_{xy} = 5.6727$ and the standard deviations are $s_x = 6.7478$ and $s_y = 2.2557$. $r = 5.6727 / (6.7478 \times 2.2557) = 0.3727$. The correlation coefficient is not high, but neither is it negligible. Is it statistically significant, that is, greater than what is likely to occur by chance? This question is answered in Section 15.10. If we had been investigating rather than illustrating, at the beginning of the exercise, we would have asked if the assumptions underlying correlation were
CHAPTER 15 Linear regression and correlation

satisfied. If we plot quick frequency distributions of temperature and age, we find temperature is approximately normal in distribution, but age is bimodal (although the two modes are rather weak) and is not far from a uniform distribution (nearly the same frequency for all age intervals). The assumptions for age are not satisfied; we should use $r_s$ rather than $r$. Upon ranking the data, then finding, squaring, and adding together the rank differences, and substituting in Eq. (15.18), we find $r_s = 0.3997$. This is rather close to the $r$ of 0.3727, so the violation of an assumption did not cause a major error in $r$.

ADDITIONAL EXAMPLE 2: HOSPITAL STAY AND INTELLIGENCE QUOTIENT FOR MENTAL PATIENTS CONTINUED

In Additional Example 2 of Section 15.3, we asked about an association between hospital stay and IQ at a mental hospital. A psychologist’s clinical experience suggests that brighter patients seem to stay in the hospital longer. If true, such a finding might result from brighter patients being harder to treat. IQ measurements ($x$) and days in the hospital ($y$) are collected for the next 50 patients treated. Means, standard deviations, and the covariance are calculated as $m_x = 100$, $m_y = 10.5$, $s_x = 10$, $s_y = 11.2$, and $s_{xy} = 39.43$, from which the use of Eq. (15.17) yields $r = 39.43/(10 \times 11.2) = 0.35$. A correlation coefficient of 0.35 is not decisive but is large enough to believe that some relationship exists. (A test of significance is performed in Section 15.10.)

Exercise 15.8

Using the descriptive statistics for the data given in Exercise 15.2, calculate the correlation coefficient between age and respiration rate. How is this correlation coefficient interpreted?

Exercise 15.9

Using the descriptive statistics for the data in Exercise 15.4, calculate the correlation coefficient. How is this correlation coefficient interpreted?

Exercise 15.10

As part of a study on circulatory responses to laryngoscopy and intubation (Exercise 15.3), the relationship between change in SBP due to laryngoscopy and SVmR was of interest. Readings were taken on $n = 26$ patients with the results: $s_{xy} = 2.2746$, $s_x = 0.1552$, and $s_y = 20.0757$. Calculate $r$. What does this $r$ indicate about the association of SBP change and SVmR? If $\sum d_i^2 = 863$, what is $r_s$? How well does $r_s$ agree with $r$?

15.9 CORRELATION AS RELATED TO REGRESSION

Both the correlation coefficient and the regression slope using the same data contain similar information about the relationship between $x$ and $y$. However, clearly they are not exactly the same since the sample slope $b_1$ is given by
\[ b_1 = \frac{s_{xy}}{s_x^2} = \frac{\text{cov}(x, y)}{sd(x)sd(x)}. \]  
\[ (15.19) \]

and the sample correlation \( r \) is given by

\[ r = \frac{s_{xy}}{s_x s_y} = \frac{\text{cov}(x, y)}{sd(x)sd(y)}. \]  
\[ (15.20) \]

Simple algebra will verify that

\[ r = b_1 \frac{s_x}{s_y}. \]  
\[ (15.21) \]

What does this difference imply? As noted in connection with Fig. 15.5, in correlation the fit arises from simultaneously minimizing the distances from each point perpendicular to the proposed line, while in regression the fit arises from simultaneously minimizing the vertical \((y)\) distances from each point to the proposed line. Thus regression makes the assumption that the \(x\) measurements are made without random error. Another difference is that regression may be generalized to curved models, whereas the correlation is restricted to straight lines only. A difference in interpretation is that correlation primarily is used to express how closely two variables agree (the width of an envelope enclosing all the additional information that 5-day level standard deviation points), whereas regression can indicate this (the amount of slope in the fit) plus provide a prediction of the most likely \(y\)-value for a given \(x\)-value. Both can provide confidence intervals and be subjected to hypothesis tests, but tests on regression parameters and models are more incisive, can be generalized to curved models, and can often be related to other statistical methods (such as analysis of variance).

**15.10 ASSESSING CORRELATION: TESTS AND CONFIDENCE INTERVALS**

**EXAMPLE POSED: ARE THEOPHYLLINE LEVEL CORRELATIONS DIFFERENT BY DAY?**

In the emphysema example, the correlation coefficient of baseline serum theophylline level with the 10 days after beginning an antibiotic was found to be \( r_{0,10} = 0.7887 \) in Section 15.8. Baseline standard deviation was found to be 3.7655. The slope of 5-day level predicted by baseline is estimated to be \( b_{0,5} = 0.8206 \). With the additional information that 5-day level standard deviation is 4.6093, the correlation coefficient between 5-day and baseline levels can be found from Eq. (15.21) as \( r_{0,5} = 0.8506 \times 3.7655 ÷ 4.6093 = 0.6949 \). Are the two correlation coefficients significantly different from 0? Is there probabilistic evidence that they differ one from the other, or is the observed difference due only to random fluctuations? What is a confidence interval on the 10-day versus baseline coefficient?
Methods for testing correlation and finding a confidence interval

The sample correlation coefficient, \( r \), estimates the population correlation coefficient, \( \rho \). It indicates how closely a scattergram of \( x, y \) points cluster about a 45 degrees straight line. A tight cluster (see Fig. 15.7) implies a high degree of association. The coefficient of determination, \( R^2 \), introduced in Section 15.4, indicates the proportion of ability to predict \( y \) that can be attributed to the model using the independent (predictor) variables. In the case of a single predictor \( x \) in a straight-line relationship with \( y \), \( R^2 \) is just the square of \( r \). It was noted that Eq. (15.10) provides a test statistic for the hypothesis that the population coefficient of determination is 0. The test may be rewritten for \( r \).

A SIGNIFICANCE TEST FOR \( \rho = 0 \)

The test of \( H_0: \rho = 0 \) is

\[
t_{(n-2)df} = \sqrt{\frac{(n-2)r^2}{1-r^2}}.
\]

(15.22)

If the calculated \( t \) is greater than a critical \( t \) from Table II, \( H_0 \) is rejected.

Suppose the correlation coefficient between two blood test measures for repeated samples of healthy people has proven to be some \( \rho_0 \), a theoretical correlation coefficient other than 0, perhaps 0.6, for example. We obtain a sample of ill patients and would like to know if the correlation coefficient between the blood tests is different for ill versus well patients.

A TEST FOR \( \rho \) OTHER THAN 0

We want to test if our sample correlation coefficient arose from a theoretical correlation coefficient \( \rho_0 \) other than 0. The (unknown) population coefficient from the sample to be tested is \( \rho \), estimated by \( r \). The null hypothesis becomes \( H_0: \rho = \rho_0 \). It has been shown mathematically that the expression

\[
m = \frac{1}{2} \ln \left( \frac{1 + r}{1 - r} \right)
\]

(15.23)

is distributed approximately normal (for larger samples, \( n > 50 \)) with mean

\[
\mu = \frac{1}{2} \ln \left( \frac{1 + \rho_0}{1 - \rho_0} \right)
\]

(15.24)

and standard deviation

(Continued)
The test is just the usual $z$ test on the standardized normal

$$Z = \frac{m - \mu}{\sigma}. \quad (15.26)$$

If the calculated $z$ from Eq. (15.26) is larger than a critical $z$ found from Table I, $H_0$ is rejected.

Suppose we have large-sample correlation coefficients between blood test measures for type 1 and type 2 diseases; we want to compare two population correlation coefficients, $\rho_1$ and $\rho_2$, using estimates from two independent samples.

A TEST OF TWO CORRELATION COEFFICIENTS

To test the null hypothesis that two $\rho$'s are equal, that is, $H_0: \rho_1 = \rho_2$, we calculate $m_1$ and $\sigma_1$ for sample 1 and $m_2$ and $\sigma_2$ for sample 2 in the forms of Eqs. (15.23) and (15.25). The test is a $z$ test conducted as with Eq. (15.26), where

$$Z = \frac{m_1 - m_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \quad (15.27)$$

No good tests have been developed for these cases for small samples.

A CONFIDENCE INTERVAL ON $\rho$

If $r$ is the estimator of the population $\rho$, we can find a confidence interval for $\rho$ as introduced in Section 8.9. However, the expression and calculation are somewhat bothersome. It is preferable to use a confidence interval on the regression $\beta_1$ if appropriate, but, if not, a few minutes with a capable calculator will provide the confidence interval given by Eq. (15.28).

$$\left(1 + r - (1 - r)e^{(2z_{1-\alpha/2})/\sqrt{n-3}}; \frac{1 + r - (1 - r)e^{-(2z_{1-\alpha/2})/\sqrt{n-3}}}{1 + r + (1 - r)e^{-(2z_{1-\alpha/2})/\sqrt{n-3}}} \right) \quad (15.28)$$
EXAMPLE COMPLETED: ARE THEOPHYLLINE LEVEL CORRELATIONS DIFFERENT BY DAY?

SIGNIFICANCE TESTS OF POPULATION CORRELATION COEFFICIENTS

In Section 15.4, addressing the prediction of the 10-day level by baseline level, we tested \( H_0: \beta_1 = 0 \) using Eq. (15.7), concluding that the \( t \) of 4.80 showed the prediction to be significant. It was noted that Eq. (15.7) gives the same result as Eqs. (15.19) or (15.22). Substitution of \( r = 0.7887 \) in Eq. (15.22) yields \( t = 4.80 \), indeed the same result. This sample \( t \) is much larger than 2.145, the two-tailed critical \( t_{0.95} \) for 14 \( df \), yielding \( p \)-value < 0.001 and indicating a significant association. Similarly, substitution of the 5-day versus baseline correlation of 0.6949 in Eq. (15.22) yields \( t = 3.62 \), also larger than the critical 2.145, with \( p \)-value = 0.001. There is strong evidence that both \( \rho_{0,5} \) and \( \rho_{0,10} \) are greater than 0.

TEST OF TWO CORRELATION COEFFICIENTS

First, we must note that we have only \( n = 16 \) in our sample, too small for a proper approximation. In addition, the two correlation coefficient estimates are not independent as they are computed from the same individuals. As such, we shall carry out the calculations only for illustration, not for a legitimate medical conclusion. Substitution of 0.7887 and 0.6949 in turn in Eq. (15.23) yields \( m_1 = 1.0680 \) and \( m_2 = 0.8574 \). From Eq. (15.25), both variances are \( 1/13 = 0.0769 \), yielding a pooled standard deviation of 0.3922. By substituting in Eq. (15.27), we find \( z = 0.5370 \). A critical \( z \), using two-tailed \( \alpha = 0.05 \), is the familiar 1.96. The calculated \( z \) is far less than the critical \( z \); we cannot reject the null hypothesis of no difference. We have insufficient evidence to establish that the two population correlation coefficients are different. The actual \( p \)-value = 0.2956.

A 95% CONFIDENCE INTERVAL ON THE POPULATION CORRELATION COEFFICIENT

Using the estimated correlation coefficient \( r = 0.7887 \), what is a 95% confidence interval for the population \( \rho \)? Substitution of \( r \), \( z_{1 - \alpha/2} = 1.96 \), and \( n = 16 \) in Eq. (15.28) yields a 95% confidence interval of (0.48, 0.92).

ADDITIONAL EXAMPLE 1: INFANT LUNG CONGESTION CONTINUED

A SIGNIFICANCE TEST OF THE CORRELATION COEFFICIENT

In the example of lung opacity in \( n = 234 \) infants,\(^2 \) the correlation coefficient between temperature and age was found to be 0.3727. Testing \( H_0: \rho = 0 \) by substitution in Eq. (15.22) yields a \( t \) of 6.12, very much larger than the critical \( t_{0.95} \) for 232 \( df \) of 1.97; \( p \)-value < 0.001. The \( x,y \) association is highly significant.

TEST OF TWO CORRELATION COEFFICIENTS

Of the 234 infants, 78 proved to have lung opacity on radiography (sample 1) and 156 not (sample 2). The correlation coefficients between temperature and age
for these groups were \( r_1 = 0.4585 \) and \( r_2 = 0.3180 \). By substituting in turn in Eq. (15.23), we find \( m_1 = 0.4954 \) and \( m_2 = 0.3294 \). Similarly, by using Eq. (15.25), \( \sigma_1^2 = 0.0133 \) and \( \sigma_2^2 = 0.0065 \). Substitution of these values in Eq. (15.27) yields \( z = 1.18 \), smaller than the 0.05 two-tailed critical \( z \) of 1.96. There is insufficient evidence to conclude a difference between the population correlation coefficients giving rise to these two samples.

**A 95% CONFIDENCE INTERVAL ON THE POPULATION CORRELATION COEFFICIENT**

What is the 95% confidence interval for the population \( \rho \)? The sample \( r = 0.3727 \), \( z_{1-\alpha/2} = 1.96 \), and \( n - 3 = 231 \). Substitution of these values in Eq. (15.28), we obtain a 95% confidence interval for \( \rho \) of (0.26, 0.48).

**ADDITIONAL EXAMPLE 2: INTELLIGENCE QUOTIENT VERSUS HOSPITAL STAY FOR MENTAL PATIENTS CONTINUED**

**A SIGNIFICANCE TEST THAT THE CORRELATION COEFFICIENT = 0**

Using \( n = 50 \) given in Section 21.3 and \( r = 0.35 \) in Section 15.7, we substitute in Eq. (15.22) to find \( t = 2.59 \), yielding \( p\text{-value}_{48 df} = 0.006 \). We reject the null hypothesis and conclude that the true correlation is greater than 0.

**A SIGNIFICANCE TEST AGAINST A CORRELATION COEFFICIENT OTHER THAN 0**

We note that 0.35 is not a very high coefficient despite being significantly greater than chance. Such a situation often occurs and is discussed further in Section 15.11. Let us say that the correlation coefficient must be at least 0.70 in order to be clinically useful. We perform a test to if the true correlation coefficient is significantly different from 0.70. By substituting \( r = 0.35 \) in Eq. (15.23), \( \rho = 0.70 \) in Eq. (15.24), and \( n = 50 \) in Eq. (15.25), we find \( m = 0.3654 \), \( \mu = 0.8673 \), and \( \sigma = 0.1459 \). By substituting these values in Eq. (15.26), we find \( z = -3.4400 \), yielding \( p\text{-value} < 0.001 \). We conclude that the true correlation coefficient is below the 0.70 that we consider minimum for clinical use.

**A 95% CONFIDENCE INTERVAL ON THE POPULATION CORRELATION COEFFICIENT**

By substituting \( r = 0.35 \), \( n = 50 \), and \( z_{1-\alpha/2} = 1.96 \) in Eq. (15.28), we obtain a 95% confidence interval for \( \rho \) of (0.08, 0.57).

**Exercise 15.11**

Let us continue with the SVmR example, in which the correlation between SVmR and SBP was 0.7295 for 26 patients. In Exercise 15.3, the hypothesis \( H_0: \beta_1 = 0 \) was tested. Repeat the test in the form of Eq. (15.22) to test \( H_0: \rho = 0 \). Suppose physiological theory posed a \( \rho \) of 0.5. Does this sample agree with the theory or differ from it?
15.11 INTERPRETATION OF SMALL-BUT-SIGNIFICANT CORRELATIONS

It is not unusual to encounter a small correlation coefficient that tests significant. Indeed, a coefficient as small as 0.06 becomes significant if we calculate it from 1000 patients, but \( r = 0.06 \) has no clinical use. Suppose we have a laboratory test designed to indicate the severity of a disease. The correlation between the test result and the disease severity is found to be statistically significantly different from 0, although the coefficient is only 0.20. Shall we use this test to designate treatment for the patient? Hardly. The interpretation of the significance may be stated as follows: \textit{the observed association between the test result and the severity would be a rare event if no association truly exists.} It does not say that the association is close enough to use one in place of the other in choosing treatment. To rely on the test result to select a treatment, we would want the coefficient to be quite high, we would hope in the 0.90s, but definitely not below 0.70. One too often sees small correlation coefficients reported in the medical literature as meaningful because they test significant. We must be wary of such reporting and be sure that the coefficient is large enough to be clinically meaningful rather than depending on its being greater than random occurrence.

REFERENCES

1. Missing sources. The sources of a few examples could not be found despite a strong effort to locate them. Such data that could not be referenced were slightly altered so as not to reflect on any investigator later appearing.
Multiple linear and curvilinear regression and multifactor analysis of variance

16.1 INTRODUCTION

Adjusting for multiple variables simultaneously

Chapter 15, Linear regression and correlation, introduced univariate regression as a representation of the dependence of one variable upon another (independent) variable. The concept may be extended to represent a variable as depending upon several (independent) variables. Furthermore, multiple variables may be used as predictors. The number of days a patient must be retained in the hospital may be predicted by body temperature, bacterial count, the presence—absence of emesis and neurologic signs, etc. In some cases, no single variable is very useful as a predictor, but several in consort may provide sufficient accuracy to be useful.

As an additional example, in an investigation of cardiovascular syncope, no one sign might be sufficient to predict its occurrence. The investigator might begin with, at least, systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate (HR) and extend the list of predictors to include perhaps the presence—absence of murmurs, clicks, vascular bruits, etc.

Finally, if our scientific goal is to estimate the association between a given predictor and some outcome, we may need to adjust for multiple other factors in the model in order to control for confounding. For example, we may be interested in the association between diabetes and SBP. Weight, however, is known to be associated with both diabetes and blood pressure. We term “weight” a confounder in the relationship between diabetes and blood pressure. In order to isolate the independent association between diabetes and blood pressure (above and beyond weight), it would be necessary to adjust for weight in a model relating the two.

Analysis by regression of a dependent variable on several independent variables is called multiple regression.

Interpretation of regression coefficients

As in Chapter 15, Linear regression and correlation, linear regression models the mean of the response in terms of predictors. Consider a simple linear regression model of the
form $E[y] = \beta_0 + \beta_1 x$, where $E[y]$ denotes the mean of the response variable $y$. ("$E$" denotes "expectation of"). We saw in Chapter 15, Linear regression and correlation, that $\beta_0$ is the mean of the response, $y$, in a population of individuals with predictor value $x = 0$. This may, or may not, have real-world meaning (e.g., if $x$ represents weight in lb, this would be the mean of the response in a population weighing 0 lb). While not always having real-world meaning, inclusion of the intercept in the model is generally necessary to provide an adequate model fit. $\beta_1$ is the difference in the mean of $y$ comparing two subpopulations differing in $x$ by 1 unit. This is an interpretable estimate of the effect of $x$ on $y$.

The abovementioned interpretations of the estimated regression coefficients change when multiple predictors are included in the model. For example, consider a multiple regression model with three predictors, $E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Here $\beta_0$ is the mean of the response, $y$, in a population of individuals with all predictor values equal to 0 (i.e., $x_1 = x_2 = x_3 = 0$). $\beta_1$ is the difference in the mean of $y$ comparing two subpopulations differing in $x_1$ by 1 unit but having the same value of $x_2$ and $x_3$. This is what we mean by the “adjustment” of $x_2$ and $x_3$. It allows us to fairly compare subpopulations differing in $x_1$ while holding the other predictor values constant. The interpretations of $\beta_2$ and $\beta_3$ are analogous.

**Curved line models**

Just as a model can be a segment of a straight line, a segment of a curved line can be used as well if the data are fit better by it or, preferably, if known physiologic or medical relationships follow a curve. The curve may be of any form: a segment of a parabola (the number of cells of a fetus or cancer may increase as the square of time units), a logarithmic form (the number of bacteria surviving antibiotic treatment may decrease as the logarithm of time), or even a cyclic pattern (aspergillosis has been seen to follow a sine wave over the seasons). Analysis by regression of a dependent variable on a curved pattern of an independent variable is called *curvilinear regression* and will be addressed in Section 16.5.

**Visualizing models in two dimensions**

We visualized a straight-line model in Section 15.3 and its algebraic representation in Eq. (15.3) as $y = b_0 + b_1 x$ (omitting the error of the fitted value for ease of exposition), a point to anchor a wheel of lines and a slope to specify the member of the wheel. As an example, let us look at the survival of rats infected with malaria. In part of the experiment giving rise to DB11, 100 infected rats were treated with RBC. Number (same as %) surviving by day for 10 days is plotted in Fig. 16.1 with a linear fit shown. The equation is $survival = 103 - 4 \times day\ number$, the negative slope showing that survival decreases with time.
Aside from a straight-line (first degree) model as seen in the preceding example, commonly seen curves are illustrated in Fig. 16.2A–F. These include (A) an upward opening parabola (second degree), (B) a third-degree curve (adding a $x^3$ term to a parabolic model), (C) a logarithmic curve, (D) an exponential curve, (E) a biological growth curve, and (F) a sine wave. Sign changes will flip the second- or third-degree
curves top for bottom in shape. The logarithmic curve looks a little like a portion of a downward opening parabola, but it never reaches a maximum as does the parabola, increasing, however, more and more slowly with increasing x. The exponential curve looks a little like a portion of the upward opening parabola but increases more rapidly. Growth curves fit many growth patterns, for example that of animal (and human) weight over time, or the volume of a cancer. Periodic curves, of which the sine wave is a simple case, are frequently seen in cardiopulmonary physiology.

**Visualizing models in three dimensions**

The case of y depending on two xs, the multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (omitting the error term for ease of explanation), can be visualized geometrically. If $y$ is SBP, $x_1$ might be age in years and $x_2$ might be minutes of vigorous exercise just prior to the measure. The $\beta$s are the parameters and might be thought of as slopes associated with their individual dimensions. To visualize the relationship, think of $y$ as the intersection of two walls in the corner of a room and the two $x$s as the intersections of these walls with the floor. The point $(y, x_1, x_2)$ is represented as a position in the space above and in front of the corner. $x_1$ and $x_2$ are readings on the independent variables (e.g., age and minutes of exercise) and $y$ is the reading on the dependent variable (e.g., SBP). Just as a line may be fit to a sample of points in two dimensions, a plane may be fit to a sample of points in three dimensions. Fig. 16.3 represents a three-dimensional space showing a single point and a plane that

![Figure 16.3](image-url)
might have been fit to a set of such points. For this figure, $\beta_0$ is the $y$-intercept (where the plane cuts the $y$-axis), $\beta_1$ is the slope of the line formed by the plane cutting the $x_1,y$-plane, and $\beta_2$ is the slope of the line formed by the plane cutting the $x_2,y$-plane.

A fit by a linear regression equation with two $x$ dimensions may be thought of as a plane in three dimensions, as depicted in Fig. 16.3.

A CURVED SURFACE IN THREE DIMENSIONS
Just as the straight line $y = \beta_0 + \beta_1 x$ may be generalized to represent a curve by adding a squared term, as $y = \beta_0 + \beta_1 x + \beta_2 x^2$, the plane $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ may be generalized to become a curved surface by adding one or two squared terms, as perhaps $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^2$. A curved surface rather than a flat surface in three dimensions may still be easily visualized.

More than three dimensions
Suppose the dependent variable $y$ depends on more than two dimensions, as if we added a measure of general cardiovascular health $x_3$. Algebraically, we have no problem. We can add as many variables in whatever powers we want, and the model may be readily used in analysis. Visualization of a relationship in four or more dimensions is difficult, if possible at all. Four-dimensional displays have been attempted using time (moving images), colors or shades, and other mechanisms, but with limited success, because depiction of the fourth dimension is necessarily different in nature from that of the first three. An algebraic expression of a four-dimensional fit would be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

The case of multiple dependent variables
Suppose we are concerned with both SBP and DBP as predicted by the several independent variables, we have $y_1$ and $y_2$ depending on the several $x$s. Such models usually are denoted multivariate. Methods exist to analyze such models, but they are beyond the scope of this book.

Analysis of variance and regression
An example of simple regression, a continuous dependent variable depending on a continuous independent variable, might be body mass index (BMI) depending on age. In Section 15.6, we looked at types of regression differing by types of dependent variable, assuming the independent variable was continuous. If we have a continuous dependent variable but categorical independent variable, we have a one-way analysis of variance (ANOVA), met in Section 11.4. An example might be BMI depending on ethnic group. In this simpler form, it may be thought of as the other side of the same coin. If our continuous dependent variable depends on
two categorical independent variables, we have a two-way ANOVA, to be addressed in Section 16.6. An example might be BMI depending on ethnic groups and sex. We can also have more than two categorical independent variables in various patterns, some of which will be met in later sections. An even more involved pattern of variables would be the dependent continuous variable depending on a combination of continuous and categorical variables, for example, BMI depending on age and ethnic group. This case could be treated by what is called analysis of covariance (ANCOVA) or by multiple regression with ethnic groups being separated into pairs by creating “dummy variables.” This is where regression and ANOVA meet.

### 16.2 MULTIPLE LINEAR REGRESSION

**EXAMPLE POSED: PREDICTING LENGTH OF HOSPITAL STAY DUE TO PSYCHOLOGICAL PROBLEMS**

A psychologist would like to be able to predict the length of hospital stay (LOS) (in days) for inpatients at time of admission. Information available prior to seeing the patient includes intelligence quotient (IQ, mean 100, standard deviation 10), age (years), and sex (0,1). Data are available for \( n = 50 \) inpatients. We might ask several questions: Can length of stay be predicted by three variables simultaneously, with each making a contribution? How much does each variable contribute to the prediction? Do any fail to contribute at all and may therefore be removed? Is the prediction significantly better than random selection? How good is the quality of the prediction?

**METHOD**

**Developing a Model**

There is no one rule for developing a model. Ideally, one *a priori* states their model depending upon the scientific question they wish to answer. This may include physiologic knowledge of the underlying process as well as all confounding variables that one deems appropriate to adjust for. This would be the way to perform a confirmatory analysis regarding an association of interest. Alternatively, the investigator may be exploring data to develop a model and formulate new hypotheses to be confirmed later and is using regression methods to identify the variables involved and their relative contribution. One approach is to test each potential predictor by univariate regression and add it to the model if it looks promising. If \( y \) is related to \( x_1 \) in a straight line, we add \( \beta_1 x_1 \) to the model. If \( y \) is related to \( x_1 \) in a second-degree curve, we add \( \beta_1 x_1 + \beta_2 x_1^2 \) to the model. Then we proceed with \( x_2 \), etc. (Components combining variables in the same term, perhaps \( \beta_3 x_1 x_2 \), but require careful interpretation of model parameters.) Another approach is to start with all potential (Continued)
variables included and eliminate those that do not contribute to the prediction. Both approaches are forms of stepwise multiple regression, discussed later in this section.

Data Input
Data are entered into the computer software model just as in simpler regressions. In univariate regression, we usually enter the patient number as a first column, the dependent variable $y$ as a second column, and the independent variable $x$ as a third. In multiple regression, we just add additional variables as additional columns.

Solving the Regression Equation
The regression curve calculated is the best-fit curve according to various mathematical criteria of "best", as appropriate to the form and assumptions. In simple regression, the criterion was least squares used for linear models (see Section 15.3). Other criteria are maximum likelihood (used more for nonlinear models), unbiasedness, minimum variance of deviations from the fit, etc., or combinations of these. For simple regression the least-squares solutions gave rise to relatively simple formulas, seen in Chapter 15, Linear regression and correlation. However, for more complicated models, sets of linear equations derived using calculus must be solved by methods of matrix algebra, which is not easily done by hand. Usually these are solved using computer programs contained in statistical software packages. It is reasonable to assume that any investigator using multiple regression will have access to such a package. This chapter does not attempt to present the mathematics of solutions but will concentrate on choosing the model, entering the data, and interpreting the results.

In order to estimate the $\beta$s that provide a best fit, the software solves a set of linear equations simultaneously. After calculating the best fit, the significance of the model is tested by $F$ and the associated $p$-value given. For a model with $n$ subjects and $k$ component variables (predictors), $F$ has $k$ (numerator) and $n - k - 1$ (denominator) df. A regression with model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ for 50 patients would have 2 and 50 \(2 - 1 = 47 \) df. Also, for each predictor, that part of its contribution that is not also provided by another predictor is tested by $t$ and the associated $p$-value given. These results are conceptually not very different from the equivalent results in simpler regressions. The coefficient of determination $R^2$ is also given. $R^2$ is $1$ – (residual variance/total variance), where total variance is $s_y^2$ and residual variance is the variance of differences of data from the model. $R^2$ estimates the proportion of total observed variation in the response that is accounted for by the model used.

Selecting and Interpreting Results
Statistical software packages usually display a number of results, many of which are needed only occasionally. The user must select those that answer the questions being asked of the data. The following types of results are the most likely to be of use: (1) estimates of the regression coefficients and corresponding inference (confidence intervals and $p$-values). The magnitude of regression coefficients provides information on clinical relevance and confidence intervals provide a measure of precision of these estimates. The coefficient estimates that are also essential if predictions are to be made from the model. (2) In validating an overall model or exploring data to identify relationships, the $p$-value of a test of the model (Continued)
(CONTINUED)

(usually an $F$ test) tells us whether the relationship between $y$ and the set of predictive components is probably real or probably just due to sampling fluctuations. (3) The coefficient of determination $R^2$ tells us whether the predictive capability of the overall model is clinically useful. The value $1 - R^2$ tells us the proportion of predictive capability attributable to causal factors not contained in the model, to a different model form, and to random effects. (4) $R^2$ also helps identify the clinically useful predictors. For a model with $k$ components, note its $R^2$.

An admonition

The $\beta$s estimating the $\beta$s used in the model depend on the units chosen. Temperature measured in degrees Fahrenheit will yield different coefficients from that measured in degrees Celsius. However, the statistical results, $F$s, $t$s, and $R^2$s, will not change.

Stepwise multiple regression

If we were to start with one predictor variable and add more variables one at a time, we would be following a form of forward stepwise regression. We would monitor the coefficient of determination $R^2$ (see Section 15.4) to see how much predictive capability each additional variable added and, given we added in the order of predictive strength ascertained from single predictor analyses, stop when additional variables added only an unimportant level of predictive capability. More frequently, the entire set of predictors under consideration is used at the start, the strategy being to eliminate the least contributive variables one by one until elimination excessively reduces the predictive capability. This form is known as backward stepwise regression. Statistical software is capable of performing this task with one command with the added benefit that at each elimination all previously eliminated variables are retried. This process corrects the error of correlation between two variables sometimes leading to removal of one that would not have been removed had the other one not been present at the outset. However, the addition—removal decisions are made on the basis of mathematics and the investigator loses the ability to inject physiological and clinical information into the decision. For example, if two correlated variables contribute only slightly different predictive ability and one should be removed, the software may remove one that occurs in every patient’s chart while leaving one that is difficult or costly to measure. In addition, it is important to note that the interpretation of the model parameters changes as other predictors are added or removed from the model. In truth, one is estimating a different association for each predictor with each change in the other predictors. In summary, performing backward stepwise regression by statistical software is done with the cost of losing control over which correlating overlapped variables are
removed and leads to changing interpretations of the model coefficient estimates at each step of the process.

If stepwise regression is to be performed, what is the criterion for removal of a variable? To perform a software-based backward stepwise regression, we must specify the cut-point significance level, say $P(r)$, for removal from versus retention in the model. Variables with $p$-values greater than $P(r)$ are removed and those with $p$-values less are retained for further consideration.

### Managing nominal variables

Independent variables used to predict the dependent variable may be used just as recorded if they are continuous measurements or if they can be put in rank order (e.g., small to large). The only type of variable that must be altered for use in multiple regression is nominal data that cannot be ordered, such as ethnic group or disease type. A dichotomy, perhaps malignant versus benign, would not require special treatment. However, a categorization with three or more classes cannot be used as is. A racial heritage categorization classified as European, African, or Asian, for example, if assigned recording codes 1, 2, and 3, respectively, would predict a disease state differently than if assigned 1, 3, and 2. The method will interpret the coding as a value and interpret the prediction to use an Asian as three times the value of a European. In this case, a strategy to use such classes as predictors is to reduce them to a set of dichotomies, that is, to create “dummy variables”. Variables are created for each class, denoting “that class” as, say, 1 and “not that class” as 2. Thus we would have a “European variable” with Europeans denoted 1 and non-Europeans 2, an “African variable” with Africans denoted 1 and non-Africans 2, etc. For $k$ classes in the categorization, we need only $k-1$ variables to replace the original variable. The last category’s variable is redundant, because all the information needed is contained in the first $k-1$ variables.

**EXAMPLE COMPLETED: PREDICTING LENGTH OF HOSPITAL STAY DUE TO PSYCHOLOGICAL PROBLEMS**

**CHOOSING THE MODEL**

The dependent variable days in hospital is $y$. The IQ, age, and sex are $x_1$, $x_2$, and $x_3$, respectively. The model becomes $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. With three variable-containing terms in the model, the $F$ for the model contains $k = 3$ and $n-k-1 = 50 - 3 - 1 = 46$ df. The above is our a priori model and should be utilized for confirmatory inference.

**DATA INPUT**

In the software package, prepare the data sheet with the list of patient identifiers as the first column. Label the next four columns as $y$, $x_1$, $x_2$, and $x_3$. For each patient, enter the datum for days in into the $y$ position, for IQ into the $x_1$ position, for age into the
CHAPTER 16 Multiple linear and curvilinear regression

$x_2$ position, and for sex into the $x_3$ position. Then select and run multiple regression from the package.

MODEL ASSESSMENT AND DATA-DRIVEN MODEL BUILDING

If we wished to explore our model fit and potentially modify the model in order to refine for future hypothesis confirmation or prediction, we might consider the following steps: (1) the computer result gives the $p$-value of the $F$ test to be 0.030. We conclude that the predictive ability of the overall model is unlikely to be due to chance. (2) $R^2 = 0.175$. This tells us that, even though the model is significant as a predictor, it represents less than 18% of the predictive capability; 82% remains for other predictors and the influence of randomness. (3) Interpolating from Table II, the critical value of $t_{0.95}$ for 46 df is 2.01. The values of $t$ calculated for the independent portion of each $x$ are IQ 2.258, age 1.624, and sex 0.244. Only IQ is greater than the critical value, but the issue is to identify the predictors that contribute to the prediction. Sex produced the smallest $t$ and appears not to be a useful predictor. Let us delete sex and recalculate the multiple regression, using the new model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. (1) The $p$-value for the $F$ test is 0.011, actually somewhat improved. (2) $R^2 = 0.174$. We lost nothing by dropping sex; it is not clinically useful. (3) The values of $t$ calculated for the independent portion of each $x$ are IQ: 2.312 and age: 1.688. IQ is still significant ($p$-value = 0.025) and age not ($p$-value = 0.098). We ask how much age contributes to the prediction. We delete age and recalculate the regression, using the reduced model $y = \beta_0 + \beta_1 x_1$. The significance of the model is little changed ($p$-value = 0.012), but $R^2 = 0.124$, a notable drop in predictive ability. We conclude that we should retain age for prediction purposes, even though it is not of significant use individually. We return to the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. (4) The computer output lists the estimates of the $\beta$s as $b_0 = -32.5938$ (constant), $b_1 = 0.3576$ (coefficient of IQ), and $b_2 = 0.1686$ (coefficient of age). The resulting predictive model is: days in hospital = $-32.5938 + 0.3576$IQ $+ 0.1686$age. Patient number 10 remained 4 days in the hospital. His IQ was 90.4 and age was 24. His predicted days in was $-32.5938 + 0.3576 \times 90.4 + 0.1686 \times 24 = 3.8$.

ADDITIONAL EXAMPLE 1: INFANT LUNG CONGESTION

Let us consider again the prediction of temperature in infants with pulmonary complications, first used in Additional Example 1 of Section 15.3. Age was found to be a significant predictor. Data on HR and pulse oximetry had also been recorded. Are they useful predictors?

CHOOSING THE MODEL

This is a case where we do not have a model in mind but seek to determine which predictors may be associated with temperature. Hence, our model building will be data-driven as opposed to seeking to confirm a given association. The model for age
As a predictor of temperature \((y)\) was \(y = \beta_0 + \beta_1 x_1\). Adding terms for HR \((x_2)\) and pulse oximetry \((x_3)\) gives the model \(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\). There are \(n = 234\) patients. The number of variable-containing terms is \(k = 3\). The model has 3 and \(n - k - 1 = 234 - 3 - 1 = 230\) df.

**DATA INPUT**

In the data spreadsheet, patient numbers are entered into the first column, the sample’s temperature values are into the second column \((y)\) position), age value into the third column \((x_1)\) position), HR value into the fourth column \((x_2)\) position), and pulse oximetry into the fifth column \((x_3)\) position).

**MODEL ASSESSMENT AND DATA-DRIVEN MODEL BUILDING**

To explore our model fit and potentially modify the model in order to refine for future hypothesis confirmation or prediction, we might consider the following steps:

1. The computer result gives the \(p\)-value of the \(F\) test to be \(< 0.001\). We conclude that the predictive ability of the model is real, not due to chance.
2. \(R^2 = 0.289\). This tells us that the model represents only about 29\% of the predictive capability (71\% remaining for other predictors and the influence of randomness), which may be clinically useful, although not definitive.
3. The values of \(t\) calculated for the independent portion of each \(x\) are age: 7.117, HR: 6.895, and pulse oximetry: 0.111. The first two are large compared to the critical \(t\) of 1.97, but pulse oximetry appears not to be a useful predictor. Let us delete pulse oximetry and recalculate the multiple regression, using the new model \(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2\).

\(1\) The \(p\)-value for the \(F\) test is still \(< 0.001\).

\(2\) \(R^2 = 0.284\). We lost only a negligible amount of the predictive ability by dropping pulse oximetry; it is not clinically useful.

\(3\) The values of \(t\) calculated for the independent portion of age and HR were significantly large, so we conclude that both are clinically useful predictors.

\(4\) The computer output lists the estimates of the \(\beta\) s as \(b_0 = 94.4413\) (constant), \(b_1 = 0.1333\) (coefficient of age), and \(b_2 = 0.0354\) (coefficient of HR). The resulting predictive model is temperature \(= 94.4413 + 0.1333(x) + 0.0354(HR)\). Infant number 1 had an observed temperature of 101.4\(^\circ\)F. Her age was 1 month, and HR was 180. Her predicted temperature was \(94.4413 + 0.1333 \times 1 + 0.0354 \times 180 = 100.9\)\(^\circ\)F, which is within half a degree.

**ADDITIONAL EXAMPLE 2: CAN WE PREDICT LENGTH OF HOSPITAL STAY DUE TO STREP INFECTION?**

**UNIVARIATE REGRESSIONS**

A sudden outbreak of *Streptococcus pneumoniae* occurred at the Marine Corps Recruit Depot in San Diego. Can LOS be predicted by clinical readings and laboratory test results? Data from 128 patients included LOS, albumin, white blood cell (WBC), age, temperature, HR, respiration rate, pulse oximetry, and rate of coughing spells. Let us
explore the data to understand it better. Performing univariate regression of LOS on each potential predictor yields Table 16.1, showing sample sizes \( n \), \( p \)-values from \( t \) tests on the slope of the regression line, and \( R^2 \) values. No variable shows both a \( p \)-value high enough and an \( R^2 \) low enough to reject it off hand.

Entering the predictors into a multiple regression yields an overall \( F_{8,74} = 9.72 \) with \( p \)-value \( < 0.001 \), a highly significant prediction of LOS, and \( R^2 = 51\% \), telling us that this set of predictors accounts for more than half the possible perfect prediction.

### Table 16.1 Results of univariate regression on potential variables in hospital length of hospital stay prediction.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( n )</th>
<th>( p )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin</td>
<td>116</td>
<td>(&lt; .001)</td>
<td>0.294</td>
</tr>
<tr>
<td>WBC</td>
<td>120</td>
<td>(&lt; .001)</td>
<td>0.124</td>
</tr>
<tr>
<td>Age</td>
<td>127</td>
<td>.215</td>
<td>0.215</td>
</tr>
<tr>
<td>Temperature</td>
<td>127</td>
<td>.181</td>
<td>0.181</td>
</tr>
<tr>
<td>Heart rate</td>
<td>121</td>
<td>(&lt; .001)</td>
<td>0.198</td>
</tr>
<tr>
<td>Respiration rate</td>
<td>107</td>
<td>(&lt; .001)</td>
<td>0.229</td>
</tr>
<tr>
<td>Pulse oximetry</td>
<td>114</td>
<td>(&lt; .001)</td>
<td>0.246</td>
</tr>
<tr>
<td>Cough frequency</td>
<td>127</td>
<td>.262</td>
<td>0.010</td>
</tr>
</tbody>
</table>

\( WBC \), White blood cell.

---

**THE EFFECT OF MISSING OBSERVATIONS ON DF**

Before we examine the multiple regression, let us ask why there were 128 patients, yet only 8,74 \( df \) in the \( F \) test! With \( k = 8 \) independent variables, we would expect the residual \( df \) to be \( n - k - 1 = 119 \). The reason is that the statistical software performs regression only on those patients having no missing readings. Albumin is missing for 12 patients and respiration rate for 21 patients, yet only 2 patients are missing both; thus 31 patients have been dropped from the regression for missing data in at least one of those two variables, dropping the \( df \) available from 119 to 88. Other variables are missing in 14 more cases, dropping the residual \( df \) to 74. We must be sure we do not lose so many \( df \) to missing data that our analysis is compromised. We must keep track of the \( df \) as we progress in the analysis. We must also be aware that, if missing data are not random but due to a specific cause, we could be introducing a sampling bias by dropping these patients. We must ask if there is anything unique to the patients whose data are missing.

### A FULL MODEL REGRESSION

The multiple regression gave a highly significant \( p \)-value \( < 0.001 \) and \( R^2 = 51\% \). Table 16.2 shows the \( p \)-values in the same order as in Table 16.1.
Cough, WBC, and age all show very high p-values, suggesting that they contribute negligibly to the prediction. Cough and age were not significant univariate predictors, but we immediately note that WBC’s univariate p-value is < 0.001 in contrast to its p-value of 0.901 after adjustment for the other model predictors. This is because WBC is correlated with other variables, especially albumin, where \( r = 0.51 \). It was noted under “stepwise regression” earlier in this section that multiple regression eliminates the overlapping effect of two correlated predictors, leaving the influence of the uncorrelated portions alone. In this case, the overlap of WBC with other variables was eliminated, showing that after adjustment for the other predictors in the model there is little association between WBC and the response.

**ELIMINATING VARIABLES**

We drop cough, WBC, and age from the model as poor predictors and rerun the multiple regression, finding the recalculated p-values as in Table 16.3. The model retains its p-value and \( R^2 \); it has lost no predictive capabilities by eliminating the three variables.

### Table 16.2 Multiple regression p-values for eight potential variables in hospital length of hospital stay prediction.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin</td>
<td>.006</td>
</tr>
<tr>
<td>WBC</td>
<td>.901</td>
</tr>
<tr>
<td>Age</td>
<td>.614</td>
</tr>
<tr>
<td>Temperature</td>
<td>.100</td>
</tr>
<tr>
<td>Heart rate</td>
<td>.005</td>
</tr>
<tr>
<td>Respiration rate</td>
<td>.128</td>
</tr>
<tr>
<td>Pulse oximetry</td>
<td>.024</td>
</tr>
<tr>
<td>Cough frequency</td>
<td>.966</td>
</tr>
</tbody>
</table>

The model’s overall \( n = 83 \), p-value < 0.0001, and \( R^2 = 51\% \). WBC, White blood cell.

### Table 16.3 Multiple regression p-values for five potential variables in hospital length of hospital stay prediction.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin</td>
<td>.006</td>
</tr>
<tr>
<td>Temperature</td>
<td>.100</td>
</tr>
<tr>
<td>Heart rate</td>
<td>.005</td>
</tr>
<tr>
<td>Respiration rate</td>
<td>.128</td>
</tr>
<tr>
<td>Pulse oximetry</td>
<td>.024</td>
</tr>
</tbody>
</table>

The model’s overall \( n = 83 \), p-value < 0.0001, and \( R^2 = 51\% \).
We now note that respiration rate is not significant, showing the probability of a false positive of nearly 13%. Eliminating respiration rate and rerunning the regression yields the results of Table 16.4. All remaining predictors are significant. \( R^2 = 47\% \), having lost 4% due to eliminating respiration rate, not a debilitating loss. We now want to know what the predicted LOS values are with the remaining model and how good that prediction is. We note that reducing some of the variables reduced some of the missing data, leaving a sample size increased to 96 out of the potential 128.

**THE PREDICTION MODEL**

The final model of the form \( y = b_0 + b_1 x_1 + \cdots + b_4 x_4 \) is given by Eq. (16.1).

\[
\text{LOS}_{\text{calc}} = 78.88 - 2.22 \times \text{albumin} - 0.41 \times \text{temperature} + 0.09 \times \text{heart rate} - 0.37 \times \text{pulse oximetry}.
\]  

(16.1)

Upon substituting the predictor values in Eq. (16.1) for the 96 readings having all data present, a calculated LOS emerges with difference from actual LOS having mean 0.002 and standard deviation 0.328. (A few predictions having small negative values were replaced by 0.)

**PERFORMING THE BACKWARD STEPWISE REGRESSION BY SOFTWARE**

In “backing out” noncontributing variables by hand, we performed a form of backward stepwise regression. To perform it by software, we must specify the significance level, \( P(r) \), that separates removal from retention. Following Table 16.3, we removed a potential predictor with \( p \)-value > 0.10 and retained one with \( p \)-value = 0.10, so let us specify \( P(r) = 0.10 \). A software-based backward stepwise multiple regression on the eight potential variables yields a solution as in Table 16.5.

Inspection shows that the results are similar to those obtained by the manual backward stepwise regression, but not quite the same. The same predictors are chosen, but the analysis started with the reduced \( df \) due to missing data and maintained the same data subset for analysis throughout, accounting for the discrepancy.
A FINAL NOTE ON MODEL VALIDATION

The previous example generated a prediction model based upon inferential testing. While this is one approach to building such a model, it does not directly address our primary goal in prediction modeling: Is our model good for predicting new independent observations? As discussed in Chapter 2, Planning analysis: how to reach my scientific objective, in order to do this we would need to validate the model (and possibly other candidates) on an independent test data set. If prediction is the primary goal of an analysis, this should always be performed.

ADDITIONAL EXAMPLE 3: CAN FASTING CHOLESTEROL BE PREDICTED WITHOUT FASTING?

Often fasting total cholesterol is wanted on a patient who failed to fast. Can nonfasting cholesterol-related measures be used to predict fasting cholesterol? Total cholesterol after fasting, and taken upon another occasion after eating, was recorded for 33 subjects along with component measures low-density lipoproteins (LDL), high-density lipoproteins (HDL), triglyceride level, and glucose level. Fasting total cholesterol was regressed on the five after-eating measures, yielding overall $p$-value < 0.001 and $R^2 = 86\%$. The $p$-values for the individual predictors appeared as in Table 16.6. It appears that only total cholesterol after eating contributes to the prediction. Upon removing the other four variables, $p$ remains < 0.001 and $R^2 = 85\%$, as in Table 16.7. In total, the other four predictors contributed only 1% of the possible perfect prediction. A backward stepwise regression with a removal probability of 0.20 yields the same result.

What, then, is the prediction of fasting total cholesterol by total cholesterol after eating? By substituting the coefficients from Table 16.7 into a linear equation similar to that of Eq. (16.1), we find

$$\text{Predicted fasting total cholesterol} = 6.325 + 0.969 \times \text{Eating total cholesterol}. \quad (16.2)$$

Upon calculating the prediction and the difference between that prediction and the measure for the 33 subjects, we find (rounded to 0.1 mg/dL) a mean difference of...
0 mg/dL with a confidence interval (CI) ± 4.9 mg/dL. This result appears rather good but note that the CI is on the mean (using $s_m$), not on an individual reading (using $s$). To use the prediction clinically as a prediction on an individual patient, we want 95% confidence bounds of

$$
\pm t_{32df, 0.025} \times s = \pm 2.04 \times 13.87 = \pm 28.3.
$$

Thus our prediction of fasting total cholesterol from eating total cholesterol may be in error by 28 mg/dL or more one time in 20. For our 33 subjects, we expect 1 or 2 (calculated 1.65) exceedances and indeed we find 1.

**Exercise 16.1**

By using the data of DB10, we may be able to predict the strength of hamstrings, quadriceps muscles, or tendons following surgery by using the strength and control of the unoperated leg.

Choosing the model. Strength following surgery is measured by the distance covered in a triple hop on the operated leg. Possible predictors are the equivalent on the unoperated leg ($x_1$) and the time to perform the hop ($x_2$). Write down the equation of the model.

Data input. How are which data to be entered into a data spreadsheet?

Results selected and their interpretation. $F$'s $p$-value $< 0.001$. $R^2 = 0.974$. $t$ Values are 11.617 for $x_1$ and 1.355 for $x_2$. By omitting $x_2$, the F test's $p$-value $< 0.001$, $R^2 = 0.965$, and $t = 12.832$. For the simple regression of $y$ on $x_1$, $b_0 = -300.868$ and $b_1 = 1.466$. Is the predictive ability of the model using both $x$'s real? Is it clinically useful? Judging from the $t$
values of each $x$, should we consider dropping $x_2$? If we do, is the reduced model significant? Clinically useful? What is the predictive equation for the reduced model? If the triple hop distance for a particular patient is 504 cm, what is the predicted distance for the operated leg? How does this agree with an observed distance of 436 cm from DB10 data?

16.3 MODEL DIAGNOSIS AND GOODNESS OF FIT

Underlying assumptions of the linear regression model

As with any statistical method, the linear regression model makes particular assumptions about the underlying data being analyzed. Assumptions about the linear regression model are placed upon the residuals of the model, or the random error term that exists after we have modeled the mean as a linear combination of variables and the regression coefficients. In the previous explanations, we have purposefully omitted the error term in order to simplify the discussion of the model and ease interpretations. In order to check the assumptions of our model, based upon a given fit, we must now focus on the fitted error terms or residuals. Specifically, we can write the multiple linear regression model as

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + e,$$

where $e$ is the model residual. In practice, the residual is estimated by subtracting the fitted model from the observed response:

$$e = y_{\text{obs}} - (b_0 + b_1 x_1 + \cdots + b_k x_k).$$

In order for the exact theory of the linear regression model to hold, we assume that the residuals are (1) independent, (2) are normally distributed, (3) have mean 0, and (4) have constant variance.

Luckily, most of these assumptions are generally ensured in practice. For (1), we can generally know whether or not are data are independent by understanding how the sampling of the data was performed. For example, if a random sample is taken from the population where one observation is obtained on each patient, then independence will hold. However, if we were to repeatedly measure each patient over and over again, then there would be correlation among the observations within each patient (e.g., blood pressure measurements obtained on one patient are more alike than blood pressure measurements obtained across different patients).

For assumption (2), it turns out that in large samples, tests and CIs for the estimated regression parameters in the linear regression model are still valid even if the assumption of normality does not hold. This is because estimated coefficients are weighted averages of the responses and the central limit theorem ensures that they will be asymptotically normally distributed under fairly technical, yet mild, conditions.
For (3), to ensure that the model residuals have mean 0, it is simply necessary to include an intercept term, \( b_0 \), in the model. Doing this means that any deviations in the response not accounted for by the terms with variables will be collapsed into the intercept. Thus it is almost always advisable to include an intercept in any linear regression model.

Assumption (4) tends to be the most violated in practice. The reason for this is because the variance of a response often changes when the mean grows large or becomes small. Since the linear regression seeks to model the mean at many different variable values, the mean will be changing and hence the variance may also be changing.

In light of the above, the two most important diagnostic procedures that should be performed are a check of normality of the residuals in small samples and a check of whether the residuals have constant variance. Beyond this, it is also important to determine if one or a handful of observations have large influence on the estimated regression coefficients.

**Normality**

To assess normality of the residuals, we can start by simply plotting a histogram of the residuals to observe whether or not they follow the classic bell-shaped normal distribution. In some cases, we may see that the distribution of the residuals is obviously skewed in one direction or another; hence, the assumption is violated. This, however, is a very subjective way of assessing normality and may not be adequately sensitive. A slightly better approach is to produce a quantile-quantile (or QQ) plot. A QQ plot simply plots the estimated cumulative distribution of the observed residuals against the theoretical cumulative distribution of the residuals that would be expected if the residuals truly were normally distributed. If the normality assumption holds, the points on the QQ plot should roughly follow the 45-degree diagonal line. Fig. 16.4 displays an example of QQ plot. We can see that for the majority of points, they fall very close to the 45-degree line. There are a handful of points at the tails of the distribution that are off of this line, but this is typically the case as one moves toward the ends of the distribution.

A formal test of whether the residuals deviate from the line can be conducted. Two of the most popular tests are the Shapiro–Wilk test and the Kolmogorov–Smirnov test, addressed in Chapter 13, Tests on variability and distributions. Both will return a \( p \)-value corresponding to the test of the null hypothesis that the residuals follow a normal distribution. However, as was stated earlier, in larger sample sizes (say 30 and above with a modest number, 2 or 3, variables in the model), the central limit theorem will ensure validity of the tests of the regression coefficients. If sample sizes are small and the normality assumption appears to be violated, transforming the response (e.g., taking the log of the response) may help. However, it should be noted
that transforming the response will also change the interpretation of the regression parameters as we are now modeling the mean of the log-response variable.

**Constant variance**

Constant variance of the residuals (remaining constant throughout the range of the variable, technically termed “homoscedasticity”) is a key assumption to ensure valid tests of regression coefficients in the linear regression model. To investigate whether the residuals have constant variance, the most useful diagnostic tool is to plot the residuals against the fitted values. If the constant variance assumption holds, then the vertical spread of the residuals should be roughly constant as we move across the x-axis of the plot. Fig. 16.5 (left) provides an example of residuals that exhibit constant variance. Fig. 16.5 (right) is an example of nonconstant variance (technically termed “heteroscedasticity”). In the case of Fig. 16.5 (right), we can see that the vertical spread grows along the x-axis, resulting in a bullhorn shape. This is a classic example of heteroscedasticity, in which the variance of the residuals grows with the mean. In Section 16.4, we discuss three common strategies for dealing with heteroscedasticity.

**Influential observations**

In statistics, an influential observation is an observation that has a larger than usual impact on the fitted model. We often think of influential points as being outlying
observations. While it is true that in order to have influence, an observation must be an outlier, not all outliers have influence. More specifically, an observation can only have high influence if (1) the predictors for the observations are quite different from the rest of the sample and (2) given the predictors, the observation’s response is quite different from what is expected. The result of both (1) and (2) occurring means that our model estimates will be highly affected by the observations. Fig. 16.6 (left) provides an example where (2) holds, but (1) does not. In this case, the slope of the fitted regression line would not change if the red observation were removed from the sample. Fig. 16.6 (right), however, is an example where both (1) and (2) hold. In this case, the slope that would be estimated would be higher if the red observation were

Figure 16.5 Example of homoscedastic (left) and heteroscedastic (right) residuals.

Figure 16.6 The left is an example of an observation that is an outlier but does not have high influence on the slope parameter. The right is an example of a highly influential point. In this case, the observation is an outlier in the $x$-space as well as not fitting the general trend of the rest of the data.
removed from the sample. Thus we say that the red observation in Fig. 16.6 (right) has high influence.

Assessing influential observations often focuses on determining if the predictors for the observations are quite different from the rest of the sample and then determining if our model estimates changes dramatically when the observation is removed from the analysis.

**Leverage**

To judge if the predictors for a given observation are quite different from the rest of the sample, we focus on what is called leverage. This is a fancy word that simply means we wish to compare the difference in the predictor values for a given observation to the mean predictor value of all of the observations in the sample. If this difference is large, then we would say that the observation has high leverage, implying that observation’s predictor values are quite different from all of the other observations in the sample. Referring back to Fig. 16.6, the red observation in the left figure does not have high leverage because the value of the predictor for this observation is near the center of the distribution of the predictor values for the rest of the sample. Conversely, the red observation in the right figure does have high leverage because the value of the predictor for this observation is appreciably higher than the mean of the predictor values for the rest of the sample. In order to be influential, an observation must have high leverage.

**Cook’s distance**

A natural measure of the influence of an observation is the amount that estimated model coefficients change when the observation is removed from the data set. If the changes in the model estimates are large when the observation is removed, this would indicate that the observation has a high influence on the model fit. Cook’s distance can be thought of as a standardized average of the change in model parameter estimates when each observation is removed. The formula for computing Cook’s distance is beyond the scope of this textbook, though it does have a fairly easy interpretation: observations with a large Cook’s distance tend to have high influence. A useful plot for identifying influential points is to plot Cook’s distance against leverage. Fig. 16.7 displays an example plot to assess Cook’s distance in relation to leverage. The numbers plotted next to each point are used to identify the observation number for that point. The horizontal reference line in the plot is two times the standard deviation of Cook’s distance. The vertical line is three times the average leverage in the data set. Observations in the upper right quadrant of the plot should be examined for their influence. We can see that for this example, no observations have great influence.
What to do about influential points?

If influential observations are identified, it begs the question as to what should be done with them. The first step is to make sure that the influential observation is not simply a data recording error. If this is an error, the answer is easy in that it should either be fixed or, if the true value of the observation is unknown, it might be removed from the analysis. If the observation is not an obvious data recording error, the answer is not as simple. We do not simply wish to “throw away” observations that do not fit our models. Doing so would likely bias our results. However, we also know that these observations are special in that they do not fit the general trend of the data. We believe that, in the case of highly influential points, best practice is to present both model estimates with and without the influential points and clearly articulate the impact of the influence to one’s audience. This allows for the educated reader to determine the weight that they would place on the estimates including or excluding the influential point(s).

16.4 ACCOUNTING FOR HETEROSCEDASTICITY

Previously, we discussed the concept of heteroscedasticity, or nonconstant variance in the residuals. If heteroscedasticity is present in the data, it is important to account for it in the analysis in order to ensure that hypothesis tests and CIs for regression parameters
are valid. There are essentially three approaches to this: (1) transforming the response, (2) differentially weighting the observations, and (3) “fixing” the estimated variance of the regression coefficient estimates.

**Transformations**

Most textbooks will advise transforming the response of linear regression model in order to obtain constant variance in the residuals. The process is simple and essentially comes down to trial and error. Common approaches would be to take the log or the square root of the response variable then determine if the residuals have roughly constant variance after the transformation (using the same diagnostic plot that we previously discussed).

The problem with transformations is that they inherently change the interpretation of the model parameters; hence, the model may no longer answer the scientific question of interest. Linear regression estimates the changes in the mean of the outcome, which are associated with changes in the predictors. This is a very natural interpretation that we are often scientifically interested in. If, however, we log-transform the response, then the interpretation of the coefficients is in terms of changes in the mean of the log of the response. This is not often what we are interested in (e.g., we do not often ask about the differences in the mean log-blood pressure between treated and untreated patients). As such, we tend to avoid using transformations as a remedy for nonconstant variance.

**Weighted least squares**

Weighted least squares is another approach to solving the problem of heteroscedasticity. In this case, if we knew (or had a good estimate) of what the variance of each residual was, then multiplying each observation (response and predictors) by the inverse of the square root of the variance for the corresponding residual would result in homoscedasticity. This approach is theoretically optimal; however, it generally does not work well in practice. The reason is that we usually do not have a good estimate of the variance of each observation. As we have learned, a sample size of one does not generally result in a very precise estimate of a population parameter! As such, unless there is strong information regarding what the variance of each residual is, weighted least squares in the case of continuous independent data is generally not practical in real-world data analyses.

**Empirical variance estimation**

We noted deficiencies with the previous two approaches. The final approach is to simply try and “fix up” the problems that heteroscedasticity causes when we try to estimate the distribution of the estimated regression parameters in order to draw inference
on them. To do this, we can simply try to estimate the variance of the estimated regression coefficients after model fitting and do so in such a way that does not assume constant variance of the residuals. While the mathematics of this approach is beyond the scope of the current text, the basic idea is to plug in estimates of the variance for each residual into the variance calculation for the regression parameters and then average over all observations. The approach works very well for moderately large sample sizes (say, greater than 50) and is implemented in nearly all modern software packages as an option. One thing to note is that different software packages, and authors, may refer to this approach differently. Specifically, it may be referred to as the empirical variance estimator, the sandwich estimator, the robust variance estimator, or the Huber–White variance estimator. When requesting this option from software, it is advised that the help file be consulted to determine what nomenclature is used. The attractiveness of the empirical variance estimator is that the interpretation of the model parameters does not change but we can still obtain valid inference on the model parameters in the face of heteroscedasticity (at least in moderately large sample sizes).

16.5 CURVILINEAR REGRESSION

Up to this point, our regression models have been straight lines, planes, and hyperplanes. Starting with a simple linear regression model, if we add a squared term, as $y = b_0 + b_1x + b_2x^2$, we have a second-degree equation (a parabola). Only a section of the parabola will fit our data, so we must be careful not to extend use of the model outside the realm of our data.

EXAMPLE POSED: A BETTER FIT TO THE SURVIVAL OF MALARIAL RATS

Let us look again at the survival of rats infected with malaria from DB11. In Fig. 16.1, we saw the straight-line regression fit for the RBC treatment. In another treatment in the experiment, infected rats were treated with hetastarch as a placebo, still sample size 100. Number (same as %) surviving by day for 10 days is plotted in Fig. 16.8. We want to fit a regression curve expressing the general pattern of survival over time.

CHOOSING THE MODEL FOR RAT SURVIVAL

Survival diminishes rapidly, then slows, and appears to be approaching a minimum survival percent toward the end of the period, looking more like a segment of an upward opening parabola (second-degree expression) might be a better fit than a line. The data suggest a second-degree model, opening upward. We select the model $y = \beta_{.0} + \beta_{.1}x + \beta_{.2}x^2$, where $y$ is percent survival and $x$ is day number. How is such a model fitted? How do we compare the advantage of a second-degree fit over a first-degree fit?
METHOD

Concept
The concept of curved (more exactly, curvilinear) regression is the same as simple regression throughout, except that the form of the model is not restricted to a straight line. We now can generally refer to the regression curve, which includes the straight line as a subordinate case. Common model forms were illustrated in Fig. 16.2. The most frequently used curve is the parabola, which is like a simple regression with a $x^2$ term added. However, any mathematical function may be appropriate.

Relationship of Curvilinear to Multiple Regression
The mathematics for higher degree regression models is similar to that for multiple regression. For a quadratic model, we start with a univariate straight-line regression ($y$ predicted by $x_1$). Instead of adding a second variable ($y$ predicted by $x_1$ and $x_2$) as in multiple regression, we use the square of the first variable in the position of the second ($y$ predicted by $x_1$ and $x_1^2$).

Choosing the Model
The method of choosing the model varies with goals of the study. If the study is being used to assess a theoretical relationship (physiology, etc.), the form of the model will arise from the theory and the regression significance will be used to validate the theory. If the study is being used to develop the predictor of an established form, that form dictates the model and the regression is used to identify the parameters (constants) used in the prediction. If the study is used to explore relationships, the form of the model will be (Continued)
suggested by shape and pattern in the data plot. The model needs to be appropriate only within the range of existing data (see “An admonition” at the end of the completed example.)

Data Input
The model that is chosen dictates the inputs. Basically, we have a \( y \) depending on some function of \( x \). We substitute \( x \)- and \( y \)-values in the forms given by the model. Where \( y \) appears, \( y \)-values are put in; where \( x^2 \) appears, \( x \)-values are squared and those squares put in; where \( \ln(x) \) appears, logarithms of \( x \)-values are found and put in; and so forth.

Conducting the Analysis
Because curvilinear regression is a particular form of multiple regression, the solution, interpretation, stepwise approaches, and treatment of nominal variables will be the same as in Section 16.2, with some subscripts changed to superscripts, for example, \( x_2 \) replaced by \( x^2 \).

EXAMPLE COMPLETED: A BETTER FIT TO THE SURVIVAL OF MALARIAL RATS
DATA INPUT
We select regression with its appropriate model in whatever statistical software package we choose. We enter percent survival into the \( y \) position and day number into \( x \). We square day number and enter these values into \( x^2 \). For most packages, identification numbers for the rats compose the first column in the spreadsheet, \( y \)-values the second, \( x \)-values the third, and \( x^2 \)-values the fourth.

RESULTS OUTPUT
Let us look at both the straight-line and curved fits. The straight-line equation turns out to be \( y = 96.59 - 7.54x \), superposed on the rat placebo data in Fig. 16.9. For a curved line fit, different statistical software packages use different results display formats, but most are labeled well enough to select the values we need. We can perform a check on the correctness of our selection by anticipating a gross approximation to the values we desire. In this example, we find the sample estimates of the \( \beta \)'s to be \( b_0 = 105.77 \), \( b_1 = -13.66 \), and \( b_2 = 0.61 \). Substitution in the model yields the equation representing the prediction of percent survival by day number as \( y = 105.77 - 13.66x + 0.61x^2 \). The curve given by this equation is shown superposed on the data in Fig. 16.9. By looking further at the results, we find the \( p \)-value for the \( F \) test of both models to be \( < 0.001 \). \( R^2 \) for the straight-line fit is 0.91 and for the curved fit is 0.97.

INTERPRETATION
The \( R^2 \) of 0.91 of the first-degree model shows it to be a rather good fit. What do we gain by using a second-degree model? The significant \( p \)-value of the second-degree
term (0.009) indicates that a real (curved) predictive relationship of percent survival by day number does exist. The very large $R^2$ of 0.97 indicates that survival percent can be even better predicted by this model, with only 3% of the survival variability left for randomness and other causal factors.

**AN ADMONITION**

This prediction is valid up to the modeled 10 days only; if we extend the model to periods outside the experiment, the model will claim that the rats come back to life and the percent survival begins to increase. This nonsensical result illustrates the necessity of inferring conclusions only within the limits of data being modeled.

**ADDITIONAL EXAMPLE: INFANT LUNG CONGESTION CONTINUED**

Let us consider again the dependence of temperature on age in infants with pulmonary complications, introduced in the Additional Example 1 of Section 15.3. At that time, our simple (straight-line) regression fit was $y = 100.18 + 0.12x$. Further, in the additional example of Section 15.4, we found that the $p$-value for a $t$ test on the regression slope was $< 0.001$ and that $R^2 = 0.1389$.

**CHOOSING THE MODEL**

Looking at Fig. 15.8, we can see that the line seems to be centered on the data in the right part of the plot but lies a bit below center in the 6- to 15-month region and
above center on the left. That causes us to conjecture that temperature, which we showed by simple regression to increase with age, increases rapidly with age in newborns and slows its increase as the infant grows older. This suggests a logarithmic fit. We would like to see if we obtain a better fit by using the model \( y = \beta_0 + \beta_1 \ln(x) \).

**DATA INPUT**

We enter our model into the software package. We calculate the (natural) log of \( x \) for each \( x \) and enter them and the \( y \)-values. We instruct the computer to calculate the regression of temperature on log age. The calculations are carried out just as in simple (straight-line) regression but use log age rather than age.

**RESULTS OUTPUT**

The \( p \)-value for the F test on the model and coefficient of determination \( R^2 \) for the two models are shown in Table 16.8 and the log fit on the data is shown in Fig. 16.10.

**Table 16.8** \( p \)-Value and \( R^2 \) for straight line and logarithmic fits to infant pulmonary data.

<table>
<thead>
<tr>
<th>Model</th>
<th>( p )-Value for F test</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straight line</td>
<td>&lt; .001</td>
<td>0.1389</td>
</tr>
<tr>
<td>Logarithmic</td>
<td>&lt; .001</td>
<td>0.1651</td>
</tr>
</tbody>
</table>

**Figure 16.10** Temperature depending on age for 234 infants with pulmonary complications, with a logarithmic regression fit shown.
We can see from the relatively low $R^2$ in both cases that, whereas age is a significant predictor, it does not account for a larger portion of the total variability in the response. However, $R^2$ has increased from about 14% to about 17%; the log fit describes the physiologic process slightly better.

**Exercise 16.2**

Let us continue the question of circulatory response to intubation during laryngoscopy (see Exercises 15.3, 15.7, 15.10, and 15.11). We found that a simple regression of SBP on $SV_{mR}$ gave the fit $y = 8.94 + 94.38x$, that the $p$-value of a significance test on the model was $< 0.001$, and that $R^2 = 0.5322$.

Choosing the model: Fig. 16.11 shows the data. Although the simple regression fit seems to be useful, a case can certainly be made for curvature, opening downward. Write the equation for a parabolic fit, using SBP and $SV_{mR}$ rather than $y$ and $x$.

Data input: What data would be entered into the computer?

Results and interpretation: The outputs are read as $b_0 = -2.92$, $b_1 = 289.00$, $b_2 = -414.41$, $p$-value $< 0.001$, and $R^2 = 0.6822$. Write the final predictive equation. Is the relationship between SBP and $SV_{mR}$ significant in both models? Is the modeled $SV_{mR}$ a major predictor in both models? Which model provides a better fit? Calculate an approximate $y$-value for the left, center, and right regions of $SV_{mR}$ and roughly sketch the curve.

**Figure 16.11** Change in SBP as depending on $SV_{mR}$ amplitude in intubated patients. SBP, Systolic blood pressure; $SV_{mR}$, skin vasomotor reflex.
## 16.6 TWO-FACTOR ANALYSIS OF VARIANCE

**ORIENTATION BY EXAMPLE: HEART RATE EXAMINED BY TWO FACTORS**

A 73-year-old male patient with elevated blood pressure complains of occasional faintness around mealtimes. Data were taken to evaluate this complaint. Blood pressure and HR were measured three times just before and 30 minutes after each meal (approximately 7:00 a.m., noon, and 6:00 p.m.) in one day. Blood pressure provided no answers, but HR was revealing. HR data are given in Table 16.9. We ask the questions: (1) Does eating affect the mean HR? (2) Does time of day affect the mean HR? (3) Does the effect of eating interact with time of day to affect the mean HR (i.e., does HR change at one time of day but not another)?

**EFFECTS OF EATING ON MEAN HEART RATE USING A T TEST**

If we had asked only question (1), we could have taken the nine observations before eating ($m_1 = 56.0000$, $s_1 = 7.4498$) and the nine after ($m_2 = 59.8889$, $s_2 = 6.5659$) and tested the difference using a two-sample $t$ test. Calculating $s_d$ and $t$ from formulas in Chapter 11, Tests of location with continuous outcomes, we find $s_d = 3.3101$ and $t = -1.1749$. From Table II, for $n_1 + n_2 - 2 = 16\ df$, a $|t|$ of 1.17 shows a two-tailed $\alpha$ exceeding 0.20. From statistical software, $p$-value = 0.257. There is insufficient evidence to conclude that there is a relationship between HR and eating.

**EFFECT OF EATING ON MEAN HEART RATE USING ONE-WAY ANALYSIS OF VARIANCE**

The $t$ test is familiar to most users. In Chapter 11, Tests of location with continuous outcomes, we noted the relationship $F_{1,k\ df} = (t_{k\ df})^2$. To make comparisons of one- and two-way ANOVAs, we want question 1 answered by a one-way ANOVA. Let us use the formulas of Table 11.9. $k = 2$, $n = 18$, and $n_1 = n_2 = 9$. Sum of squares for means ($SSM) = 68.0559$. The sample variance for all $n = 18$ observations is $s^2 = 50.4085$ (which is also MST). Sums of squares for total ($SST) = (n - 1)\ s^2 = 856.9445$. Sum of squares for error ($SSE) = SST - SSM = 788.8886$. Mean square

<table>
<thead>
<tr>
<th>Pre- (1) or postmeal (2)</th>
<th>HR at meal 1</th>
<th>HR at meal 2</th>
<th>HR at meal 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>57</td>
<td>60</td>
</tr>
<tr>
<td>1</td>
<td>48</td>
<td>55</td>
<td>73</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>55</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>71</td>
<td>56</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>64</td>
<td>58</td>
<td>50</td>
</tr>
</tbody>
</table>
of means (MSM) = SSM, because \( k - 1 = 1 \). Mean square of error (MSE) = SSE/\( (n - k) \) = 49.3055. Finally, \( F_{1,16} \) = MSM/MSE = 1.3803. From Table V the critical value for 0.05 \( \alpha \) for 1,16 \( df \) is 4.49. Because 1.38 < 4.49, we insufficient evidence to conclude that there is a relationship between HR and eating. From statistical software, \( p \)-value = 0.257, the same as using \( t \). We note that \( (-1.1749)^2 = 1.3804 \), the same to three decimal places.

\[ F = \frac{MSM}{MSE} = 1.3803 \]

\[ p\text{-value} = 0.257 \]

\[ (-1.1749)^2 = 1.3804 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]

\[ 3.68 \]

\[ 0.1711 \]

\[ 0.844 \]

\[ 0.50 \]

\[ 50.4085 \]
METHOD FOR TWO-FACTOR ANALYSIS OF VARIANCE (ANOVA)

The Basic Datum and Some Definitions

Let us designate the basic observation or reading as \( x \). We have two factors and several readings for each pairing of factor indicators, so that we need three subscripts to identify each reading, say \( x_{ijk} \). In the example, we have the factors eating \((i)\) and time of day \((j)\), with three replicated readings \((k)\) for eating-state–time-state pair. Let us denote the numbers of categories as \( r \) for the first and \( c \) for the second, and the number of repeatedly taken readings for each \( i,j \)-pair, that is, repetitions, as \( w \). We calculate the \( ij \)th mean using these \( w \) readings. These definitions can be summarized as follows:

\[
\begin{align*}
&i = 1, 2, \ldots, r \\
&j = 1, 2, \ldots, c \\
&k = 1, 2, \ldots, w. \\
&m_{ij} = \frac{\sum_{k=1}^{w} x_{ijk}}{w}
\end{align*}
\]

A Means Table

The first step is to set up the means we want to compare. We create a table with one of our two factors represented by rows and the other by columns. We can denote the row factor by \( R \) and the column factor by \( C \). (Hence, \( r \) categories in \( R \) and \( c \) in \( C \).) As in Table 10.1, where we used a dot subscript to indicate the subscript summed over, let us use \( m_{.} \) to indicate the mean over \( j \), that is, \( m_{.} = (m_{1.} + \cdots + m_{c.})/c \), etc. The table will appear as Table 16.11.

Table 16.11 Components of a two-factor means table.

<table>
<thead>
<tr>
<th>( R ) (rows)</th>
<th>1</th>
<th>\ldots</th>
<th>( c )</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( m_{1.1} )</td>
<td>\ldots</td>
<td>( m_{1.} )</td>
<td></td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td></td>
<td>\vdots</td>
<td></td>
</tr>
<tr>
<td>( r )</td>
<td>( m_{r.1} )</td>
<td>\ldots</td>
<td>( m_{r.} )</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>( m_{.1} )</td>
<td>\ldots</td>
<td>( m_{.c} )</td>
<td>( m_{.} )</td>
</tr>
</tbody>
</table>

An “Adjustment” Term

In Table 11.9 the formulas for sums of squares (SS) for mean (SSM) and SST included the overall mean subtracted from each element in the sum. By using algebra, we could have carried this subtraction outside the parentheses and used a single “adjustment” or “correction” term for each sum of squares. The concept is easier to understand as the table is given, but the calculations are easier with a single adjustment term. Because two-factor and higher factor ANOVAs are complicated to calculate, the adjustment term, \( A \), will be used. It is just the total sample size times the square of the overall mean.

(Continued)
Formulas for Calculation

The basic sums of squares (SS) that appear in the ANOVA table appear in Table 16.12.

### Table 16.12 Formulas for components in a two-factor analysis of variance.

<table>
<thead>
<tr>
<th>Component</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total SST</td>
<td>( A = rwc \times m^2 )</td>
</tr>
<tr>
<td>Sum of squares for total (SST)</td>
<td>( \sum r \sum c \sum w \chi^2_{ijk} - A )</td>
</tr>
<tr>
<td>Sum of squares for rows (SSR)</td>
<td>( cw \sum \sum r \chi^2_i - A )</td>
</tr>
<tr>
<td>Sum of squares for columns (SSC)</td>
<td>( rw \sum \sum c \chi^2_j - A )</td>
</tr>
<tr>
<td>Sum of squares for interaction row by column (SSI)</td>
<td>( w \sum \sum r \sum c \chi^2_{ij} - A - SSR - SSC )</td>
</tr>
<tr>
<td>Sum of squares for error (or residual) (SSE)</td>
<td>( SST - SSR - SSC - SSI )</td>
</tr>
</tbody>
</table>

*a* The symbolism \( \sum_{i} \) implies “sum over \( i \) from 1 to \( r \),” and equivalently for other indicator symbols.

Analysis of Variance Table

The remaining calculations and results of the experiment are displayed in an ANOVA table, shown in Table 16.13. Such a table provides the basis for interpreting the outcome of the experiment and is often presented in its entirety when the experiment is published.

### Table 16.13 Two-factor analysis of variance table.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>( F )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>SSR from Table 16.12</td>
<td>( r - 1 )</td>
<td>MSR = SSR/df</td>
<td>MSR/MSE</td>
<td>MSC/MSE</td>
</tr>
<tr>
<td>Columns</td>
<td>SSC from Table 16.12</td>
<td>( c - 1 )</td>
<td>MSC = SSC/df</td>
<td>MSI = SSI/df</td>
<td>MSI/MSE</td>
</tr>
<tr>
<td>Interaction ((R \times C))</td>
<td>SSI from Table 16.12</td>
<td>((r - 1)(c - 1))</td>
<td>MSI = SSI/df</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (residual)</td>
<td>SSE from Table 16.12</td>
<td>( rc(w - 1) )</td>
<td>MSE = SSE/df</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SST from Table 16.12</td>
<td>( rwc )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Interpretation of the analysis of variance table

The significance level \( \alpha \) was chosen at the outset of the experiment. Most often \( \alpha = 0.05 \) by convention. The critical values are found from the \( F \) table, Table V. Numerator \( df \) is the \( df \) for the factor, that is, the source of variability, being tested. Denominator \( df \) is the \( df \) for error. For example, the critical value for the rows factor is the Table V entry for \( r - 1 \) along the top and \( rc(w - 1) \) along the side. The actual \( p \)-value can be found using statistical software. As with other tests of significance, a significant \( p \)-value implies that it would have been a rare event to have observed the results we did if there truly were no difference in the population means. When a significant factor contains three or more means, we question which difference(s) between members of mean pairings account for the significance. This question can be answered using methods of multiple comparisons as addressed in Section 11.4.
**Table 16.14** Two-factor heart rate means.

<table>
<thead>
<tr>
<th></th>
<th>Breakfast</th>
<th>Lunch</th>
<th>Dinner</th>
<th>Across mealtimes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before eating</td>
<td>49.3333</td>
<td>55.6667</td>
<td>63.0000</td>
<td>56.0000</td>
</tr>
<tr>
<td>After eating</td>
<td>67.6667</td>
<td>57.3333</td>
<td>54.6667</td>
<td>59.8889</td>
</tr>
<tr>
<td>Across eating</td>
<td>58.5000</td>
<td>56.5000</td>
<td>58.8333</td>
<td>57.9444</td>
</tr>
</tbody>
</table>

**Table 16.15** Two-factor heart rate sums of squares.

\[
A = 18(57.9444)^2 = 60,435.9630
\]

\[
SST = 50^2 + 48^2 + \cdots + 58^2 + 50^2 - 60,435.963 = 856.9444
\]

\[
SSR = 9(62^2 + 59.8889^2) - 60,435.963 = 68.1601
\]

\[
SSC = 6(58.5^2 + 56.5^2 + 58.8333^2) - 60,435.963 = 19.1801
\]

\[
SSI = 3(49.3333^2 + \cdots + 54.6667^2) - 60,435.963 - 68.1601 - 19.1801 = 544.3777
\]

\[
SSE = 856.9444 - 68.1601 - 19.1801 - 544.3777 = 225.2265
\]

**Table 16.16** Two-factor analysis of variance table for heart rates.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>F</th>
<th>Critical F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before versus</td>
<td>68.1601</td>
<td>1</td>
<td>68.1601</td>
<td>3.63</td>
<td>6.55</td>
<td>.081</td>
</tr>
<tr>
<td>after eating</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mealtimes</td>
<td>19.1801</td>
<td>2</td>
<td>9.5901</td>
<td>0.51</td>
<td>5.10</td>
<td>.613</td>
</tr>
<tr>
<td>Interaction</td>
<td>544.3777</td>
<td>2</td>
<td>272.1888</td>
<td>14.50</td>
<td>5.10</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>225.2265</td>
<td>12</td>
<td>18.7689</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>856.9444</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TWO-FACTOR HEART RATE EXAMPLE COMPLETED**

The first step is to generate the means table. The data are as in Table 16.14.

We note that \( r = 2 \), \( c = 3 \), and \( w = 3 \). Using Table 16.12, we find the SS as in Table 16.15.

Completing the ANOVA table, as in Table 16.13, we find the results as in Table 16.16.

**INTERPRETATION**

A column has been added showing critical values of \( F \) from Table V. The first entry is \( F_{0.95} \) for 1,12 \( df \). The second and third entries are \( F_{0.95} \) for 2,12 \( df \). We can see that the calculated \( F \) is less than critical for the two main or primary factors, as we found in the one-way ANOVAs, but is greater than critical for the interaction term. The significant interaction indicates that the eating factor has a pattern of mean differences that
changes according to mealtime. We can see from the means table that mean HR increases remarkably over breakfast but decreases over dinner, while it is little different at lunchtime. The increase at breakfast time and decrease at dinnertime obscured each other when pooled over time, so that no significance appeared.

**NOTE THE IMPROVEMENT OVER THE T TEST AND THE ONE-WAY ANALYSIS OF VARIANCE**

Not only are we able to answer all three questions with one analysis, keeping our \( p \)-value from accumulating by repeated testing, but also the tests of questions (1) and (2) are more sensitive and control for the other factors in the model. Using two-way ANOVA, we removed identifiable causes of variability from SSE present in the one-way ANOVAs, with resulting increases in \( F \) and decreases in the \( p \)-value. Going from one- to two-way ANOVAs reduced the \( p \)-value from 0.257 for the eating factor to 0.081 and from 0.884 for the mealtime factor to 0.613.

**ADDITIONAL EXAMPLE: COOLING KIDNEYS PRIOR TO SURGERY**

DB15 provides data on two methods of cooling kidneys to prevent necrosis during surgery. Six anesthetized pigs were opened and their kidneys cooled, one by infusing cold saline (treatment 1) and the other by packing in ice (treatment 2). Temperature (°C) was measured at baseline (time 1) and at 5 (time 2), 10 (time 3), and 15 (time 4) minutes. Because one kidney was cooled from the outside and the other from the inside, the depth measurement would be expected to be relevant. However, for the purpose of illustrating a two-factor ANOVA, temperatures at the two depths will be taken as just further replications, so that 2 depths on 6 pigs allows 12 replications (\( w \)) from which to calculate means for each treatment and time combination. This database will be analyzed again using depth as a factor in Section 16.8 by three-factor ANOVA.

The first step is to generate the means table, as in Table 16.17.

We note that \( r = 2, \ c = 4, \) and \( w = 12 \). Using Table 16.12, we find SS as in Table 16.18.

Completing the ANOVA table, as in Table 16.13, we find the results as in Table 16.19.

All sources of mean differences are highly significant. (The \( F \) values are huge.) We see that the means of kidney cooling are different for the two treatments, that the kidneys cool significantly over time regardless of the treatment, and that the pattern of kidney cooling over time is different for the two treatments. Looking at the means

<table>
<thead>
<tr>
<th></th>
<th>0 min</th>
<th>5 min</th>
<th>10 min</th>
<th>15 min</th>
<th>Across time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold saline</td>
<td>34.4000</td>
<td>31.9417</td>
<td>28.7000</td>
<td>27.9583</td>
<td>30.7500</td>
</tr>
<tr>
<td>Ice</td>
<td>36.1167</td>
<td>16.9500</td>
<td>10.3833</td>
<td>7.5000</td>
<td>17.7375</td>
</tr>
<tr>
<td>Over treatments</td>
<td>35.2583</td>
<td>24.4458</td>
<td>19.5417</td>
<td>17.7292</td>
<td>24.2438</td>
</tr>
</tbody>
</table>
CHAPTER 16 Multiple linear and curvilinear regression

Table 16.18 Sums of squares for kidney cooling.

\[ A = 95(24.2438)^2 = 56,424.6710 \]

\[ \text{SST} = 36.7^2 + 33.2^2 + \cdots + 16.6^2 + 9.6^2 - 56,424.6710 = 11,531.3562 \]

\[ \text{SSR} = 48(30.75^2 + 17.7375^2) - 56,424.6710 = 4064.0365 \]

\[ \text{SSC} = 24(35.2583^2 + 24.4458^2 + 19.5417^2 + 17.7292^2) - 56,424.6710 = 4462.0673 \]

\[ \text{SSI} = 12(34.4^2 + \cdots + 7.5^2) - 56,424.6710 - 4064.0365 - 4462.0673 = 1826.4673 \]

\[ \text{SSE} = 11,531.3562 - 4064.0365 - 4462.0673 - 1826.4673 = 1178.7851 \]

Table 16.19 Analysis of variance table for kidney cooling.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>F</th>
<th>Critical F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments (rows)</td>
<td>4064.0365</td>
<td>1</td>
<td>4064.0365</td>
<td>303.39</td>
<td>3.96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time (columns)</td>
<td>4462.0673</td>
<td>3</td>
<td>1487.3558</td>
<td>111.04</td>
<td>2.72</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Interaction</td>
<td>1826.4673</td>
<td>3</td>
<td>608.8224</td>
<td>45.45</td>
<td>2.72</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>1178.7851</td>
<td>88</td>
<td>13.3953</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11,531.3562</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table, we see that the kidneys infused with saline lower in temperature, but not to the required 15°C, while those in ice do.

Exercise 16.3

Effectiveness of a drug for patients with lower right quadrant pain. Historically, results on effectiveness of a certain drug were mixed. An ER specialist designed a two-way ANOVA to assess effectiveness for kidney stones versus other causes. Randomizing patients into drug versus placebo groups, he collected visual analog pain scale (mm) data on 14 kidney-stone and 14 other-cause patients. Data, giving the change in pain scale readings at presentation minus 20 minutes after drug administration, were as in Table 16.20.

Conduct the two-way ANOVA with interaction and interpret the results. SS were 24.14286 for drug, 531.57143 for stones, 1032.14286 for drug X stones interaction, and 4218.42857 for total.

16.7 ANALYSIS OF COVARIANCE

Purpose

It was said earlier that one purpose of experimental design is to control statistically factors that cannot be controlled physically. In the preceding sections, we saw how to control factors that fall into specific categories, perhaps modes of treatment. But what do we do if one factor is continuous rather than categorical, perhaps patient age? Such
a factor has been called a covariate and the method of analyzing the variability of mixed categorical and continuous factors is called ANCOVA.

**ORIENTATION BY EXAMPLE: DO SEX AND/OR AGE AFFECT THEOPHYLLINE LEVELS OF EMPHYSEMA PATIENTS?**

**DATA**

In DB3, serum theophylline was measured on emphysema patients prior to and during a course of azithromycin. The clinical expectation was that the antibiotic would raise theophylline levels. Let us denote the change in level as the outcome variable $x = \text{level at day 5} - \text{level at day 0}$, sex as $g$ (for group), and age as the continuous independent variable $u$. Did the sex and/or age of the patient affect $x$? The resulting data set is as in Table 16.21.

Let us denote subscript indicators and numbers of observations as follows. Group (sex) denotes $g_i$, $i = 1, 2$, 1 is female; $j$ denotes number within group, $j = 1, \ldots, n_i$, where $n_1 = 6$, $n_2 = 9$, and $n = n_1 + n_2 = 15$; change is $x_{ij}$; and age is $u_{ij}$.

**MEANS**

The mean, $m_{x.}$, of change in level $x_{ij}$ overall is $-1.11$, $-3.33 = m_{x1}$ for females and $0.37 = m_{x2}$ for males. The mean, $m_{u.}$, of age $u_{ij}$ overall is $65.20$, $63.33 = m_{u1}$ for females and $66.44 = m_{u2}$ for males. Does the fact that males are older affect change in

<table>
<thead>
<tr>
<th>Patient number</th>
<th>Drug ($0 = n$, 1 = y)</th>
<th>Stones ($0 = n$, 1 = y)</th>
<th>Pain change (mm)</th>
<th>Patient number</th>
<th>Drug ($0 = n$, 1 = y)</th>
<th>Stones ($0 = n$, 1 = y)</th>
<th>Pain change (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>-9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>-7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>16</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>20</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>22</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>28</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>35</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>-5</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>-3</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>-3</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>-2</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>5</td>
<td>27</td>
<td>1</td>
<td>1</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>6</td>
<td>28</td>
<td>1</td>
<td>1</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

aData were simplified to facilitate the example.
theophylline level? If age were categorized, perhaps as $<70$ versus $\geq 70$, we would have a $2 \times 2$ two-factor ANOVA. However, because age is continuous and not categorical, we cannot include it in an ANOVA.

**WHAT TO DO ABOUT THE CONTINUOUS VARIABLE**

If age is an influence and we do not adjust for it, it may bias the outcome. In addition to reducing potential bias, removing the variability of another influencing factor from the error mean square will increase sensitivity and precision of the remaining analysis. With age continuous, we must use ANCOVA, an ANOVA with outcome adjusted for the continuous variable. Change $x$ is potentially dependent on age $u$. We can investigate this dependency by the regression of $x$ on $u$, as in Chapter 15, Linear regression and correlation. While any sort of regression model is possible to use, the explanation here will be restricted to a straight-line regression. The slope of the regression line, $b$, is assumed to be the same for both sexes. (We could test this assumption if we are unsure.) However, the different means for males and females imply a regression line for each: $m_{x1} + b(u_{1j} - u_{1})$ for females and $m_{x2} + b(u_{2j} - u_{2})$ for males.

Each $x_{ij}$ is adjusted by its respective regression for the deviation of the group means from the overall mean to become $x_{ij}^{(a)}$, say. We find that the adjusted female mean is $m_{x1}^{(a)} = -3.12$ and the adjusted male mean is $m_{x2}^{(a)} = 0.23$.
ANALYSIS OF COVARIANCE (ANCOVA) TABLE

A sum of squares for the categorical variable sex adjusted for age and a sum of squares for the covariate age as a factor itself are calculated. (These are more involved than the clinical reader will care to follow.) Let us use \( s^2 \) for variance with a letter subscript to denote its variable and with a 1 or 2 attached to the subscript to indicate sex; let us use \( s \) with two subscript letters to denote the covariance of the two corresponding variables. (The symbol \( s \) with a single subscript indicates the standard deviation for the variable corresponding to that subscript.) The error (or residual) sum of squares appears as

\[
\text{SSE} = s_{x_1}^2 + s_{x_2}^2 - \frac{(s_{y_1} + s_{y_2})^2}{s_{y_1}^2 + s_{y_2}^2} = 127.16.
\]

The SSE has \( n_1 + n_2 - 3 \) df. The ANCOVA table, looking much like an analysis of variance (ANOVA) table, thus generated is as in Table 16.22.

Table 16.22  Analysis of covariance for sex and age as factors in theophylline levels.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>( F )</th>
<th>Critical ( F )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>42.5901</td>
<td>1</td>
<td>42.5901</td>
<td>4.02</td>
<td>4.75</td>
<td>.068</td>
</tr>
<tr>
<td>Age</td>
<td>2.2092</td>
<td>1</td>
<td>2.2092</td>
<td>0.21</td>
<td>4.75</td>
<td>.656</td>
</tr>
<tr>
<td>Error</td>
<td>127.6541</td>
<td>12</td>
<td>10.5970</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>178.6573</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If we had tested sex unadjusted for age with a \( t \) test or a one-way ANOVA, we would have found \( p \)-value = 0.044, below a significance cut point of 0.05. However, when adjusted for age, sex falls above 0.05, losing significance. Thus age is not a significant influence on theophylline level by itself (\( p \)-value = 0.68), nor is level significantly different between sexes (\( p \)-value = 0.063), but age interacted with sex enough to alter theophylline level significance by sex.

ADDITIONAL EXAMPLE: COMFORT OF BANDAGE LENSES FOLLOWING EYE SURGERY

Part of the procedure for photorefractive keratometry (PRK) correction of vision is to remove the corneal membrane. To protect the eye and reduce pain, a soft contact lens is worn as a bandage. A study compared comfort level between two types of lens for 100 post-PRK eyes for each type. A \( t \) test of mean comfort level between the lens types two days after surgery yielded \( p \)-value = 0.011, indicating a significant difference. However, the defect size, a continuous variable, could not be controlled, because the defect heals at different rates for different patients. Let us denote the lens types by 1 and 2. The comfort level means are 1.93 overall, 2.06 for lens 1, and 1.79 for lens 2. The day-2 defect sizes are 2.87 overall, 3.62 for lens 1, and 2.12 for lens 2. Does the defect size interact with lens type to alter the significance of comfort level difference by lens? An ANCOVA table appears as Table 16.23.
When adjusting each lens’ comfort level for defect size, the difference between lenses is slightly magnified. This is supportive evidence that the difference in defect size is not a major confounder in the relationship between comfort level and lens type.

16.8 THREE-WAY AND HIGHER WAY ANALYSIS OF VARIANCE

ORIENTATION BY EXAMPLE: COOLING KIDNEYS EXTENDED TO ALL FACTORS

In the additional example of Section 16.6, the kidney-cooling comparison was addressed using a two-way ANOVA by treating readings on the factor of depth measurement within the kidney as replications. Because one treatment cooled from the outside and the other from the inside, depth is very relevant. Now we will reexamine the data using the depth measure in three-way ANOVA. Three-way ANOVA is conceptually a simple extension of two-way. We have three main effect tests rather than two, three two-factor interactions rather than one, and the addition of a three-factor interaction. The main effect and two-factor interaction results are interpreted one by one in the same way as they are for two-way ANOVA. What is difficult about three-way ANOVA is means tabulation and the interpretation of the three-factor interaction. Let us provide the example and discuss these issues in context.

METHOD BY EXAMPLE

The SS, df, and mean squares of treatment, time, and their interaction are identical with those of a two-way analysis. We create a means table for treatment by depth, from which we calculate the sums of squares (SS) for the depth main effect and the treatment-by-depth interaction following the patterns in Table 16.12. The number of df for depths is the number of depths less one, or \(2 - 1 = 1\), the number of df for the treatment-by-depth interaction is \(df \times df \text{ for depth, or } 1 \times 1 = 1\). We create a means table for time by depth, from which we calculate the SS for the time-by-depth interaction following the two-factor interaction pattern in Table 16.12. \(df = 1 \times 3 = 3\). We create a more extended means table that gives two factors along one axis and the third factor along the other axis, for (Continued)
(CONTINUED)

example, each depth with each treatment pairing along the top and time down the side, from which we calculate the SS for the treatment-by-depth-by-time interaction as the product of number of elements in each main factor \((2 \times 2 \times 4)\) times the raw sum of squares of treatment \(\times\) time \(\times\) depth means \(- A - SS\) for all main effects and second-order interactions. \(df\) for the third-order interaction is the product of \(df\) for main effects, that is, \(1 \times 1 \times 3 = 3\).

The reader is not carried through the actual computations, because no one is likely to perform three-factor ANOVA manually. The issue is to understand how the calculations found in a software-generated table came about. Table 16.24 is such a software-generated table.

Table 16.24 Three-way analysis of variance table for kidney cooling experiment.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>(F)</th>
<th>Critical (F)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4063.8037</td>
<td>1</td>
<td>4063.8037</td>
<td>455.30</td>
<td>3.96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time</td>
<td>4461.8704</td>
<td>3</td>
<td>1487.2901</td>
<td>166.63</td>
<td>2.72</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Depth</td>
<td>145.0417</td>
<td>1</td>
<td>145.0417</td>
<td>16.25</td>
<td>3.96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Treatment (\times) time</td>
<td>1826.6405</td>
<td>3</td>
<td>608.8224</td>
<td>68.22</td>
<td>2.72</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Treatment (\times) depth</td>
<td>201.8400</td>
<td>1</td>
<td>201.8400</td>
<td>22.61</td>
<td>3.96</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time (\times) depth</td>
<td>43.0092</td>
<td>3</td>
<td>14.3364</td>
<td>1.61</td>
<td>2.72</td>
<td>.194</td>
</tr>
<tr>
<td>Treatment (\times) time (\times) depth</td>
<td>75.1142</td>
<td>3</td>
<td>25.0381</td>
<td>2.81</td>
<td>2.72</td>
<td>.045</td>
</tr>
<tr>
<td>Error (residual)</td>
<td>714.0366</td>
<td>88</td>
<td>8.9255</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11,531.3562</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**INTERPRETATION**

We note that MSE has reduced to two-thirds of its two-factor size due to removal of the variability due to the third factor. This, of course, increases all the \(F\) values, making the tests on the original factors more sensitive. Time, treatment, and time-by-treatment interaction are all highly significant as before with the same interpretations. Depth is highly significant, as we anticipated. The treatment–depth interaction is also highly significant. By examining a means table, we could see that the depth effect is less in the cooled saline than in the ice treatment. The time–depth interaction is not significant, indicating that the pattern of reduction in mean temperature over time is little different for the two depths. Finally, the third-order interaction is just barely significant. The interpretation of third-order and higher order interactions is usually not obvious and takes thought and understanding of the processes. In this case, it might best be explained by saying that the treatment–depth interaction changes over time.

**16.9 CONCEPTS OF EXPERIMENTAL DESIGN**

In this chapter, models have been getting increasingly involved, so we should take a look at some ideas of experimental design. The need for more complex statistical
models is usually driven by the need to control for other variables or to attend to deficiencies in the data that we sample. Many of these issues, however, could have been avoided by considering the experimental design used to collect our data in order to answer a scientific question. Of course, retrospectively conducted studies, where data have already been collected, are out of our hands in this regard, but sometimes we can find or develop a suitable design. If, however, we are to prospectively collect data, many of the issues that lead to complicated statistical modeling can be avoided by implementing fairly simple design principles.

**Purpose of design**

In too many medical studies, especially retrospective ones, we analyze as best we can the data that fate supplies. There are a great many uncontrolled factors influencing the outcome we are measuring. The history of physics as a science has been one of the controlling factors so that only the one under study influences the outcome. By contrast, the discipline of economics, which is not an experimental science, is the victim of myriad influences that cannot be controlled. Biology, including medicine, has lain somewhere in between: some factors can be controlled, others not. The subject matter of experimental design is that of setting up a data-producing study that controls physically those factors that can be controlled physically, controls statistically those factors that cannot be, and allows the analysis of both. In the following sections, we describe some of the most common types of historical designs. While not meant to be an exhaustive treatment of these designs, the reader is encouraged to further read up on the topics before attempting to implement them in practice.

**Factors influencing outcomes**

Suppose we want to know what influences blood loss in surgery. Using the methods of Chapter 11, Tests of location with continuous outcomes, we could compare the mean loss (dependent variable) according to a type-of-surgery factor (independent variable) containing two or three types of surgery. However, we could all name a number of other influencing factors that we would like to measure and analyze simultaneously, such as length and invasiveness of surgery, leading to the use of regression and ANOVA methods met earlier in this chapter.

**Assumptions and balance in multifactor designs**

**Definition**

A balanced experimental design contains the same number of data readings in each cell of the categorical independent variables, that is, in each possible pairing of categories, so that every possible combination is represented an equal number of times. Such
balanced designs provide the most efficient statistical control and assessment of the effect of influences on an experiment’s outcome. Lack of balance implies a loss of efficiency.

**EXAMPLE: TWO TYPES OF INFLUENCE ON BLOOD FLOW AFTER ORTHOPEDIC CASTING**

Suppose an orthopedic experiment consisted of comparing the mean rate of blood flow permitted by two modes of casting, each with two types of pinning, forming a two-factor ANOVA with interaction. (This could be designated a $2 \times 2$ design for the two categories in each factor. If there were three modes of casting, it would be a $3 \times 2$ design.) The first main effect results compare mean blood flow for the cast modes averaged over the pinnings and the second, for the pinning types averaged over the castings. The interaction effect compares the pinning results for each mode of cast to learn if the effect of pinning type is conditional on mode. If the orthopedist obtained 6 readings on each pinning type for one mode of cast, but only 4 on the other, the design would be unbalanced.

**BALANCE IS ASSUMED IN THIS CHAPTER**

In the formulas given in this chapter, balance is assumed. Balanced designs are easier to understand and the formulas are simpler. The interpretation of ANOVA results for unbalanced designs will be similar to that from the result had balance been present. If the user cannot design a balanced experiment, in most cases the statistical software adjusts the calculations for lack of balance and provides accurate $df$.

**Missing observations**

If a very small proportion of observations in a balanced design is lost, perhaps one to two in a small sample or a few in a large sample, the missing observation(s) may be estimated and degrees of freedom reduced accordingly. Specifically, this estimation must follow one of a few technically defined methods, denoted *imputation*. Such a case might occur if an animal dies or a patient is lost to follow up. Statistical software exists for imputation of missing observations. Lacking software, one approach (usually not the best) is to substitute the average of the other observations on that variable in that cell. If many observations are missing, techniques for unbalanced designs exist but are not simple; the help of a biostatistician should be sought.

**More specialized designs and techniques**

Many specialized designs have been developed to meet unusual constraints on data availability, indeed more than could be discussed in any single book. The statistical literature abounds with them. A few major classes of specialized designs are described in this section, with no pretension of being of direct use. If the reader recognizes a match
to data to be analyzed, these are the names to seek in more specialized books or to discuss with a biostatistician.

**FACTORIAL DESIGNS**

Factorial designs are essentially multifactor designs that are balanced, so that all variable combinations are represented and have equal numbers. All the examples in the two- and three-way ANOVAs in this chapter are balanced in this way and are factorial designs. The kidney-cooling experiment, containing two treatments at two depths over four times was a $2 \times 2 \times 4$ three-way factorial ANOVA. There were $2 \times 2 \times 4 = 16$ readings on each pig; $6 \text{ pigs} \times 16 \text{ readings} = 96 \text{ readings used}$.

**SPARSE-DATA DESIGNS**

In some situations, the investigator has considerable control over experimental variables, but data are very difficult or very expensive to obtain. In such a circumstance, we want to maximize the information to be obtained from every datum. Consider an experiment in which we want to investigate cancer patient survival based on combinations of radiation, chemotherapy, and surgery. We have three levels of radiation to examine, three dosages of chemotherapy, and three types of surgery. We might display a design of all combinations of the first two variables as in *Table 16.25*.

It can be seen that each chemotherapy dose ($C$) is paired with each radiation level ($R$), forming nine combinations. If we were to pair each of the surgery options ($S$) with each of these pairings, we would have a $3 \times 3 \times 3$ (or $3^3$) factorial ANOVA, requiring 27 patients per replication (randomized into 27 treatment combinations). If, however, we overlay the two-way pattern of numbers with three letters, referring to three surgery options, balanced by each occurring once and only once in each row and column, we then have a design in which the main effect results would be available with only nine patients. The design might look as in *Table 16.26*.

There are a number of different patterns satisfying the balance requirement and the pattern can be chosen randomly. This design is named a *Latin square*, because of the (Latin) letters paired with the numerical combinations.

There are numerous other sparse-data designs. The *Graeco-Latin square*, in which a pattern of Greek letters, balanced with both the number pairings and the Latin letters,

---

**Table 16.25** Combinations of chemotherapy and radiation.

<table>
<thead>
<tr>
<th>Radiation</th>
<th>Chemotherapy</th>
<th>Dose 1</th>
<th>Dose 2</th>
<th>Dose 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>Level 2</td>
<td>Level 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
overlays a square, yields a design in which four main effects may be tested from the square. An incomplete block design is a design that lacks some of its balance. It usually arises from an inability to apply all treatments in one block. If three types of eye surgery are being compared on rabbits, we are limited to only two eyes per rabbit, so we might use an incomplete block design. A split-plot design is one in which data providing results for the most important effects occur in balance, but that for some lesser effects occur only within one or another primary effect. Some information is lost, but the lesser effects are not abandoned.

REFERENCES


Table 16.26 Latin square design combining chemotherapy, radiation, and surgery type.

<table>
<thead>
<tr>
<th>Radiation Level</th>
<th>Chemotherapy Dose 1</th>
<th>Chemotherapy Dose 2</th>
<th>Chemotherapy Dose 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>Level 2</td>
<td>C</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Level 3</td>
<td>B</td>
<td>C</td>
<td>A</td>
</tr>
</tbody>
</table>
Logistic regression for binary outcomes

17.1 INTRODUCTION

In Chapter 9, Tests on categorical data, the concept of a contingency table was introduced. Two-way contingency table analyses consider testing whether an association exists between two categorical variables. A special case of general contingency tables is the $2 \times 2$ table, where each variable takes on one of two levels. For example, consider a study where we randomly sample $n_1$ female subjects ($FEMALE = 1$) and $n_0$ male subjects ($FEMALE = 0$). Here $FEMALE$ represents a binary variable taking on the value of 1 to indicate that the subject is a female and 0 if male. Suppose we then follow each subject for up to 30 years, recording whether or not the subject was diagnosed with coronary heart disease ($CHD = 0,1$) during that time period. The resulting data could then be organized into a $2 \times 2$ contingency table as given in Table 17.1. Scientifically, we would most likely be interested in comparing the probability of CHD ($CHD = 1$) for each of the two sexes ($FEMALE = 0$ and $FEMALE = 1$). That is, if we let $\pi_0 = \Pr[CHD = 1 | FEMALE = 0]$ (read “the probability of CHD given male sex”) and $\pi_1 = \Pr[CHD = 1 | FEMALE = 1]$, then we would like to determine if $\pi_0 = \pi_1$. Note that $\pi_0 = \pi_1$ is equivalent to $\pi_1 - \pi_0 = 0$ (the risk difference is equal to 0), $\pi_1/\pi_0 = 1$ (the relative risk is equal to 1), and $(\pi_1/(1 - \pi_1))/(\pi_0/(1 - \pi_0)) = 1$ (the odds ratio is equal to 1). In addition, though less obvious, $\pi_0 = \pi_1$ also implies that the rows of the contingency table given in Table 17.1 are independent of the columns, or that $CHD$ and $FEMALE$ are independent variables.

In light of the previous connections, if one were interested in testing the null hypothesis $H_0 : \pi_0 = \pi_1$, this could be accomplished by testing any of the following equivalent hypotheses:

$H_0 : \pi_1 - \pi_0 = 0$

$H_0 : \pi_1/\pi_0 = 1$
CHAPTER 17 Logistic regression for binary outcomes

Table 17.1 2 × 2 Contingency table depicting the observed data relating coronary heart disease (CHD) to sex.

<table>
<thead>
<tr>
<th></th>
<th>CHD = 0</th>
<th>CHD = 1</th>
<th>Marginal total</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE = 0</td>
<td>n₀₀</td>
<td>n₀₁</td>
<td>n₀.</td>
</tr>
<tr>
<td>FEMALE = 1</td>
<td>n₁₀</td>
<td>n₁₁</td>
<td>n₁.</td>
</tr>
<tr>
<td>Marginal total</td>
<td>nᵦ₀</td>
<td>nᵦ₁</td>
<td>nᵦ.</td>
</tr>
</tbody>
</table>

\[ H_0: \frac{\pi_1}{1 - \pi_1} = \frac{\pi_0}{1 - \pi_0} = 1 \]

\[ H_0: \pi_{ij} = \pi_{i.} \times \pi_{.j}, \text{ for } i = 0, 1 \text{ and } j = 0, 1, \]

where \( \pi_{ij} = \Pr[FEMALE = i, \ CHD = j], \ \pi_{i.} = \Pr[FEMALE = i], \) and \( \pi_{.j} = \Pr[CHD = j]. \) Note that the latter hypothesis implies that CHD and FEMALE are independent. As was seen in Chapter 9, Tests on categorical data, either the Pearson chi-squared test for independence or the two-sample test of binomial proportions could be used to test the abovementioned hypotheses. However, these methods do not allow us to adjust for other variables that may confound the relationship between sex and the risk of CHD. In Chapter 16, Multiple linear and curvilinear regression and multifactor ANOVA, multiple linear regression was introduced as a method to relate a continuous dependent variable to an independent variable while adjusting for (or attempting to hold constant) other factors. While this approach works well when the outcome is continuous, it does not work well when the outcome is binary, as in the case of our example. In this chapter, we will consider logistic regression as an approach to estimate the association between a binary outcome variable and multiple independent variables.

17.2 EXTENSIONS OF CONTINGENCY TABLE ANALYSES

SIMPLE LOGISTIC REGRESSION

We first consider formulating a regression model to estimate and test the association between a binary outcome variable and a binary independent variable, as considered in the example from the previous section. The simple logistic regression model is given by

\[ \ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x \]  

(17.1)

where \( \pi = \Pr[CHD = 1 | FEMALE = x]. \) Note that the response in the model shown by Eq. (17.1) is the natural logarithm of the odds (or log-odds) of the
17.2 Extensions of Contingency Table Analyses Simple Logistic Regression

probability that $CHD = 1$ when $FEMALE$ takes on value $x$. Further, since this probability can theoretically take on values ranging from 0 to 1, the odds can take on values ranging from 0 to $\infty$. This implies that the log-odds can take on values ranging from $-\infty$ to $\infty$. In more intuitive terms the log transform of the odds scatters the expected probability of the response onto the interval $(-\infty, \infty)$ so that the transformed model can be calculated and interpreted much like simple linear regression.

To be a useful model for our purposes, it is necessary that the model in Eq. (17.1) be interpretable to us in a scientific sense. That is, the coefficients in the regression model ($\beta_0$ and $\beta_1$) should have meaning to us. A primary benefit of the logistic regression model is that one can easily interpret the regression coefficients in terms of the odds of the response. To see this, consider “plugging in” a value of $FEMALE = 0$ (indicating a male subject) into Eq. (17.1). By recalling our notation $\pi_0 = \Pr[CHD = 1 | FEMALE = 0]$ and $\pi_1 = \Pr[CHD = 1 | FEMALE = 1]$, we have

$$\ln \left( \frac{\pi_0}{1 - \pi_0} \right) = \beta_0, \text{ or equivalently } e^{\beta_0} = \frac{\pi_0}{1 - \pi_0}. \tag{17.2}$$

Thus $e^{\beta_0}$ in the simple logistic regression model represents the odds of CHD when the independent variable, $FEMALE$, takes on a value of 0. Further, by recalling a basic property of logarithms that $\ln(a/b) = \ln(a) - \ln(b)$, we can use the form of Eq. (17.1) to obtain

$$\ln \left( \frac{\pi_1}{1 - \pi_1} / \frac{\pi_0}{1 - \pi_0} \right) = \ln \left( \frac{\pi_1}{1 - \pi_1} \right) - \ln \left( \frac{\pi_0}{1 - \pi_0} \right) = \beta_0 + \beta_1 - \beta_0 = \beta_1. \tag{17.3}$$

Exponentiating both sides implies that the odds ratio comparing females ($FEMALE = 1$) to males ($FEMALE = 0$) is equal to $e^{\beta_1}$. Therefore to test the equivalent null hypotheses stated in Section 17.1, we could alternatively test the hypothesis

$$H_0 : e^{\beta_1} = 1, \text{ or equivalently } H_0 : \beta_1 = 0. \tag{17.4}$$

EXAMPLE: ESTIMATING THE PROBABILITY OF CORONARY HEART DISEASE IN THE FRAMINGHAM STUDY

The Framingham study was a cohort study that initially enrolled $N = 5209$ subjects from the small town of Framingham, Massachusetts in 1948. Subjects agreed to undergo biennial exams where biomarkers such as blood pressure, serum cholesterol, weight, and height were measured at each visit. Here, we will consider data on subjects for up to 30 years of enrollment into the study. Our interest will be in quantifying the probability of a subject experiencing CHD over the course of the follow-up,
with CHD being defined either via a clinical diagnosis of CHD or death from a CHD-related event such as heart attack. For illustrative purposes an example of data for five randomly sampled patients is given in Table 17.2. Covariates available in the data set include FEMALE (1 if female sex, 0 otherwise), SBP (systolic blood pressure), DBP (diastolic blood pressure), CHD (1 if the subject had CHD, 0 otherwise), FOLLOW-UP (the total number of months the subject was followed for), AGE (the age of the subject at baseline in years), BMI (body mass index of the subject), and ID (a randomly assigned unique identifier for the subject).

Based upon these data, one could estimate the association between sex and the risk of CHD. To do this, we could consider a logistic regression model of the form

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \text{FEMALE}$$

(17.5)

where $\pi$ denotes the probability that $\text{CHD} = 1$ and FEMALE is a binary indicator for female sex. Using standard statistical software, the estimates of $\beta_0$ and $\beta_1$ based upon the Framingham data are given by $b_0 = -0.399$ and $b_1 = -0.725$, respectively. So, from Eq. (17.2), we estimate that the odds of CHD among males (when FEMALE = 0) are $e^{-0.399} = 0.671$. In addition, from Eq. (17.3), we estimate that the odds ratio comparing females to males is $e^{-0.725} = 0.484$. In other words, we estimate that the odds of CHD are 51.6% $(1 - 0.484)$ lower among females as compared to males.

**Exercise 17.1**

In the Framingham dataset, there are a total of $N = 2049$ males and $N = 2650$ females. In addition, 823 males and 650 females had CHD at the 30-year follow-up. Using these data and the odds and odds ratio estimators given in Chapter 9, Tests on categorical data, estimate the odds of CHD for males and the odds ratio for CHD comparing females to males. Conclude that these estimates are the same as those obtained from the logistic regression model given in the Framingham example. (Note: This equivalence is true in general).
From the abovementioned example, we can see that the probability and odds ratio estimates that we obtain from the logistic regression model with a single binary predictor are the same as those found using the contingency table approaches presented in Chapter 9, Tests on categorical data. So why do we have an entire chapter devoted to the logistic regression model? The answer is that the logistic regression model is a more generalizable method when compared to the approaches of Chapter 9, Tests on categorical data. This generalization provides two primary advantages of the logistic regression model when compared to the contingency table approaches presented earlier. First, the logistic regression model allows us to estimate the association between a continuous predictor, such as age or SBP, and the odds of an event. The second is that the logistic regression model allows us to control for other covariates such as confounding factors when estimating the association between a predictor of interest and the odds of an event. In the Framingham example, it could be that males tend to be older than females in the sample. Because age is positively associated with the risk of CHD, this could then explain the reason why females are estimated to have lower odds of CHD when compared to males. Thus to make a “fair” comparison, we would want to compare males and females of a similar age. This could be accomplished by including age in the regression model given in Eq. (17.5).

We first begin by considering the interpretation of the logistic regression model parameters when a continuous predictor is included in the model. Suppose now that $X$ is a continuous predictor such as $SBP$ and consider formulating the logistic regression model given by

$$\ln\left(\frac{\pi_x}{1 - \pi_x}\right) = \beta_0 + \beta_1 x$$  \hspace{1cm} (17.6)

where $\pi_x = \Pr[Y = 1 \mid X = x]$. As before, we can interpret the intercept in this model, $\beta_0$, by plugging in a value of $X = 0$ into model. In this case, we have

$$\ln\left(\frac{\pi_0}{1 - \pi_0}\right) = \beta_0$$  \hspace{1cm} (17.7)

so that $\pi_0/(1 - \pi_0) = e^{\beta_0}$ gives the odds of the outcome among a population with $X = 0$. If $X$ denotes SBP and we are considering the outcome of CHD as in the previous example, we can see that the intercept really has no real-world interpretation in this case, since the exponentiated intercept represents the odds of CHD among individuals with a SBP of 0 mmHg and there are no such individuals in our sample! Despite this, there would still be meaning to $\beta_1$, the coefficient associated with $X$. Specifically, $e^{\beta_1}$ represents the odds ratio comparing two subpopulations differing in $X$ by one unit.
Thus in the case of \( X \) denoting SBP for the Framingham data, \( e^{\beta_1} \) would be the relative difference in the odds comparing subpopulations differing in SBP by 1 mmHg.

To visually see how the underlying probability of the outcome is impacted by the coefficient in the logistic regression model, Fig. 17.1 provides a graphical representation of the probability as a continuous covariate \( x \) changes in the logistic regression model. Fig. 17.1A depicts the probability response curve for a positive coefficient associated with \( x \), specifically \( \ln\left(\frac{\pi_x}{1 - \pi_x}\right) = 1 + 0.2 \, x \), with positive coefficient 0.2. Fig. 17.1B depicts the probability response curve for a negative coefficient associated with \( x \), specifically \( \ln\left(\frac{\pi_x}{1 - \pi_x}\right) = 1 - 0.2 \, x \), with negative coefficient \(-0.2\). What can be seen from the figures is that the logistic regression model bounds the probability of an event between 0 and 1 (as we would hope!). The other obvious point is that the probability does not change linearly with the covariate \( x \). Indeed, in the logistic regression model given in Eq. (17.6), it is the log-odds that changes linearly with the covariate \( x \). Finally, we can see that for a positive coefficient, the probability of the outcome increases with increasing values of \( x \). When the coefficient is negative, the probability of the outcome decreases with increasing values of \( x \).

The other major benefit of the logistic regression model is the ability to adjust for other variables when estimating the association between a predictor and a binary outcome variable. Previously, we noted that one would likely wish to control for age when estimating the association between sex and the probability of CHD in the Framingham data. Such a logistic regression would take the form

\[
\ln\left(\frac{\pi_{\text{FEMALE,AGE}}}{1 - \pi_{\text{FEMALE,AGE}}}\right) = \beta_0 + \beta_1 \, \text{FEMALE} + \beta_2 \, \text{AGE}. \tag{17.8}
\]
The regression given in Eq. (17.8) now allows us to compare the relative difference in odds of CHD comparing females to males of a similar age. To see this, suppose that we wished to compare the odds CHD between females and males who are 53 years of age. In this case, we would have

\[
\ln\left( \frac{\pi_{1,53}}{1 - \pi_{1,53}} \right) = \ln\left( \frac{\pi_{0,53}}{1 - \pi_{0,53}} \right) = \beta_0 + \beta_1 + 53 \times \beta_2 - (\beta_0 + 53 \times \beta_2) = \beta_1.
\]  

(17.9)

Exponentiating both sides as we did before implies that the odds ratio for CHD comparing females to males, aged 53 years, is equal to \( e^{\beta_1} \). Although the abovementioned example considered a specific age (53 years), it is easy to see that we could have used any value and it would have canceled in the difference. The implication of this is that \( e^{\beta_1} \) actually represents the odds ratio for CHD comparing females to males of the same age (regardless of what that age is).

The previous concept can easily be extended to multiple adjustment variables. For example, we may wish to estimate the association between sex and the odds of CHD while adjusting for both age and BMI, another potential confounder. In general, we could consider many adjustment variables giving rise to a logistic regression of the form

\[
\ln\left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
\]

(17.10)

where \( \pi = \Pr[Y = 1 | X_1 = x_1, \ldots, X_k = x_k] \). Following the same logic used in the age adjustment example, we can interpret the coefficients of the logistic regression model in terms of the odds ratio associated with each predictor.

**INTERPRETATION OF LOGISTIC REGRESSION PARAMETERS**

For the logistic regression model

\[
\ln\left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
\]

(17.11)

\( e^{\beta_1} \) is interpretable as the odds ratio comparing two subpopulations differing in \( x_1 \) by 1 unit that have similar values for \( x_2, \ldots, x_k \).

**Exercise 17.2**

In the Framingham dataset\(^1\) a logistic regression model fits adjusting for sex and age results in the following estimates

\[
\ln\left( \frac{p}{1 - p} \right) = -2.22 - 0.761 \times FEMALE + 0.040 \times AGE.
\]

(17.12)
CHAPTER 17 Logistic regression for binary outcomes

From this equation, what is the estimated odds ratio comparing females to males of a similar age? What is the estimated odds ratio comparing individuals differing in age by 1 year who have the same sex?

Exercise 17.3
Again, consider the Framingham dataset. A logistic regression model fits adjusting for sex, age, and BMI results in the following estimates

\[
\ln \left( \frac{p}{1-p} \right) = -3.80 - 0.740 \times FEMALE + 0.035 \times AGE + 0.069 \times BMI. \quad (17.13)
\]

From this, what is the estimated odds ratio comparing females to males of a similar age and BMI?

As with the linear regression model, the logistic regression model can also be used to obtain predicted probabilities for a given set of independent covariate values. In order to do this, we must solve for \( \pi \) in Eq. (17.10). This is done by first exponentiating both sides of (17.10), then solving through. To obtain a predicted probability from a logistic regression model, we solve for the probability of the response as follows:

\[
\ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad \Rightarrow \quad \frac{\pi}{1 - \pi} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k} \quad \Rightarrow \quad \pi = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} \quad (17.14)
\]

OBTAINING PREDICTED PROBABILITIES
After software is used to obtain estimates \( b_0, b_1, \ldots, b_k \), the predicted probability of the outcome for a subject with covariate values \( x_1, \ldots, x_k \) is given by

\[
p = \frac{e^{b_0 + b_1 x_1 + \cdots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + \cdots + b_k x_k}} \quad (17.15)
\]

EXAMPLE: PREDICTING THE PROBABILITY OF CORONARY HEART DISEASE IN THE FRAMINGHAM STUDY
For the Framingham example,\(^1\) we can use the abovementioned formula to estimate the probability of CHD for a given sex and age. Recall from Eq. (17.12) that the
estimated model for CHD was

\[
\ln\left(\frac{p}{1-p}\right) = -2.22 - 0.761 \times FEMALE + 0.040 \times AGE. \tag{17.16}
\]

This implies that the probability of CHD, as determined by sex and age, is estimated to be

\[
p = \frac{e^{-2.22+-0.761 \times FEMALE+0.040 \times AGE}}{1 + e^{-2.22+-0.761 \times FEMALE+0.040 \times AGE}}. \tag{17.17}
\]

Now consider the estimated probability of CHD for a randomly sampled 47-year-old male. Plugging these values into Eq. (17.17), we obtain

\[
p = \frac{e^{-2.22+-0.761 \times 1+0.040 \times 47}}{1 + e^{-2.22+-0.761 \times 1+0.040 \times 47}} = 0.250, \tag{17.18}
\]

so that the predicted probability of CHD for a 47-year-old male is 25%.

**Exercise 17.4**

Using the model estimates given in Eq. (17.12), find the predicted probability of CHD for a 67-year-old female and an 83-year-old male.

**Exercise 17.5**

A better prediction would probably account for a person’s BMI. Using the model estimates given in Eq. (17.13), find the predicted probability of CHD for a 67-year-old female with a BMI of 27.2 and an 83-year-old male with a BMI of 22.4.

#### 17.4 Inference for Association Parameters

**Test of association**

It is commonly of interest to determine whether an association between an independent variable and a dependent variable exists. In Section 17.3 we saw that for the logistic regression model specified in Eq. (17.8), \(e^{\beta_1}\) represents the odds ratio for CHD comparing females to males of the same age. Thus if \(e^{\beta_1} = 1\), or equivalently if \(\beta_1 = 0\), it would imply that the odds of CHD are the same comparing females to males of the same age. This would imply that no association exists between sex and CHD. Conversely, if \(\beta_1 \neq 0\), this would imply that the odds of CHD differ between females and males of the same age. Thus it may be of interest to test the null hypothesis \(H_0: \beta_1 = 0\) versus the alternative hypothesis \(H_1: \beta_1 \neq 0\). In order to carry out this hypothesis test, we require a test statistic for which we know its distribution under the null hypothesis, \(H_0\).
Consider the general logistic regression model
\[
\ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.
\] (17.19)

Statistical software should be used to obtain the rather complicated estimates of the parameters \(\beta_0, \beta_1, \ldots, \beta_k\). Let \(b_0, b_1, \ldots, b_k\) denote these estimates. Furthermore, software should also be used to calculate the rather complicated estimates of the standard errors of \(b_0, b_1, \ldots, b_k\), which we denote as \(se(b_0), se(b_1), \ldots, se(b_k)\).

**TESTS FOR A SINGLE LOGISTIC REGRESSION PARAMETER**

A test of the null hypothesis \(H_0: \beta_j = 0\) versus the alternative hypothesis \(H_1: \beta_j \neq 0, j = 0, \ldots, k\), can be conducted with the following test statistic
\[
z = \frac{b_j}{se(b_j)} \sim N(0, 1).
\] (17.20)

The value of \(z\) can be looked up in Table I and the resulting \(p\)-value is the probability of observing a test statistic as big, or bigger, in magnitude than that observed if the null hypothesis were true.

**EXAMPLE: TESTING THE ASSOCIATION BETWEEN SEX AND CORONARY HEART DISEASE IN THE FRAMINGHAM STUDY**

Returning to the Framingham study, we obtain logistic regression estimates for the model, including sex, age, and BMI as in Table 17.3.

From Table 17.3, we can see that the \(z\) statistic for testing whether sex is associated with the odds of CHD is given by
\[
z = \frac{-0.74035}{0.06539} = -11.32.
\] (17.21)

<table>
<thead>
<tr>
<th></th>
<th>(b)</th>
<th>se((b))</th>
<th>(z)</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.80376</td>
<td>0.26145</td>
<td>-14.55</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.74035</td>
<td>0.06539</td>
<td>-11.32</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>AGE</td>
<td>0.03475</td>
<td>0.00388</td>
<td>8.95</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>BMI</td>
<td>0.06937</td>
<td>0.00801</td>
<td>See Ex</td>
<td>See Ex</td>
</tr>
</tbody>
</table>

*BMI, Body mass index.*
The \( p \)-value corresponding to the \( z \) statistic for testing that the log-odds ratio corresponding to sex is less than 0.001, as seen in Table I. If conducting a level 0.05 test, we may say that the \( p \)-value is significant. The actual \( p \)-value associated with the test as calculated from a software package is \(< 0.0001\). We have strong evidence from the sample to conclude that the odds of CHD differ by sex after adjustment for both age and BMI.

**Exercise 17.6**

Using the model estimates given in Table 17.3, conduct a level 0.05 test of the null hypothesis that BMI is not associated with CHD after adjustment for both sex and age.

**Confidence intervals for the odds ratio**

The test statistic given in Eq. (17.20) provides us with a binary decision regarding the association between a predictor of interest and the probability of an event. However, it is also important to quantify our uncertainty in the odds ratio associated with the predictor. A confidence interval can be formulated from the estimated regression parameter, the estimated standard error, and the appropriate quantile from a standard normal distribution.

**CONFIDENCE INTERVAL FOR THE ODDS RATIO**

A \( 100 \times (1 - \alpha)\% \) confidence interval for the odds ratio associated with predictor \( x_j \) is given by

\[
\left( e^{b_j - z_{1 - \alpha/2} \times se(b_j)}, e^{b_j + z_{1 - \alpha/2} \times se(b_j)} \right)
\]

where \( z_{1-\alpha/2} \) is the \( 1 - \alpha/2 \) quantile of the standard normal distribution. This value is 1.96 for a 95% confidence interval \((\alpha = 0.05)\).

**EXAMPLE: CONFIDENCE INTERVAL FOR THE ODDS RATIO CORRESPONDING TO SEX IN THE FRAMINGHAM STUDY**

Using the output in Table 17.3, the estimated odds ratio associated with sex (after adjustment for age and BMI) is equal to \( e^{-0.74} = 0.48 \). A 95% confidence interval for the odds ratio is

\[
\left( e^{-0.74-1.96 \times 0.065}, e^{-0.74+1.96 \times 0.065} \right) = (0.42, 0.54).
\]

**Exercise 17.7**

Using the model estimates given in Table 17.3, compute a 95% confidence interval for the odds ratio associated with BMI after adjustment for both sex and age.
17.5 MODEL DIAGNOSTICS AND GOODNESS-OF-FIT

When we fit a regression model with the goal of predicting the probability of a response, it is good practice to assess how well the model fits our observed data and to determine if particular observations in the data set have undue influence on our predictions. Model diagnostics help us diagnose where problems might exist in the model. For example, in the context of the Framingham data where we may be interested in predicting the risk of CHD for subjects based upon their characteristics, it may be that \( SBP \) is a strong predictor of CHD but that this relationship is not linearly (or even monotonically) increasing. It may also be that we have omitted some variables from the model that are important predictors of CHD, and in doing so, we fail to correctly predict the occurrence of CHD in our observed data adequately. Model diagnostics will help us to assess these issues.

Assessing functional form of predictors

The \textit{functional form} of a predictor gives us information on how the predictor should be transformed in a regression model to best predict our outcome of interest. The goal of regression model diagnostics in assessing functional form is to examine how best to transform a predictor with the goal of fitting our observed data. In the linear regression model (Chapter 15: Linear regression and correlation), this was done by plotting the model residuals against the predictor of interest. The resulting shape of the plot would then reveal the appropriate transformation of the predictor. A similar approach can be utilized with the logistic regression model, but we must do something a bit fancier to “scatter” the residuals in order to see patterns (recall that in logistic regression, our outcome takes on only one of two values).

To generate this scatter, we utilize what is known as the partial residual in the statistical literature. Consider the logistic regression model

\[
\ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
\]

and suppose that we wish to determine whether or not the predictor \( x_1 \) needs to be transformed (e.g., by squaring it for a quadratic effect that would indicate both low and high values of \( x_1 \) lead to an increased probability of the response).

\[
\text{THE PARTIAL RESIDUAL FOR LOGISTIC REGRESSION}
\]

The \textit{partial residual} for \( x_1 \) corresponding to observation \( i \) is computed as

\[
r_{\text{part},i} = (y_i - \pi_i) \times \frac{1}{\pi_i(1 - \pi_i)} + x_i \hat{b}_1
\]

(17.25)

(Continued)
where $p_i$ is the predicted response for observation $i$ and $x_{1i}$ is the value of predictor $x_1$ for observation $i$.

Why is $r_{part,i}$ called the partial residual? First, note that the first term of Eq. (17.25) contains $y_i - p_i$, the observed minus fitted value for the $i$th observation. This is the usual residual. The factor $1/(p_i(1 - p_i))$ helps to better approximate the log-odds. The term “partial” comes from the fact that we are adding back on the contribution of $x_{1i}b_1$ to the residual. This additional term allows us to isolate the proper transformation that should be used for $x_1$.

To assess how $x_1$ may need to be transformed, we can then plot $r_{part,i}$ against $x_{1i}$ for all of the observations in our data set. The shape of this plot suggests how best to transform $x_1$. If the underlying shape is a straight line, this would indicate that $x_1$ does not need to be transformed at all. Because these plots can be hard to read, it is often useful to add a running average line through the points to better visualize the shape.

**EXAMPLE: PREDICTION OF NODAL INVOLVEMENT IN PROSTATE CANCER PATIENTS**

Here we consider the problem of predicting whether or not prostate cancer has spread to the lymph nodes among a sample of newly diagnosed prostate cancer patients. The data in Table 17.4 consist of multiple variables, including the age of the patient at the time of diagnosis ($AGE$), the amount of serum acid phosphatase measured in the patient at the time of diagnosis ($ACID$), a binary indicator of x-ray findings indicating nodal involvement ($X\text{-RAY}$), the size ($SIZE$; large vs small) and grade ($GRADE$; advanced vs non-advanced) of the primary tumor, and indicator of lymph node involvement based upon biopsy ($LNODE$). The goal of the study is to use the covariates available to predict

<table>
<thead>
<tr>
<th>PATIENT</th>
<th>AGE</th>
<th>ACID</th>
<th>X-RAY</th>
<th>SIZE</th>
<th>GRADE</th>
<th>LNODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>0.48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>66</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>56</td>
<td>0.52</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>0.49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>0.46</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>60</td>
<td>0.62</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>49</td>
<td>0.55</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
whether or not a patient has nodal involvement, in the absence of an invasive biopsy. As such, the logistic regression model is a natural approach to the problem.

Suppose that after some testing of different covariates in the data set, we obtain a fitted logistic regression model for predicting the probability of nodal involvement that involves each of the predictors listed in Table 17.4. However, it may still be unclear as to whether or not \textit{ACID} (serum acid phosphatase) should be transformed in the model to obtain a better prediction. To assess this, we might start by naively plotting our observed response versus the predictor \textit{ACID}. The resulting plot is given in Fig. 17.2. To help with seeing the plot, we have also added a running average (the sample proportion) of subjects with nodal involvement (red line) as well as the predicted values that our model would suggest for all of the values of \textit{ACID} (blue line). What is immediately clear from this plot is that it is difficult to judge just how \textit{ACID} should be transformed. This is because the value of the response takes on only two values (nodal involvement: yes or no).

A better strategy is to plot the partial residuals for \textit{ACID} versus the variable \textit{ACID} itself so that we can “scatter” the outcome a bit. This plot is given in Fig. 17.3A. After adding a running average to the plot, we can see that the shape does not appear linear. This indicates that some transformation is needed. We can see that the running average tends to underestimate the partial residuals for low values of \textit{ACID}, overestimate them for middle values of \textit{ACID}, and underestimate them again for high values of \textit{ACID}.

![Figure 17.2](image)

\textbf{Figure 17.2} Plot of observed response (indicator for spread of cancer into lymph nodes) versus serum acid phosphatase. The red line is the running average (proportion) of subjects with nodal involvement. The blue line is the fitted lines from regressing nodal involvement onto serum acid phosphatase as a linear term in the logistic regression model.
This suggests that a logarithmic transformation of ACID may need to be added to the model. After adding this term to the model and replotting the resulting partial residuals versus the variable ACID we obtain Fig. 17.3B. Now we can see that the running average more closely matches the straight line (added for reference), indicating the squared term for ACID does indeed help with the prediction of nodal involvement.

Assessing influential observations

When trying to predict a response, it is important to determine if a single observation has, or a small number of observations have, a disproportionate amount of influence on our model’s predictions. An observation can have high influence only if (1) the predictors for the observations are quite different from the rest of the sample and (2) given the predictors, the observation’s response is quite different than what is expected. The result of both (1) and (2) occurring means that our model estimates will be highly affected by the observations. Thus assessing influential observations often focuses on determining if the predictors for the observations are quite different from the rest of the sample and then determining if our model estimates change dramatically when the observation is removed from the analysis.

To judge if the predictors for a given observation are quite different from the rest of the sample, we focus on what is called leverage. This is a fancy word that simply means we wish to compare the difference in the predictor values for a given observation to the mean predictor value of all other observations in the sample. If this difference is large, we would say that the observation has high leverage, implying that observation’s predictor values are quite different from those of all other observations in the sample.

EXAMPLE: PREDICTION OF NODAL INVOLVEMENT IN PROSTATE CANCER PATIENTS

Let us return to the problem of predicting nodal involvement in our sample of prostate cancer patients. Now let us consider whether any subject has high leverage in the
data set and may therefore be overly influential on our predictions. Fig. 17.4 depicts the leverage for each subject in the data set. We have added the subject ID numbers for each point to easily identify which subject(s) may have high leverage. We can see that most of the points tend to cluster together near the bottom of the plot (indicating low leverage). While there are some points scattering upward, we can see that one subject tends to stand out. This is subject 24 who appears to have much higher leverage than any of the other subjects in the data set. In order for this to happen, subject 24 must have quite a different value for at least one predictor.

To further investigate this, let us consider the predictor values of subject 24 and compare them to the summarized predictor values in the rest of the data set. Subject 24’s predictor values are as appear in Table 17.5. We can see that subject 24 was 64 years of age, had serum acid phosphatase value of 1.87, had a negative x-ray, a small tumor size, a low-grade tumor, and no nodal involvement. Now, consider the summaries of the predictor values for the overall sample of patients given in Table 17.6.

![Figure 17.4 Plot of leverage versus patient ID. From the plot we can see that patient 24 has relatively higher leverage than all other patients, indicating that his/her covariate values differ in some way from the rest of the study sample. The red horizontal line in the plot indicates three times the average leverage in the sample. ID, Identifier.](image)

**Table 17.5** Data for patient identifier 24.

<table>
<thead>
<tr>
<th>PATIENT</th>
<th>AGE</th>
<th>ACID</th>
<th>X-RAY</th>
<th>SIZE</th>
<th>GRADE</th>
<th>LNODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>64</td>
<td>1.87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
What is immediately obvious is that the \textit{ACID} value for subject 24 is very high (actually the highest in the sample by far). In fact, the 75th percentile of \textit{ACID} values in the sample is only 0.78 (less than half that of patient 24). At this point, one should investigate whether this is a data entry error and fix it if necessary. It turns out that a serum acid phosphatase value of 1.87 (while very large) is plausible and was the correct value for this subject. As such, it would be best to leave this observation in the model, but to further assess how much this leverage has on the resulting model estimates and predictions.

\begin{table}[h]
\centering
\caption{Summaries of covariates for all patients in the prostate cancer study.}
\begin{tabular}{lccccccc}
\hline
\textit{PATIENT} & \textit{AGE} & \textit{ACID} & \textit{X-RAY} & \textit{SIZE} & \textit{GRADE} & \textit{LNODE} \\
\hline
Mean & 59.4 & 0.69 & 0.28 & 0.51 & 0.38 & 0.38 \\
Median & 60.0 & 0.65 & 0 & 1 & 0 & 0 \\
SD* & 6.16 & 0.26 & 0.46 & 0.51 & 0.49 & 0.49 \\
\hline
\end{tabular}
\footnotesize{*SD = standard deviation.}
\end{table}

\textbf{Cook’s distance}

A natural measure of the influence of an observation is the amount that the estimated model coefficients change when the observation is removed from the data set. If the change in the model estimates is large when the observation is removed, this would indicate that the observation has a high influence on the model fit. \textit{Cook’s distance} can be thought of as a standardized average of the change in model parameter estimates when each observation is removed. The formula for computing Cook’s distance is beyond the scope of this textbook, though it does have a fairly easy interpretation: observations with a large Cook’s distance tend to have high influence. A useful plot for identifying influential points is to plot Cook’s distance against leverage. \textit{Fig. 17.5} displays this plot for the prostate cancer data. The horizontal reference line in the plot is two times the standard deviation of Cook’s distance. The vertical line is three times the average leverage in the data set. Observations in the upper right quadrant of the plot should be examined for their influence. We can see that for this example, no observations have great influence.

\textbf{Assessing goodness-of-fit: The Hosmer—Lemeshow test}

The concept of \textit{goodness-of-fit} speaks to how well our model explains the variation in the response. In the case of the logistic regression model, one measure of goodness-of-fit is to quantify how well the predicted likelihood of any observation (based upon their covariate values) matches the actual outcome that was observed in a patient. One difficulty with assessing goodness-of-fit for the logistic regression model is that our predictions are probabilities (ranging from 0 to 1), while the outcome for each observation is either 0 or 1. To overcome this the \textit{Hosmer—Lemeshow goodness-of-fit} test
groups observations that are similar with respect to their probability of an event then compare the observed number of events for each group with the expected number of events as given by the logistic regression model. The procedure works as follows:

**THE HOSMER–LEMESHOW GOODNESS-OF-FIT PROCEDURE**

1. Fit the model
2. Generate the predicted probabilities for each observation as
   \[ p_i = \frac{e^{b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}}}{1 + e^{b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}}}, i = 1, \ldots, n \]  
   \[ (17.26) \]

3. Order the predicted probabilities
4. Group the data by quantiles of the predicted probabilities.
   The number of groups, \( g \), should allow for sufficient numbers of expected events in each group (at least 5)
5. For each group \( j \), compute:
   a. The observed number of successes, \( o_j \), equals the sum of outcomes for all observations in group \( j \)
   b. The expected number of successes, \( e_j \), equals the sum of predicted probabilities for all observations in group \( j \)
6. Compare the observed to expected number of successes using the following statistic:
   \[ \chi^2_{HL} = \sum_{j=1}^{g} \frac{(o_j - e_j)^2}{e_j(1 - (e_j/n_j))} \]  
   \[ (17.27) \]

   where \( n_j \) is the number of observations in group \( j \)

(Continued)
7. Test the hypothesis

\( H_0: \) The model fits the observed data adequately

versus

\( H_A: \) The model does not fit the observed data adequately

by comparing \( X^2_{HL} \) to a chi-squared distribution with \( g - 2 \) degrees of freedom

Note that the comparison subtracts the expected number of successes from the observed number of successes, squares, then standardizes the differences before summing across groups. If the model fits well, each difference will be small, and hence the overall statistic will be close to 0. It has further been noted that if the fitted model is appropriate, then \( X^2_{HL} \) will have an approximate chi-square distribution with \( (g - 2) \) degrees of freedom, allowing for the computation of a \( p \)-value for testing how well the model fits the observed data.

EXAMPLE: ESTIMATING THE PROBABILITY OF CORONARY HEART DISEASE IN THE FRAMINGHAM STUDY

Let us return to the Framingham study example where our goal is to build a model for the risk of CHD. First, let us consider a model that includes patient sex, age, and SBP. Thus we have a logistic regression model of the form:

\[
\ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 \text{FEMALE} + \beta_2 \text{AGE} + \beta_3 \text{SBP}.
\] (17.28)

After fitting the model using statistical software and obtaining the predicted probabilities from the model for each subject, we divided the data set into 10 groups (see step 4 in the Procedure) using the ordered predicted probabilities. The total number of observed and expected CHD cases was computed for each of the 10 groups and is presented in Table 17.7. We can see that the number of observed CHD cases generally increases with the group ordering. Further, the expected number of cases as predicted by the model is fairly close for some groups but tends to overestimate for lower groups and overestimate for higher groups. For example, in the lowest risk group we observed 55 cases but predicted 67.5 cases. In the highest risk group, we observed 225 cases but predicted 258.0 cases. Computing the Hosmer–Lemeshow statistic in Eq. (17.27), we obtain

\[
X^2_{HL} = \frac{(55 - 67.5)^2}{67.5 \left(1 - \frac{67.5}{472}\right)} + \cdots + \frac{(225 - 258.0)^2}{258.0 \left(1 - \frac{258.0}{470}\right)} = 25.38.
\] (17.29)

Because \( X^2_{HL} \) approximately follows a chi-squared distribution with \( 10 - 2 = 8 \) degrees of freedom, from Table III we can see that the \( p \)-value for testing the
null hypothesis that our model adequately fits the observed data is < 0.01. Thus we would reject the null hypothesis at level 0.05 and conclude that the model does not provide an adequate fit to the data. The actual $p$-value computed from software is 0.0013.

We have rejected the abovementioned test, indicating that the logistic regression model with only age, gender, and SBP does not provide an adequate fit to the observed data. However, it is well known that BMI is also a strong predictor of one’s propensity to develop CHD. Allow us to consider the model that is adjusted for BMI as well as age, gender, and SBP. Thus we now consider a model of the form

$$
\ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 \text{FEMALE} + \beta_2 \text{AGE} + \beta_3 \text{SBP} + \beta_4 \text{BMI}.
$$

After fitting the model using software, we obtain estimates for the regression parameters as in Table 17.8.
We can see that BMI is statistically significant indicating that there is an association with the risk of CHD, but does it substantially improve the overall fit of the model for predicting CHD? To determine this, we again appeal to the Hosmer–Lemeshow goodness-of-fit test, obtaining the predicted probabilities from the model, ordering them and again forming 10 groups. Doing this results in the observed and expected CHD cases given in Table 17.9.

Now we can see that the observed and expected values are much closer than before, across all groups. Computing the Hosmer–Lemeshow statistic in the model now adjusting for BMI we obtain

\[ X^2_{HL} = \frac{(50-57.9)^2}{57.9 \left(1 - \frac{57.9}{469}\right)} + \cdots + \frac{(247-256.3)^2}{256.3 \left(1 - \frac{256.3}{469}\right)} = 5.86. \] (17.31)

Again, \( X^2_{HL} \) approximately follows a chi-squared distribution with \( 10-2 = 8 \) degrees of freedom. From Table III, we can see that the \( p \)-value for testing the null hypothesis that our model adequately fits the observed data is far greater than 0.10. Indeed, a software calculation yields \( p \)-value = 0.663. Thus we fail to reject the null hypothesis at level 0.05 and conclude that the model does provide an adequate fit to the data.

**REFERENCES**

1. Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. <https://biolincc.nhlbi.nih.gov/teaching/>.

### Table 17.9

Statistics needed for computing the Hosmer–Lemeshow goodness-of-fit statistic using a model that includes sex, age, and systolic blood pressure.

<table>
<thead>
<tr>
<th>Group</th>
<th>No. observations</th>
<th>Obs. no. CHD</th>
<th>Exp. no. CHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>469</td>
<td>50</td>
<td>57.9</td>
</tr>
<tr>
<td>2</td>
<td>469</td>
<td>74</td>
<td>75.6</td>
</tr>
<tr>
<td>3</td>
<td>469</td>
<td>98</td>
<td>94.1</td>
</tr>
<tr>
<td>4</td>
<td>469</td>
<td>119</td>
<td>117.5</td>
</tr>
<tr>
<td>5</td>
<td>469</td>
<td>150</td>
<td>138.1</td>
</tr>
<tr>
<td>6</td>
<td>469</td>
<td>144</td>
<td>155.7</td>
</tr>
<tr>
<td>7</td>
<td>469</td>
<td>173</td>
<td>173.3</td>
</tr>
<tr>
<td>8</td>
<td>469</td>
<td>197</td>
<td>191.3</td>
</tr>
<tr>
<td>9</td>
<td>469</td>
<td>220</td>
<td>212.2</td>
</tr>
<tr>
<td>10</td>
<td>469</td>
<td>247</td>
<td>256.3</td>
</tr>
</tbody>
</table>

After model fitting the predicted probabilities were ordered, then subjects were divided into 10 roughly equal groups. CHD, Coronary heart disease.
18.1 INTRODUCTION

In many biomedical studies we are interested in comparing the mean of a count variable or the rate of occurrences of an event over a known period of observation. Examples of this might be a comparison of the number of seizures recorded for epileptic patients over a given period of time, the number of vascular access replacements over time among patients undergoing hemodialysis, or the number of words that can be stated that begin with the letters F, A, or S over a 60 second period among subjects undergoing neuropsychological testing. In each of the above-mentioned examples, the outcome is discrete (taking on whole number values) and nonnegative. In such cases, it is common to assume that the response might follow a Poisson distribution. In this chapter, we will further discuss the Poisson distribution and consider the use of this distribution for estimating the rate of a count response and comparing rates of events across patient subpopulations.

18.2 THE POISSON DISTRIBUTION

In Chapter 4, Distributions, we briefly introduced the Poisson distribution, named after the French mathematician, Siméon Denis Poisson, who published a theory on it in 1837. It was noted there that the Poisson distribution arises in cases in which there are many opportunities for an event to occur, but a very small chance of occurrence on any one trial. There are a few defining characteristics that give rise to a random variable following a Poisson distribution, which we can state a bit more formally. First, a random variable that follows a Poisson distribution is always nonnegative and takes on only whole number values. For this reason, it is commonly regarded as a natural choice for modeling counts of an event over time, such as the count of the number of seizures that an epileptic patient experiences over 2 weeks of observation. However, a few other properties must also be satisfied in order to yield a Poisson distribution. The first is that the rate at which the event occurs remains constant over all time. For example, if we followed an epileptic patient for multiple 2-week intervals, this would imply that the patient’s rate of seizures does not increase or decrease from 1–2-week interval to the
next. The next property that must be satisfied in order to give rise to the Poisson distribution is that the occurrence of an event does not change the likelihood that another event will occur. In the context of the seizure example, this would imply that if the patient had a seizure, this would not impact the future risk of them having another seizure. This is sometimes referred to as the “memoryless” property of the Poisson distribution, meaning that the system can “forget” the history of events since the risk of an event occurring remains unchanged regardless of whether an event has or has not already previously occurred. The final property that gives rise to a Poisson distribution is that if we were to shrink the observation time to be very small, the risk of an event occurring during that time would tend toward 0. In the context of the seizure example, if we only followed patients for 1 second rather than 2 weeks, we would expect that the observed number of seizures over that 1 second time period would be essentially 0.

Suppose that we have an outcome \( Y \) that denotes the number of seizures observed for a patient over a 2-week time period and suppose that \( Y \) follows a Poisson distribution. Further, suppose that the mean of \( Y \) is \( \mu \). In this case we would write \( Y \sim \text{Poisson}(\mu) \). A key feature of the Poisson distribution is that the variance of a random variable that follows a Poisson distribution is actually equal to the mean. Thus if the mean of \( Y \) is \( \mu \), then the variance of \( Y \) is also \( \mu \). This characteristic of a Poisson random variable can be useful for statistical inference because once we have estimated the mean of a Poisson random variable, we already have an estimate of the variance.

### 18.3 MEANS VERSUS RATES

The Poisson distribution is commonly regarded as a natural choice for modeling counts of an event over time. If we observe a count of the number of events that have occurred over time, we can focus on estimating the mean of the count, but it is often advantageous to instead standardize the count relative to the time of observation, yielding a rate of the occurrence of the event. To better illustrate this, in the seizure example, suppose that \( Y \) represents the observed number of seizures for a patient over a 2-week interval. If the mean of \( Y \) is denoted by \( \mu \), we could also say that the rate of seizures for the patient is \( \mu \) per 2 weeks. Alternatively, we may wish to translate this into the rate of seizures over a full-year period. In this case, we could conclude that the rate of seizures is \( 26 \times \mu \) per year, since there are 52 weeks in a year and \( \mu \) is the expected number of seizures observed over any 2-week period (because a defining characteristic of the Poisson distribution is that the rate of the event does not change over any additional 2-week follow-up periods, we can simply multiply the two-week rate by 26 to obtain the expected number of seizures over 1 year).

In notation, suppose that we observed a count, \( Y \), after following a subject for a time period of \( A \). In our seizure example, \( A \) would be 2 weeks. The observed rate of
the event over 1 unit of $A$ is given by $Y$ divided by $A$. Further, if $\mu$ is the mean of $Y$, then the (true) rate of $Y$ per 1 unit of time is given by $\mu/A$, which we will denote by $\lambda$. Thus we have the relationship $\lambda = \mu/A$ or equivalently, $\mu = \lambda \times A$. Note that we typically choose the length of time for standardization to be clinically relevant. For example, it would be relevant to discuss the rate of seizures for a patient over 1 month or 1 year, but it wouldn’t be relevant to discuss the rate of seizures over 1 second (since this number would be practically zero and hence clinically irrelevant).

**EXAMPLE: ESTIMATING THE RATE OF SEIZURES FOR EPILEPTIC PATIENTS**

Suppose that the number of seizures recorded for a patient over a 2-week time period is 3. In this case, $Y = 3$ seizures and $A = 2$ weeks. The estimated rate of seizures over 1 week is $3/2 = 1.5$ seizures per week.

**Exercise 18.1**

Suppose that the number of seizures recorded for a patient over a 2-week time period is 3. What is the observed rate of seizures over 1 day? 1 year? 5 years?

**18.4 INFERENCES FOR THE RATE OF A POISSON RANDOM VARIABLE**

We previously saw how one could estimate the rate of an event by observing the count of events over a given period of time. We may also wish to test whether the true rate is equal to some hypothesized value. That is, we may wish to test the null hypothesis

$$H_0: \lambda = \lambda_0$$

versus the alternative hypothesis

$$H_1: \lambda \neq \lambda_0.$$ (18.2)

If $Y$ denotes the count of events over follow-up time period $A$ and $Y \sim \text{Poisson}(\mu)$, then $Y$ can be regarded as the sum of counts over many small intervals of time that add to $A$. For example, in the case of the seizure study, we observe $Y$ seizures over a 2-week time period. This could be obtained by adding the count of seizures each day for a total of 14 days, where the rate of seizures on any given day is $\mu/(7A) = \mu/14$. Because of this relationship, we can appeal to the central limit theorem to approximate the distribution of $Y$. Specifically, the distribution of $Y$ is approximately normally distributed with mean $\mu$ and variance $\mu$. We denote this as

$$Y \sim \text{N}(\mu, \mu),$$ (18.3)
where the variance is equal to the mean as a result of the fact that \( Y \) is truly Poisson distributed. This approximation becomes better as the value of \( \mu \) grows. In addition, if we estimated the rate, \( \lambda \), by \( l = Y/A \), then the central limit theorem implies that

\[
l = \frac{Y}{A} \approx N\left(\frac{\mu}{A}, \frac{\mu}{A^2}\right) = N\left(\lambda, \frac{\lambda}{A}\right).
\] (18.4)

While the previous approximation does hold for large \( \mu \), when \( \mu \) is small the approximation is not as good. This intuitively makes sense because we know that \( l \geq 0 \) and yet we are approximating it with a normal distribution that can take on any value (positive or negative). A better result can be obtained by approximating the distribution of the natural logarithm of \( l \), \( \ln(l) \). This is because the natural logarithm of a positive quantity can take on any values (negative or positive).

From the central limit theorem, it can be shown that

\[
\ln(l) = \ln\left(\frac{Y}{A}\right) \approx N\left(\ln\left(\frac{\mu}{A}\right), \frac{1}{\mu}\right) = N\left(\ln(\lambda), \frac{1}{\lambda A}\right).
\] (18.5)

In this case, if we are interested in testing

\[ H_0: \lambda = \lambda_0 \text{ versus } H_1: \lambda \neq \lambda_0, \] (18.6)

then this is equivalent to testing

\[ H_0: \ln(\lambda) = \ln(\lambda_0) \text{ versus } H_1: \ln(\lambda) \neq \ln(\lambda_0), \] (18.7)

and hence our test can be based upon the statistic \( \ln(l) = \ln\left(\frac{Y}{A}\right) \), the distribution of which can be approximated by (18.5). Further, it follows that under the null hypothesis above

\[
\frac{\ln(l) - \ln(\lambda_0)}{\sqrt{1/\lambda A}} \approx N(0, 1).
\] (18.8)

Of course, in practice, we do not know the true value of \( \lambda \), which is both the mean and the variance above. Estimating it by substituting \( l = Y/A \) leads to the most commonly used test, the Poisson Wald test.

The one-sample Poisson Wald test (denoted here as \( PW_1 \)) uses the estimate of the mean \( \lambda \) as \( l = Y/A \) and tests \( H_0: \lambda = \lambda_0 \) by

\[
PW_1 = \frac{\ln(l) - \ln(\lambda_0)}{\sqrt{1/Y}} \approx N(0, 1).
\] (18.9)

where \( Y \) denotes the count of events over the follow-up time period \( A \). The \( p \)-value resulting from this test is obtained from the entry in Table 1’s “area in both tails” column of \( z = PW_1 \).
Consider the example where \( \lambda \) is the number of side effects in a week (\( A = 1 \) week, therefore is 1), \( \lambda_0 = 1 \), and we observe \( Y = 3 \) effects. We want to test whether the true rate of side effects is greater than 1 per week. \( l = Y/A = 3 \) and \( PW_1 = 1.9 \) (\( z \) in Table 1). The resulting \( p \)-value is 0.054.

**EXAMPLE: HYPOTHESIS TEST FOR THE RATE OF SEIZURES AMONG EPILEPTIC PATIENTS**

Suppose that we followed an epileptic patient for 10 years and observed 73 seizures over that time period. Further suppose we wish to test whether the true rate of seizures per year in this patient is equal to 6. Then we wish to test

\[
H_0: \lambda = 6 \text{ versus } H_1: \lambda \neq 6,
\]

or equivalently,

\[
H_0: \ln(\lambda) = \ln(6) \text{ versus } H_1: \ln(\lambda) \neq \ln(6).
\]

In this case, \( Y = 73 \) and \( A = 10 \). Further, our estimate of the rate of seizures per year in this patient is \( l = Y/A = 73/10 = 7.3 \) seizures per year. To test the hypothesis of interest, we then compute

\[
PW_1 = \frac{\ln(7.3) - \ln(6)}{\sqrt{1/73}} = 1.68.
\]

The \( p \)-value of \( PW_1 \) for testing that the rate of seizures per year is equal to 6 is greater than 0.05 and just less than 0.10, as seen in Table I. If conducting a level 0.05 test, we would fail to reject \( H_0 \). The actual \( p \)-value associated with the test as calculated from a software package is 0.093.

**18.5 COMPARING POISSON RATES FROM TWO INDEPENDENT SAMPLES**

In clinical studies, it may be of interest to compare the rate of an event between two populations. For example, suppose we were investigating the ability of a new epilepsy treatment aimed at reducing the rate of seizures in patients. In this case, we might randomize patients to the new treatment or to a placebo, follow the patients for a given period of time, and then compare the observed rates of seizures among treated and untreated patients. If we denote \( \lambda_{Trt} \) and \( \lambda_{Plac} \) as the true rate of seizures among treated and untreated patients, we would likely be interested in testing the hypothesis

\[
H_0: \lambda_{Trt} = \lambda_{Plac} \text{ versus } H_1: \lambda_{Trt} \neq \lambda_{Plac},
\]
or equivalently,

\[ H_0 : \frac{\lambda_{Trt}}{\lambda_{Plac}} = 1 \text{ versus } H_1 : \frac{\lambda_{Trt}}{\lambda_{Plac}} \neq 1. \]  

(18.14)

Recalling from the properties of logarithms that \( \ln(a/b) = \ln(a) - \ln(b) \), we then have that the hypotheses given in Eq. (18.13) are equivalent to

\[ H_0 : \ln(\lambda_{Trt}) - \ln(\lambda_{Plac}) = 0 \text{ versus } H_1 : \ln(\lambda_{Trt}) - \ln(\lambda_{Plac}) \neq 0. \]  

(18.15)

Now, suppose that we observe \( Y_{Trt} \) and \( Y_{Plac} \) seizures among the treated and untreated patients, where treated patients were followed for \( A_{Trt} \) weeks and untreated patients were followed for a total of \( A_{Plac} \) weeks. Then we would estimate the rate of seizures per week in each group, \( \lambda_{Trt} \) and \( \lambda_{Plac} \), by \( l_{Trt} = Y_{Trt} / A_{Trt} \) and \( l_{Plac} = Y_{Plac} / A_{Plac} \), respectively. From Section 18.1, we know that

\[ \ln(l_{Trt}) = \ln \left( \frac{Y_{Trt}}{A_{Trt}} \right) \sim N \left( \ln(\lambda_{Trt}), \frac{1}{\lambda_{Trt}A_{Trt}} \right), \]  

(18.16)

and

\[ \ln(l_{Plac}) = \ln \left( \frac{Y_{Plac}}{A_{Plac}} \right) \sim N \left( \ln(\lambda_{Plac}), \frac{1}{\lambda_{Plac}A_{Plac}} \right). \]  

(18.17)

Now, because the two groups are independent, we can then approximate the distribution of \( \ln(\lambda_{Trt}) - \ln(\lambda_{Plac}) \) as

\[ \ln(l_{Trt}) - \ln(l_{Plac}) \sim N \left( \ln(\lambda_{Trt}) - \ln(\lambda_{Plac}), \frac{1}{\lambda_{Trt}A_{Trt}} + \frac{1}{\lambda_{Plac}A_{Plac}} \right). \]  

(18.18)

From this, we can test the null hypothesis in Eq. (18.15) for formulating the statistic

\[ \frac{\ln(l_{Trt}) - \ln(l_{Plac})}{\sqrt{\left(1/\lambda_{Trt}A_{Trt}\right) + \left(1/\lambda_{Plac}A_{Plac}\right)}} \sim N(0, 1). \]  

(18.19)

As before, we do not know the true values of \( \lambda_{Trt} \) and \( \lambda_{Plac} \), so we estimate them by \( l_{Trt} = Y_{Trt} / A_{Trt} \) and \( l_{Plac} = Y_{Plac} / A_{Plac} \), respectively. This yields the two-sample Poisson Wald statistic.

The two-sample Poisson Wald statistic (denoted here as \( PW_2 \)) replaces \( \lambda_1 \) and \( \lambda_2 \) by their estimates, \( l_1 = Y_1 / A_1 \) and \( l_2 = Y_2 / A_2 \), and tests \( H_0 : \lambda_1 = \lambda_2 \) versus \( H_1 : \lambda_1 \neq \lambda_2 \) by

\[ PW_2 = \frac{\ln(l_1) - \ln(l_2)}{\sqrt{\left(1/l_1A_1\right) + \left(1/l_2A_2\right)}} \sim N(0, 1). \]  

(18.20)
EXAMPLE: TESTING THE RATE OF SEIZURES AMONG TREATED AND CONTROL EPILEPTIC PATIENTS

Consider the results from a randomized clinical trial designed to test the efficacy of a new therapy for reducing the rate of seizures among epileptic patients. Specifically, \(N = 31\) patients were randomized to the experimental treatment and \(N = 28\) patients were randomized to placebo. All patients were treated for 8 weeks, and the total number of seizures each patient experienced between weeks 6 and 8 was recorded. Thus each patient was observed for seizures for 2 weeks so that \(A_{\text{T}} = 2 \times 31 = 62\) weeks and \(A_{\text{Pl}} = 2 \times 28 = 56\) weeks. Further, suppose that \(Y_{\text{T}} = 208\) seizures were observed in the treated group and \(Y_{\text{Pl}} = 223\) seizures were observed in the placebo group. Based upon these data, we estimate that the rate of seizures per week among treated patients is \(l_{\text{T}} = \frac{Y_{\text{T}}}{A_{\text{T}}} = \frac{208}{62} = 3.36\); and that the rate of seizures among placebo patients is \(l_{\text{Pl}} = \frac{Y_{\text{Pl}}}{A_{\text{Pl}}} = \frac{223}{56} = 3.98\).

Consider testing the null hypothesis that the true rate of seizures among treated and control patients is equal versus the alternative hypothesis that the two rates differ. Thus we wish to test

\[ H_0: \lambda_{\text{T}} = \lambda_{\text{Pl}} \text{ versus } H_1: \lambda_{\text{T}} \neq \lambda_{\text{Pl}}, \quad (18.21) \]

or equivalently,

\[ H_0: \ln(\lambda_{\text{T}}) - \ln(\lambda_{\text{Pl}}) = 0 \text{ versus } H_1: \ln(\lambda_{\text{T}}) - \ln(\lambda_{\text{Pl}}) \neq 0. \quad (18.22) \]

To test the hypothesis of interest we then compute

\[ PW_2 = \frac{\ln(3.36) - \ln(3.98)}{\sqrt{(1/3.36 \times 62) + (1/3.98 \times 56)}} = -1.76. \quad (18.23) \]

The \(p\)-value associated with \(PW_2\) for testing that the rate of seizures differs between treatment and placebo is greater than 0.05 and just less than 0.10, as seen in Table I. If conducting a level 0.05 test, we would fail to reject \(H_0\). The actual \(p\)-value associated with the test as calculated from a software package is 0.079.

18.6 THE SIMPLE POISSON REGRESSION MODEL

We first consider formulating a regression model to estimate and test the association between a count outcome variable and a binary independent variable. Specifically, we will consider the case of comparing the rate of seizures between two treatment groups (\(TRT = 1\) for treated patients and \(TRT = 0\) for untreated patients). The simple Poisson regression model is given by

\[ \ln(\lambda) = \beta_0 + \beta_1 x \quad (18.24) \]
where \( \lambda \) is the true rate (per 1 unit of exposure time) when \( TRT = x \), with \( TRT = 1 \) indicating that treatment was given and \( TRT = 0 \) indicating that placebo was given. Note that the response in model (18.24) is the natural logarithm of the rate of the count response for the specified value of \( x \). However, it is commonly written in terms of the mean of the observed count by recalling that \( \lambda = \mu / A \), so that (18.24) can be rewritten as

\[
\ln(\lambda) = \ln\left(\frac{\mu}{A}\right) = \ln(\mu) - \ln(A) = \beta_0 + \beta_1 x, \tag{18.25}
\]

or equivalently

\[
\ln(\mu) = \beta_0 + \beta_1 x + \ln(A). \tag{18.26}
\]

It is worth noting that the term \( \ln(A) \) in (18.26) does not have a coefficient in front of it. As such, it is commonly referred to as an offset term.

To be a useful model for our purposes, it is necessary that the simple Poisson regression model be interpretable to us in a scientific sense. That is, the coefficients in the regression model (\( \beta_0 \) and \( \beta_1 \)) should have meaning to us. A primary benefit of the Poisson regression model is that one can easily interpret the regression coefficients in terms of the rate of the response. To see this, consider “plugging in” a value of \( TRT = 0 \) (indicating a subject treated with placebo) into (18.24). Then we have

\[
\ln(\lambda_{\text{Plac}}) = \beta_0, \text{ or equivalently } e^{\beta_0} = \lambda_{\text{Plac}}. \tag{18.27}
\]

Thus \( e^{\beta_0} \) in the simple Poisson regression model represents the rate of seizures when the independent variable, \( TRT \), takes on a value of 0. To obtain the rate of seizures among treated patients, we would plug in \( TRT = 1 \), yielding

\[
\ln(\lambda_{\text{Trr}}) = \beta_0 + \beta_1, \text{ or equivalently } e^{\beta_0 + \beta_1} = \lambda_{\text{Trr}}. \tag{18.28}
\]

A nice feature of the Poisson regression model is that we can also make a relative comparison of the rates between two groups. From (18.24) we have that

\[
\ln\left(\frac{\lambda_{\text{Trr}}}{\lambda_{\text{Plac}}}\right) = \ln(\lambda_{\text{Trr}}) - \ln(\lambda_{\text{Plac}}) = \beta_0 + \beta_1 - \beta_0 = \beta_1 \tag{18.29}
\]

Exponentiating both sides implies that the relative rate ratio comparing treated patients (\( TRT = 1 \)) to untreated patients (\( TRT = 0 \)) is equal to \( e^{\beta_1} \). Therefore to test the null hypotheses stated in (18.15), we could alternatively test the hypothesis

\[
H_0: e^{(\beta_1)} = 1, \text{ or equivalently } H_0: \beta_1 = 0. \tag{18.30}
\]
EXAMPLE: COMPARISON OF THE RATE OF WORD REPETITIONS AMONG COGNITIVELY NORMAL AND DEMENTED PATIENTS

A clinical diagnosis of dementia is generally performed by a trained neuropsychologist and is based upon a collection of standardized neuropsychological tests as well as subjective complaints from the patient, family, and caregivers. One easy-to-administer neuropsychological test that is routinely given to subjects is the Controlled Oral Word Association Test (COWAT)-FAS test in which patients are asked to list as many different words starting with the letters F, A, or S in a 60 second time period. After completion of the test the total number of words is recorded along with the number of times repetitions occurred (the subject stated a word that was previously stated) and the number of errors, implying that a stated word does not begin with the letters F, A, or S. The COWAT-FAS test is thought to measure both the language and executive functioning domains of the brain, thereby discriminating between cognitively normal and impaired individuals.

The Alzheimer’s Disease Research Center (ADRC) at the University of California Irvine (UCI) campus routinely follows a cohort of volunteers overtime and performs annual neuropsychological testing on the participants with the hopes of better understanding the early predictors and consequences of cognitive impairment, dementia, and Alzheimer’s disease. Participants range from cognitively normal to clinically diagnosed dementia.

The COWAT-FAS test is regularly performed on participants in the UCI ADRC. Table 18.1 displays data for one visit on each of \( N = 182 \) participants comprised of 80 cognitively normal participants, 71 participants diagnosed with mild cognitive impairment (MCI), and 31 participants diagnosed with dementia. These data are contained in DB16. Variables listed below are each participant’s unique ID number (ID), total number of repeated words and words stated on the COWAT-FAS (FAS.TR and FAS.TS, respectively), the dementia diagnosis category for the participant (DEM.CAT, recorded as 0 = normal, 1 = MCI, and 2 = dementia), total years of education (EDUC), age in years (AGE), and the sex of the participant (SEX, recorded as 0 for males and 1 for females).

<table>
<thead>
<tr>
<th>ID</th>
<th>FAS.TR</th>
<th>FAS.TS</th>
<th>DEM.CAT</th>
<th>EDUC</th>
<th>AGE</th>
<th>SEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>37</td>
<td>0</td>
<td>16</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>48</td>
<td>0</td>
<td>12</td>
<td>86</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>62</td>
<td>0</td>
<td>14</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>61</td>
<td>0</td>
<td>18</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>42</td>
<td>0</td>
<td>16</td>
<td>77</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>47</td>
<td>0</td>
<td>18</td>
<td>85</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 18.1 First six rows from the University of California Irvine Alzheimer’s Disease Research Center Controlled Oral Word Association Test-FAS data set.
One comparison of interest is to determine if the rate of repeated words differs between cognitively normal subjects and subjects with dementia. In this case, the rate would be relative to the total number of words stated. That is, we wish to know if the expected number of words repeated per word given differs between normal participants and demented participants. In this example, the count \( Y \) is given by \( FAS.TR \) and the denominator for standardizing the rate, \( A \), is given by the total number of words given, \( FAS.TS \). To make the comparison of interest, we could fit a simple Poisson regression model of the form

\[
\ln(\lambda) = \beta_0 + \beta_1 I_{\text{DEMCAT}=2},
\]

or equivalently

\[
\ln(\mu) = \beta_0 + \beta_1 I_{\text{DEMCAT}=2} + \ln(FAS.TS).
\]

where \( \lambda \) denotes the rate of repeated words per word stated (i.e., \( \lambda = \mu / FAS.TS \), the true mean of the number of repeats divided by the total number of words stated, \( FAS.TS \)) and \( I_{\text{DEMCAT}=2} \) is a binary indicator for whether a participant has dementia (vs normal cognitive status). Using standard statistical software, the estimates of \( \beta_0 \) and \( \beta_1 \) based upon the ADRC data are given by \( b_0 = -3.03 \) and \( b_1 = 0.93 \), respectively. Given that \( I_{\text{DEMCAT}=2} = 0 \) implies cognitively normal subjects, the true rate of repeats per word given among cognitively normal participants is given by \( \lambda_{\text{Norm}} = e^{b_0} \), and hence our estimate based upon the observed data \( l_{\text{Norm}} = e^{b_0} = e^{-3.03} = 0.048 \). Thus we estimate that among cognitively normal participants the expected number of repeats per word stated is approximately 0.048. As previously noted, it is often useful to scale the rate for interpretability. In this case, we could scale the rate relative to 50 words (near the average number of words stated in the sample) and yield an estimate of \( 50 \times 0.048 = 2.40 \) repeats for every 50 words stated. For an estimate of the repeat rate among demented participants, we must plug in a value of 1 for \( I_{\text{DEMCAT}=2} \). In this case the rate is \( \lambda_{\text{Dem}} = e^{b_0 + b_1} \), and hence our estimate based upon the observed data \( l_{\text{Dem}} = e^{b_0 + b_1} = e^{-3.03 + 0.93} = 0.122 \). Thus we estimate that among demented participants the expected number of repeats per word stated is approximately 0.122. Scaling this to 50 words we estimate that the expected number of repeats per 50 words stated is \( 50 \times 0.122 = 6.12 \). It is clear that the rate of repeats among demented participants is higher than that of normal subjects. To quantify this, we can use the ratio of rates,

\[
\ln\left( \frac{\lambda_{\text{Dem}}}{\lambda_{\text{Norm}}} \right) = \frac{e^{b_0 + b_1}}{e^{b_0}} = e^{b_1}.
\]

Thus our estimate of the rate ratio comparing demented participants to normal participants is \( e^{b_1} = e^{0.93} = 2.54 \). In other words, we estimate that for a fixed set of stated words, demented participants will state approximately 2.54 times more repeats than normal participants.
Exercise 18.2
In the UCI ADRC data set, there were a total of $A_{\text{Norm}} = 3887$ words stated across all $N = 80$ normal participants. There were a total of $A_{\text{Dem}} = 852$ words stated across all $N = 31$ demented participants. Further, there were $Y_{\text{Norm}} = 187$ and $Y_{\text{Dem}} = 104$ repeats observed among normal and demented participants, respectively. Using these observed statistics, estimate the rate of repeats among normal and demented participants, as well as the rate ratio comparing the two populations. Conclude that these estimates are the same as those obtained from the simple Poisson regression model fit given in the previous example.

18.7 MULTIPLE POISSON REGRESSION: MODEL SPECIFICATION AND INTERPRETATION

In the previous section, we saw that the rate and rate ratio estimates that we obtain from the simple Poisson regression model with a single binary predictor are the same as those found using the two-sample comparison techniques shown in Section 18.5. Multiple Poisson regression extends the simple Poisson regression model to allow for the incorporation of continuous independent variables and to allow for the adjustment of other variables (analogous to the multiple linear regression model for continuous outcomes and the multiple logistic regression model for binary outcomes).

We first begin by considering the interpretation of the Poisson regression model parameters when a continuous predictor is included in the model. Suppose now that $X$ is a continuous predictor like age and consider formulating the Poisson regression model given by

$$\ln(\lambda) = \beta_0 + \beta_1 x$$

where $\lambda$ is the rate of the outcome of interest when $X = x$. As before, we can interpret the intercept in this model, $\beta_0$, by plugging in a value of $X = 0$ into model. In this case, we have

$$\ln(\lambda) = \beta_0$$

so that $\lambda = e^{\beta_0}$ is the rate of the outcome among a population with $X = 0$. If $X$ denotes age and we are considering the outcome of the rate of repetitions as in the previous cognitive testing example, we can see that the intercept really has no real-world interpretation in this case since the exponentiated intercept represents the rate of repetitions among participants with an age of 0 years. Of course, there are no such individuals in our sample. However, there is still meaning to $\beta_1$, the coefficient associated with $X$. Specifically, $e^{\beta_1}$ represents the rate ratio comparing two subpopulations differing in $X$ by 1 unit. Thus in the case where $X$ denotes age in the cognitive testing
data, $e^{\beta_1}$ would be the relative difference in the rate of repetitions comparing subpopulations differing in age by 1 year.

The other major benefit of the multiple Poisson regression model is the ability to adjust for other variables when estimating the association between a predictor of interest and a count outcome variable. For example, in the case of the cognitive testing data where we wished to compare the rate of repetitions between normal and demented participants, it would be desirable to adjust for age and education since both of these variables have been shown to relate to memory and functioning and hence may represent potential confounders in the relationship between cognitive diagnosis and the rate of repetitions. In addition, our data set also includes individuals diagnosed with MCI. The multiple regression model allows us to simultaneously include them in the comparison by simply adding another indicator for this group.

A Poisson regression model comparing normal, MCI, and demented participants while adjusting for age and years of education would take the form

$$\ln(\lambda) = \beta_0 + \beta_1 I_{IDEMCAT=1} + \beta_2 I_{IDEMCAT=2} + \beta_3 AGE + \beta_4 EDUC$$

or equivalently

$$\ln(\mu) = \beta_0 + \beta_1 I_{IDEMCAT=1} + \beta_2 I_{IDEMCAT=2} + \beta_3 AGE + \beta_4 EDUC + \ln(FAS.TS).$$

where $\lambda$ denotes the rate of repeats per word stated (i.e., $\lambda = \mu / FAS.TS$, the true mean of the number of repeats divided by the total number of words stated, $FAS.TS$), and $I_{IDEMCAT=1}$ is a binary indicator of whether a participant has MCI, $I_{IDEMCAT=2}$ is a binary indicator whether a participant has dementia, $AGE$ is the age of the participant in years, and $EDUC$ is the number of years of education. The regression given in (18.36) now allows us to compare the relative difference in the rate of repeats between cognitive status groups who are of a similar age and education level. To see this, suppose that we wished to compare the rate of repeats between MCI and cognitively normal participants that are 62 years of age with 12 years of education. In this case, we would have

$$\ln \left( \frac{\lambda_{MCI, 62,12}}{\lambda_{Norm, 62,12}} \right) = \ln(\lambda_{MCI, 62,12}) - \ln(\lambda_{Norm, 62,12})$$

$$= \beta_0 + \beta_1 + 62 \times \beta_3 + 12 \times \beta_4 - (\beta_0 + 62 \times \beta_3 + 12 \times \beta_4) = \beta_1.$$

Exponentiating both sides as we did before implies that the rate ratio for repeats comparing MCI participants to cognitively normal participants, aged 62 years with 12 years of education, is equal to $e^{\beta_1}$. Although the above mentioned
example considered a specific age and education level (62 and 12 years, respectively), it is easy to see that we could have used any value and it would have canceled in the difference. The implication of this is that $e^{\beta_1}$ actually represents the rate ratio for repeats comparing MCI participants to cognitively normal participants of the same age and education level (regardless of what that age and education level is).

In general, we could consider an arbitrary number of adjustment variables giving rise to a Poisson regression model of the form

$$\ln(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

or equivalently

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \ln(A).$$

Then following the same logic used in the previous Alzheimer’s disease example we can interpret the coefficients of the Poisson regression model in terms of the rate of events associated with each predictor.

**INTERPRETATION OF POISSON REGRESSION PARAMETERS**

For the Poisson regression model

$$\ln(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

$e^{\beta_1}$ is interpretable as the rate ratio comparing two subpopulations differing in $x_1$ by 1 unit that have similar values for $x_2, \ldots, x_k$.

**Exercise 18.3**

Using the UCI ADRC data set, a Poisson regression model fit of (18.36) results in the following estimates

$$\ln(l) = -3.48 + 0.30 \times I_{\text{DEMCAT}=1} + 0.86 \times I_{\text{DEMCAT}=2} + 0.02 \times \text{AGE} - 0.05 \times \text{EDUC}.$$  

From this, answer the following:

a. What is the estimated rate ratio comparing MCI ($\text{DEMCAT}=1$) to cognitively normal ($\text{DEMCAT}=0$) participants of a similar age and education level?

b. What is the estimated rate ratio comparing demented ($\text{DEMCAT}=2$) to cognitively normal ($\text{DEMCAT}=0$) participants of a similar age and education level?

c. What is the estimated rate ratio comparing demented ($\text{DEMCAT}=2$) to MCI ($\text{DEMCAT}=1$) participants of a similar age and education level?
What is the estimated rate ratio comparing participants differing in age by 1 year who have a similar cognitive status and the same number of years of education?

18.8 OBTAINING PREDICTED RATES

As with the linear and logistic regression model, the Poisson regression model can also be used to obtain predicted rates for a given set of independent covariate values. In order to do this, we must solve for $\lambda$ in Eq. (18.41). This is done by exponentiating both sides of (18.41) as follows:

$$\ln(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$$\Rightarrow \lambda = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k} \quad (18.43)$$

As we have previously noted, $\lambda$ represents the rate per 1 unit of how observation time was measured (4). In many cases, it is desirable to scale the rate over a length of observation time that is clinically and scientifically relevant. For example, if we wish to predict the average count over 10 units of observation time we would multiply $\lambda$ by 10.

**EXAMPLE: PREDICTING THE RATE OF REPETITIONS IN NEUROPSYCHOLOGICAL TESTING**

For the UCI ADRC cognitive testing example, we can use the abovementioned formula to estimate the rate of repeats for a given cognitive diagnosis, age, and education group. Recall from (18.42) the estimated model for the rate of repetitions was

$$\ln(l) = -3.48 + 0.30 \times I_{\text{DEM Cat}=1} + 0.86 \times I_{\text{DEM Cat}=2} + 0.02 \times \text{AGE} - 0.05 \times \text{EDUC}. \quad (18.45)$$

This implies that the rate of repetitions, as determined by cognitive status, age, and education level, is estimated to be

$$l = e^{-3.48+0.30 \times I_{\text{DEM Cat}=1}+0.86 \times I_{\text{DEM Cat}=2}+0.02 \times \text{AGE}-0.05 \times \text{EDUC}}. \quad (18.46)$$
Now consider the estimated rate of repetitions for a randomly sampled MCI patient who is 67 years old and went through 16 years of education. Plugging these values into Eq. (18.46) we have

\[ l = e^{-3.48 + 0.30 \times 67 + 0.02 \times 16 + 0.05 \times 0.071} = 0.071, \]  

so that the predicted rate of repetitions for a 67-year-old MCI participant with 16 years of education is 0.071 repetitions per word stated. As previously noted, it is often useful to scale the rate for interpretability. We could scale the rate relative to 50 words (near the average number of words stated in the sample), yielding an estimate of \( 50 \times 0.071 = 3.57 \) repeats for every 50 words stated.

**Exercise 18.4**

Using the model estimates given in Eq. (18.46), find the predicted rate of repeats of CHD for a 67-year-old cognitively normal participant with 16 years of education, and a 72-year-old demented participant with 14 years of education. In each case, scale your answers to an observation time of 50 stated words.

### 18.9 Inference for Association Parameters

#### Test of association

In Section 18.7, we saw that for the Poisson regression model specified in Eq. (18.36), \( e^{\beta_1} \) represents the rate ratio for repeats comparing MCI participants to cognitively normal participants of the same age and education level. Thus if \( e^{\beta_1} = 1 \), or equivalently if \( \beta_1 = 0 \), then this would imply that the rate of repeats is the same comparing MCI participants to cognitively normal participants of the same age and education. This would imply that no association exists between cognitive status and the rate of repeats among normal and MCI participants. Conversely, if \( \beta_1 \neq 0 \) then this would imply that the rate of repeats differs between MCI and cognitively normal participants of the same age and education level. Thus it may be of interest to test the null hypothesis \( H_0 : \beta_1 = 0 \) versus the alternative hypothesis \( H_1 : \beta_1 \neq 0 \). In order to carry out this hypothesis test, we require a test statistic for which we know its distribution under the null hypothesis, \( H_0 \).

Consider the general Poisson regression model

\[ \ln(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \]  

Statistical software can be used to obtain estimates of the parameters \( \beta_0, \beta_1, \ldots, \beta_k \). Let \( b_0, b_1, \ldots, b_k \) denote these estimates. Further, software can be used to estimate the standard error of \( b_0, b_1, \ldots, b_k \), which we denote as \( se(b_0), se(b_1), \ldots, se(b_k) \).
TESTS FOR A SINGLE POISSON REGRESSION PARAMETER

A test of the null hypothesis $H_0 : \beta_j = 0$ versus the alternative hypothesis $H_1 : \beta_j \neq 0$, $j = 0, \ldots, k$, can be conducted with the following test statistic

$$z = \frac{b_j}{se(b_j)} \sim N(0, 1).$$  \hspace{1cm} (18.49)

The value of $z$ can be looked up in Table I and the resulting $p$-value is the probability of observing a test statistic as big or bigger in magnitude than the one that we did if the null hypothesis were true.

EXAMPLE: TESTING THE ASSOCIATION BETWEEN COGNITIVE STATUS AND THE RATE OF REPEATS IN NEUROPSYCHOLOGICAL TESTING

Returning to the UCI ADRC study, we obtain from Table 18.2 Poisson regression estimates for the model, including cognitive status, age, and education.

From the above table, we can see that the $z$ statistic for testing whether MCI and cognitively normal participants differ with respect to the rate of repeats is given by

$$z = \frac{0.30}{0.15593} = 1.95.$$  

The value of $z$ for testing that the rate of repeats comparing MCI participants to cognitively normal participants is just below 1.96. If conducting a level 0.05 test, we may say that the $p$-value is on the edge of statistical significance since it is just below 0.05, as seen in Table I. The actual $p$-value associated with the test as calculated from a software package is 0.051. Note that this does not imply that an association does not exist. It simply reflects the precision we have for our estimate. Despite this result being on the edge of statistical significance, one should still consider the magnitude or clinical significance of the estimated association, reporting this, the $p$-value, and a corresponding confidence interval (discussed next) together.

Table 18.2 Poisson regression model estimates for modeling the rate of repeats in the Controlled Oral Word Association Test-FAS test using the University of California Irvine Alzheimer’s Disease Research Center data.

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>$se(b)$</th>
<th>$z$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-3.48</td>
<td>0.72280</td>
<td>-4.81</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>$I_{DEMCAT=1}$ (MCI)</td>
<td>0.30</td>
<td>0.15593</td>
<td>1.95</td>
<td>0.051</td>
</tr>
<tr>
<td>$I_{DEMCAT=2}$ (DEMENTIA)</td>
<td>0.86</td>
<td>0.18727</td>
<td>4.58</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>AGE</td>
<td>0.02</td>
<td>0.00798</td>
<td>2.11</td>
<td>See Ex</td>
</tr>
<tr>
<td>EDUC</td>
<td>-0.05</td>
<td>0.02473</td>
<td>-2.04</td>
<td>0.041</td>
</tr>
</tbody>
</table>
Exercise 18.5
Using the model estimates given in Table 18.2, conduct a level 0.05 test of the null hypothesis that the rate of repeats comparing demented participants to cognitively normal participants is the same after adjustment for age and education level.

Confidence intervals for the rate ratio
The test statistic given in Eq. (18.49) provides us with a binary decision regarding the association between a predictor of interest and the rate of the response. However, it is also important to quantify our uncertainty in the rate ratio associated with the predictor. A confidence interval can be formulated from the estimated regression parameter, the estimated standard error, and the appropriate quantile from a standard normal distribution.

CONFIDENCE INTERVALS FOR THE RATE RATIO
A \((1 - \alpha)\)% confidence interval for the rate ratio associated with predictor \(x_j\) is given by

\[
\left( e^{b_j - z_{1-\alpha/2} \times se(b_j)}, e^{b_j + z_{1-\alpha/2} \times se(b_j)} \right)
\]

(18.50)

where \(z_{1-\alpha/2}\) is the \(1 - \alpha/2\) quantile of the standard normal distribution. This value is 1.96 for a 95\% confidence interval \((\alpha = 0.05)\).

EXAMPLE: CONFIDENCE INTERVAL FOR THE RATE RATIO COMPARING MILD COGNITIVE IMPAIRMENT TO COGNITIVELY NORMAL PATIENTS IN NEUROPSYCHOLOGICAL TESTING
Using the output in Table 18.2, the estimated rate ratio comparing MCI subjects to cognitively normal subjects (after adjustment for age and education) is equal to \(e^{0.30} = 1.35\). A 95\% confidence interval for the rate ratio is

\[
\left( e^{0.30 - 1.96 \times 0.156}, e^{0.30 + 1.96 \times 0.156} \right) = (0.99, 1.83).
\]

(18.51)

Exercise 18.6
Using the model estimates given in Table 18.2, compute a 95\% confidence interval for the rate ratio comparing subjects with dementia to cognitively normal subjects, after adjustment for both age and education.

REFERENCE
Analysis of censored time-to-event data

19.1 SURVIVAL CONCEPTS

A broad use of the term “survival”

“Survival” as a term is used more broadly here than just “nondying.” It includes success of a treatment (nonrecurrence of disease), success of an orthopedic device (nonfailure), satisfaction of a patient with treatment (nondissatisfaction), correct diagnosis by a radiologist (nonerror), and so forth. Indeed, the logistic regression methods in Section 17.4 can be used to try to predict the outcome of any binary-dependent variable.

Survival versus failure to survive

Survival may be investigated as a binary variable: the subject survives or not usually recorded in a data sheet as 1 or 0, respectively. Association of binary survival may be assessed by contingency tests (Fisher’s exact or $\chi^2$) for categorical variables (see Chapter 9: Tests on categorical data) by the rank-sum test for rank-order variables (see Chapter 11: Tests of location with continuous outcomes) or by the $t$ test for continuous variables (see Chapter 11: Tests of location with continuous outcomes). The prediction of binary survival by potentially associated variables is assessed as logistic regression, addressed in Section 17.4.

Time-dependent survival

Survival may be investigated as a time-dependent variable: the subject has survived for a certain length of time. Time-dependent data are common in medicine. Many monitors in surgery, including a variety of cardiac and pulmonary measures, are time-dependent data. Cardiac stress measures are time-dependent data. Electroencephalograms are time-dependent data. Many epidemiological measures can be time dependent, including survival time per se. Methods for estimating survival times, typically quantiles of the survival distribution, are addressed in Section 19.3, along with tables and graphs as display formats. Methods for testing survival time estimates, such as testing for possible change
CHAPTER 19 Analysis of censored time-to-event data

from a historic mean survival time or comparing survival times from two or more treatments, are addressed in Section 19.4. The association of survival time with multiple variables is assessed as Cox proportional hazards (PH) regression, addressed in Section 19.5.

Time-series and survival

A large number of observations recorded through time on the same variable are usually called a time-series in statistics. Survival time data sets usually have fewer observations than a true time-series, but occasionally survival can be addressed using time-series methods. However, time-series methods usually address other types of issues and therefore are addressed separately in Chapter 20, Analysis of repeated continuous measurements over time.

19.2 CENSORING

Most survival time data include some patients who are lost to follow up before death. Their data are designated censored. We do not have a survival time for these cases because they have not yet experienced an event when observation stops, but we do not want to lose the information that they did survive until lost. Therefore we include them while they are known to be alive and remove them from the database when they are lost. Of course, removing them at that point implies that we adjust the baseline number as well to obtain proper proportions of survival after censoring.

When considering survival as a time-dependent variable, it is often of interest to model the time from a fixed or designated origin to the occurrence of some event of interest (e.g., death or progression of disease). Censoring is a commonly encountered problem in the analysis of time-to-event data. Generally speaking, censoring occurs when the time to the event of interest is not known exactly but is only known to have occurred within a broader time interval. Survival analysis methods allow us to include censored patients in our analysis. This is a crucial component in most survival analysis methods.

Multiple types of censoring may exist. Here, we briefly define the most frequently encountered forms of censoring. For further discussion on the forms of censoring that can occur, we refer the reader to the text of Klein and Moeschberger.\(^1\) Perhaps, the most common form of censoring that is encountered in practice is right censoring. This form of censoring describes the situation where subjects enter into a study at some defined origin and who are at risk for the event of interest at the time of their entry. A subject may then become right censored if they are lost to follow up prior to the event of interest having occurred. Thus a subject may become right censored for a variety of reasons, including the study follow-up ending at some
prespecified time point that occurs prior to the subject’s true event time (this is commonly termed administrative right censoring), the subject choosing to discontinue follow-up prior to the event occurring for some reason (commonly termed random right censoring), or another event occurring that precludes the observation of the event of interest (termed a competing risk). Most survival methods, including those discussed throughout the remainder of this chapter, make an assumption that the censoring time is independent of the true survival time for all subjects (conditional upon adjusted covariates that may be collected on the subjects). This critical assumption is referred to as the noninformative censoring assumption. In the case of administrative right censoring, noninformative censoring is generally a reasonable assumption, since the end of follow-up is prespecified. However, for random censoring or in the presence of competing risks, careful consideration for the possibility of informative censoring must be made. In the presence of informative censoring, most standard survival analysis methods, including those discussed in this chapter, can lead to biased and inconsistent parameter estimates.

Three additional types of censoring can occur. Left censoring occurs when we do not know the exact event time for subjects who experienced the event before a certain time; specifically, if the event happens before the start of the study, we may not know precisely when it happened. For example, suppose that patients of ages 65 and older are observed for the development of Alzheimer’s disease (AD). If they develop the disease before they turn 65 years old, we will only know that it happened before they turned 65.

Interval censoring occurs when we only know that an event occurred within a specific time interval. We again refer to AD studies for an example. Patients at the UC Irvine Alzheimer’s Disease Research Center are seen every year. Each year, they are given a series of tests to determine their cognitive status. If a patient is observed to have dementia, we do not know the exact time at which they developed the disease, only that it occurred sometime between the current visit and the previous year’s visit. Such data are said to be interval censored with the intervals of length specified, in this example approximately 1 year.

Finally, double censoring occurs when there is a combination of left and right censoring. As an example, double censoring could occur if a 3-year study was conducted in which patients were followed for the development of dementia. Patients could be enrolled regardless of whether or not they had dementia, but if they developed the disease before the start of the study we only know that it happened sometime before, and if they become afflicted by the disease after the study has ended, researchers only know that the event occurred at some point after the end of follow-up.

The remainder of the chapter will focus on statistical methods that assume noninformative right censoring.
19.3 SURVIVAL ESTIMATION: LIFE TABLE ESTIMATES AND KAPLAN–MEIER CURVES

Life table estimates of survival

Basic survival time data are times to death (or, more generally, failure) of members of a cohort (demographic group). It has been noted that “survival” is not restricted to remaining alive but may represent other terminal events, for example, surviving without the recurrence of cancer, but we shall use the remaining alive connotation for convenience. The proportion of a cohort’s survival to successive time points can be calculated from the raw data, providing the historic life table. If no patients are lost to follow up, the life table is simply the proportion surviving (surviving number divided by initial number) at each measured point in time. In most cases a life table includes other influences, such as censored data.

Data and calculations required for a life table

EXAMPLE POSED: WHAT IS THE SURVIVAL PATTERN OF DIABETICS

We pose time to death of 319 men and 370 women in Rochester, MN, having adult-onset diabetes mellitus who were older than 45 years at onset during 1980–90. We want to generate life tables and survival curves for these cohorts and compare survival patterns. (The example is continued through the following sections.)

A sample life table is given as Table 19.1 for the male cohort. Basic data for a life table on \( n \) patients can be described using the columns in which we enter data in the spreadsheet: Column 1 Interval: time intervals; Column 2 Begin: the number at the beginning of each time interval; Column 3 Died: the number dying in each time interval; and Column 4 Lost: the number lost to follow up in each time interval. The rest of the table is calculated from these basic data. The method used here is the simplest in that it assumes the time of death or loss to occur at the end of the time interval.

Table 19.1 Survival data of 319 men in Rochester, MN, having adult-onset diabetes mellitus who were older than 45 years at onset during 1980–90.

<table>
<thead>
<tr>
<th>Interval (years)</th>
<th>Begin</th>
<th>Died</th>
<th>Lost</th>
<th>End</th>
<th>S (survived)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (onset)</td>
<td>319</td>
<td>0</td>
<td>0</td>
<td>319</td>
<td>1.0000</td>
</tr>
<tr>
<td>&gt;0—2</td>
<td>319</td>
<td>16</td>
<td>0</td>
<td>303</td>
<td>0.9498</td>
</tr>
<tr>
<td>&gt;2—4</td>
<td>303</td>
<td>19</td>
<td>0</td>
<td>284</td>
<td>0.8902</td>
</tr>
<tr>
<td>&gt;4—6</td>
<td>284</td>
<td>19</td>
<td>0</td>
<td>265</td>
<td>0.8306</td>
</tr>
<tr>
<td>&gt;6—8</td>
<td>265</td>
<td>8</td>
<td>0</td>
<td>257</td>
<td>0.8055</td>
</tr>
<tr>
<td>&gt;8—10</td>
<td>257</td>
<td>0</td>
<td>2</td>
<td>255</td>
<td>0.7865</td>
</tr>
</tbody>
</table>
Other methods giving sophisticated adjustments exist. For convenience a second line is entered for a time interval having censored data. Calculation for a *Column 5 End:* entry is $\text{end} = \text{begin} - \text{died} - \text{lost}$. Calculation for *Column 6 S (survived):* it is the proportion surviving given by $S$ for last period $\times (\text{end for this period}/\text{end for last period})$. The reason for this comes from the Multiplicative Law of Probability, Eq. (3.4): the probability that both of two independent events occur is the probability that the first event occurs times the probability that the second event occurs, given the first has occurred. In this case the chance of surviving to the end of the current interval is the chance of surviving to the beginning of the interval multiplied by the chance of surviving during the interval, given survival to the beginning of the interval.

**Calculations for the life table for men with diabetes mellitus**

Table 19.1 provides basic and calculated data for the life table on the survival of 319 men. In the first period ($>0$–2 years), $\text{died} = 16$ and $\text{lost} = 0$, so $\text{end} = 319 - 16 - 0 = 303$. $S = 1.00 \times (303/319) = 0.9498$. In the second period, $S = 0.9498 \times (284/303) = 0.8902$. In the third period, $S = 0.8902 \times (265/284) = 0.8306$. In the fourth period, eight died and two were lost to follow up, which were separated on two lines, $\text{died}$ first. $S$ for that period is $0.8306 \times (257/265) = 0.8055$. At the end of that period, we subtracted the 2 $\text{lost}$, leaving 255. Since we assumed that they survived to the end of the period, they did not reduce the survival, but they are removed when calculating survival in the next period. At the end of 10 years, about 79% of the men remain alive. This may also be interpreted as, “the probability that a man randomly chosen at the outset will remain alive longer than 10 years is estimated to be 0.79.”

**Kaplan–Meier estimator of the survival curve**

The commonly used estimate and graphical display of survival information was developed by E.L. Kaplan and P. Meier in 1958. The Kaplan–Meier estimator can be thought of as an infinitesimal version of the life table estimator. Specifically, it is equivalent to shrinking the life table intervals so small that only one death occurs within each interval. It is more accurate than a life table and should be used when statistical software is available. Lacking that, or for the purpose of understanding the concepts, a simple mode of display is just to graph the survival data from the life table against the time intervals. A survival datum stays the same for the period of an interval, dropping at the end, which produces a stepped pattern. The survival curve for Table 19.1 is shown as Fig. 19.1. Note that the number lost to follow up (censored) is shown as a small integer over the line for the period in which they were lost, distinguishing those lost from those dying, and keeping the viewer informed as to the diminishing sample size.
Confidence intervals on survival estimates

The survival curve shows estimates of a population’s survival pattern based on the data from a sample. How confident of that estimate are we? As with many other estimates, we can find confidence intervals. In this case, there will be a confidence interval on each survival proportion that leads to confidence curves enclosing the survival curve. The confidence curves have a similar stepped appearance. Different methods exist to calculate the confidence intervals on survival proportions. The most general are evolved from a method originated by M. Greenwood in 1926, but these are difficult to calculate. They are good to use if statistical computer software offers the capability. Otherwise, a much simpler method by Peto et al. may be used, so long as the user remembers that they are rougher approximations. The pattern of a confidence expression is the same for confidence intervals on other types of estimates, namely, that given as Eq. (8.1). (The desired probability that the specified interval contains a population statistic yields the level of confidence interval is obtained by considering the distribution of the estimate of the population statistic.) It was stated in Section 8.6 that a sample proportion is distributed approximately normal for moderate to large samples.

CALCULATION OF CONFIDENCE INTERVAL ON A SURVIVAL ESTIMATE

If \( S_i \) denotes the estimate of the survival proportion at the end of interval \( i \) and \( \text{SEE} \) denotes the standard error of the estimate, the probability is \( 1 - \alpha \) that the true survival proportion is bracketed by \( S_i \pm z_{1-\alpha/2} \times \text{SEE} \). For 95% confidence, \( z_{1-\alpha/2} = 1.96 \). \( \text{SEE} \), as for proportions met in Section 8.6, depends only on \( S_i \) and the number at the beginning of each time period, let us say \( n_{i-1} \). \( n_{i-1} \) replaces \( \text{begin} \) in order to keep track of the interval involved. \( \text{SEE} \) is calculated as in the following equation:

(Continued)
19.3 Survival Estimation: Life Table Estimates and Kaplan–Meier Curves

(CONTINUED)

\[ SEE = \sqrt{\frac{S_i(1 - S_i)}{n_{i-1}}} \]  \hspace{1cm} (19.1)

The 95% confidence interval on the true survival proportion for each time period becomes

\[ S_i \pm 1.96 \times SEE = S_i \pm 1.96 \times \sqrt{\frac{S_i(1 - S_i)}{n_{i-1}}} \]  \hspace{1cm} (19.2)

Since these confidence intervals are not very exact, we have to use common sense to avoid letting them exceed 1 or fall below 0 in reporting or graphing them.

<table>
<thead>
<tr>
<th>Period (years)</th>
<th>Begin ((ni - 1))</th>
<th>(S_i)</th>
<th>SSE</th>
<th>95% Confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (outset)</td>
<td>319</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;0–2</td>
<td>319</td>
<td>0.9498</td>
<td>0.0119</td>
<td>0.9265</td>
</tr>
<tr>
<td>&gt;2–4</td>
<td>303</td>
<td>0.8902</td>
<td>0.0169</td>
<td>0.8571</td>
</tr>
<tr>
<td>&gt;4–6</td>
<td>284</td>
<td>0.8306</td>
<td>0.0203</td>
<td>0.7908</td>
</tr>
<tr>
<td>&gt;6–8</td>
<td>265</td>
<td>0.8055</td>
<td>0.0218</td>
<td>0.7628</td>
</tr>
<tr>
<td>&gt;8–10</td>
<td>255</td>
<td>0.7865</td>
<td>0.0228</td>
<td>0.7418</td>
</tr>
</tbody>
</table>

![Figure 19.2](image)

**Figure 19.2** Survival curve of Fig. 19.1 enclosed by the 95% confidence intervals given in Table 19.2.
CHAPTER 19  Analysis of censored time-to-event data

Table 19.3  Survival data of 274 women in Rochester, MN, having adult-onset diabetes mellitus who were older than 45 years at onset during 1970–80.

<table>
<thead>
<tr>
<th>Interval (years)</th>
<th>Begin ( (n_{i-1}) )</th>
<th>Died</th>
<th>Lost</th>
<th>End ( (n_i) )</th>
<th>( S_i ) (survived)</th>
<th>Confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (outset)</td>
<td>274</td>
<td>0</td>
<td>0</td>
<td>274</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>&gt;0–2</td>
<td>274</td>
<td>14</td>
<td>0</td>
<td>260</td>
<td>0.9222</td>
<td></td>
</tr>
<tr>
<td>&gt;2–4</td>
<td>260</td>
<td>13</td>
<td>0</td>
<td>247</td>
<td>0.8413</td>
<td></td>
</tr>
<tr>
<td>&gt;4–6</td>
<td>247</td>
<td>14</td>
<td>0</td>
<td>233</td>
<td>0.7708</td>
<td></td>
</tr>
<tr>
<td>&gt;6–8</td>
<td>233</td>
<td>18</td>
<td>0</td>
<td>215</td>
<td>0.7103</td>
<td></td>
</tr>
<tr>
<td>–</td>
<td>215</td>
<td>0</td>
<td>1</td>
<td>214</td>
<td>0.6611</td>
<td></td>
</tr>
<tr>
<td>&gt;8–10</td>
<td>214</td>
<td>19</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EXAMPLE: MEN WITH DIABETES MELLITUS

Table 19.2 shows confidence intervals calculated using Eq. (19.2) for the 319 diabetic men of Table 19.1. Fig. 19.2 shows the life table survival curve of Fig. 19.1 with the confidence intervals as calculated using Eq. (19.2) and shown in Table 19.2. The user may note that the confidence limits associated with a particular period in Table 19.2 are drawn about the succeeding period in Fig. 19.2. The reason is that the survival proportion is related to the end of the period and is maintained until the next death figure, at the end of the following period. If we had data giving the exact time of death rather than a period during which it occurred, this pictorial lag would not occur.

Exercise 19.1

Life table data for survival of diabetic women in period 1970–80 is shown in Table 19.3. Complete the life table. What is the estimate of probability that a woman with diabetes survives 10 years? Graph a survival curve. Calculate and graph the 95% confidence intervals on this survival curve.

19.4 SURVIVAL TESTING: THE LOG-RANK TEST

Testing the difference among survival curves

EXAMPLE POSED: SURVIVAL OF MEN VERSUS WOMEN WITH DIABETES

Fig. 19.3 superposes estimated survival curves for 319 men (Fig. 19.1 from the data of Table 19.1) and 370 women with diabetes mellitus during the 1980–90 decade (from the data of Table 19.4). We see a difference by inspection, but is there evidence for a true difference in the survival curves between men and women?
One statistical procedure that addresses the difference between two or more survival curves is the log-rank test. This test uses a chi-square statistic based on the difference between the observed survival and the survival that would be expected if the curves were not different, in the same way that a chi-square goodness-of-fit test [Eq. (13.7)] uses the sum of squares of weighted differences between the observed and expected curves. However, the log-rank test’s \( \chi^2 \) is more complicated to calculate. It uses matrix algebra, multiplying vectors of differences for the time periods and the matrix of variances and covariances. A statistical software package should be used for this calculation. The result of the calculation is a \( \chi^2 \) statistic that may be compared with a \( \chi^2 \) critical value from Table III (see Tables of probability distributions) for \( df = \) number of survival curves – 1. When two curves, as in Fig. 19.3, are

![Survival curves for 319 men and 370 women in Rochester, MN, having adult-onset diabetes mellitus who were older than 45 years at onset during the decade 1980–90.](image)

**Figure 19.3** Survival curves for 319 men and 370 women in Rochester, MN, having adult-onset diabetes mellitus who were older than 45 years at onset during the decade 1980–90.

**Table 19.4** Survival data of 370 women in Rochester, MN, having adult-onset diabetes mellitus who were older than 45 years at onset during 1980–90.

<table>
<thead>
<tr>
<th>Interval (years)</th>
<th>Begin</th>
<th>Died</th>
<th>Lost</th>
<th>End</th>
<th>( S ) (survival rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (outset)</td>
<td>370</td>
<td>0</td>
<td>0</td>
<td>370</td>
<td>1.0000</td>
</tr>
<tr>
<td>&gt;0–2</td>
<td>370</td>
<td>33</td>
<td>0</td>
<td>337</td>
<td>0.9108</td>
</tr>
<tr>
<td>&gt;2–4</td>
<td>337</td>
<td>20</td>
<td>0</td>
<td>317</td>
<td>0.8567</td>
</tr>
<tr>
<td></td>
<td>317</td>
<td>0</td>
<td>3</td>
<td>314</td>
<td></td>
</tr>
<tr>
<td>&gt;4–6</td>
<td>314</td>
<td>18</td>
<td>0</td>
<td>296</td>
<td>0.8076</td>
</tr>
<tr>
<td>&gt;6–8</td>
<td>296</td>
<td>22</td>
<td>0</td>
<td>274</td>
<td>0.7476</td>
</tr>
<tr>
<td></td>
<td>274</td>
<td>0</td>
<td>2</td>
<td>272</td>
<td></td>
</tr>
<tr>
<td>&gt;8–10</td>
<td>272</td>
<td>20</td>
<td>0</td>
<td>252</td>
<td>0.6926</td>
</tr>
</tbody>
</table>
The log-rank test compared to other tests of survival curves

Two alternative tests that might be considered for use are the Mantel–Haenszel test and the Cox PH test (Section 19.5). The Mantel–Haenszel test is almost the same as the log-rank test. Indeed, both Mantel and Haenszel contributed to the theory of the log-rank test. However, the Mantel–Haenszel test is restricted to two curves, whereas the log-rank test may use more than two. Therefore the log-rank test is recommended. The Cox PH test can be extended to allow for the adjustment of other covariates via a regression model. Cox's regression model is discussed further in Section 19.5.

EXAMPLE COMPLETED: SURVIVAL OF MEN VERSUS WOMEN WITH DIABETES

We use the log-rank test to compare survival for men and women diabetes patients. From Table III the critical $\chi^2$ for 1 df is 3.84. The log-rank test yields $\chi^2 = 7.2$, greater than 3.84. We conclude that men have a significantly better survival than women. (The actual log-rank p-value = 0.007.)

Figure 19.4 Kaplan–Meier survival curves of 44 patients with advanced cancer. Treatment 0 implies no treatment. Treatments 1 and 2 are experimental treatments of unknown efficacy.
ADDITIONAL EXAMPLE: TESTING THREE CANCER SURVIVAL CURVES

Fig. 19.4 shows survival curves for patients with advanced cancer simulated for classroom use by a radiation oncologist. In this example, treatment 0 implies no treatment was given to the patient and treatments 1 and 2 refer to experimental treatments of unknown efficacy. Because three curves are being compared in the log-rank test, \( df = 2 \). The critical \( \chi^2 \) for 2 \( df \) from Table III is 5.99. A software package yields \( \chi^2 = 6.35 \). Since the calculated \( \chi^2 \) is larger than the critical \( \chi^2 \), the null hypothesis of no difference is rejected. (The actual \( p \)-value = 0.042.) It appears that treatment 1 is worse for survival than no treatment and treatment 2 may be better than no treatment but the estimated survival curve for treatment group 2 is highly variable as indicated by the size of the steps in the estimated survival curve.

Exercise 19.2

The survival table for 274 diabetic women during the decade 1970–80 is given in Table 19.5. The equivalent data on 370 women for the decade 1980–90 appears in Table 19.4. A log-rank test of the difference between the two curves yields \( \chi^2 = 0.63 \). Find the critical value, compare the \( \chi^2 \) values, and interpret the resulting decision.

### Table 19.5

Completion of Table 19.3, survival data of 274 women in Rochester, MN, having adult-onset diabetes mellitus who were older than 45 years at onset during 1970–80.

<table>
<thead>
<tr>
<th>Interval (years)</th>
<th>Begin ( (n_{i-1}) )</th>
<th>Died</th>
<th>Lost</th>
<th>End ( (n_i) )</th>
<th>( S_i ) (survived)</th>
<th>Confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (outset)</td>
<td>274</td>
<td>0</td>
<td>0</td>
<td>274</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>&gt;0–2</td>
<td>274</td>
<td>14</td>
<td>0</td>
<td>260</td>
<td>0.9489</td>
<td>0.9228 0.9750</td>
</tr>
<tr>
<td>&gt;2–4</td>
<td>260</td>
<td>13</td>
<td>0</td>
<td>247</td>
<td>0.9015</td>
<td>0.8662 0.9368</td>
</tr>
<tr>
<td>&gt;4–6</td>
<td>247</td>
<td>14</td>
<td>0</td>
<td>233</td>
<td>0.8504</td>
<td>0.8082 0.8926</td>
</tr>
<tr>
<td>&gt;6–8</td>
<td>233</td>
<td>18</td>
<td>0</td>
<td>215</td>
<td>0.7847</td>
<td>0.7360 0.8334</td>
</tr>
<tr>
<td>&gt;8–10</td>
<td>215</td>
<td>19</td>
<td>0</td>
<td>195</td>
<td>0.7150</td>
<td>0.6615 0.7685</td>
</tr>
</tbody>
</table>

### 19.5 ADJUSTED COMPARISON OF SURVIVAL TIMES: COX REGRESSION

Survival methods that address the binary event of surviving or not surviving, specifically logistic regression, were met in Chapter 17, Logistic regression for binary outcomes. However, survival time analysis includes models and methods that use...
CHAPTER 19 Analysis of censored time-to-event data

time-to-event data as its primary concern. The issue is not just whether the patient survived or not but also how long the patient survived. The inclusion of time adds to the sensitivity of the model but comes with a restriction: it cannot be just the amount of time that has passed, but the time until the key event, such as death, inception of disease, or protocol completion, has occurred (although censored observations are permitted).

Cox PH (proportional hazard) regression includes in the model the time a subject is at risk. (The following explanation is somewhat simplified for the level of this book.) The Cox PH model focuses on the hazard function as the response. The hazard, or instantaneous risk, can be thought of as the probability of a subject failing within a very small interval of time given that the subject has survived up to the time of that interval, divided by the length of the interval. Thus a high hazard indicates a higher risk of failing and a low hazard indicates a lower risk of failing. A mathematical connection can further be made between the hazard function and the survival function, allowing for comparisons of the time to an event. Cox regression differs from logistic regression in that in logistic regression the binary-dependent variable is associated with the event of survival or failure to survive, not the length of survival time as in Cox regression.

EXAMPLE POSED: TIME FOR ACQUIRED IMMUNODEFICIENCY SYNDROME PATIENTS TO CONTRACT SYPHILIS

In DB23, we have a database of 390 acquired immunodeficiency syndrome (AIDS) patients who have contracted syphilis, including the length of time from contracting AIDS until contracting syphilis. We conjecture that certain comorbidities contribute to susceptibility. We ask if the presence of hepatitis B or hepatitis C shortens the time to infection, that is, increases the hazard of infection.

Method for Cox regression

TERMINOLOGY

Predictor: independent variable (medical treatment, previously acquired disease, demographic variables, etc.) with the potential to explain or shed light on an issue in question. Endpoint: the occurrence of a response (death, failure, etc.) on the issue in question; the time-varying-dependent variable that will be regressed on the predictors; the survival measure. Hazard: the instantaneous risk, or rate, of the endpoint (response event) occurring, given that it has not occurred up to that time. The hazard is denoted \( H(t) \), where \( t \) denotes time since inception. Baseline hazard: the hazard that would occur in the absence of the potential predictors; denoted \( H_0(t) \). Hazard ratio: ratio of the hazard in the presence of a predictor to the hazard without the predictor, that is \( H(t)/H_0(t) \). PH: assumption that predictors affect a hazard
proportionally (e.g., increase it 20%) regardless of the value of \( t \); proportional hazards may be abbreviated PH.

**THE MODEL**

The model assumes that hazards for two groups, say a treatment group and a control group, are proportional regardless of the baseline hazard or the length of time after baseline, that is, the hazard ratio does not depend on time. This implies that the effect of the influencing factor may be calculated without knowing the probability distribution of the individual hazards. Cox’s development starts with the model

\[
H(t) = H_0(t) \times e^{\beta_1 x_1 + \cdots + \beta_k x_k},
\]

in which the exponent gives the linear regression form of the predictors \( x_1, \ldots, x_k \). Dividing by the baseline hazard and taking natural logarithms yield

\[
\ln \left( \frac{H(t)}{H_0(t)} \right) = \beta_1 x_1 + \cdots + \beta_k x_k.
\]

The left side is the logarithm of the hazard ratio relative to the baseline hazard and the right side is the same as in ordinary multiple regression. Thus we can calculate the regression of the log hazard ratio on the set of predictors.

As in multiple regression, statistical software can estimate the \( \beta \)s and provide the Cox proportional hazards regression as in Eq. (19.4). After the \( \beta \)s are estimated the survival probability for any point in time, \( S(t) \), is provided in some software as

\[
S(t) = e^{-H_0(t) \times (\beta_1 x_1 + \cdots + \beta_k x_k)}.
\]

where \( H_0(t) \) represents the estimated cumulative baseline hazard at time \( t \), or the baseline hazard function summed up to time \( t \). The crucial items in the software display to use are the overall \( p \)-value for the model, the \( p \)-value for the parameter associated with each predictor, and the hazard ratio and its confidence interval associated with each predictor.

An additional advantage accrues from taking logarithms. A decrease in a hazard ratio must lie in the small range between 0 and 1, while an increase may take on any value from 1 to \( \infty \). If we take logarithms, positive and negative changes have equal magnitude.

While PH may approximately hold in many medical studies, the assumption does not always hold and the investigator should verify the appropriateness. There exist advanced methods to use when the variables do not follow PH.

It should be noted that Cox PH regression assumes fully continuous (not discrete) readings of time, so that no ties in observed survival times exist. Two forms of adjustment, one by Breslow and one by Efron, exist if the data are discrete and form a few ties. The user can find documentation on their use in the statistical software.
EXAMPLE COMPLETED: TIME UNTIL ACQUIRED IMMUNODEFICIENCY SYNDROME PATIENTS CONTRACT SYPHILIS

The endpoint is contracting syphilis and time is recorded as the time from contracting AIDS until the endpoint. Hepatitis B and hepatitis C are the potential predictors. The model significance is calculated as $\chi^2$ for 2 df, given by the software as 32.81, which yields $p$-value $< 0.001$. Individually, for hepatitis B, $p$-value $< 0.001$ and the hazard ratio = 0.55 with confidence interval 0.45–0.68; for hepatitis C, $p$-value = 0.897 and the hazard ratio = 0.96 with confidence interval 0.55–1.68. The overall model is statistically significant; there is some predictive capability among the predictors. Clearly, the predictive capability likely lies in hepatitis B and not in hepatitis C. We would interpret the hazard ratio as estimating that the instantaneous risk of acquiring syphilis for a patient with hepatitis B is just over half that for a patient without hepatitis B. Indeed, the mean length of time a hepatitis patient has AIDS before contracting syphilis is 12.7 years versus 8.8 years for one without hepatitis. This result appears somewhat counterintuitive and begs further investigation. It could be that many of the patients with hepatitis are ill enough to be disinterested in sexual activity. It could be that hepatitis patients are under closer medical care. It could even be a false positive result.

Exercise 19.3

Laryngeal cancer treated by surgery, radiation, or both may fail. A possible salvage is total laryngectomy. DB32 contains recordings for 63 laryngeal cancer patients so treated, 52 receiving the salvage therapy and 11 not. Time until death is recorded for those who died and time until end of study (censored) for those surviving at that time. Those receiving salvage therapy survived an average of 4.8 years, while those not receiving it survived an average of 6.3 years. We would like to know if the event of receiving or not receiving salvage therapy can predict survival time. Performing a Cox PH regression of salvage therapy or not on years of survival yields a hazard ratio of 1.83 with confidence interval 0.64–5.26. The test on salvage as a predictor yields $z = 1.12$ with $p$-value $= 0.261$. (While the test on the overall model is $\chi^2$, the test on each predictor is a $z$ test.) Interpret these results.

REFERENCES

Analysis of repeated continuous measurements over time

20.1 INTRODUCTION

Repeated measures of biomedical outcomes are often obtained on the same individual or a cluster of individuals. Examples of such data include blood pressure measurements taken periodically over time during routine medical checkups. In preclinical Alzheimer’s disease (AD) clinical trials, we may assess performance of patients on cognitive or functional tests at predefined points in time (e.g., every 6 months for 3 years). The data described previously, which are collected over time on multiple individuals, are termed longitudinal data. Sometimes we have repeated measures on an individual that are not observed over time. For example, we may draw blood on a patient at one point in time, then run a particular assay on the sample multiple times. In this case, we simply refer to the data as repeated measures data.

To this point, we have been primarily concerned with analyzing independent data, meaning that the measured outcome for one observation does not impact the outcome we observe for another observation in our dataset. When repeated measures are obtained on the same individual the data are, generally speaking, no longer independent. This is because measurements taken on the same person are usually more similar to one another than measurements taken on different individuals. Thus repeated measures data (whether taken repeatedly over time or taken at the same time) give rise to correlated data, where correlation exists within individuals.

Repeated measures data provide multiple scientific opportunities that are not available when only a single measurement is available on each individual. The biggest benefit is that we gain information on within-subject variability. This allows us to compare variation within a person to the variation between persons. Further, when longitudinal data are obtained, we can estimate the impact of covariates on the outcome within a subject and changes in the outcome over time on the same subject. These gains in the ability to answer new scientific questions do, however, give rise to statistical challenges that must be dealt with in order to draw valid statistical inference from repeated measures data. Because the data are no longer independent (both within and between individuals), the regression methods we have previously
introduced are generally not valid in this setting. Specifically, we must account for
the correlation within individuals in order to appropriately estimate the variance of
our statistical estimators. That is the primary concern of statistical methods for corre-
lated data.

20.2 DISTINGUISHING LONGITUDINAL DATA FROM TIME-
SERIES DATA

The terms longitudinal data and time-series data are sometimes confused in the scien-
tific literature. This is not surprising because both titles refer to repeated measures
obtained on a system over time, and the statistical methods for analyzing longitudinal
data and time-series data are all concerned with properly accounting for the correla-
tion that arises between observations measured on the same entity over time. There
are, however, clear differences between longitudinal data and time-series data that give
rise to differences in how we analyze the two.

In plain words the term “longitudinal data” generally describes the setting where a
fairly small number of repeated observations are taken on many different entities (e.g.,
patients) over time. For example, we may take annual cholesterol measurements on
100 patients for a total of 10 years, giving rise to 10 measurements for each of the 100
subjects. In contrast the term time-series data generally describes the setting where
many (sometimes hundreds, thousands, or even hundreds of thousands) of observations
are taken on a single or maybe a couple of different entities. A classic example of
time-series data are stock market prices for a particular index fund. In the context of
health data the use of personal fitness devices has given rise to more individual time-
series data such as heart rate (HR) or actigraphy data measured continuously through-
out the day.

So why are the methods for analyzing longitudinal data and time-series data dif-
ferent? The reason is that by having data on many different (independent) individ-
uals, longitudinal data allows us to borrow information across individuals when
estimating the correlation structure of the data for each person. We generally do this
by assuming that the same correlation structure exists within each person and then
averaging our estimates of the correlation over people. In the context of time-series
data, we usually only have one entity, so we cannot average over entities to estimate
correlation. This means that we need to make an assumption on the correlation and,
through diagnostics, try to assure ourselves that our assumption is correct. We do,
however, have a lot of data on the single entity in time-series data which can allow
us to effectively answer questions about when trends change over time (so called
change point detection).
20.3 ANALYSIS OF LONGITUDINAL DATA

Correlation due to repeated measures: Repeated measures analysis of variance

ORIENTATION BY EXAMPLE: HEART RATE BY DISEASE AND BY TIME OF DAY

One of two elderly patients (no. 1) suffered from a mitral valve prolapse, the other (no. 2) from high blood pressure. HRs were taken from each upon arising, at midday, and in the evening (times 1, 2, and 3, respectively).\(^1\) Are their mean vital signs different? Do these signs vary over time? Do these signs vary differently over time for each patient? (Simple data were chosen to exemplify the method; it is not suggested that they have research or clinical importance.)

At first glance, it might appear that a two-factor analysis of variance (ANOVA) would be in order. However, note that data taken through time are on the same patient. The through-time data contain influences from cardiac characteristics, while patient-to-patient data contain both that and person-to-person influences. If we use the same estimate of error variability (mean square of error) to divide by, we introduce a bias and change the sensitivity of the test. We need two estimates of random variability: one containing person-to-person differences and the other not.

DATA AND MEANS

The data are shown in Table 20.1 (means symbols will be used later).

We will need an additional set of means that does not appear in a two-way means table: those across the repeated measure (time) for each replication, which appear as the last column in Table 20.1. These “among-times” means will allow a within-times sum of squares to be calculated. A means table of the sort seen for two-factor ANOVA is shown in Table 20.2.

The calculations appear after some needed formulas are presented.

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Replication no.</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Means (m_{rk})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>80</td>
<td>84</td>
<td>78</td>
<td>80.6667</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>77</td>
<td>74</td>
<td>78</td>
<td>76.3333</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>71</td>
<td>81</td>
<td>75</td>
<td>75.6667</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>50</td>
<td>57</td>
<td>60</td>
<td>55.6667</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>50</td>
<td>55</td>
<td>73</td>
<td>59.3333</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>48</td>
<td>55</td>
<td>56</td>
<td>53.0000</td>
</tr>
</tbody>
</table>
METHOD FOR REPEATED MEASURES ANALYSIS OF VARIANCE (ANOVA)

Goal
As with other multifactor ANOVAs, the goal is to obtain mean squares and then \( F \) values for the main and interaction effects. However, we have one set of effects with variability caused by the treatment factor plus a case-to-case (e.g., patient-to-patient) factor and another set having variability without case-to-case differences, because it is repeated with each case. For example, suppose we are comparing mean pain levels reported by patients randomized to two types of anesthesia immediately post-op and 4 hours post-op. Pain readings between anesthetics contain patient-to-patient differences, but pain readings between 0 and 4 hours post-op do not. Thus to calculate \( F \), we need one mean square for error that includes random variability influences of measures not repeated, and one that includes random variability influences of the repeated-measure-to-measure influences.

Two SSE Terms
The sums of squares for the main effects and interaction are the same as for a two-factor ANOVA, seen in Table 16.12. The difference, so far as calculation is concerned, is finding two sums-of-squares-due-to-error (SSE) terms. One is a within-repeated-measures sum of squares due to error, or SSE(W), that estimates error variability for causal terms involving the repeated factor. In the preceding example, this SSE(W) contains the influence of measures over the same patient. The other is a between-repeated-measures sum of squares due to error, or SSE(B), that estimates error variability for causal terms not involving the repeated factor, that is, contains the influence of the independent, or nonrepeated, factor.

Row and Column Designations
The references to row and column in Section 16.4 must be clarified. In order to be consistent with many software packages the means table will present repeated measures across the rows and independent measures down the columns. Thus sum of squares for rows (SSR) will denote the SS for the independent measure, and sum of squares for columns (SSC) will denote the SS for the repeated measure.

Similar to the preceding section on two-factor ANOVA, \( c \) denotes the number of repeated columns (repeated measures), \( r \) the number of independent measures, and \( w \) the number of replications. Subscript designations appear in Table 20.3.

Sum of Squares for Error Calculations
The first step is to find an interim “sum of squares across repeated measures”, or SSAcross. This SS is the sum of squares for the means of each replication across the repeated measures, as shown in the rightmost column of the data table, \( m_{ijk} \), as in Table 20.1. As before, \( A \) (Continued)
is \( n \) times the squared sum of all observations in the experiment, or alternatively \( rcw \times m \). Table 16.4 gave formulas for calculating SST, SSR, SSC, and SSI. Table 20.4 supplements Table 16.4 with the additional formulas required for repeated measures ANOVA.

**Analysis of Variance Table**

Table 20.5 provides the repeated-measures (two-factor) ANOVA table.

**Interpretation**

Now that we have adjusted for the repeated measures, we can interpret the three \( F \)-values (and their \( p \)-values) in the same way as the ordinary two-factor ANOVA.

**An admonition**

Recall that an assumption is the independence, that is, lack of correlation, among observations. Usually the repeated measure observations are correlated. An example might be pain ratings arising from the same patient; one patient may be more sensitive and rate pain more highly across the board than another patient. In such a case an adjustment to reduce the significance level of the repeated measure is in order. Several statistical software packages contain such adjustments that can be designated when choosing the repeated measures ANOVA option.

**Table 20.3** Subscript designations.

\[
i = 1, 2, \ldots, r \\
j = 1, 2, \ldots, c \\
k = 1, 2, \ldots, w \\
m_{ij} = \sum_k x_{ijk} / w
\]

**Table 20.4** Formulas supplemental to those in Table 16.12 for components in a repeated measures analysis of variance.

\[
\text{SSAcross} = \sum_j m_{i, k}^2 - A \\
\text{SSE(B)} = \text{SSAcross} - \text{SSC} \\
\text{SSE(W)} = \text{SST} - \text{SSAcross} - \text{SSR} - \text{SSI}
\]

*SS Across*, Sum of squares across repeated measures; *SSC*, sum of squares for columns; *SSE(B)*, between-repeated-measures sum of squares for error; *SSE(W)*, within-repeated-measures sum of squares for error; *SSI*, sum of squares for interaction and row-by-column; *SSR*, sum of squares for rows; *SST*, sum of squares for the total.
Table 20.5 Repeated measures (two-factor) analysis of variance table.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent groups</td>
<td>SSR (Table 16.12)</td>
<td>$r - 1$</td>
<td>$MSR = SSR / df$</td>
<td>$MSR / MSE(B)$</td>
<td></td>
</tr>
<tr>
<td>Error between</td>
<td>SSE(B) (Table 20.4)</td>
<td>$(w-1)$</td>
<td>$MSE(B) = SSE(B) / df$</td>
<td>$MSC / MSE(W)$</td>
<td></td>
</tr>
<tr>
<td>Repeated measures</td>
<td>SSC (Table 16.12)</td>
<td>$c - 1$</td>
<td>$MSC = SSC / df$</td>
<td>$MSC / MSE(W)$</td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>SSI (Table 16.12)</td>
<td>$(r - 1)(c - 1)$</td>
<td>$MSI = SSI / df$</td>
<td>$MSI / MSE(W)$</td>
<td></td>
</tr>
<tr>
<td>Error within</td>
<td>SSE(W) (Table 20.4)</td>
<td>$r(c - 1)(w - 1)$</td>
<td>$MSE(W) = SSE(W) / df$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SST (Table 16.12)</td>
<td>$rw - 1$</td>
<td>$MSE$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*MSE*, Mean square of error; *MSC*, mean square for columns; *MSI*, mean square for interaction; *MSR*, means square for rows; *SSC*, sum of squares for columns; *SSI*, sum of squares for interaction and row-by-column; *SSR*, sum of squares for rows; *SST*, sum of squares for the total.
EXAMPLE COMPLETED: HEART RATE BY DISEASE AND BY TIME OF DAY

In the cardiac example, we have \( r = 2 \) patients, \( c = 3 \) times per patient, and \( w = 3 \) replications for each patient-time. The adjustment term \( A = rcw = 18 \times 66.77782 = 80 266.94 \).

\[
SSR = cw \sum m_{ij}^2 - A = 9(77.5556^2 + \cdots) - A.
\]

\[
SSC = w \sum m_{ij}^2 - A = 6(62.6667^2 + \cdots) - A.
\]

\[
SSI = w \sum m_{ij}^2 - A - SSR - SSC = 3(76^2 + \cdots) - A - SSR - SSC.
\]

\[
SSA\text{cross} = c \sum m_{ik}^2 - A = 3(80.6667^2 + \cdots) - A.
\]

These calculations lead to the ANOVA table (Table 20.6).

There is a highly significant difference between the two patients, not surprising in light of their different medical maladies. The times of day at which readings are taken are not significantly different, nor is the interaction significant, indicating that time-of-day variability for one patient is not different from that for the other.

ADDITIONAL EXAMPLE: FASCIOTOMY AND ANTIVENIN TO RELIEVE SNAKEBITE EDEMA

**Experiment and data**

A widely used part of the treatment for rattlesnake bite is fasciotomy to relieve the pressure from edema. An ER specialist\(^2\) questioned its benefit. He injected the hind legs of 12 anesthetized pigs with rattlesnake venom. He treated six pigs, chosen randomly, with antivenin, and the other six without, performing fasciotomy on one leg but not the other. The first independent variable is antivenin treatment or not, which includes pig-to-pig influence as well as influence due to the treatment and fasciotomy factors. The second independent variable is fasciotomy, which includes the treatment and fasciotomy influences, but not pig-to-pig influence, because fasciotomy versus not-fasciotomy was performed on the same pig. As the outcome (dependent) is variable, he measured the percent tissue necrosis at 8 hours. The data are shown in Table 20.7.
CHAPTER 20 Analysis of repeated continuous measurements over time

Table 20.7 Fasciotomy and antivenin data.\(^a\)

<table>
<thead>
<tr>
<th>Pig number</th>
<th>Antivenin treatment</th>
<th>8-h % necrosis</th>
<th>Means across fasciotomy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>With fasciotomy</td>
<td>Without fasciotomy</td>
</tr>
<tr>
<td>1</td>
<td>No</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>57</td>
<td>52</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>25</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>Yes</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Yes</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>Yes</td>
<td>40</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>Yes</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

\(^a\) Data were slightly simplified to facilitate the example.

Table 20.8 Fasciotomy and antivenin means.

<table>
<thead>
<tr>
<th>Fasciotomy</th>
<th>Yes</th>
<th>No</th>
<th>Across fasciotomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antivenin treatment</td>
<td>No</td>
<td>26.3333</td>
<td>17.0000</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>17.8333</td>
<td>9.1667</td>
</tr>
<tr>
<td>Across antivenin</td>
<td></td>
<td>22.0833</td>
<td>13.0833</td>
</tr>
</tbody>
</table>

MEANS TABLE
The across-repeated-measures means, specifically across fasciotomy or not, are appended in the rightmost column of the data table. The means used to calculate the main- and interaction-effect sums of squares appear in Table 20.8.

CALCULATIONS
Substitution in the formulas of Tables 16.4 and 20.4 yields the sums of squares shown in Table 20.9.

ANALYSIS OF VARIANCE TABLE
Substitution in the formulas of Table 20.4 yields the \(df\), mean squares, and \(F\)-values appearing in the ANOVA table (Table 20.10).
No significant effects are seen at all. The antivenin treatment and the fasciotomy have not been shown to be effective. Certainly, the near-zero $F$ for the interaction indicates that the effect of fasciotomy is not different with antivenin treatment than without it.

**Exercise 20.1**

Comparing treatment of severe migraine headaches. An ER specialist wanted to compare the pain-relieving effects of two drugs, diazepam and prochlorperazine, for patients suffering from acute migraine headaches. He randomly selected 19 of 38 patients for one drug, the other 19 for the other, measuring pain levels between 0 and 100 by a visual analog scale at baseline, and 30 and 60 minutes. The pain by drug contains patient-to-patient variability, while the pain by time is measured within each patient. Thus a repeated measures ANOVA is appropriate. The first few data are shown in Table 20.11.

Table 20.9 Fasciotomy and antivenin sums of squares.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>$F$</th>
<th>Crit. $F$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antivenin treatment</td>
<td>400.2122 (SSR)</td>
<td>1</td>
<td>400.2122</td>
<td>1.143</td>
<td>4.96</td>
<td>0.310</td>
</tr>
<tr>
<td>Between-pig error</td>
<td>3502.6493 [SSE(B)]</td>
<td>10</td>
<td>350.2650</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasciotomy</td>
<td>486.0000 (SSC)</td>
<td>1</td>
<td>486.0000</td>
<td>2.294</td>
<td>4.96</td>
<td>0.161</td>
</tr>
<tr>
<td>Trtmnt. $\times$ fasc. interaction</td>
<td>0.6353 (SSI)</td>
<td>1</td>
<td>0.6353</td>
<td>0.003</td>
<td>4.96</td>
<td>0.956</td>
</tr>
<tr>
<td>Within-pig error</td>
<td>2118.3647 [SSE(W)]</td>
<td>10</td>
<td>211.8365</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6507.8615 (SST)</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 20.10 Repeated measures analysis of variance (ANOVA) table for fasciotomy and antivenin.**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>$F$</th>
<th>Crit. $F$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antivenin treatment</td>
<td>400.2122 (SSR)</td>
<td>1</td>
<td>400.2122</td>
<td>1.143</td>
<td>4.96</td>
<td>0.310</td>
</tr>
<tr>
<td>Between-pig error</td>
<td>3502.6493 [SSE(B)]</td>
<td>10</td>
<td>350.2650</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fasciotomy</td>
<td>486.0000 (SSC)</td>
<td>1</td>
<td>486.0000</td>
<td>2.294</td>
<td>4.96</td>
<td>0.161</td>
</tr>
<tr>
<td>Trtmnt. $\times$ fasc. interaction</td>
<td>0.6353 (SSI)</td>
<td>1</td>
<td>0.6353</td>
<td>0.003</td>
<td>4.96</td>
<td>0.956</td>
</tr>
<tr>
<td>Within-pig error</td>
<td>2118.3647 [SSE(W)]</td>
<td>10</td>
<td>211.8365</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6507.8615 (SST)</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**INTERPRETATION**

No significant effects are seen at all. The antivenin treatment and the fasciotomy have not been shown to be effective. Certainly, the near-zero $F$ for the interaction indicates that the effect of fasciotomy is not different with antivenin treatment than without it.
CHAPTER 20  Analysis of repeated continuous measurements over time

Table 20.11  Migraine data.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Drug</th>
<th>Pain at 0 min</th>
<th>Pain at 30 min</th>
<th>Pain at 60 min</th>
<th>Mean across</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Prochlorper</td>
<td>80</td>
<td>46</td>
<td>17</td>
<td>47.6667</td>
</tr>
<tr>
<td>2</td>
<td>Prochlorper</td>
<td>76</td>
<td>14</td>
<td>0</td>
<td>30.0000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>37</td>
<td>Diazepam</td>
<td>43</td>
<td>6</td>
<td>1</td>
<td>16.6667</td>
</tr>
<tr>
<td>38</td>
<td>Diazepam</td>
<td>61</td>
<td>68</td>
<td>56</td>
<td>61.6667</td>
</tr>
</tbody>
</table>

Table 20.12  Migraine Means.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Pain 0 min</th>
<th>Pain 30 min</th>
<th>Pain 60 min</th>
<th>Across pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diazepam</td>
<td>69.8421</td>
<td>59.8421</td>
<td>55.3158</td>
<td>61.6667</td>
</tr>
<tr>
<td>Prochlorper</td>
<td>76.0526</td>
<td>32.7368</td>
<td>17.4210</td>
<td>42.0702</td>
</tr>
<tr>
<td>Across drug</td>
<td>72.9474</td>
<td>46.2895</td>
<td>36.3684</td>
<td>51.8684</td>
</tr>
</tbody>
</table>

\[ A = 306 \ 697.7200, \quad SST = 136 \ 171.0400, \quad \text{and} \quad SS\text{Across} = 82 \ 177.4640. \quad \text{The means table is shown in Table 20.12.} \]

Perform the repeated measures analysis of variance and interpret it.

Covariance structures

The repeated measures ANOVA design is used for comparing means across a discrete number of groups when there exist repeated measures on each subject. In order to assess the relationship between a continuous covariate and a response that is repeatedly measured over time, we can turn to a regression model. However, in order to appropriately account for the fact that correlation exists between observations, we will need to specify and utilize a plausible covariance structure, meaning that we need to specify the covariance for any two observations made on the same individual.

Before considering the commonly assumed covariance structures, it is useful to review the definition of the covariance between two random variables and to highlight how it is related to correlation. Suppose that \( Y_{ij} \) and \( Y_{ik} \) are the \( j \) and \( k \) measurements taken on individual \( i \). For example, they might be blood pressure measurements made at time \( t_j \) and time \( t_k \). The covariance between \( Y_{ij} \) and \( Y_{ik} \) is then defined as

\[
\text{Cov}(Y_{ij}, Y_{ik}) = \rho_{jk} \sigma_j \sigma_k
\]  

(20.1)

where \( \sigma_j^2 \) and \( \sigma_k^2 \) denote the variance of \( Y_{ij} \) and \( Y_{ik} \), respectively, and \( \rho_{jk} \) is the correlation between \( Y_{ij} \) and \( Y_{ik} \). From this it is easy to see that the correlation can be expressed as a function of the covariance and the variances. Namely,
Common choices of covariance structures

When repeated measures are taken on the same individual, it is more likely than not that the observations taken on the same person will be correlated. To estimate and draw inference on the association between a covariate and a response in this case requires that we account for the correlation in our estimation and inference procedures. To make this practical, it is typically necessary to assume some sort of structure for the form of the covariance between any two observations on an individual. The most commonly utilized assumptions for the covariance structure are defined next:

**Exchangeable:** An exchangeable covariance structure assumes that the covariance between any two observations on the same person is the same and that the variance for each observation is also constant. Thus an exchangeable covariance structure assumes that

\[
\text{Cov}(Y_{ij}, Y_{ik}) / \sigma_j \sigma_k = \rho \sigma^2; \quad \text{for all } i, \text{ and } j \neq k.
\]  

Notice that the right side of the above does not involve subscripts \(i, j,\) or \(k,\) implying that the covariance for any two observations on the same subject is the same regardless of how far apart in time the two measurements may have been.

**Autoregressive:** An autoregressive covariance structure assumes that the correlation, and hence the covariance, between any two observations on the same person tends to decrease as the distance between the two measurements increases. Specifically, an autoregressive covariance structure assumes that

\[
\text{Cov}(Y_{ij}, Y_{ik}) = \rho |k-j| \sigma^2, \quad \text{for all } i.
\]  

where \(\rho\) represents the correlation between two observations taken at adjacent visits (i.e., \(|k-j| = 1\)). Since \(\rho\) is bounded between \(-1\) and \(1,\) we can see that the correlation and hence the covariance will decrease as the distance between \(j\) and \(k\) grows larger. This is a natural assumption in many settings, since we expect observations taken closer in time to be more alike than those taken farther apart in time.

**Continuous autoregressive:** The autoregressive covariance structure only considers the measurement number in defining distance. If a study is designed to have an equal amount of time between measurements for all individuals (e.g., if blood pressure is measured exactly every 30 days for all individuals). However, in clinical studies, this rarely happens since patients cannot always schedule visits at exact intervals in time. Thus the time between measurements may vary both within an individual and between individuals. Assuming that observations \(Y_{ij}\) and \(Y_{ik}\) are the \(j\) and \(k\)
measurements taken on individual \( i \) at times \( t_j \) and time \( t_k \), the continuous autoregressive covariance structure assumes that

\[
\text{Cov}(Y_{ij}, Y_{ik}) = \rho |t_k - t_j| \sigma^2. \tag{20.5}
\]

where \( \rho \) represents the correlation between two observations taken one unit in time apart (i.e., \(|t_k - t_j| = 1\)).

**Unstructured:** A final covariance structure that is sometimes used when data are both balanced and complete (meaning that number of repeated measurements on each individual is the same and no missing values are present) is an unstructured covariance matrix that makes no assumptions on the form of the covariance between two observations on the same subject. That is, we assume

\[
\text{cov}(Y_{ij}, Y_{ik}) = \sigma_{jk}. \tag{20.6}
\]

In the equation, note that the only assumption is that the covariance between the \( j \)th and \( k \)th observations is the same across all subjects, but it changes for all \( j \) and \( k \). This means we would need to estimate for \( \sigma_{jk} \) for every unique pair of observations \((j,k)\). In order to do so, this covariance structure requires a large amount of data on all pairs of observations, hence the need for balanced and complete data.

**METHOD FOR INVESTIGATING THE COVARIANCE STRUCTURE**

When presented with longitudinal data, we must investigate the data to determine a plausible covariance structure. The simplest way to do this is to compute the sample pairwise correlation and variance estimates in the data. However, it is important to first remove the systematic component of any other covariates that we will eventually wish to associate with the response. We can do this by using by using the linear regression model (for independent data) that was introduced in Chapter 15, Linear regression and correlation. The procedure works as follows:

1. Suppose that we observe responses \( y_{i1}, \ldots, y_{in} \) measured at times \( t_{i1}, \ldots, t_{in} \), for subject \( i, i = 1, \ldots, n \). Further suppose we also observe a covariate \( x_{i1}, \ldots, x_{in} \) at each time.

2. We wish to remove the mean trend of time and the covariate from the response, which can be done with a linear regression model of the form

\[
y_j = \beta_0 + \beta_1 t_j + \beta_2 x_j + \epsilon_j. \tag{20.7}
\]

Using ordinary least squares as presented in Chapter 15, Linear regression and correlation, we can obtain estimates \( b_0, b_1, \) and \( b_2 \) of \( \beta_0, \beta_1, \) and \( \beta_2 \), respectively. From this, we can estimate the residual for each observation as (Continued)
These residuals will then have mean 0.

4. To estimate the variance, pairwise correlation, and pairwise covariance of the residuals over time, we can divide time into multiple intervals. For example, if measurements are taken roughly once a year over multiple years, we might divide time into 1-month intervals. For each interval, we can plot the residuals for each subject against one another as shown in the example of Fig. 20.1. In the figure, we can see that for years close to one another (on the off-diagonal) the correlation between observations tends to be high and this remains true even for observations up to 10 years apart. This suggests a fairly constant (i.e., exchangeable) correlation structure in the data.

The general linear model

**ORIENTATION BY EXAMPLE: RETEST EFFECTS IN ALZHEIMER’S DISEASE**

AD is a type of dementia which is typically only diagnosed definitively at death upon identification of the characteristic plaques and tangles found in the brain. Consequently, diagnosis and treatment of the disease in the living relies heavily upon...
neuropsychological testing, imaging, and subjective clinical judgment including established guidelines for diagnosis. These guidelines were decided upon by large interdisciplinary research groups and professional organizations. Built into these guidelines are a pattern of deficits that include dysfunction in certain cognitive areas which must be documented by accepted neuropsychological testing as well as documented decline over time. However, a typical problem of longitudinal cognitive testing is that repeated testing may yield increases in score upon subsequent testing times. These increases over time due to repeated testing are known as retest effects and may mask the very effect that is supposed to provide evidence of illness.

Longitudinal data have been collected on $N = 10,900$ unique subjects, with each subject having at least two cognitive tests and a maximum of 12 visits. Cognitive testing visits were scheduled to occur approximately once a year; however, there is variation in the times at which subjects returned for visits. The cognitive test we focus on here is the Logical Memory test. The Logical Memory test is a subset of the Wechsler Memory Scale, with total scores ranging from 0 to 25. For this test the items/questions do not change from one visit to the next. Because of this, it is of interest to determine if there exist retest effects in which patients scores may actually increase with increased number of visits (despite increasing age between visits). Specifically, it is of interest to determine if retest effects exist and are different depending upon whether a subject has been diagnosed as clinically normal, mild cognitive impairment (MCI), or demented. As such, we wish to use these longitudinal data to determine if test scores actually rise over time in subjects among any of the three diagnostic groups. Given that one would expect scores to decrease naturally with aging, a positive slope in scores over time would indicate a retest effect.

**SUMMARY OF AVAILABLE DATA**

Patient characteristics are given in Table 20.13 and available in DB16. The majority of patients have a normal diagnosis ($N = 6583$) followed by MCI ($N = 2723$) and AD ($N = 1594$). The mean age is roughly the same across diagnosis groups, though the proportion of females goes down as the diagnosis severity increases. In addition, individuals diagnosed with AD tended to have the fewest years of education. As expected, the baseline cognitive score (Logical Memory test score) is lowest in individuals with AD and highest among those with a cognitively normal diagnosis.

Table 20.14 shows the number of patients that have a given number of follow-up visits (and hence test scores). Patients had between 3 and 12 visits. Fig. 20.2 displays the outcome of these visits. The upper left plot displays the mean cognitive test scores over time by diagnosis group. The remaining plots depict the observed cognitive scores for a random sample of 25 patients within each diagnosis group. From the upper left plot, we can see that, when averaging across patients at given time points, the average cognitive test score increases for cognitively normal and MCI subjects.
This suggests a retest effect, but our goal is to estimate the average within-patient change in test scores over time by group and to compare these estimates across the diagnostic groups. The general linear model is one appropriate tool for this.

**WRITING DOWN THE GENERAL LINEAR MODEL**

We can extend the usual linear model presented in Chapter 16, Multiple linear and curvilinear regression and multifactor analysis of variance, to the case of correlated observations fairly easily once we have specified the covariance structure. Suppose again that we have observed responses \( y_{i1}, \ldots, y_{in} \) measured at times \( t_{i1}, \ldots, t_{in} \), for subject \( i, i = 1, \ldots, n \). Further suppose we observe covariates \( x_{i1,1}, \ldots, x_{i1,n_i} \) to \( x_{ip,1}, \ldots, x_{ip,n_i} \) at each time as well. A model for the response can then be written as

\[
y_{ij} = \beta_0 + \beta_1 x_{i1,j} + \ldots + \beta_p x_{ip,j} + \epsilon_{ij}. \tag{20.9}
\]

In the above general linear model the intercept \( \beta_0 \) represents the mean of the response when all independent variables are equal to zero. \( \beta_k \) represents the mean difference in the response comparing subpopulations differing in the \( k \)th independent variable by 1-unit. Time can also be included as one of the independent variables in the above

---

**Table 20.13** Characteristics of patients included in the cognitive testing study.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Normal (( N = 6583 ))</th>
<th>MCI (( N = 2723 ))</th>
<th>AD (( N = 1594 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at first visit (years)</td>
<td>72.03 (9.56)</td>
<td>73.36 (8.84)</td>
<td>73.61 (9.38)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2216 (33.7%)</td>
<td>1416 (52%)</td>
<td>919 (57.7%)</td>
</tr>
<tr>
<td>Female</td>
<td>4367 (66.3%)</td>
<td>1307 (48%)</td>
<td>675 (42.3%)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>5386 (82.1%)</td>
<td>2256 (83.6%)</td>
<td>1365 (86.3%)</td>
</tr>
<tr>
<td>Black</td>
<td>872 (13.3%)</td>
<td>298 (11%)</td>
<td>122 (7.7%)</td>
</tr>
<tr>
<td>Native American</td>
<td>16 (0.2%)</td>
<td>7 (0.3%)</td>
<td>9 (0.6%)</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>2 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Asian</td>
<td>109 (1.7%)</td>
<td>55 (2%)</td>
<td>33 (2.1%)</td>
</tr>
<tr>
<td>Other</td>
<td>176 (2.7%)</td>
<td>82 (3%)</td>
<td>52 (3.3%)</td>
</tr>
<tr>
<td>Education (years)</td>
<td>15.74 (2.91)</td>
<td>15.43 (3.15)</td>
<td>14.93 (3.14)</td>
</tr>
<tr>
<td>Baseline score (logical memory)</td>
<td>13.62 (3.85)</td>
<td>9.66 (4.2)</td>
<td>6.21 (3.92)</td>
</tr>
</tbody>
</table>

**Table 20.14** Total number of patients by number of follow-up visits.

<table>
<thead>
<tr>
<th>No. of visits</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>3301</td>
<td>2280</td>
<td>1464</td>
<td>1237</td>
<td>939</td>
<td>717</td>
<td>544</td>
<td>264</td>
<td>124</td>
<td>30</td>
</tr>
</tbody>
</table>

\( AD \), Alzheimer’s disease; \( MCI \), mild cognitive impairment.
In this case the coefficient associated with time represents the expected difference in the response associated with a 1-unit increase in time, or the slope of the response over time.

**EXAMPLE: INTERPRETATION OF COEFFICIENT RELATED TO TIME IN THE COGNITIVE TESTING STUDY**

In the cognitive testing study, we are interested in the change of the logical memory score as a function of time (or the slope of the logical memory score over time). To address this question in the total population, we would include the time of each test as a covariate (independent variable) in the general linear model. Thus we would consider a model given by

\[
\text{Logic. Mem}_{ij} = \beta_0 + \beta_1 t_{i1,j} + \epsilon_{ij},
\]

(20.10)

In this case, \(\beta_1\) represents the mean change in Logical Memory test score that is associated with a 1-unit change in time (or 1 year in the case of our example). Thus \(\beta_1\) is the slope of the Logical Memory test score over time.
EXAMPLE: COMPARING SLOPES OVER TIME ACROSS DIAGNOSTIC POPULATIONS

In the cognitive testing study we are also interested in comparing whether the slope of the Logical Memory test score differs by diagnostic group (Normal, MCI, or AD). The hypothesis is that cognitively normal patients would have a higher slope. We would include the time of each test as a covariate (independent variable) as well as an indicator of diagnostic group and the interaction between them in the general linear model. Thus we would consider a model given by

\[
\text{Logic. Mem}_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 I_{i,MCI} + \beta_3 I_{i,AD} + \beta_4 t_{ij} \times I_{i,MCI} + \beta_5 t_{ij} \times I_{i,AD} + \epsilon_{ij}. \tag{20.11}
\]

Now, \(\beta_1\) represents the slope of the Logical Memory test score over time among cognitively normal patients, \(\beta_1 + \beta_4\) represents the slope of the Logical Memory test score over time among MCI patients, and \(\beta_1 + \beta_5\) represents the slope of the Logical Memory test score over time among AD patients.

Fitting the model

The general linear model can be fit using most common statistical software packages. The procedure involves using weighted least squares (see Chapter 16: Multiple linear and curvilinear regression and multifactor analysis of variance) where the weights are specified via the covariance structure that is assumed for the data. In general, one must specify the response, independent variables, a unique subject ID to identify which measurements belong to which subjects, and the covariance structure that is to be assumed for the model fitting.

Tests of association

It is commonly of interest to determine whether an association between an independent variable and a dependent variable exists. In Eq. (20.9), \(\beta_k\) represents the mean difference in the response comparing subpopulations differing in the \(k\)th independent variable by 1-unit. Thus if \(\beta_k = 0\) this would imply that the mean of the outcome is the same comparing subpopulations differing in the \(k\)th independent variable by 1-unit. This would imply that no association exists between the outcome and \(x_k\). Conversely, if \(\beta_k \neq 0\) then this would imply that the mean of the response is associated with a change in \(x_k\). Thus it may be of interest to test the null hypothesis \(H_0: \beta_k = 0\) versus the alternative hypothesis \(H_1: \beta_k \neq 0\). In order to carry out this hypothesis test, we require a test statistic for which we know its distribution under the null hypothesis, \(H_0\).

Consider the general linear model

\[
y_{ij} = \beta_0 + \beta_1 x_{i1,j} + \ldots + \beta_p x_{ip,j} + \epsilon_{ij}. \tag{20.12}
\]
Statistical software can be used to obtain estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_p$. Let $b_0, b_1, \ldots, b_p$ denote these estimates. Further, software can be used to estimate the standard error of $b_0, b_1, \ldots, b_p$, which we denote as $se(b_0), se(b_1), \ldots, se(b_p)$.

A test of the null hypothesis $H_0: \beta_k = 0$ versus the alternative hypothesis $H_1: \beta_k \neq 0$, $k = 0, \ldots, p$, can be conducted with the following test statistic

$$z = \frac{b_k}{se(b_k)} \sim N(0, 1). \tag{20.13}$$

The value of $z$ can be looked up in Table I and the resulting $p$-value is the probability of observing a result as or more indicative of the alternative hypothesis if the null hypothesis were true.

**EXAMPLE CONTINUED: COMPARING SLOPES OVER TIME ACROSS DIAGNOSTIC POPULATIONS**

To test the hypothesis that cognitively normal patients have a higher slope than MCI patients using a model of the form

$$Logic. Mem_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 I_{i,MCI} + \beta_3 I_{i,AD} + \beta_4 t_{ij} \times I_{i,MCI} + \beta_5 t_{ij} \times I_{i,AD} + \epsilon_{ij}, \tag{20.14}$$

first recall that $\beta_1$ represents the slope of the Logical Memory test score over time among cognitively normal patients, while $\beta_1 + \beta_4$ represents the slope of the Logical Memory test score over time among MCI patients. Thus if $\beta_4 = 0$, that would imply that the slope of the cognitive score is the same among cognitively normal and MCI subjects. Hence to compare the slopes between these groups, we would want to test the hypothesis $H_0: \beta_4 = 0$ versus the alternative hypothesis $H_1: \beta_4 \neq 0$.

**EXAMPLE COMPLETED: RETEST EFFECTS IN ALZHEIMER’S DISEASE**

**Assessment of the covariance/correlation structure**

For the retest effects analysis, we must first investigate and decide upon a covariance structure that will be used in the model fitting process of the general linear model. For this, we consider the estimated correlation between residuals on the same individual as given in Table 20.15. Note that these estimates are simply the correlation on the residuals shown in Fig. 20.1 of “Interpretation” section. As was noted from the figure, we can see from the table that the correlation remains fairly constant regardless of the distance in time between tests. This would suggest an exchangeable correlation structure as defined in “Interpretation” section.
Table 20.15 Estimated correlation between cognitive test scores for the same subject across years.

<table>
<thead>
<tr>
<th>Year of measurement</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.63</td>
<td>0.6</td>
<td>0.58</td>
<td>0.56</td>
<td>0.56</td>
<td>0.51</td>
<td>0.56</td>
<td>0.57</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>0.63</td>
<td>1</td>
<td>0.67</td>
<td>0.66</td>
<td>0.56</td>
<td>0.55</td>
<td>0.57</td>
<td>0.56</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>0.67</td>
<td>1</td>
<td>0.67</td>
<td>0.64</td>
<td>0.63</td>
<td>0.6</td>
<td>0.58</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>0.58</td>
<td>0.66</td>
<td>0.67</td>
<td>1</td>
<td>0.67</td>
<td>0.66</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.59</td>
</tr>
<tr>
<td>5</td>
<td>0.56</td>
<td>0.6</td>
<td>0.64</td>
<td>0.7</td>
<td>1</td>
<td>0.68</td>
<td>0.63</td>
<td>0.63</td>
<td>0.64</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>0.56</td>
<td>0.62</td>
<td>0.63</td>
<td>0.66</td>
<td>0.68</td>
<td>1</td>
<td>0.67</td>
<td>0.63</td>
<td>0.66</td>
<td>0.58</td>
</tr>
<tr>
<td>7</td>
<td>0.51</td>
<td>0.55</td>
<td>0.6</td>
<td>0.6</td>
<td>0.63</td>
<td>0.67</td>
<td>1</td>
<td>0.71</td>
<td>0.65</td>
<td>0.54</td>
</tr>
<tr>
<td>8</td>
<td>0.56</td>
<td>0.57</td>
<td>0.58</td>
<td>0.6</td>
<td>0.63</td>
<td>0.63</td>
<td>0.71</td>
<td>1</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>9</td>
<td>0.57</td>
<td>0.56</td>
<td>0.55</td>
<td>0.59</td>
<td>0.64</td>
<td>0.66</td>
<td>0.65</td>
<td>0.72</td>
<td>1</td>
<td>0.64</td>
</tr>
<tr>
<td>10</td>
<td>0.61</td>
<td>0.58</td>
<td>0.55</td>
<td>0.55</td>
<td>0.58</td>
<td>0.58</td>
<td>0.54</td>
<td>0.63</td>
<td>0.64</td>
<td>1</td>
</tr>
</tbody>
</table>
**CHAPTER 20 Analysis of repeated continuous measurements over time**

Table 20.16 General linear model estimates for modeling the slope of cognitive test scores over time.

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>se(b)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$: (Intercept)</td>
<td>12.78319</td>
<td>0.0432601</td>
<td>295.4961</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$b_1$: Time (years)</td>
<td>0.20815</td>
<td>0.0059485</td>
<td>34.9917</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$b_2$: MCI</td>
<td>−1.98803</td>
<td>0.0559392</td>
<td>−35.5391</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$b_3$: AD</td>
<td>−4.30652</td>
<td>0.0781516</td>
<td>−55.1047</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$b_4$: Time $\times$ MCI</td>
<td>−0.12054</td>
<td>0.0125132</td>
<td>−9.6332</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>$b_5$: Time $\times$ AD</td>
<td>−0.44781</td>
<td>0.0155522</td>
<td>See Ex</td>
<td>See Ex</td>
</tr>
</tbody>
</table>

**MODEL ESTIMATES**

Table 20.16 displays the estimates that come from fitting the general linear model given in Eq. (20.11) to our data. Here the exchangeable covariance structure was assumed as was indicated above. In this case, $b_1$ represents the slope of Logical Memory test scores among cognitively normal individuals. Thus we estimate that, on average, Logical Memory test scores rise by approximately 0.21 points each year the test is given to cognitively normal participants. A test of $H_0: \beta_1 = 0$ versus the alternative hypothesis $H_1: \beta_1 \neq 0$ yields a test statistic of 34.99 and the resulting $p$-value for this test is < 0.001. Hence, we reject the null hypothesis and conclude that, on average, test scores are significantly rising among cognitively normal participants, indicating a retest effect. In addition the estimated slope for MCI subjects is given by $b_1 + b_4 = 0.21 - 0.12 = 0.09$. Hence our estimate of the slope among MCI patients is also positive. To test whether this is different than the slope of cognitively normal patients, we test $H_0: \beta_4 = 0$ versus the alternative hypothesis $H_1: \beta_4 \neq 0$. This results in a test statistic of −9.63 and the resulting $p$-value for this test is < 0.001. Thus we conclude that MCI patients have a significantly lower slope when compared to cognitively normal patients. In other words, there does still appear to be a learning effect in MCI patients, but it is not as strong as in cognitively normal patients.

**Exercise 20.2**

Retest effects among AD patients. Using the general linear model estimates provided in Table 20.16, answer the following questions:

**a.** What is the slope of Logical Memory test scores over time among AD patients?

**b.** Based upon the above estimate, does there appear to be a retest effect among AD patients?

**c.** Is the slope of Logical Memory test scores over time among AD patients significantly different than that of cognitively normal patients?

**Linear mixed effects models**

Another popular approach to analyzing longitudinal data is to use random effects to account for correlation among repeated measurements on the same individual.
This means that we allow for each individual to have their own regression parameter value. In order to do this, we assume that the population parameter value is the norm, and that the subjects in our data set vary around that norm. In the context of the cognitive testing example, we might allow each patient to have their own intercept, realizing that some patients will naturally tend to have higher or lower test scores at the first visit when compared to the average of the sample. This is called a random intercept model. Here we will focus on random intercept models, but we can also consider the coefficient associated with time to vary by patient. In the context of the cognitive testing example, this would allow for the fact that retest effects may vary from one person to the next.

Method for the random intercept model

Suppose that we have observed responses $y_{i1}, \ldots, y_{in_i}$ measured at times $t_{i1}, \ldots, t_{in_i}$, for subject $i$, $i = 1, \ldots, n$. Further suppose we also observe covariates $x_{i1,1}, \ldots, x_{i1,n_i}$ to $x_{ip,1}, \ldots, x_{ip,n_i}$ at each time. A random intercept model for the response can then be written as

$$y_{ij} = \beta_0 + \beta_1 x_{i1,j} + \ldots + \beta_p x_{ip,j} + \beta_0 i + \epsilon_{ij}. \quad (20.15)$$

In the above model the intercept $\beta_0$ represents the mean of the response across the population when all independent variables are equal to zero, while the subject-specific deviation from the population intercept is given by $\beta_0 i$. Thus if $\beta_0 i$ is positive, it would mean that subject $i$ has a higher intercept than the rest of the population, and if $\beta_0 i$ is negative it would mean that subject $i$ has a lower intercept than the rest of the population. As in the general linear model, $\beta_k$ still represents the mean difference in the response comparing subpopulations differing in the $k$th independent variable by 1-unit. $\beta_0, \ldots, \beta_p$ are generally referred to as marginal effects or fixed effects.

In order to facilitate parameter estimation the random intercept model assumes that $\beta_0 i$ is normally distributed with mean 0 (i.e., on average we have the population intercept) and some variance which we will denote by $\tau_0^2$. We also assume that $\epsilon_{ij}$ is normally distributed, is independent of $\beta_0 i$ and that conditional upon $\beta_0 i$, all residuals within the subject are independent of one another. Under these assumptions, estimates of $\beta_0, \ldots, \beta_p$ and $\beta_0 i, i = 1, \ldots, n$ can be obtained through software. The estimates of $\beta_0 i$ are obtained through a method called empirical Bayes estimation.

As with the general linear model, time can also be included as one of the independent variables in the above model. In this case the coefficient associated with time represents the expected difference in the response associated with a 1-unit increase in time, or the slope of the response over time. A random intercept model in this case would be given by

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_0 i + \epsilon_{ij}. \quad (20.16)$$
EXAMPLE: INTERPRETATION OF COEFFICIENT RELATED TO TIME IN THE COGNITIVE TESTING STUDY

In the cognitive testing study, we are interested in the change of the logical memory score as a function of time (or the slope of the logical memory score over time). To address this question in the total population, we would include the time of each test as a covariate (independent variable) in the general linear model. Thus we would consider a model given by

\[ \text{Logic. Mem}_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_{0i} + \epsilon_{ij}. \]  

(20.17)

In this case, \( \beta_0 \) represents the population mean score at time 0 and \( \beta_1 \) represents the population mean change in Logical Memory test score that is associated with a 1-unit change in time (or 1 year in the case of our example). Thus \( \beta_1 \) is the population slope of the Logical Memory test score over time. \( \beta_0 + \beta_{0i} \) represents the mean score at time 0 for subject \( i \).

EXAMPLE: COMPARING SLOPES OVER TIME ACROSS DIAGNOSTIC POPULATIONS

Recall that in the cognitive testing study, we are also interested in comparing whether the slope of the Logical Memory test score differs by diagnostic group (Normal, MCI, or AD). The hypothesis is that cognitively normal patients would have a higher slope. We would include the time of each test as a covariate (independent variable) as well as an indicator of diagnostic group and the interaction between them in the random intercept model. Thus we would consider a model given by

\[ \text{Logic. Mem}_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 I_{i,MCI} + \beta_3 I_{i,AD} + \beta_4 t_{ij} \times I_{i,MCI} + \beta_5 t_{ij} \times I_{i,AD} + \beta_{0i} + \epsilon_{ij}. \]  

(20.18)

In this model, \( \beta_1 \) represents the slope of the Logical Memory test score over time among cognitively normal patients, \( \beta_1 + \beta_4 \) represents the slope of the Logical Memory test score over time among MCI patients, and \( \beta_1 + \beta_5 \) represents the slope of the Logical Memory test score over time among AD patients. Again, the model allows for a different intercept for each patient. Specifically, the mean test score at time 0 for patient \( i \) is given by \( \beta_0 + \beta_{0i} \).

Tests of association

As with the general linear model, in Eq. (20.12), \( \beta_k \) represents the mean difference in the response comparing subpopulations differing in the \( k \)th independent variable by 1-unit. Thus if \( \beta_k = 0 \) this would imply that the mean of the outcome is the same comparing subpopulations differing in the \( k \)th independent variable by 1-unit. This means that no association exists between the outcome and \( x_k \). Conversely, if \( \beta_k \neq 0 \) then this would imply that the mean of the response is associated with a change in \( x_k \).
Thus it may be of interest to test the null hypothesis $H_0: \beta_k = 0$ versus the alternative hypothesis $H_1: \beta_k \neq 0$. In order to carry out this hypothesis test, we require a test statistic for which we know its distribution under the null hypothesis, $H_0$.

Consider the random intercepts model given by

$$y_{ij} = \beta_0 + \beta_1 x_{i1,j} + \ldots + \beta_p x_{ip,j} + \beta_{0i} + \epsilon_{ij}. \quad (20.19)$$

Statistical software can be used to obtain estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_p$ (as well as $\beta_{0i}$). Let $b_0, b_1, \ldots, b_p$ denote these estimates. Further, software can be used to estimate the standard error of $b_0, b_1, \ldots, b_p$, which we denote as $\text{se}(b_0), \text{se}(b_1), \ldots, \text{se}(b_p)$.

A test of the null hypothesis $H_0: \beta_k = 0$ versus the alternative hypothesis $H_1: \beta_k \neq 0$, $j = 0,\ldots, p$, can be conducted with the following test statistic

$$z = \frac{b_k}{\text{se}(b_k)} \sim N(0, 1). \quad (20.20)$$

The value of $z$ can be looked up in Table I and the resulting $p$-value is the probability of observing a result as or more indicative of the alternative hypothesis if the null hypothesis were true.

**EXAMPLE CONTINUED: COMPARING SLOPES OVER TIME ACROSS DIAGNOSTIC POPULATIONS**

To test the hypothesis that cognitively normal patients have different slope than MCI patients using a model of the form

$$\text{Logic. Mem}_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 I_{i,\text{MCI}} + \beta_3 I_{i,\text{AD}} + \beta_4 t_{ij} \times I_{i,\text{MCI}} + \beta_5 t_{ij} \times I_{i,\text{AD}} + \beta_{0i} + \epsilon_{ij}, \quad (20.21)$$

first recall that $\beta_1$ represents the slope of the Logical Memory test score over time among cognitively normal patients, while $\beta_1 + \beta_4$ represents the slope of the Logical Memory test score over time among MCI patients. Thus if $\beta_4 = 0$, that would imply that the slope of the cognitive score is the same among cognitively normal and MCI subjects. Hence to compare the slopes between these groups, we would want to test the hypothesis $H_0: \beta_4 = 0$ versus the alternative hypothesis $H_1: \beta_4 \neq 0$.

**EXAMPLE COMPLETED: RETEST EFFECTS IN ALZHEIMER’S DISEASE**

Table 20.17 displays the estimates that come from fitting the random intercepts model given in Eq. (20.21) to our data. $b_1$ represents the slope of Logical Memory test scores among cognitively normal individuals. Thus we estimate that, on average, Logical Memory test scores rise by approximately 0.21 points with each year the test is given to cognitively normal patients. A test of $H_0: \beta_1 = 0$ versus the alternative hypothesis $H_1: \beta_1 \neq 0$ yields a test statistic of 34.74 and the resulting $p$-value for this test is $< 0.001$. Hence, we reject the null hypothesis and conclude that, on average, test
scores are significantly rising among cognitively normal patients, indicating a retest effect. In addition the estimated slope for AD subjects is given by $b_1 + b_4 = 0.21 - 0.45 = -0.24$. Hence our estimate of the slope among AD patients is negative, implying that retest effects may be less of an issue in this subpopulation. To test whether the slope in AD subjects is different from cognitively normal patients, we test $H_0: \beta_5 = 0$ versus the alternative hypothesis $H_1: \beta_5 \neq 0$. This results in a test statistic of $-28.36$ and the resulting p-value for this test is $<0.001$. Thus we conclude that AD patients have a significantly lower slope when compared to cognitively normal patients. Notice that all estimates are very close to those obtained from the general linear model. This is because it can be shown that the random intercept model also induces an exchangeable correlation structure like was used in the general linear model example.

**Exercise 20.3**

Retest effects among MCI patients. Using the random intercept model estimates provided in Table 20.17, answer the following questions:

a. What is the slope of Logical Memory test scores over time among MCI patients?

b. Based upon the above estimate, does there appear to be a retest effect among MCI patients?

c. Is the slope of Logical Memory test scores over time among MCI patients significantly different than that of cognitively normal patients?

d. Compare your results to those given in Table 20.16, obtained using the general linear model with an exchangeable covariance structure.

**20.4 TIME-SERIES:**

Detecting patterns

“Long-series” data appear most often as a sequence of data through time, so that most analytic methods for long-series data are found under the time-series heading. Types of long sequences other than time-dependent are also important and will be illustrated by examples, but time will be referred to as the independent variable for convenience.
Some examples of time-dependent data from medical literature are incidence of nosocomial infection in a hospital; wave-form analysis of neurologic potentials in infants; the course of Ménière’s disease through time; course of opioid use during bone marrow transplantation; seasonal variation in hospital admissions for specific diseases; critical care monitoring; analysis of respiratory cycles; velocity of eye movements during locomotion; trends in juvenile rheumatoid arthritis; movement effects in MRI; and assessing the level of anesthesia. Indeed, time-series are encountered in almost every field of medicine.

**QUESTIONS ADDRESSED BY TIME-SERIES**

The questions most commonly asked of time-series are as follows: (1) is there a trend in the event through time? Identifying the best-fit regression curve will give some information about trend through time (Chapter 15: Linear regression and correlation). (2) Is the event cyclic through time? Autocorrelation, the correlation of a data set over a period with a time-offset of itself, may detect cycles and describe their nature. (3) Is the event correlated with other events in time? Cross-correlation, the correlation between two data sets over time, may detect other events that follow related patterns through time. (4) Is there a point in time at which the event changes its pattern? Change-point estimation will be addressed in the “Testing Patterns” subsection.

**The time-series methods introduced here are basic**

As with much of statistics, there exist more sophisticated methods than will be introduced here. For example, a term sometimes encountered in contemporary medical articles is autoregressive moving average (ARMA), a weighted moving mean adjusted for the influence of autoregression. This chapter will have accomplished its purpose if the reader understands the basic ideas of time-series and can appreciate what can be done.

**The need for smoothing processes**

Very often, a potential pattern of a sequential event is obscured by variability, a classic case of the metaphor, “You can’t see the forest because of the trees.” The pattern is better discerned if the variability about the pattern is reduced, or “smoothed.” Smoothing is not used for very small samples, because patterns are not perceptible anyway, whatever the method. Smoothing becomes useful when used on sequences with dozens, or hundreds, of time-dependent data.

**EXAMPLES POSED: PROSTATE-SPECIFIC ANTIGEN, PROSTATE VOLUME, AND AGE**

Sequential data from DB1 are shown in Fig. 20.3A and B. In addition to the major ideas, these examples illustrate long-series data that behave like time-series but in which the independent variable is not time.
**RELATION OF PROSTATE VOLUME TO PROSTATE-SPECIFIC ANTIGEN**

Fig. 20.3 A shows the prostate volumes of 301 men presenting with urological problems as related to the rank order of prostate-specific antigen (PSA) levels in increasing sequence. Any pattern of change in volume with PSA is rather well obscured by the variability.
RELATION OF PROSTATE-SPECIFIC ANTIGEN TO AGE

In Fig. 20.3B, PSA levels from the same men are shown in the sequence of the men’s age. Again, any pattern of change is obscured by variability.

Method for moving averages

MOVING SAMPLES

Smoothing is based on the concept of a moving sample. Consider a sequence composed of many data, perhaps 100, and subsets of \( n \) data, say 10. We take the first 10 data and calculate some statistic, perhaps an average. Then we drop off the 1st datum, add the 11th datum, and recalculate the average. We continue, dropping the 2nd and adding the 12th, dropping the 3rd and adding the 13th, etc. We have a sample moving through the sequence, always with size \( n \), providing a statistic moving through the sequence. We may name this sequence of samples a moving sample and also apply “moving” to the statistic, as a moving average. The moving average mutes the variability that obscures patterns.

Types of moving average

There are two relevant moving averages: the moving mean and the moving median. The moving mean subdues the variability, bringing extremes closer to the overall path through time, but allowing them some influence. In a moving mean of 20 in a sequence of 1000 visual evoked potential readings, a “hump” of 50 beats due to a stimulus image being presented will still appear, although it will be somewhat muted. However, a single extreme reading due to a blink will also appear as a slightly enlarged mean. Patterns extending longer through time are reduced less than shorter patterns. The moving median, even one as small as three readings, completely eliminates single outliers, as the greatest reading in the three will always be replaced by the middle one. A moving median of five will eliminate paired outliers. Thus we think of a moving mean to allow causal influences to appear and of a moving median to remove outliers. Other more subtle benefits will appear to the user with practice.

Calculation of moving average

A moving mean or median is just the ordinary mean or median of the moving sample. However, because the sample drops the left-most reading and adds one to the right end, most of the last calculated average will be the same; it needs only to be adjusted for the two readings changed. As an example, we have readings 2, 4, 3, 5, 1, 6, and want a moving mean of 3. Let us name it \( m \) with a subscript as the rank of the middle digit. Then \( m_2 \) would be \( (2 + 4 + 3)/3 = 3 \) and \( m_3 \) would be \( 3 + (5 - 2)/3 = 4 \). The values of a moving median of 3 would be 3, 4, 3, 5. There are missing elements at the
beginning unless they are supplied by specially defined values. For example, in the
moving mean of 3 from the sequence just illustrated, we might define a first reading
\( m_1 \) as twice the first value plus the second value divided by 3, or \( (2 \times 2 + 4)/3 = 2.67 \).
Some statistical software packages and even Microsoft Excel provide moving mean
capability. Usually these packages place the moving average at the center of the mov-
ing sample. The scheme used here will aid in testing, addressed in the next section.
To plot this scheme with means at the center of the moving average of \( n \) points, just
plot \( m_k \) at position \( k - (n - 1)/2 \).

Serial correlation
Other types of patterns we might wish to discern may be detectable using time-
dependent correlation, technically named serial correlation. A relationship between two
variables through time may be identified by cross-correlation, and the repetition of a
pattern, or periodicity, by autocorrelation. These concepts are useful in epidemiology
(see Chapter 23: Epidemiology) as well as in clinical medicine.

Cross-correlation
If two matching data sets are taken sequentially through time, the correlation between
them is termed cross-correlation. It tells us how closely related the two variables are
through time. Blood pressure and cardiac output through a sequence of exercise and
rest would be expected to be crosscorrelated. The correlation is based on the relation-
ship between them, not on their individual behavior through time. Thus if they both
rise and fall together, the correlation is high whatever the time-dependent pattern
may be.

The calculation may be easily understood if we think of lining the two data sets in
adjacent columns of a table and calculate the ordinary correlation coefficient between
the columns. Also, cross-correlation can be calculated with one of the sets lagged
behind the other. For example, the appearance of symptoms of a disease having a
2-week incubation period can be correlated with exposure to the disease, where expo-
sure observations are paired with the symptom observations that occurred 2 weeks
later. By varying the lag, we may be able to find the incubation period. For example,
bacterial vaginosis has been found to facilitate AIDS infection. We would expect to
see a cross-correlation between incidences of the two diseases, although not a high
one because each occurs alone. In this case the issue is not the size of the coefficient
but the amount of lag at which it is maximum, which might provide insight into the
nature and timing of exposures. Again, the calculation may be thought of as finding
the ordinary correlation coefficient between variables listed in two columns of a table,
but now one column is slid down the table a number of rows equal to the desired lag.
Autocorrelation

Observations through time may be correlated with a lagged version of themselves. If the autocorrelation coefficient retreats from 1.00 as the lag increases but then returns to nearly 1.00, we know that we have a periodically recurring disease, as in symptoms of malaria. If the observed lag is 12 months, the disease is seasonal. In Chapter 16, Multiple linear and curvilinear regression and multifactor analysis of variance, we noted that aspergillosis is a seasonal phenomenon and found that mean incidence by season fitted a sine wave with a period of 1 year. We ask if older patients are more susceptible in winter. If we calculate the autocorrelation of mean age of infected patients per season with mean age 1 year later, we find the coefficient to be $-0.22$, which verifies what we expected, that the age at which patients are infected does not vary seasonally. We also ask if percent infection by sex is seasonal, which we do not expect. The autocorrelation for percent male with a 1-year lag yields an autocorrelation coefficient of $-0.14$, telling us that infection rate by sex is not seasonal. Interestingly, the cross-correlation coefficient of age and percent male is 0.58, telling us that older patients tend to be male.

EXAMPLES COMPLETED: PROSTATE-SPECIFIC ANTIGEN, PROSTATE VOLUME, AND AGE

Relation of prostate volume to prostate-specific antigen level

We need to smooth the data in order to perceive a pattern. We create a moving average (more precisely, a moving mean) of $n = 17$. The choice of 17 is somewhat arbitrary here; too small an $n$ leaves too much variability (too little smoothing of the pattern) and too large an $n$ obscures shorter aspects of the pattern and may leave too few observations in the time-series. The size of a moving sample is a compromise. The graph of that moving mean is superposed on the data in Fig. 20.4A. The resulting graph still varies up and down in what seems to be random cycles, but a pattern begins to emerge. The prostate volume seems to be increasing with PSA. Also, we might suspect that the increase is not quite in a straight line, but might be slightly curved, concave downward. We could fit an ARMA model to the data, but we will retain the simplicity of regression methods already met in Chapter 15, Linear regression and correlation. The fit of a first-degree (straight-line) model yields $p$-value $< 0.001$; we have strong evidence that volume increases with PSA throughout the entire PSA spectrum. However, $R^2 = 0.185$, telling us that, although PSA is a significant and important predictor of volume, it is certainly not a decisive predictor. Further, we test the proposal of a curvilinear fit. We fit a second-degree regression curve to it, using the model $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (see Section 16.5). The resulting $p$-value is $< 0.001$, as before, but $R^2$ remains at 0.185, indicating that the second-degree model prediction of the smoothed prostate volume by PSA rank adds nothing to the first-degree. We conclude that prostate volume increases with PSA over the entire PSA range.
A number of very large PSA values obscures any pattern that might be present. Some investigators might suggest deleting the large levels, but doing so would no longer leave the same data set; we might even be obscuring the pattern we wish to see. Can we mute the large values without destroying the pattern? We apply a moving median of 3, which is a wide enough moving sample to reduce the most extreme levels, since

---

**Figure 20.4** (A) Prostate volumes of 301 men depending on increasing levels of PSA with a moving mean of 17 superposed. (B) A moving median (of 3) of 301 men’s PSA levels depending on age followed by a moving mean (of 51) on the moving median result superposed on the data.

**Relation of Prostate-Specific Antigen to Age**

A number of very large PSA values obscures any pattern that might be present. Some investigators might suggest deleting the large levels, but doing so would no longer leave the same data set; we might even be obscuring the pattern we wish to see. Can we mute the large values without destroying the pattern? We apply a moving median of 3, which is a wide enough moving sample to reduce the most extreme levels, since
most large levels do not occur next to each other. Then we apply a moving mean of 51, chosen somewhat arbitrarily. The result of a moving median on the raw data followed by a moving mean on the moving median data is shown superposed in Fig. 20.4B. Note that the vertical axis has a very much smaller scale, allowing the local variation to be perceived. Inspection of the shape of the smoothed PSA by age reveals that it is not far from a straight line, but slightly curved, opening downward. Upon fitting an ordinary second-degree regression model, we find the $F$ statistic = 401.93, $p$-value < 0.001, and $R^2 = 0.730$, a rather good fit. We might interpret the result as PSA increasing with age until the late 60s, at which time it becomes stable.

**CROSS-CORRELATION OF PROSTATE-SPECIFIC ANTIGEN AND PROSTATE VOLUME AS DEPENDING ON AGE**

We noted that smoothed volume appears to depend on PSA and smoothed PSA appears to depend on age. Let us consider both smoothed PSA across age (as in Fig. 20.4B) and smoothed volume across age (a newly created moving average of 51). Do they have similar patterns of behavior as age increases? Fig. 20.5 shows a plot of the smoothed paths of volume and PSA plotted against increasing values of age. Both curves seem to increase at first, until age reaches the mid-to-late 60s, and then level out. The cross-correlation between these curves is calculated to be 0.727, indicating that the perceived similarity of pattern is real.

**ADDITIONAL EXAMPLE: PURITY OF DENTAL WASH WATER**

A large dental clinic was concerned about the purity of water being used to wash drilled teeth and small open wounds. Dental unit water tubing harbors bacteria-laden
biofilms, which often contain pathogens. Although technicians are instructed to use purified water and purge the lines with compressed air at scheduled intervals, they do not always follow these procedures with sufficient care. A 12-week study compared eight technician groups. The contamination level of water ejected from the spray head was examined by a pathologist and rated according to the number of colony-forming bacterial units found. Technician group and date were recorded. Fig. 20.6 shows data smoothed by a moving mean of 9 (initial data overlapped too much to perceive a pattern) in the order recorded (A) and further smoothed by a moving mean of 21 (B). Examination of the data in the vicinity of the peaks shows that group seven was using tap water, a discovery that might not have been made from the raw data. In addition, it appears that the curve averages a constant level at first but begins a steady rise about halfway through the study. Methods to test for a point of change will be examined in the following section.

**AUTOCORRELATION: IS THE CONTAMINATION PERIODIC?**

The tap-water peaks look as if they might be periodic rather than haphazard. The autocorrelation coefficient was calculated for various lags. The greatest coefficient, appearing at a lag of about 50, was 0.18. We conclude that tap-water usage is not periodic.

**Exercise 20.4**

The systolic time ratio (STR) indicates the strength and efficiency of the heart. Does STR relate to HR? Fig. 20.7A displays STR graphed against 228 patients’ HRs in increasing order. A linear regression shows no relationship, which just says that a straight-line fit does not have a significant slope, not that no pattern of relationship exists. Perhaps a smoothing process will reveal some relationship. A moving mean of 51 is superposed on the raw data. Fig. 20.7B shows the
moving mean with the vertical axis enlarged so that the pattern can be perceived. What hypothesis for further testing may be posed from (B)?

**Testing patterns**

In time-series, testing is mostly concerned with verifying a model form and with identifying change points. Testing the nature of a model has been addressed in sections on regression; this section will focus on testing for a point of change from one model to another.
Methods of change-point identification

Most change-point methods require a baseline against which to compare evolving data. In baseline-based methods a statistic triggered by a change from baseline in the longitudinal process must be calculated. In one exception a current data-based method a change in the functional form of a smoothed sample must be detected. One convenient categorization of most frequently used change point tests is by type of sampling scheme from which this triggering statistic is calculated. These may be denoted as (1) periodic samples, (2) accumulating samples, (3) smoothed samples, and (4) moving samples. Each has some advantages and some disadvantages.

Periodic samples

Although periodic or sporadic sampling usually costs less than complete sampling and it allows destructive sampling, it will detect a change but fails to locate it.

Accumulating samples

In cumulative sampling the concept is to accumulate errors until the sum becomes improbably large. This form of analysis has memory: an unusual deviation from the expected path is never forgotten, but neither is the influence of an outlier. An accumulation of no-longer-relevant events, or of outliers, may trigger an indication of change where there is none. The dominant methods of this type are named CUSUM, in which the probability of the observed cumulative sum is found, and EWMA, a control chart scheme based on an exponentially weighted (geometric) moving average. CUSUM appears to be the most frequently applied change-point method, although EWMA weights data less as they recede into the past, so that it is less limited by memory of irrelevant events. Comparisons show little difference in the decision outcomes of the two methods.

Smoothed samples

Current-data methods detect patterns that emerge from smoothing by a moving-sample regression fit. Independence from baseline data is the advantage of this method. However, it detects only a limited set of change types, for example, jumps in the moving average but not changes in variance or the nature of the model, nor does it detect outliers.

Moving samples

Moving samples, or more exactly, a moving sample compared to a baseline sample, are of two distinct types, Markov chains and Moving F.

Markov chain methods were addressed in Chapter 20 of the third edition of this book and are touched on in Sections 28.6 to 28.8 in this fourth edition. Its more
complete coverage is omitted from this edition because it is so much more complicated than other methods given as to be inappropriate and seldom used by medical researchers. Some time-series can be imbedded in Markov chains, posing and testing a likelihood model. The sophistication to Markov chain Monte Carlo (MCMC) addresses the widest variety of change-point issues of all methods and will solve a great many problems other than change-point identification. On the other hand, not all time-series are susceptible to MCMC analysis, and it is a method that requires considerable mathematical ability and sophisticated software, coupled with the intuition for good model building, a talent more rare than we would wish. We deem it too complicated to use for simple change-point identification.

By contrast, Moving \( F^{7,8} \) is a rather simple test, but one that seems to work well for most change-point purposes. In this test the variance of a moving sample is divided by the variance of the baseline sample, creating a Moving \( F \) statistic. The point at which this Moving \( F \) first exceeds a critical value as it moves through time both detects and locates a change point.

The Moving \( F \) tests for a change on every move from each datum to the ensuing datum. When the test succeeds, its success generates the estimate of the point of change. Thus it was initially termed testimation when Riffenburgh developed it in the 1960s.\(^9,10\)

In contrast to other change point detectors, Moving \( F \) will detect any sort of change in a time-series, including jumps (trauma has changed a physiological process), changes in slope (a disease begins), changes in variability (laboratory test results change standard deviation when a disease worsens), and changes in model (a lab test result changes from approximately constant to an exponential increase when cancer starts). In addition, outliers are obvious. For example, a decimal point may be left out of a lab test result, yielding a spurious datum.

Because Moving \( F \) is the simplest of the change-point identifiers with wide generality for the types of change identified, this method and several examples will be detailed here.

**EXAMPLE POSED: LEVEL OF MEDICAL TECHNICAL SUPPORT**

The US Navy Hospital Corps is composed of enlisted technicians who support medical care operations. Corpsmen provide basic medical service aboard ships, to marines in land action, and at ashore bases. The size of the Hospital Corps through time presents an interesting commentary on the politico-military history of the United States. Fig. 20.8 shows the strength of the Corps during the 20th Century.\(^11\) Growth spurts can be seen at the times of WWI, WWII, and the Korean War, after each of which the Corps was maintained at higher manning than before. We hardly need tests. However, were the seeming increases at the times of the Vietnam War, the military buildup of the 1980s (what might be thought of as an economic war against the
Soviet Union), and the First Gulf War real increases? If not, we would model post–Korean War manning level as a constant with random fluctuations. Let us detect and locate any change points in that period.

Moving F method

MOVING SAMPLE DESIGNATIONS

A moving sample is composed of a sequence of (mostly overlapping) values, say $x$-values. We need to keep a track of which member of the sequence of samples we are dealing with at any moment. Let us designate by $k$ the rightmost value in a moving sample of size $n$, so that the sample elements are $x_k - n + 1$, $x_k - n + 2$, $\ldots$, $x_k$. The moving mean is $(x_k - n + 1 + x_k - n + 2 + \cdots + x_k)/n = m_k$. The first moving mean is $m_1$, the second $m_2$, etc. $m_k$ is just the mean of $n$ data ending with number $k$. Similarly, we could think of a moving variance $s_k^2$ as just the sample variance of the same $n$ data.

A reader might perceive the methods more easily by visualizing an application such as monitoring exhaled nitric oxide (eNO) to indicate bronchoconstriction in asthmatics.

CONCEPT OF THE SIMPLE MOVING F TEST

Suppose we start a time-series with a stable baseline of $b$ data, denoting data positions as 1 through $b$. This baseline has variance $s_b^2$ (naming it for its rightmost datum, $b$). Recall that an $F$ test is just the ratio of two variances (assuming the data are approximately normal). If we start with datum $b + 1$ and calculate the sequence of moving variances $s_k^2$ of size $n$, dividing each by $s_b^2$, we have a Moving $F$ statistic traveling through the time-series. Upon comparing (Continued)
this with a critical value of $F$ from Table V (see Tables of Probability Distributions) for $df$ $(b - 1, n - 1)$, usually for $\alpha = 0.05$, we can see where the variability exceeds baseline variability so much that it is unlikely to have occurred by chance. This is a Moving $F$ test.

The Moving $F$ test will detect a sudden decrease in eNO (a beginning of bronchoconstriction), a change in the rate of decrease of eNO (bronchoconstriction is worsening), or a change in the variability (higher highs and lower lows).

The reason for denoting a moving sample by its leading member

A moving statistic does not occur at a point on the time axis but is calculated from an interval. As we move through the series, the first point at which a new model form has changed the moving sample enough to trigger significance will be taken as the point of change. Thus if the moving statistic becomes significant at leading point $k$, $k$ is the change point.

CALCULATION OF MOVING $F$ USING SOFTWARE

The Moving $F$ method is considered relatively simple because it can be handled by users outside the field of statistics using widely available software. For example, a medical investigator can perform a Moving $F$ in a few steps on Microsoft Excel software or on a statistics package, such as Stata (Stata Corporation, College Station TX).

1. Start with a spreadsheet having its first column as time (or other independent variable) $t_i$ and its second column as readings $x_i$. Choose a baseline sample of size $b$, cutting and pasting these $b$ data into the third column, leaving the remainder as blanks.

2. Regress these first $b$ $x$'s on $t$ to find the functional form of the model say $f$, perhaps a constant or a sloped line, and the baseline variance about this model, $s^2_b$. $s^2_b$ is usually designated as the residual mean square in the software’s regression display.

3. Make a fourth column of $f$ (for “function”) values, $f_i$, such as by substituting $t_i$ in the regression equation for all $i$.

4. Make a fifth column $(x_i - f_i)^2/(n - 1)s^2_b$. This is the standardized squared deviation of a reading from the predicted function. If we denote it as a deviation column, we now have columns of values for time, readings, baseline, predicted function, and deviation.

5. Use the software’s moving average capability to create the Moving $F$ as moving average of the deviation column. The Moving $F$ values should appear as a sixth column that can be used for graphing as well as making decisions.
A two-tailed test

We may be concerned not only with an increase in blood pressure but also a decrease. We could split the \( \alpha \), for example 5\%, into two 2.5\% areas and identify an upper and a lower critical \( F \), the lower value having 97.5\% of the \( F \) distribution greater than this value and the upper value having 2.5\% greater. (A more complete \( F \) table than given in this book or software able to calculate probability functions would be required.) Then we could detect either significant increases or decreases.

Multiple causes of variability are possible

The Moving \( F \) considered to this point will test for any change or combination of changes in the process:

- a change in the shape of the process (the form of the model)
- a shift in average
- a change in the variability

Often, we need to identify which influence is causing the significant \( F \). Consider an electronic monitoring instrument. Its readings may have become more variable. Alternatively, it could have shifted its setting while maintaining the same variability about the setting. And as yet another possibility, it could have changed from a straight line to a curve. Any of these events would trigger an increase in \( F \) in comparison with the baseline. How are these different sources of variability separated and identified? The numerator of the moving sample’s variance is the sum of squares (SS) of deviations of \( x \) values from the baseline mean, say \( m_b \) for simplicity, or \( SS_k = \sum (x_i - m_b)^2 \). We can add and subtract the moving mean value \( m_k \) from \( x_i \) to obtain \( \sum (x_i - m_k + m_k - m_b)^2 \), which can be shown mathematically to be \( \sum (x_i - m_b)^2 + n (m_k - m_b)^2 \), a component due to randomness at the moving sample’s position in the time-series, and a component due to the difference between the baseline mean and the mean at the moving sample’s position. When divided by \( df \), they become variances. Taken in ratio to the baseline variance, we have one Moving \( F \) due to mean shift and one due to random variability. Each may be compared with critical \( F \)-values to form a test. (It is also possible to separate out and test a component due to the shape of the process, but that subtlety will not be pursued here.)
EXAMPLE COMPLETED: LEVEL OF MEDICAL TECHNICAL SUPPORT

Let us choose a moving sample of size $n = 7$, long enough for smoothing but not so long as to use too much of the time-series. The 7-year mean level 1955–61 is taken as baseline ($6\ deg\ f$). Its mean is $m_b = 23,310$ and variance $s_b^2 = 457,410.8$ (not much different from the regression’s residual variance). We begin a moving sample of $7\ (6\ deg\ f)$ with 1962, dividing its variance by the residual variance to create a Moving $F$ statistic. From Table V, we see that a critical value of $F$ at $\alpha = 0.05$ and $6.6\ deg\ f$ is 4.28. Fig. 20.9 shows the Moving $F$ for the years 1962–98. The increase in manning level becomes apparent in 1964, a year after President Johnson assumed the presidency. The buildup and decline of military support for the Vietnam War can be followed year-by-year. Manning level returns to baseline for 1975 under President Ford and remains there during President Carter’s administration. It increases in the 1980s under President Reagan’s guidance during the arms race with the former Soviet Union, becoming significant in 1987. It remains significant during the period of collapse of the Soviet Union, which runs into the First Gulf War. Manning level declines at the end of that war, losing significance in 1993 and declines again to baseline during President Clinton’s administration.

ADDITIONAL EXAMPLE: HEART RATE BY TYPE OF ANESTHETIC IN TRAUMA

Morphine given intravenously (IV) to a severely injured patient with extensive blood loss impairs the already traumatized vascular system. It was hypothesized that intrathecal (IT) administration of morphine (injection under the nerve-covering sheath) would mitigate that impairment. An experiment\(^7\) used a sample of two pigs, one assigned to...
IV and the other to IT injection. A number of variables were measured for a 30-
minute baseline and a 60-minute simulated hemorrhage. The hemorrhage was stopped
and morphine injected. Measurements were continued for 180 minutes more.

Figure 20.10 (A) Heart rates for IV and IT morphine at baseline, during hemorrhage, and posthemorrhage. (B) Moving F of HR for IV and IT morphine over same periods. Both subjects show significantly increased HR during hemorrhage but IV morphine at the end of hemorrhage exacerbates HR even further, whereas IT morphine allows HR to return to normal. HR, Heart rate; IT, intrathecal; IV, intravenously.

IV and the other to IT injection. A number of variables were measured for a 30-
minute baseline and a 60-minute simulated hemorrhage. The hemorrhage was stopped
and morphine injected. Measurements were continued for 180 minutes more.  

Fig. 20.10A shows HR for the two pigs during this period. The variance of the first
15 prehemorrhage HRs averaged for the two pigs was taken at baseline. The moving
sample also was taken as size 15. From Table V (interpolated), the 95% critical value
of $F$ for 14,14 $df$ is 2.48. Fig. 20.10B shows the plots of Moving $F$ for IV and IT
treatments.
Although both curves rise above the critical value during hemorrhage, the $F$ for IV morphine rises even further after the end of hemorrhage and the $F$ for IT morphine drops back to the vicinity of baseline. The evidence indicates that IT morphine is better for the vascular system of trauma patients following hemorrhage.

**Exercise 20.5**

*Harold Shipman was a rural physician in Britain who was discovered in 1998 to have murdered a large number of his patients. A recent article* provided data on the cumulative number of death certificates by sex that he signed in excess of the local average, totaling 224. Starting from 1978 with the first excess, data were read from the graph, sexes pooled, and total number of excess certificates counted by year. The first four (1978—1981) were used as baseline and a constant baseline was assumed. The baseline variance was used as the denominator of the Moving $F$. A Moving $F$ of three data, calculated for 1982—98, is shown in Fig. 20.11. Find the critical $F$-value and identify where the number of excess death certificates first becomes significant in probability. How many years of murdered patients would have been prevented if Dr. Shipman had been stopped then?

**REFERENCES**

Sample size estimation

21.1 ISSUES IN SAMPLE SIZE CONSIDERATIONS

Why are we concerned with sample size?

How large a sample do we need? Speaking statistically, *the larger the sample, the stronger the statistical conclusions*. Larger samples provide better estimates, more confidence, and smaller test errors. From a statistical view the best sample size is one that takes all the data our time, money, support facilities (hospital, animal lab), and ethics of patient use will permit. Then why all the attention to methods for estimating the minimum sample size required? There are two primary purposes.

One purpose is to *verify that we will have enough data* to make the study worthwhile. If we can manage only 50 subjects and the sample size requirement methods show we need 200, we should not undertake the study.

Another purpose arises in cases in which cost (funds and/or resources) or subject ethics mandate that we minimize sample size: *to ensure that we do not sample many more subjects than required to answer the question the study poses*. For example, if 100 subjects, split between a trial drug and a placebo, are enough to show that the new drug is effective, a sample of 200 would deny the benefit of the drug to the additional 50 placebo patients.

Concept of estimating the minimum required sample size

Estimation of the minimum sample size required for a decision is not a single unique method, but the concepts underlying most methods are similar. Let us look at the method for a test of means. Fig. 21.1 shows distributions for null and alternative hypotheses for samples of size \( n \) (A) and size \( 4n \) (B). The size of the difference, \( \delta \), between mean treatment outcomes that will answer the clinical question being posed is often termed clinical significance or, better, *clinical relevance*. For such a clinically important difference between means, the sizes of error probabilities (\( \alpha \): the risk of a false-positive, and \( \beta \): the risk of a false-negative) are illustrated by the shaded areas under the curves.
When the sample is quadrupled (moving from A to B), the standard error of the mean is halved, so that the curves become more slender (smaller standard error of the mean—SEM or $\sigma_m$), overlap less, and consequently yield smaller error probabilities.

Figure 21.1 Distributions for null and alternate hypotheses in a test of means (A) with sizes of associated error probabilities indicated by shaded areas (see Section 7.2). If the sample size is quadrupled,

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

shrinks to half, yielding an equivalent diagram (B) with more slender curves. Note how much $\beta$ has shrunk for a fixed $\alpha$. The method of estimating a minimum sample size to yield specified error sizes for detecting a given distance between means is to start with those error sizes and that distance and “back-solve” the relationships to find the associated $n$.

When the sample is quadrupled (moving from A to B), the standard error of the mean is halved, so that the curves become more slender (smaller standard error of the mean—SEM or $\sigma_m$), overlap less, and consequently yield smaller error probabilities.
The method of estimating a minimum sample size is to specify the clinically relevant
difference between means and the required chances of error and then to find the \( n \)
that satisfies these specifications.

**The term power analysis**

A note on terminology: \( 1 - \beta \) is called “power.” As \( \beta \) is the probability of falsely fail-
ing to reject the null hypothesis, power is the probability of correctly rejecting the
null hypothesis. This approach to estimating the minimum sample size required is
often termed *power analysis* because power is one of the specifications it depends on—
along with \( \alpha \), \( \delta \), and standard error.

**Value of very small samples**

Much is said about small samples in medicine. Although derision directed toward gen-
eralizing from “samples of size one” (anecdotal data) is usually deserved, nevertheless
the first datum encountered in a situation of complete ignorance provides the greatest
amount of information from one datum an investigator will encounter. Consider
the appearance of a new disease for which there is no knowledge whatsoever about
the incubation period. The first patient for whom we know the time of exposure
and the time of appearance of the disease syndrome provides useful information. This
first incubation period datum gives us an order of magnitude, improves our study plan-
ing, and suggests many hypotheses about disease patterns such as infectious sources.

**Effect of increasing the sample size**

Let us see what happens in estimating the minimum sample size as we gradually
increase our sample size from 1 to its final \( n \). Intuitively we observe that the informa-
tive value per datum decreases as the number of data increase. Consider how much
more is learned from data 1–5 than from data 101–105. There are, in fact, theoretical
reasons to say that the amount of information increases as the square root of the sam-
ple size, that is, \( \sqrt{n} \). For example, to double the information about an event, we must
quadruple the data.

**Convergence**

At some point, enough data are accumulated that the sample distribution only negligi-
bly differs from the population distribution, and we can treat the results as if they
came from the population; the characteristics of the sample are said to *converge* on
those of the population as the sample size approaches the population size.
Clinical relevance and patient care
The clinical relevance, often denoted as $\delta$ or $\Delta$, usually is the statistical parameter that most influences the sample size. It may also affect patient care. Because a larger difference will require a smaller $n$, the temptation exists to maximize the difference to allow for a small enrollment with subsequent early closure. This clearly is statistical tampering, as the choice is made on statistical rather than clinical grounds and begs an ethical question: if the proposed new therapy really is so much better than the current one, how may the researcher in good faith not offer the patient the superior course? The difference should be chosen with patient care as part of the consideration, that is, on clinical as well as statistical grounds.

Sequential analysis in relation to sample size estimation
The concept of sequential analysis (see Chapter 22, Clinical trials and group sequential analysis) is to test the hypothesis successively with each new datum. The result of the test falls into the classes: accept hypothesis, reject hypothesis, or continue sampling. At first, due to a very small sample size, the test result falls into the continue-sampling class. Sampling is continued until it falls into one of the other classes, at which point it is stopped. The purpose is to minimize the sample size required in a study. While it does serve this purpose, it does not give the investigator any advance idea of a required fixed sample size. On the contrary, sample size is random.

Interim testing in relation to sample size estimation
Lying between the concepts of sequential analysis and a single final test is interim testing, that is, testing once or several times during the course of the study. Interim analysis will be explored in Chapter 22, Clinical trials and group sequential testing. Estimating a minimum required sample size for interim testing is rather involved and will not be pursued here.

21.2 IS THE SAMPLE SIZE ESTIMATE ADEQUATE?
The power analysis’s sample size is just a (rather poor) estimate
Whatever method we use to estimate the sample size required for a study, in an ideal world this sample size would be estimated on the results of that study. Thus we cannot know the sample size until we do the study, and we do not want to do the study until we know the sample size. The usual solution to this “Catch 22” is to use inputs to the sample size equation drawn from other sources, such as a pilot study or parameter estimates quoted in the literature. Because these are not the actual data from our study, there is a wrong data source of possible error in our estimate. Furthermore, sample size is estimated on the basis of data that are subject to randomness, so that there is also a randomness source of possible error in the estimate. And, of course, the difference
δ to be detected arises from clinical judgment, so that a judgment error is possible. Finally, α and β are usually designated by convention rather than by the needs of the study. Thus because all of the inputs into a power analysis are dubious or subject to error, the estimated sample size is dubious. We must take it as an order of magnitude, not as a precise number.

A “safety factor” is advisable

Because the estimated sample size represents the very minimum allowable, whatever the method used, we should add a “safety factor” to the estimated required sample size to account for the uncertainty of the predicted sample requirement. The size of this safety factor is an educated guess. The more uneasy we are about how well the data used in the sample size estimate will agree with the ultimate study data, the larger the safety factor we should allow. The resulting estimate of sample size is a mixture of guess and statistics, a number that may be taken only as a rough indicator.

21.3 THE CONCEPT OF POWER ANALYSIS

The logic behind the method

An emergency medicine physician knows from large-sample historical evidence that mean heart rate (HR) from a particular healthy population is μ = 72 bpm (beats per minute) with standard deviation σ = 9.1 bpm. She wants to know whether mean HR from a population of patients who have just been exposed to a particular type of toxin is the same or greater. She plans to collect data from the next sample of patients who present with that toxic exposure and test the sample mean against the population mean. She needs to estimate the minimum sample size she will need. She designates the unknown population mean μs (subscript s for “sample’s distribution”). The hypotheses tested will be

\[ H_0 : \mu_s = \mu \text{ versus } H_1 : \mu_s > \mu. \]

From the data that will be forthcoming the value of \( \mu_s \) will be estimated by the sample mean \( m \). Thus to decide whether \( \mu_s > \mu \), our test statistic will compare \( m \) to \( \mu \). Fig. 21.2 shows the two distributions involved. Because they are standardized, the horizontal axis quantities are given by mean HR differences divided by the standard error. The left distribution is the null, with its mean \( \mu \) standardized to 0 indicated by a vertical line. The right is a possible alternative, with its mean \( \mu_s \), estimated by \( m \), indicated by a vertical line at about 3. σ is the standard deviation from the population data, so \( \sigma_m \), the standard error of the mean (the standard deviation of distribution shown in the figure) is \( \sigma / \sqrt{n} \), where \( n \) is the sample size we ultimately seek. To derive the sample size we use a form very much like a test for a significant difference between population means, \( \mu_s - \mu \), that would focus on the difference \( m - \mu \). There are a few
important exceptions, however. First, since we do not yet have data to calculate \( m \), we replace this with a postulated clinically relevant difference, \( \delta \). Second, the error risks are inputs and the sample size \( n \) is the output. For this discussion, we will use \( \alpha = 5\% \) and \( \beta = 20\% \) (power is \( 1 - \beta = 80\% \)). The critical value (here \( \mu + 1.645 \sigma/\sqrt{n} \)) is the position separating the two types of error, shown in Fig. 21.2 as the number of standard errors above \( \mu \) that yields a 5\% \( \alpha \) (area under the tail of the null distribution; 1.645 is the value from the normal table, Table I, for \( \alpha = 5\% \)). Similarly, \( \beta \) is the area under the tail of the alternative distribution and is specified by the number of standard errors below \( \mu_s \) that the critical value lies, or \( \mu_s - 0.84 \sigma/\sqrt{n} \) (0.84 is the value from Table I for 20\% in the tail area). Because the axis value of these two expressions both equal the critical value, we set them equal to each other, or \( \mu + 1.645 \sigma/\sqrt{n} = \mu_s - 0.84 \sigma/\sqrt{n} \). Solving the equation for \( n \) yields.

\[
\begin{align*}
n &= \frac{(1.645 + 0.84)^2 \sigma^2}{(\mu_s - \mu)^2} = \frac{(1.645 + 0.84)^2 \sigma^2}{\delta^2}.
\end{align*}
\]
Other formulas for minimum required sample size follow equivalent logic. In particular, the formula for the case of two means (in Section 21.6) uses a formula recognizably similar to Eq. (21.1).

**A list of inputs needed for a power analysis**

Summarizing from the preceding paragraphs, a power analysis needs values for (1) the risk of a false-positive, \( \alpha \); (2) the risk of a false-negative, \( \beta \); (3) the difference, \( \delta \), between the parameter values we want the detect (usually a value less than which is clinical indifference); and (4) the standard deviation of the test statistic.

**Choosing test sidedness**

We should note that sample size estimation may be based on a one- or two-sided alternative hypothesis, just as is the test for which the sample size is being estimated. In the methods of this chapter the more commonly used two-sided form of the normal tail area (\( z_{1-\alpha/2} \)) is given. For one-sided cases, just replace \( z_{1-\alpha/2} \) by \( z_{1-\alpha} \) wherever it appears. For example, replace the two-tailed 5% \( \alpha \)'s \( z = 1.96 \) by the one-tailed 5% \( \alpha \)'s \( z = 1.645 \). The effect on the patient should be considered in selecting sidedness. When a two-sided test is appropriate, a one-sided test doubles the error rate assigned to the chosen tail, so that too large a number of healthy patients and too small a number of ill patients will be classified as ill. Choice of a two-sided test when a one-sided test is appropriate creates the opposite errors.

**Exercise 21.1**

Should minimum sample size estimation be based on a one-sided or a two-sided test, assuming the underlying distributions are normal, of (a) a difference in nausea scores between ondansetron hydrochloride and placebo in DB2 and (b) the difference between assay types in DB8?

**Choosing test parameters**

Although \( \alpha = 5\% \) and power = 80\% \( (\beta = 20\%) \) have been the most commonly selected error sizes in the medical literature, 90\% power \( (\beta = 10\%) \) is becoming more frequently used. Furthermore, a \( \beta/\alpha \) ratio of 4/1 may affect the patient. When the false-positive is worse for the patient than the false-negative, as in a case of testing a drug used for a non-life-threatening disease but that has severe side effects, the common choices of \( \alpha = 5\% \) and \( \beta = 20\% \) are not unreasonable. On the other hand, in testing treatments for cancer, failing to treat the cancer, which has rate \( \beta \) (false-negatives), is more serious for the patient than unnecessary cancer tests, which has rate \( \alpha \) (false-positives), and the ratio \( \beta/\alpha \) should be decreased.
CHAPTER 21 Sample size estimation

Exercise 21.2
Assess the clinical effects of false-positive and false-negative outcomes and specify whether $\beta$ should be the usual $4\alpha$ in estimating minimum sample size for a test of (a) the effect of ondansetron hydrochloride on nausea following gall bladder surgery in DB2 and (b) a change in plasma silicone following implant in DB5.

21.4 SAMPLE SIZE METHODS
Each test has its own formula
The method discussed here used normal distribution theory and the population or large-sample $\sigma$. Small sample theory would use the $t$ distribution. A test of two means would have to use different formulas that combine the standard deviations. A test of proportion is based on the binomial distribution, that has a known relationship between $\mu$ and $\sigma$, so the standard error need not be specified.

The point is that each test has its own form of power analysis. We cannot estimate the minimum required sample size before we know the form of analysis to be used.

Sample size methods addressed
Minimum sample size methods are discussed for the following statistical techniques in Sections 21.5—21.17, respectively.

Continuous data
- Test on one mean (normal distribution) (Section 21.5)
- Test on two means (normal distributions)
- Test on means from poorly behaved distributions
- Test on means in the presence of clinical experience but no objective prior data
- Confidence interval on a mean

Categorical data
- Test of a sample proportion against a theoretical proportion
- Test of two sample proportions (equivalent to using a $2 \times 2$ contingency table)
- Confidence interval on a proportion
- Test of a correlation coefficient

Ranked data
- Variance tests, analysis of variance (ANOVA), and regression
- Equivalence tests
- Number Needed to Treat (NNT)

NNT or Number Needed to Benefit (NNB) is a related concept as in the number of mammograms required to detect one breast cancer that screening would otherwise have missed.
Minimum sample size estimation methods for other statistical techniques exist, some developed with considerable thoroughness and rigor and others not. Several require specialized software. New approaches appear in the statistical literature from time to time. This book attempts to address only the most frequently used methods.

### 21.5 TEST ON ONE MEAN (NORMAL DISTRIBUTION)

**EXAMPLE POSED: HORMONAL THERAPY FOR BENIGN PROSTATIC HYPERPLASIA**

We want to determine if benign prostatic hyperplasia (BPH) patients treated with an experimental hormonal therapy have larger prostates than patients without BPH. The prostate volume (milliliters) of our 296 patients without BPH (see DB1) has mean $\mu = 35.46$ and standard deviation $\sigma = 18.04$. What is the smallest sample size that can detect a difference in the true mean of the experimental group at least 10 mL larger than $\mu$ (implying a one-sided alternative hypothesis)? We choose $\alpha = 0.05$ and $\beta = 0.20$, as is common in medical applications.

**METHOD**

What size of sample do we need in order to decide if $\mu_S$ is different from $\mu$? (More exactly, we are trying to determine whether or not the mean of the sample’s population as estimated by $\mu_S$ is different from $\mu$.) Eq. (21.2) is derived in the same manner as Eq. (21.1), but with risks for any size $\alpha$ and $\beta$.

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\delta^2}\right)^2 \sigma^2$$

Choose the smallest distance $\delta$ between $\mu_S$ and $\mu$, that is, $\delta = \mu_S - \mu$, that you want to detect with statistical significance. Choose the risk you will accept of concluding there is a difference when there is not ($\alpha$) and of concluding there is no difference when there is ($\beta$, or 1 – power). Look up the $z$-values in Table I associated with these two risks, $z_{1-\alpha/2}$ and $z_{1-\beta}$. Substitute the $z$-values and $\delta$ along with the standard deviation $\sigma$ in Eq. (21.2) to find $n$, the minimum sample size required.

**EXAMPLE COMPLETED: HORMONAL THERAPY FOR BENIGN PROSTATIC HYPERPLASIA**

The selection of $\alpha = 0.05$ and $\beta = 0.20$ yields $z_{1-\alpha} = 1.645$ and $z_{1-\beta} = 0.84$. Substitution in Eq. (21.2) yields $n = \frac{(1.645 + 0.84)^2 \times 18.04^2}{10^2} = 6.175 \times 325.44/100 = 20.1$; we must round up to 21. In line with Section 21.2, we should choose a slightly larger $n$. 
**ADDITIONAL EXAMPLE: TREATMENT OF EMERGENCY DYSPEPSIA**

An emergency medicine physician wants to test the effectiveness of a “GI cocktail” (antacid plus viscous lidocaine) to treat emergency dyspeptic symptoms as measured on a 1−10 pain scale. The standard deviation without treatment has been scored for a large number of patients as $\sigma = 1.73$. He considers a reduction in pain rating of $\delta = 1.5$ points as clinically meaningful. He believes that the treatment cannot increase pain; therefore the test will be one-sided. By taking $\alpha = 0.05$ and $\beta = 0.20$ (power = 80%), $z_{1-\alpha} = 1.645$ and $z_{1-\beta} = 0.84$. By substituting in Eq. (21.2), he finds

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\delta^2} = \frac{(1.645 + 0.84)^2 \times (1.73)^2}{1.5^2} = 8.21.$$ 

He sees that he needs only nine patients to be able to detect significance. However, he anticipates that the post treatment standard deviation might be larger, and he wants a safety factor in line with Section 21.2; therefore he decides to sample 14 patients.

**Exercise 21.3**

An emergency medicine physician wants to know if mean HR following a particular type of trauma differs from the healthy population rate of 72 bpm. He considers a mean difference of 6 bpm clinically meaningful. He takes $\sigma$ to be the 9.1 bpm reported in a large study. How many patients will he need? Use $\alpha = 0.05$ and power = 0.80.

**21.6 TEST ON TWO MEANS (NORMAL DISTRIBUTION)**

**EXAMPLE POSED: RANGE OF MOTION IN ARTIFICIAL KNEE**

Two types of artificial knee are to be compared for range of motion (measured in degrees). Theoretically, either could give a greater range, so a two-sided alternative hypothesis is appropriate. The hypotheses to be tested become $H_0$: $\mu_1 = \mu_2$ versus $H_1$: $\mu_1 \neq \mu_2$. A journal article on the first type of knee gave $m_1 = 112^\circ$ with $s_1 = 13^\circ$, and an article on the second gave $m_2 = 118^\circ$ with $s_2 = 11^\circ$. If we want to carry out a prospective randomized clinical trial to have adequate power to detect a $6^\circ$ difference in mean range of motion between types, what is the minimum number of patients receiving each knee we must record?

**METHOD**

We ask whether $\mu_1$ (estimated by $m_1$) is different from $\mu_2$ (estimated by $m_2$). Choose the smallest distance $\delta$ (clinically relevant difference) between $m_1$ and $m_2$, that is, $\delta = \mu_1 - \mu_2$, to be detected with statistical significance. From the medical literature or pilot data, find $\sigma_1^2$
and $\sigma_2^2$, estimated by $s_1^2$ and $s_2^2$, if necessary. Choose the risk required of an erroneous rejection of $H_0$ ($\alpha$) and of a correct rejection of $H_0$ (power, or $1 - \beta$). Look up the $z$-values in Table I, $z_{1-\alpha/2}$ and $z_{1-\beta}$, for these two risks. Substitute the $z$-values and $\delta$ along with the standard deviations in Eq. (21.3) to find $n_1$ ($= n_2$), the minimum sample size required in each sample. [Note how similar Eq. (21.3) is to Eq. (21.2).] For a one-sided test, substitute $z_{1-\alpha}$ for $z_{1-\alpha/2}$.

$$n_1 = n_2 = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}. \quad (21.3)$$

EXAMPLE COMPLETED: RANGE OF MOTION IN ARTIFICIAL KNEE

We chose $\delta = \mu_1 - \mu_2 = -6^\circ$ and found the values $\sigma_1^2 = 169$ and $\sigma_2^2 = 121$ from the literature. We choose $\alpha$ as 5% and $1 - \beta$ (the power) as 80%. From Table I (with interpolation), $z_{1-\alpha/2} = 1.96$ and $z_{1-\beta} = 0.84$. Substitution in Eq. (21.3) yields

$$n_1 = n_2 = \frac{(1.96 + 0.84)^2(169 + 121)}{2^2} = \frac{2.8^2 \times 290}{6^2} = 63.16.$$

The required minimum sample size is 64. For the reasons given in Section 21.2, a few more would be advisable.

ADDITIONAL EXAMPLE: TESTING TWO TREATMENTS OF DYSPEPSIA

An emergency medicine physician wants to test the relative effectiveness of two treatments, a “GI cocktail” (antacid plus viscous lidocaine) (treatment 1) versus IV ranitidine hydrochloride (treatment 2) to treat emergency dyspeptic symptoms as measured on a 1–10 pain scale.¹ Not having data on the pain ratings of the treatments, he estimates $s_1$ and $s_2$ as the standard deviation without treatment $\sigma = 1.73$. He considers a reduction in pain rating of $\delta = 1.5$ points as clinically meaningful. Either treatment can be more effective, so the test will be two-sided. Taking $\alpha = 0.05$ and $\beta = 0.20$ (power = 80%), $z_{1-\alpha/2} = 1.96$ and $z_{1-\beta} = 0.84$. By substituting in Eq. (21.3), he finds

$$n_1 = n_2 = \frac{(1.96 + 0.84)^2(1.73^2 + 1.73^2)}{1.5^2} = 20.86.$$

He requires a minimum of 21 per sample to show significance. However, he suspects that sample standard deviations will be larger than the untreated $\sigma$ and plans a larger sample size.
Exercise 21.4
An emergency medicine physician wants to know if mean HR following two particular types of trauma is different. He considers a mean difference of 6 bpm clinically meaningful. From pilot data, he finds $s_1 = 6.13$ bpm and $s_2 = 6.34$ bpm, which he uses to estimate $\sigma_1$ and $\sigma_2$. How many patients will he need in each group? Use $\alpha = 0.05$ and power $= 0.80$.

Exercise 21.5
DB7 contains very small samples of bone density measures for men and women. On the basis of the standard deviations of those samples, estimate the minimum sample size required to detect a difference of 10 units of bone density between the means for men and women with $\alpha = 0.05$ and power $= 0.80$ ($\beta = 0.20$).

Exercise 21.6
In DB12 the extent (centimeters) of the carinal resection appears to affect survival. Using the standard deviations given for surviving (Died = 0) and dying (Died = 1) patients, estimate the minimum sample size required for a test between means of resection extent of 0.37 for those two groups with $\alpha = 0.05$ and power $= 0.80$ ($\beta = 0.20$).

Exercise 21.7
Suppose we could take a set of INR readings from the clinic and another set from the lab, as in DB13, but not paired. We want to test for a two-sided difference between means, where a difference of 0.25 INR is clinically relevant. We take $\alpha = 0.05$ and power $= 0.90$. $\sigma_{\text{clinic}} = 0.54$ INR and $\sigma_{\text{lab}} = 0.63$ INR. How many readings do we need?

Exercise 21.8
Suppose we did not have the EIB and eNO data as in DB14, but found a 5-minute difference standard deviation of 7 ppb for both normals and asthmatics. We want to test for a difference in means between these two groups using $\alpha = 0.05$ and power $= 0.80$, where we take the clinically relevant difference to be 5 ppb. Assuming equal numbers of normals and asthmatics, what minimum sample size do we require?

Exercise 21.9
In DB3 the mean for the 16 patients’ differences between baseline and 5-day serum theophylline level is 0.993, with standard deviation 3.485. For such a difference to be significant with $\alpha = 0.05$ and power $= 0.80$, $n = 97$ pairs would be required. Generating three random samples of 97 on a computer from a normal distribution with the same mean and standard deviation, we find t tests yielding p-values 0.040, 0.042, and 0.072. Comment on the adequacy of the estimated minimum sample size.
21.7 TESTS WHEN DISTRIBUTIONS ARE NONNORMAL OR UNKNOWN

Sometimes, when data follow distributions other than normal or there is no information about the distribution, a rank test is followed. A comment on minimum sample size estimation for rank tests appears in Section 21.14. This section presents an alternative method.

One relationship that is sometimes helpful in sizing samples needed to detect a difference between the mean for a population, as estimated by \( m \), and a postulated value \( \mu \) arises in an inequality named for the Russian mathematician P. L. Chebychev (and other spellings, e.g., Tchebysheff). The relationship was developed by Bienaymé in 1835; Chebychev discovered it independently a bit later. The inequality is known as the Law of Large Numbers. Of interest to us is a form of the inequality relating sample size \( n \) to the deviation of the sample mean from the population mean. This relationship is useful when an underlying data distribution is very poorly behaved or when nothing is known about the distribution. It is also useful when there are no data at all to estimate the variability. As this inequality is solved without any information as to the distribution of the statistic involved, it is a rather gross overestimate of the sample size that would be required were the distribution known and well behaved. The sample size given by Chebychev’s inequality will certainly be enough in any case, and is, therefore, a conservative sample size, but it is more than required for most applications.

EXAMPLE POSED: TESTING DRUG EFFECT ON INTRAOCULAR PRESSURE

Suppose we want to know the required minimum sample size to find a difference between mean intraocular pressure (IOP in mmHg) between patients who have been treated with a new drug and those who have not. The standard deviation is 4 mmHg. We decide as a clinical judgment that we want to detect a 2 mmHg decrease in IOP. We choose \( \alpha = 0.05 \).

METHOD

Choose \( k \), the difference you want to detect between the true mean of the distribution giving rise to the sample and a postulated population mean, expressed as the number of standard deviations apart they are. The form of the Chebychev Inequality useful in this case is

\[ P\left[ -k \leq \mu_s - \mu \leq k \right] \geq 1 - \frac{\sigma^2}{k^2n}, \tag{21.4} \]

(Continued)
where \( \mu_s \) denotes the true mean of the distribution giving rise to the sample, \( \mu \) the hypothesized population mean, and \( \sigma \) the population standard deviation. Choose \( \alpha \), the risk of error you are willing to accept. The left side of Eq. (21.4) will be \( 1 - \alpha \) so that

\[
1 - \alpha = 1 - \frac{\sigma^2}{k^2n},
\]

which reduces to

\[
n = \frac{\sigma^2}{\alpha k^2}.
\]

Substitution for \( k \) in terms of \( \sigma \) allows the \( \sigma^2 \)'s to cancel and provide a number for \( n \). The value of \( \sigma \) is not even required if the user can express the difference \( k \) in standard deviation units.

EXAMPLE COMPLETED: TESTING DRUG EFFECT ON INTRAOCULAR PRESSURE

The units of \( k \) must be standard deviations, that is, a standard deviation is one unit. In this case, we want to detect a difference of 2 and \( \sigma = 4 \) mmHg, so that \( k = 0.5\sigma \). \( \alpha = 0.05 \). Using these quantities, Eq. (21.5) becomes

\[
n = \frac{\sigma^2}{\alpha k^2} = \frac{\sigma^2}{0.05 \times 0.25\sigma^2} = 80.
\]

21.8 TEST WITH NO OBJECTIVE PRIOR DATA

If we have neither data nor experience with a phenomenon being studied, we have no way to guess a required sample size. However, if we have some experience with the quantitative outcome of the variable to be measured, but no objective data at all, we have recourse to a very rough idea of needed sample size as a starting point.

EXAMPLE POSED: EFFECTIVENESS OF AN HERBAL REMEDY IN TREATING Colds

The husband of an internist has been treating his common colds with an herbal remedy for 3 years and claims that it reduces the number of days to disappearance of symptoms. His best time was 8 days and the worst 15. The internist decides to conduct a prospective randomized double-masked study to evaluate the remedy’s efficacy. How many data should she take? She decides that a reduction of \( \delta = 1 \) day would be clinically meaningful. She chooses \( \alpha = 5\% \) and power = 80\%. 

546

CHAPTER 21 Sample size estimation
METHOD
From experience, guess the smallest and largest values of the variable you have noticed. Take the difference between them as a guess of the interval: mean ± two standard deviations. This is equivalent to assuming that the data are roughly normal and your experience covers about 95% of the possible range. Then $\sigma$ is estimated as $0.25 \times (\text{largest} - \text{smallest})$. Use this value of $\sigma$ in the method of Case 1. Clearly, this “desperation” estimate is of extremely low confidence but may be better than picking a number out of the air.

EXAMPLE COMPLETED: HERBAL REMEDY
The assumption that the remedy cannot increase the time to symptom disappearance implies a one-sided test. The chosen $\alpha$ and $\beta$ yield one-tailed $z$-values of 1.645 and 0.84, respectively. In the absence of objectively recorded data the investigator guesses $\sigma$ from the observed range: $\sigma \approx 0.25 \times (\text{largest} - \text{smallest}) = 0.25 \times (15 - 8) = 1.75$. (The symbol “$$\approx$$” is often used to indicate “approximately equal to”. ) Substitution of these values in Eq. (21.2) yields

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\delta^2} = \frac{(1.645 + 0.84)^2 \times 1.75^2}{1^2} = 18.91.$$  

Nineteen in each group (the experimental group and the placebo-treated group) is indicated as a minimum. She chooses a large safety factor (see Section 21.2) in light of the poor estimation of $\sigma$ and plans 25 data in each group.

21.9 CONFIDENCE INTERVALS ON MEANS

EXAMPLE POSED: EXTENT OF CARINAL RESECTION
In DB12 the extent of carinal resection (centimeters) is distributed approximately normal and the sample size is large enough to use the estimated standard deviation 1.24 as $\sigma$. Suppose that in a new study we will require enough precision to ensure that the total width of the resulting 95% confidence interval for $\mu$, the population mean extent of resection, is no more than 1.0 cm, or that the width of the half interval is no more than 0.5 cm. How large a sample will we need to obtain?

METHOD
In Chapter 8, Tolerance, prediction, and confidence intervals, on confidence intervals the end values of a $1 - \alpha$ confidence interval on the mean $\mu$ of a normal distribution with known $\sigma$ was given by Eq. (8.5) as $m \pm z_{1-\alpha/2}\sigma_m$. As before, the form using $\sigma$ rather than $s$ is
employed regardless of sample size. Thus the width of the half interval is
\[ w = z_{1-\alpha/2} \sigma_m. \]
We square throughout and substitute \( \sigma^2/n = \sigma_m^2 \) to obtain
\[ w^2 = (z_{1-\alpha/2} \sigma)^2/n, \]
or
\[ n = \frac{z^2_{1-\alpha/2} \sigma^2}{w^2}. \]  \hspace{1cm} (21.6)

The \( n \) calculated in Eq. (21.6) is a very minimum; it would be wise to take a
slightly larger sample for the reasons discussed in Section 21.2. For 95% confidence,
we need only replace \( z_{1-\alpha/2} \) by 1.96.

\( \beta \) is absent from confidence interval sample size estimation
Confidence intervals are estimation tools. There is no alternative hypothesis and,
therefore, no type II error, the risk of which is \( \beta \). While sample size estimation for a
test involves limiting the two error risks jointly, sample size estimation for a confidence
interval addresses only the lack of confidence in the estimate at issue.

EXAMPLE COMPLETED: EXTENT OF CARINAL RESECTION
The need for 95% confidence implies that \( z_{1-\alpha/2} = 1.96 \) (the frequently seen entry
from Table I). By substituting that value and the values of \( \sigma \) and \( w \) in Eq. (21.6), we
find that
\[ n = \frac{z^2_{1-\alpha/2} \sigma^2}{w^2} = \frac{1.96^2 \times 1.24^2}{0.5^2} = 23.63. \]
We need a sample of at least 24 patients.

ADDITIONAL EXAMPLE: EXTENT OF CARINAL RESECTION CONTINUED
In the example from DB12 just completed, suppose we required greater confidence,
say 99%, and greater accuracy, say width of the half interval \( w = 0.25 \) cm. \( \sigma \) was taken
as 1.24, derived from the data already at hand. In Table I the 0.990 entry in the col-
umn under two-tailed \( 1 - \alpha \) lies in the row for \( z_{1-\alpha/2} = 2.576 \). By substituting in
Eq. (21.6), we find that
\[ n = \frac{z^2_{1-\alpha/2} \sigma^2}{w^2} = \frac{2.576^2 \times 1.24^2}{0.25^2} = 163.25. \]
We require a minimum of 164 patients.
Exercise 21.10

In DB7 the distribution of bone density is not far from normal. From these data we calculate \( m = 154 \) and \( s = 24 \). A new study is to be conducted to estimate the population mean. Using these pilot data estimates, determine \( n \) such that the half width of a 95\% confidence interval for the population mean is no more than \( w = 10 \).

21.10 TEST OF ONE PROPORTION (ONE RATE)

Methods for testing categorical data are treated in Chapter 9, Tests on categorical data. From categorical data, proportions (which can also be thought of as rates) can always be obtained; minimum sample size is calculated for such proportions. Values used in such calculation are the error risks \( \alpha \) and \( \beta \), the proportions involved, and the difference between the proportions. This difference, here denoted \( \delta \), is the value to be tested. It should be chosen as the difference clinically important to detect. The minimum sample size depends more on this difference than on the other inputs. The sample size grows large quickly as this difference grows small.

Contingency tables

Methods for the estimation of minimum required sample size are not well developed for contingency tests, but we can use the method for tests of proportion. The cell entries over their marginal totals (the totals for each category) provide proportions. If the contingency table is bigger than \( 2 \times 2 \), an ad hoc approach is to calculate the required sample size for the various pairs by using the method for two proportions repeatedly and then accept as the estimate the largest sample size which emerges.

Test of one proportion

The true probability of success that gives rise to a sample is tested against a theoretical or previously established probability. The observed number of successes in the sample follows the binomial distribution, or can be approximated by a Poisson distribution depending on whether the proportion is central (is not near 0 or 1) or extreme (is near 0 or 1), respectively. An estimated proportion from an at least moderately sized sample is distributed approximately normal. The proportion is a sample mean for either, leaving the only difference in the method as the mode of estimating the standard deviation. Formulas for the binomial form, proportion \( \pi \) known or estimated by \( p \), respectively, are

\[
\sigma = \sqrt{\frac{\pi(1 - \pi)}{n}}, \quad s = \sqrt{\frac{p(1 - p)}{n}},
\]

(21.7)
and for the Poisson form are
\[ \sigma = \sqrt{\frac{\pi}{n}}, \quad s = \sqrt{\frac{p}{n}}. \] (21.8)

Thanks to the normal approximation, the required sample size estimation follows the concept and logic depicted in Fig. 21.2.

**EXAMPLE POSED: RATE OF POSITIVE PROSTATE CANCER BIOPSIES**

Suppose we want to test an observed 30% rate of positive prostate cancer biopsies from DB1 against a smaller theoretical proportion \( \pi = 0.25 \). How many patients do we need to yield a one-tailed level \( \alpha = 0.05 \) test with power = 0.80 if the true proportion giving rise to the sample is \( \pi_s = 0.30 \)? (Note that we have a one-sided alternative hypothesis.)

**METHOD**

We assume normal theory, with mean difference \( \pi_s - \pi \) divided by the appropriate standard deviation from Eqs. (21.7) or (21.8), which gives rise to the sample size Eqs. (21.9) and (21.10). If \( \pi \) is central (not near 0 or 1), we use a binomial form, Eq. (21.9). If \( \pi \) is extreme (near 0 or 1), we can use a Poisson form, Eq. (21.10). We choose the risk required of a false-positive (\( \alpha \)) and a false-negative (\( \beta \), or 1 – power). We look up the z-values in Table I associated with these two risks (areas in the tails of the normal curve), \( z_{1-\alpha/2} \) and \( z_{1-\beta} \). We then substitute the z-values, \( \pi \), and \( \pi_s \) in Eqs. (21.9) or (21.10) to find \( n \), the minimum sample size required.

\[
\begin{align*}
n = & \left[ \frac{z_{1-\alpha/2} \sqrt{\pi(1-\pi)} + z_{1-\beta} \sqrt{\pi_s(1-\pi_s)}}{\pi_s - \pi} \right]^2 \quad (21.9) \\
n = & \left[ \frac{z_{1-\alpha/2} \sqrt{\pi} + z_{1-\beta} \sqrt{\pi_s}}{\pi_s - \pi} \right]^2. \quad (21.10)
\end{align*}
\]

We should note that \( \pi_s - \pi \) is a major determinant in the sample size estimation and, therefore, \( \pi_s \) should be chosen carefully in line with its clinical implications. For a one-sided test, replace \( z_{1-\alpha/2} \) by \( z_{1-\alpha} \).

**EXAMPLE COMPLETED: RATE OF POSITIVE PROSTATE CANCER BIOPSIES**

From Table I the z-values are 1.645 and 0.84, respectively. Because \( \pi \) is not near 0 or 1, we use the binomial form Eq. (21.9). By substituting the z-values, \( \pi \), and \( \pi_s \), we obtain \( n = \frac{[1.645 \times \sqrt{0.25 \times 0.75} + 0.84 \times \sqrt{0.31 \times 0.69}]}{(0.31 - 0.25)^2} = 336.6 \). We require at least 337 biopsies.
ADDITIONAL EXAMPLE: RATE OF SCHISTOSOMIASIS

A Navy specialist in internal medicine, sent to an African nation, must decide if female residents of a particular rural region lying along an often stagnant river have a prevalence of schistosomiasis greater than the national average, implying the hypotheses $H_0$: $\pi_s = \pi$ (subscript $s$ for “sample”) and $H_1$: $\pi_s > \pi$ (a one-sided alternative). \(^4\) He plans an informal study to compare the local mean to the national mean and must estimate how many patients he needs to sample. Clinical judgment leads him to believe he needs to detect a difference of 2%. He finds an article in the literature that quotes a prevalence of 24% over 1600 patients examined. He chooses one-tailed $\alpha = 5\%$ and power $= 80\%$. The $z$-values from Table I are 1.645 and 0.84. By substitution in Eq. (21.9), he finds

\[
 n = \left[ \frac{z_{1-\alpha} \sqrt{\pi(1-\pi)} + z_{1-\beta} \sqrt{\pi_s(1-\pi_s)}}{\pi_s - \pi} \right]^2 = \left[ \frac{1.645 \sqrt{(0.24)(0.76)} + 0.84 \sqrt{(0.26)(0.74)}}{0.02} \right]^2 = 2867.63
\]

He will require nearly 2900 patients to be able to detect a 2% difference.

Exercise 21.11

At one time, radial keratotomy was performed by residents in ophthalmology in a certain hospital. A review of records showed that 18% of residents required enhancements in more than one-fifth of their cases.\(^5\) This frequency of enhancements was thought to arise from the surgery learning curve. Training in the surgery by a computer simulation may shorten the learning curve. How many residents would have to be monitored to detect a decrease of 6% (i.e., a drop from 18% to 12%) in the number of residents requiring enhancements in more than one-fifth of their cases? Use $\alpha = 5\%$ and power $= 80\%$.

21.11 TEST OF TWO PROPORTIONS (TWO RATES)

A test of two proportions (see also Section 9.6) can be used in place of a contingency test (see also Sections 9.2—9.4), but is usually less convenient because it does not appear in many software packages. We use it here reformatted for sample size estimation.

EXAMPLE POSED: RATE OF PERSONALITY DISORDER IN CRIMINALS

A psychiatrist wants to know if the proportion of people having a personality disorder is the same for those committing violent crimes ($\pi_1$) and those committing nonviolent crimes ($\pi_2$).\(^1\) Theoretically, either $\pi$ could be the larger, so he chooses a two-sided alternative hypothesis. He examines a few of his past records to serve as a pilot survey and estimates $\pi_1$ as 0.06 and $\pi_2$ as 0.02. How many patients does he need to detect a difference $\pi_1 - \pi_2 = 0.04$ significant at two-tailed $\alpha = 0.05$ and power $= 0.80$?
METHOD
We assume normal theory. We conjecture the proportions $\pi_1$ and $\pi_2$ from our data and then $\pi_m$ ($m$ for mean) as the average of $\pi_1$ and $\pi_2$:

$$\pi_m = \frac{\pi_1 + \pi_2}{2}$$

For $\pi_m$ arising from a common event ($\pi_m$ not near 0 or 1), we use Eq. (21.11), derived from the binomial form:

$$n_1 = n_2 = \left[ \frac{z_{1-\alpha/2} \sqrt{2 \pi_m (1 - \pi_m)} + z_{1-\beta} \sqrt{\pi_1 (1 - \pi_1) + \pi_2 (1 - \pi_2)}}{\pi_1 - \pi_2} \right]^2. \quad (21.11)$$

For $\pi_m$ arising from a rare event ($\pi_m$ near 0 or 1), we can use Eq. (21.12), derived from the Poisson form:

$$n_1 = n_2 = \left[ \frac{(z_{1-\alpha/2} + z_{1-\beta}) \sqrt{\pi_1 + \pi_2}}{\pi_1 - \pi_2} \right]^2. \quad (21.12)$$

We choose $\alpha$, the risk of a wrong rejection of $H_0$, and $\beta$ (1 - power), the risk of a wrong acceptance of $H_0$. We look up the z-values in Table I associated with these two risks (areas in the tails of the normal curve), $z_{1-\alpha/2}$ and $z_{1-\beta}$. We substitute the z-values and the $\pi$'s in Eqs. (21.11) or (21.12) to find $n$, the minimum sample size required in each group. If the test is one-sided, replace $z_{1-\alpha/2}$ by $z_{1-\alpha}$.

Estimates on the borderline between binomial and Poisson
If $\pi$ is close to the borderline for using the Poisson approximation, a correction of the normal approximation is appropriate, especially if $n_1$ is small. The corrected $n$ for each group will be $n_{corr}$ from:

$$n_{corr} = \frac{n_1}{4} \left[ 1 + \sqrt{1 + \frac{4}{n_1 \times |\pi_1 - \pi_2|}} \right]^2. \quad (21.13)$$

EXAMPLE COMPLETED: RATE OF PERSONALITY DISORDER IN CRIMINALS
Because $\pi_m$ is near 0, the Poisson form Eq. (21.12) is can be used. From Table I the z-values are 1.96 and 0.84, respectively. $\pi_m$, the mean $\pi$, is $(\pi_1 + \pi_2)/2 = 0.04$.

$$n_1 = n_2 = \left[ \frac{(z_{1-\alpha/2} + z_{1-\beta}) \sqrt{\pi_1 + \pi_2}}{\pi_1 - \pi_2} \right]^2 = \left[ \frac{(1.96 + 0.84) \sqrt{0.08}}{0.04} \right]^2 = 392.$$  

He will need a very minimum of 392 patients in each group.
ADDITIONAL EXAMPLE: SEX DIFFERENCE IN FEVER REPORTING
An emergency medicine specialist finds that 15% of patients report having a fever. She wants to know if there is a difference between reporting rates of men and women. She does not know which group would have the greater rate, which implies a two-tailed test. How many patients would she need to monitor to find a difference of 6% (e.g., 12% for one group and 18% for the other) with $\alpha = 0.05$ and power $= 0.80$? $z_{1-\alpha/2} = 1.96$ and $z_{1-\beta} = 0.84$. $\pi_m = 0.15$, not near 0 or 1, so the binomial form Eq. (21.11) is used.

$$n_1 = n_1 = \left[ \frac{z_{1-\alpha/2} \sqrt{2\pi_m(1-\pi_m)} + z_{1-\beta} \sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}}{\pi_1 - \pi_2} \right]^2$$

$$= \left[ \frac{1.96 \sqrt{2 \times 0.15 \times 0.85} + 0.84 \sqrt{0.12 \times 0.88 + 0.18 \times 0.82}}{0.12 - 0.18} \right]^2 = 554.16.$$

The minimum sample size per group is 555 patients.

Exercise 21.12
At one time an eye surgeon was performing radial keratotomies using hand surgery and considered a laser device. His hand surgery record showed that 23% of eyes required surgical enhancement. He planned a prospective study in which patients would be randomized into the hand and laser groups. On how many eyes per method would he have had to operate to detect an improvement of 10% (i.e., a reduction in enhancement rate from 23% to 13%)? He believes the laser device can only improve the enhancement rate, so the test is one-sided. Use $\alpha = 5\%$ and power $= 80\%$.

21.12 CONFIDENCE INTERVALS ON PROPORTIONS (ON RATES)
EXAMPLE POSED: RATE OF PATIENTS SATISFIED WITH ORAL SURGERY ANESTHESIA
In an example from Chapter 8, Tolerance, prediction, and confidence intervals, oral surgery patients were anesthetized by a combination of propofol and alfentanil and 89.1% of patients rated the anesthetic as highly satisfactory. We placed a 95% confidence interval on $\pi$ as approximately 83%–96%. The width of the half interval, $w$, is 6.5%. How many patients would we need to reach $w = 5\%$?

METHOD
Conceptually, we may think of estimating the sample size needed for a confidence interval as specifying the length of interval (or half interval) we require and back-solving the (Continued)
equation for \( n \). Note that power, that is, \( 1 - \beta \), is not involved; see the comment on this issue in Section 21.9. However, there are yet other considerations. From Eq. (8.8) (ignoring the continuity correction) the width of the half interval for a 95% confidence interval will be 
\[
 w = 1.96 \sqrt{p(1-p)/n},
\]
where we do not have \( p \) yet. Because we need \( p \) to estimate \( \sigma \), but will not have it until data have been gathered, we must use some rough indicator of \( \pi_s \), such as the proportion found in a pilot study or a similar study found in the literature. As a result of these uncertainties, the \( n \) obtained is not an accurate sample size, but just a rough idea of the size. We would be better assured of reaching our target confidence if we take a slightly larger \( n \), although there is no way to know just how much larger. Solving one side of the confidence interval in Eqs. (8.8) and (8.9) for \( n \) for 95% confidence, we find
\[
 n = \frac{1.96^2 \pi_s (1-\pi_s)}{w^2}.
\]  
(21.14)

for the case of central \( \pi_s \) (not near 0 or 1; for a common event), and
\[
 n = \frac{1.96^2 \pi_s}{w^2}.
\]  
(21.15)

for extreme \( \pi_s \) (near 0 or 1; for a rare event). Confidence levels other than 95% can be found by replacing the 1.96 with the appropriate probability from Table I. [Note that, because the maximum value of \( \pi_s(1-\pi_s) = 0.25 \), the numerator of Eq. (21.14) cannot exceed 0.9649. The numerator of Eq. (21.15) cannot exceed 3.842.]

**EXAMPLE COMPLETED: SATISFACTION WITH ORAL SURGERY ANESTHESIA**
\( \pi_s \) is not near 0 or 1, so we use Eq. (21.14).
\[
 n = \frac{1.96^2 \pi_s (1-\pi_s)}{0.05^2} = \frac{1.96^2 \times 0.891 \times 0.109}{0.05^2} = 149.2.
\]

We would need a very minimum of 150 and would be advised to take a few more.

Suppose we wanted to be 98% confident, that is to risk 1% chance of error on each tail. From Table I the 0.98 two-tailed \( 1 - \alpha \) yields a corresponding \( z \) of 2.326. Replacing the 1.96 in the calculation just above by 2.326, we obtain
\[
 n = \frac{2.326^2 \pi_s (1-\pi_s)}{w^2} = 210.2.
\]

**ADDITIONAL EXAMPLE: EFFICACY OF A DERMATOLOGICAL TREATMENT**
A dermatologist is studying the efficacy of tretinoin in treating women’s postpartum abdominal stretch marks. Tretinoin will be used on a randomly chosen side of the
abdomen and a placebo on the other. Neither patient nor investigator will know which side was medicated. The dermatologist will then rate one side or the other as better where he could make a distinction and afterward the code will be broken and the treated side identified. Prior studies of a similar design have observed that the treated side was chosen in 9 of 13 abdomens for an observed proportion of 0.69. If the treatment were of no value, the theoretical proportion $\pi$ would be 0.5. How many patients would be needed to have 95% confidence to rule out no value of the treatment. In this case we require that $w = |\pi_s - \pi| = 0.19$, where $\pi_s$ is assumed to be 0.69 from prior data? $\pi_s$ is not near 0 or 1, so she uses Eq. (21.14). By substituting, she finds

$$n = \frac{1.96^2 \pi_s (1 - \pi_s)}{w^2} = \frac{3.8416 \times 0.69 (1 - 0.69)}{(0.69 - 0.5)^2} = 22.8.$$ 

Hence, she requires a minimum of 23 patients.

**Exercise 21.13**

*A pediatric surgeon is studying indicators of patient condition following pyloromyotomy (correction of stenotic pylorus) in neonates.* A past study demonstrated that 14 out of 20 infants, that is, a proportion of 0.7, experienced emesis after surgery. How large a sample will he require to ensure that the width of half of a 95% will be 0.1 or less?

## 21.13 TEST ON A CORRELATION COEFFICIENT

**EXAMPLE POSED: REPAIRED ANKLE PLANTAR FLEXION CORRELATED WITH AGE**

In a study on broken ankle repair, an orthopedist found the correlation coefficient between age and plantar flexion of a repaired ankle to be 0.1945. How large a sample would he have required to be able to yield adequate power to conclude a relationship between flexion and age if the true population correlation were 0.1945?

**METHOD**

When a correlation coefficient between two variables is small enough to have occurred by chance alone, we cannot infer evidence of a relationship between these variables. What sample size will provide sufficient precision to determine if the true correlation, $\rho$, differs from 0? More precisely, how large a sample is required for the 95% confidence interval to exclude $\rho = 0$? When $\rho = 0$ is true and the correlated variables are approximately normal, a transformation of the sample correlation coefficient, $r$, to the $t$ distribution can be made, namely,

(Continued)
A little algebra will solve this equation for $n$ and provide Table 21.1. When solving for sample size, however, we do not yet know $r$, and hence replace it with a postulated value $\rho$. The last lines of Table 21.1 carry the sample size smaller (12 or so) than would be wise to use in practice. It is given to show the pattern of reduction in $n$. What is done if $\rho$ is postulated to be negative? A symmetry property of the $t$ distribution allows the same $n$ to emerge as if $\rho$ were positive, so just drop the minus sign for sample size purposes. Note that power, that is, $1 - \beta$, is not involved; see the comment on this issue in Section 21.9.

Table 21.1 Samples sizes required to infer a $\rho$ larger than would occur due to chance at 95% confidence.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$n$</th>
<th>$\rho$</th>
<th>$n$</th>
<th>$\rho$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>—</td>
<td>0.21</td>
<td>88</td>
<td>0.40</td>
<td>25</td>
</tr>
<tr>
<td>0.03</td>
<td>4269</td>
<td>0.22</td>
<td>80</td>
<td>0.41</td>
<td>23</td>
</tr>
<tr>
<td>0.04</td>
<td>2401</td>
<td>0.23</td>
<td>74</td>
<td>0.42</td>
<td>22</td>
</tr>
<tr>
<td>0.05</td>
<td>1538</td>
<td>0.24</td>
<td>68</td>
<td>0.43</td>
<td>21</td>
</tr>
<tr>
<td>0.06</td>
<td>1068</td>
<td>0.25</td>
<td>63</td>
<td>0.44</td>
<td>20</td>
</tr>
<tr>
<td>0.07</td>
<td>785</td>
<td>0.26</td>
<td>58</td>
<td>0.45</td>
<td>19</td>
</tr>
<tr>
<td>0.08</td>
<td>601</td>
<td>0.27</td>
<td>54</td>
<td>0.46</td>
<td>19</td>
</tr>
<tr>
<td>0.09</td>
<td>475</td>
<td>0.28</td>
<td>50</td>
<td>0.47</td>
<td>18</td>
</tr>
<tr>
<td>0.10</td>
<td>385</td>
<td>0.29</td>
<td>47</td>
<td>0.48</td>
<td>17</td>
</tr>
<tr>
<td>0.11</td>
<td>319</td>
<td>0.30</td>
<td>44</td>
<td>0.49</td>
<td>17</td>
</tr>
<tr>
<td>0.12</td>
<td>268</td>
<td>0.31</td>
<td>41</td>
<td>0.50</td>
<td>16</td>
</tr>
<tr>
<td>0.13</td>
<td>228</td>
<td>0.32</td>
<td>39</td>
<td>0.55</td>
<td>13</td>
</tr>
<tr>
<td>0.14</td>
<td>197</td>
<td>0.33</td>
<td>37</td>
<td>0.60</td>
<td>11</td>
</tr>
<tr>
<td>0.15</td>
<td>172</td>
<td>0.34</td>
<td>34</td>
<td>0.65</td>
<td>9</td>
</tr>
<tr>
<td>0.16</td>
<td>149</td>
<td>0.35</td>
<td>32</td>
<td>0.70</td>
<td>8</td>
</tr>
<tr>
<td>0.17</td>
<td>134</td>
<td>0.36</td>
<td>30</td>
<td>0.75</td>
<td>7</td>
</tr>
<tr>
<td>0.18</td>
<td>120</td>
<td>0.37</td>
<td>29</td>
<td>0.80</td>
<td>6</td>
</tr>
<tr>
<td>0.19</td>
<td>107</td>
<td>0.38</td>
<td>27</td>
<td>0.85</td>
<td>5</td>
</tr>
<tr>
<td>0.20</td>
<td>97</td>
<td>0.39</td>
<td>26</td>
<td>0.90</td>
<td>4</td>
</tr>
</tbody>
</table>

**STATISTICAL SIGNIFICANCE AND CLINICAL MEANING**

A look at the first few rows of Table 21.1 will show the reader that an investigator can find a correlation coefficient to be “larger than chance” just by taking a large enough sample. The fact of being larger than chance statistically may not imply that the correlation is clinically useful. For example, a 0.10 correlation coefficient between a disease and a marker may be significant, but it does not suggest the inference of the disease from the marker. Increasing sample size is useful up to the value of the...
coefficient that is clinically meaningful; beyond that, further increases are just a statistical exercise that should be avoided as the result could be misleading.

**EXAMPLE CONCLUDED: PLANTAR FLEXION AND AGE**
From Table 21.1, \( \rho = 0.1945 \) corresponds to an \( n \) between 97 and 107; he would have needed a few more than 100 patients.

**ADDITIONAL EXAMPLE: CORRELATION BETWEEN SURGICAL DIFFICULTY AND DURATION**
An anesthesiologist wanting to predict the requirement for deep sedation during surgery found a correlation coefficient between surgical difficulty and surgical duration of 0.17 on a small sample of patients. If this sort of coefficient holds, how many patients would she need to conclude that surgical difficulty is a factor in predicting surgical duration? From Table 21.1, \( \rho = 0.17 \) corresponds to \( n = 134 \).

**Exercise 21.14**
An ophthalmologist suspects that the effect of a beta-blocker (timolol) on IOP diminishes with age. He takes a pilot sample of \( n = 30 \) patients, recording their ages and their reduction in IOP 4 weeks after initiation of treatment. He calculates the sample correlation coefficient as \( r = -0.23 \). How large a sample would a subsequent study need to conclude that the population \( \rho \) is truly less than 0, based on his pilot result?

**21.14 TESTS ON RANKED DATA**
Methods for testing ranked data are treated in Chapter 11, Tests of location with continuous outcomes. Not much theoretical development has occurred for estimating minimum required sample sizes for rank-order (nonparametric) data, because the largest application is for cases in which the distributions are unusual or “poorly behaved.” Sample size methods depend on these probability distributions that are generally unknown. For extremely deviant cases of this sort the method of Section 21.7 may be used, although this method is very conservative and tends to overestimate the minimum required sample size.

There is a fall back approach. The concept in theoretical statistics of efficiency of a test deals, roughly speaking, with the quality of the test’s performance in accomplishing its purpose. Asymptotic relative efficiency (ARE) represents how well one test performs relative to another as the sample size grows increasingly large. Thus we could say that the required sample size for a test with no available power analysis is the size given by a power analysis for a test with an equivalent purpose divided by the ARE for the two. For example, \( n_{rs} \) for a rank-sum test would be the \( n_{t} \) from the power analysis for a two-sample normal means test/ARE.
CHAPTER 21 Sample size estimation

But what is the value for this ARE? In the case of data following a normal distribution the ARE for a rank test compared to a $t$ test with similar goal is 0.955. In this case a rank-sum test would require $1/0.955 = 1.047$ times the sample size of a two-sample normal test to achieve the same efficiency, or a 4.7% greater sample. However, we would not use a rank-sum test when the data followed a normal distribution. As the distribution deforms, the ARE drops. At some point the deformation is great enough that we would abandon the use of a normal test in favor of the rank-sum. What is the ARE for this level of deformation? We do not know. What we do know is that mathematical statistics shows the minimum ARE, that is the worst possible, is 0.864. Therefore we will be safe in using a power analysis to obtain $n_n$ for a normal test and planning our sample to use a rank-sum test as $n_{rs} = n_n/0.864$, 15.7% larger. The same figures occur for the paired $t$ test and its rank analog, Wilcoxon’s signed-rank test.

What of other rank tests? We do not have AREs for every rank test. However, the results would be of the same order of magnitude. Sample size estimation is based on judgmental inputs combined with data other than that from the actual sample to be obtained and are, therefore, very approximate. We would be safe in increasing the parametric test’s sample size from a power analysis by 20% for the corresponding nonparametric test.

21.15 VARIANCE TESTS, ANALYSIS OF VARIANCE, AND REGRESSION

Estimation of sample size $n$ for tests on these methods is not easy. The $\chi^2$ and $F$ probabilities used to estimate $n$ depend on $n$. A few such estimates exist as tables, but they are based only upon the type I error and are, therefore, not dependable. It is possible to develop estimates based on both types of error, but this entails using the difficult mathematical distributions known as noncentral $\chi^2$ and noncentral $F$, beyond the level of this text. Recently, software in some statistical packages has appeared that treats one or another of these problems. The use of this software depends on “effect size.” The concept of effect size is the outcome result contrast, say between two arms of the experiment, in ratio to the variability of the sampling distributions of these arms. This ratio controls the sizes of the risks $\alpha$ and $\beta$. Unfortunately, in much of the software, effect size is poorly defined or even undefined or requires a spate of judgmental inputs over and above the judgment-based values of $\alpha$, $\beta$, and $\delta$ already required. The best advice for an investigator who requires such sample sizes is to seek the assistance of an accomplished statistician.

21.16 EQUIVALENCE TESTS

As with ANOVA and regression, minimum sample size estimation for equivalence (including noninferiority) means tests depend on a difficult distribution, in this case the
noncentral \( t \) distribution. The investigator should find specialty software or a statistician for assistance. However, suppose such access is lacking. Recall the approximate nature of power analysis results. We note that minimum required sample sizes for equivalence tests are not dramatically different from those for difference tests. A conservative ad hoc strategy would be to find the solution for a difference test and increase the size by, perhaps, 15%.

### 21.17 NUMBER NEEDED TO TREAT OR BENEFIT

**Number needed to treat: Screening for disease**

The simplest case of number needed to treat (NNT), given no cases are detected by other means, is the number of people screened in order to detect one case to treat. For example, how many mammograms must be run on randomly chosen women of a certain age in order to find one case of breast cancer? How many vaccinations for hepatitis C are required to prevent one case?

**EXAMPLE POSED: SCREENING FOR LUNG CANCER IN BALTIMORE**

A population at risk, or catchment in epidemiological terminology, was composed of male smokers in Baltimore in 1980. A number of 10,387 members were screened and 47 cases of advanced lung cancer were found. How many members of the catchment had to be screened to find one case?

**METHOD FOR NUMBER NEEDED TO TREAT (NNT)**

Let us denote by \( n_s \) the number of subjects in the population or catchment screened and by \( n_d \) the number of disease cases discovered by screening. The ratio \( p_d = n_d/n_s \) is the proportion of disease cases discovered by screening. If all disease cases are found by screening, \( p_d \) also estimates the disease prevalence. NNT is the reciprocal of the proportion discovered, or

\[
\text{NNT} = \frac{1}{p_d} = \frac{n_s}{n_d}.
\]

**EXAMPLE COMPLETED: SCREENING FOR LUNG CANCER IN BALTIMORE**

The prevalence is \( p_d = 47/10387 = 0.0045 \). NNT = \( 1/0.0045 = 222 \). The authorities must screen 221 subjects without advanced lung disease for each one they discover.

**ADDITIONAL EXAMPLE: TREATING ORAL LEUKOPLAKIA**

Oral leukoplakia, which frequently progresses to overt squamous cell carcinoma, was treated with 13-\(cis\)-retinoic acid in \( n_s = 44 \) patients. \( n_d = 24 \) patients responded histologically. NNT = \( n_s/n_d = 1.83 \) patients treated per responder.
NUMBER NEEDED TO TREAT FOR ADDITIONAL DETECTIONS
In most screening programs, some of the disease cases would be discovered through ordinary medical practice. In the Baltimore screening, some of the 47 cases found by screening would have been discovered through patient visits to their primary care providers. Let us denote by $p_m$ the proportion of cases that would be found through common medical practice. Then $p_d - p_m$ is the proportion of additional detections due to the screening and

$$\text{NNT} = \frac{1}{p_d - p_m}. \quad (21.18)$$

**Exercise 21.15**
In a study\textsuperscript{13} on radiographic screening in a correctional facility in New York City, the rate of tuberculosis among entering inmates was 0.00767. A number of 4172 entering inmates were screened. How many were detected by the radiographic screening? Of those detected, 25 had entered with a prior diagnosis of tuberculosis. What is the NNT for those newly diagnosed by the screening?

Cost of number needed to treat
An essential ancillary issue is the cost of the resources (time, facilities, personnel, and money) expended for one detection and trading off this cost against the gain of treating that patient. The gain is usually intangible and the cost effectiveness becomes a matter of comparing events from two different value bases. This might be likened to comparing the value of coins from two different money systems: we have to find or develop a medium of exchange, that is, a measure of effectiveness, that will compare the cost of the NNT with the gain to the patient.

Examining only monetary costs, let us denote by $c_s$ the cost for the screening program, that is, the cost to screen $n_s$ people. Then the cost per subject screened is $c_s/n_s$ and the cost per detection $c_d$ is

$$c_d = \frac{c_s}{n_s} \times \text{NNT}. \quad (21.19)$$

If Number Needed to Treat is the number needed per additional detection, $c_d$ is the cost per additional detection.
Detection compared to other methods of efficacy, such as mortality

Is detection the final criterion of screening efficacy? It certainly is not the only one. Already we have, in addition to NNT, the number detected by the screening program \( (n_d) \) and the cost per detection \( (c_d) \). However, even a statistically significant portion of the population being detected by the screening is not evidence that the screening is beneficial. Suppose more diseased patients are detected, but that does not help them. In a German study, 41,532 men born between 1907 and 1932 were screened for lung cancer with chest fluorography every 6 months for 10 years and compared with age-matched men screened similarly every 18 months. No significant reduction in lung cancer mortality or in overall mortality was found. When mortality rather than detection rate was taken as the measure of effectiveness, tripling the screening rate provided no improvement. An investigator, and certainly a reader of medical articles, should carefully consider the method of efficacy used.

Number needed to benefit: Assessing the benefit of a new treatment

A change in treatment or mode of diagnosis often requires new instrumentation, laboratory tests, training, etc. Is the benefit worth the changeover effort and cost? One measure of the benefit is NNB, that is, the number of patients needed to be treated or diagnosed with the new procedure in order to obtain one additional successful outcome. A confidence interval on NNB would be an additional help in making an administrative decision about a changeover in treatment or diagnosis.

**EXAMPLE POSED: NUMBER NEEDED TO BENEFIT FROM SWITCHING TATTOO INK**

The majority of people with occasional tattoos get their tattoos as young adults and want removal in middle age. The commonly used titanium ink is difficult to remove. How many patients must be switched to nontitanium ink in order to facilitate one additional successful removal (NNB), and what would be a 95% confidence interval on this NNB? From DB6, 35 removals of titanium ink tattoos were attempted, with five successes. Of nontitanium ink tattoos, 15 removals were attempted, with 8 successes.

**METHOD FOR NUMBER NEEDED TO BENEFIT (NNB)**

Let us denote as \( p_1 \) the observed proportion of successful outcomes using the established treatment or mode of diagnosis, and by \( p_2 \) that of the new treatment or mode. Then \( d = p_2 - p_1 \) is the difference in success rates.

\[
NNB = \frac{1}{d}
\]

(21.20)
The confidence interval requires two steps. A confidence interval is a probability statement that an interval calculated from a sample bounds the theoretical parameter being estimated. Let us denote by \( \delta \) the theoretical difference in success rates being estimated by \( d \), and by \( \text{NNB} \) the theoretical Number Needed to Benefit being estimated by NNB. A confidence interval on the true difference, \( \delta \), estimated by \( d \), is found and then the components are inverted (as in Eq. 20.5) to provide confidence on the theoretical \( \text{NNB} \), estimated by NNB. The standard error of the difference, say SED, is as follows:

\[
SED = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}. 
\] (21.21)

Using SED, the \( 100 \times (1 - \alpha)\% \) confidence interval on \( \delta \) is given by Eq. (21.22).

\[
(d - z_{1-\alpha/2}SED, d + z_{1-\alpha/2}SED). 
\] (21.22)

To convert Eq. (21.22) to confidence on \( \text{NNB} \) the components inside the brackets are inverted. However, recall from algebra that such inversion reverses the inequality signs. The confidence interval on \( \text{NNB} \) becomes as follows:

\[
\left( \frac{1}{d + z_{1-\alpha/2}SED}, \frac{1}{d - z_{1-\alpha/2}SED} \right) 
\] (21.23)

**EXAMPLE COMPLETED: NUMBER NEEDED TO BENEFIT FROM SWITCHING TATTOO INK**

\( p_1 = 5/35 = 0.14 \). \( p_2 = 8/15 = 0.53 \). The difference between the \( p \)'s is \( d = 0.39 \). SED = \( \sqrt{0.14 \times 0.86/35 + 0.53 \times 0.47/15} = 0.14 \). The 95% confidence interval is \((0.39 - 1.96 \times 0.14, 0.39 + 1.96 \times 0.14) = (0.115, 0.664)\). Taking reciprocals within the brackets, changing the inequality signs, and then reversing the order of components to return the inequalities to their usual position yields the confidence on a theoretical \( \text{NNB} \) as \((1.51, 8.7)\). Our estimated NNB is 3, and the 95% plausible values for NNB are between 2 and 9.

**REFERENCES**

1. Missing Sources. The sources of a few examples could not be found despite a strong effort to locate them. Such data that could not be referenced were slightly altered so as not to reflect on any investigator later appearing.
3. Noecker, RS; Dirks, MS; Choplin, NT; Bernstein, P; Batoosingh, AL; Whitcup, SM A six-month randomized clinical trial comparing the intraocular pressure-lowering efficacy of bimatoprost and latanoprost in patients with ocular hypertension or glaucoma. Am. J. Ophthalmol. 2003, 135 (1), 55–63.


Clinical trials and group sequential testing

22.1 INTRODUCTION

The goals of medical research can broadly be classified into (1) identifying risk factors for disease, (2) identifying treatments for disease, (3) identifying strategies for the prevention of disease, and (4) basic science to understand the biological and physiological pathways. The chronology of scientific investigation for identifying new treatments for disease and strategies for the prevention of disease often begins with anecdotal observations from case studies that lead to the generation of new hypotheses. However, such observations provide little convincing data for treatments in general as they do not stem from a solid experimental design that minimizes the probability of the findings stemming from chance. Indeed, perhaps the most impactful statistician of all time, Sir Ronald A. Fisher (1890–1962), was famous for commenting to a colleague that was quoting the observations from a case study, “That’s not an experiment you have there, that’s an experience.” The next stage in the chronology of scientific investigation are preclinical experiments consisting of laboratory and animal experiments seeking to identify mechanistic pathways, as well as toxicology studies. Designed observational studies, including case—control studies and cohort studies, represent a third stage in the chronology of investigation. However, such studies are limited to only establishing associations between potential treatments and strategies, as opposed to establishing causality. Controlled clinical trials, representing the next stage of investigation, have been accepted as the gold standard for establishing cause-and-effect in the investigation of new treatments for disease and strategies for the prevention of disease.

The abovementioned chronology can also be thought of as establishing differing levels of scientific evidence. The US Preventive Services Task Force defines three levels of evidence in the establishment of new treatments for disease or strategies for the prevention of disease.¹ These range from at least one properly designed randomized clinical trial (Level I) to opinions by respected authorities in the field (Level III).
Why the need for clinical trials?

Clinical trials are expensive, cumbersome, and time-consuming. So why are they necessary when existing observational data often exist, or can be obtained, to estimate the relationship between a new investigative treatment or strategy? The answer is that even well-designed observational studies can suffer from unmeasured or unadjusted confounding that can provide misleading estimates of the effect a treatment. Indeed, there are numerous examples of well-designed and well-analyzed observational studies whose results were either not replicated or, even worse, reversed when a randomized controlled clinical trial was performed. Two of the highest profile examples of these are in the investigation of vitamin C for the prevention of coronary heart disease (CHD) and the use of hormone replacement therapy for reducing cardiovascular disease in postmenopausal women.

Multiple observational studies have reported a protective effect of antioxidant vitamins on cardiovascular disease. For example, Khaw et al. used data from the European Prospective Investigation into Cancer and Nutrition (EPIC) study to estimate 30% and 37% decrease in mortality from CHD at 5 years among male and female vitamin C users, respectively. These results, adjusted for many of the largest predictors of cardiovascular disease, including age, systolic blood pressure, body mass index, cholesterol, smoking, and diabetes, were both statistically significant leading one to conclude that vitamin C does indeed decrease the risk of mortality from CHD [relative risk among males 0.70, 95% confidence interval (CI): 0.51, 0.95; relative risk among females 0.67, 95% CI: 0.45, 0.90]. One year after the Khaw manuscript was published, the Heart Protection Study (HPS) of antioxidant vitamin supplementation was published (again in The Lancet, 2002). The HPS was a placebo-controlled randomized study to test the hypothesis that antioxidant vitamin supplementation could reduce mortality from congestive heart disease. In contrast to the observational study, the HPS estimated that the relative risk of 5-year mortality attributable to CHD was
1.06 (95% CI: 0.95, 1.16) comparing groups differing in vitamin C plasma concentration by 15.7 µmol/L. Thus the randomized clinical trial estimated a 6% increase in the 5-year risk of mortality attributable to CHD, though this estimate was not significantly different from zero. This begs the question: Why were results from the observational EPIC study so different from those of the randomized clinical trial? As noted by Lawlor et al., the most likely reason for this discrepancy is due to unmeasured confounding. Indeed, Lawlor considered those factors associated with vitamin C use and found that multiple socioeconomic and lifestyle factors that are associated with reduced rates of cardiovascular disease were also associated with vitamin C use. Failure to adjust for these factors when estimating the association between vitamin C and mortality from CHD in the EPIC study may have been the driving reason for the observed difference between the observational study results and the randomized clinical trial results (where vitamin C users and nonusers were balanced with respect to socioeconomic and lifestyle factors due to randomized allocation to treatment).

Another high-profile example of observational study results differing greatly from those of a subsequent randomized clinical trial is the association between hormone replacement therapy and cardiovascular disease. Observational research on postmenopausal hormone therapy suggests a 40%—50% reduction in CHD incidence among women using these preparations. To confirm these results the Women’s Health Initiative (WHI) conducted a placebo-controlled randomized clinical trial. One of the arms of the WHI consisted of treatment with daily use of 0.625 mg of conjugated equine estrogen and 2.5 mg of medroxyprogesterone acetate (referred to hereafter as E + P for estrogen plus progesterone). Despite prior observational studies finding that E + P is protective against cardiovascular disease, WHI-randomized trial stopped early by the data safety monitoring board for the study partly due to increased risk of CHD and stroke in E + P arm. Indeed, at the completion of the study, it was estimated that the risk of CHD was approximately 21% higher among women in the E + P arm compared to those in the placebo arm. However, in observational data where women were not randomized to treatment (also collected by the WHI), there was an estimated 50% reduction in the risk of CHD among women taking E + P compared to controls. Similar discrepant results were also found for stroke and venous thrombosis. Why were results so different? Certainly, part of the difference could be due to unmeasured confounding factors, as with the previous vitamin C and CHD example. However, there is compelling evidence that the difference was actually due to survivor bias occurring in the observational data. This has been conjectured because the prior exposure to E + P among users was much longer at the start of the observational study compared to exposure prior to the randomized clinical trial, where most women were naïve to E + P at the time of randomization. This can be seen in Table 22.1 as adapted from Prentice et al. Further, it was found in randomized clinical trials that the increased risk of CHD associated with E + P use was highest just after starting E + P, but the risk decreases over time. This means that the observational
studies likely observed “survivors” that remained cardiovascular disease free after a prolonged time of use of E + P, making E + P look protective relative to no use. Indeed, when the relative risk of cardiovascular heart disease is stratified by the duration of the use of E + P, the results in both the randomized clinical trial and the observational study generally agree. This can be seen in Table 22.2 (again adapted from Prentice et al.4), where in the clinical trial there is an estimated 68% increase in the risk of CHD associated with E + P during the first 2 years of use [heart rate (HR) = 1.68], but a 34% decrease in the risk after 5 years use of E + P (HR = 0.66).

The key to both the vitamin C and the hormone replacement therapy examples is that we likely would never have estimated the associations appropriately without the implementation of controlled clinical trials where patients were randomly assigned to treatment.

### 22.2 FUNDAMENTALS OF CLINICAL TRIAL DESIGN

Clinical trials, by definition, represent experimentation to investigate new treatments, devices, or preventative agents. Generally speaking, clinical trials seek to quantify (1)
the safety profile of the intervention being considered, (2) the efficacy of intervention, and (3) the effectiveness of the intervention. In assessing safety, we consider the risk-to-benefit profile of the intervention, asking whether the adverse effects associated with a new experimental treatment clearly outweigh any potential benefit. Efficacy of a treatment is a statement about the mechanistic pathways of a disease process and our questions of efficacy hence focus on whether the experimental treatment can alter the disease process in a beneficial way. Effectiveness, though related to efficacy, is different in that the effectiveness of a treatment is measured by whether adoption of the treatment as a standard will positively impact morbidity and mortality in the population. Thus effectiveness considers the net public health impact of adopting a treatment into practice given all of the caveats that go with the use of the treatments, including lack of compliance and off-label use. Because effectiveness considers the “real-world” use of a treatment, as opposed to the setting of a clinical trial, it is difficult to assess in a well-controlled experimental setting.

In order to adequately quantify the safety and efficacy profile of an intervention, fundamental elements of a clinical trial must be fully specified. These elements include a precise definition of the target population to be treated, a precise definition of the treatment or intervention, the outcome that will be used to determine efficacy, the treatment strategy that will be used to compare safety and efficacy of the treatment, and the statistical criteria that will be used to determine whether a treatment is efficacious or not.

**Defining the target population**

In defining the target population to be recruited into a clinical trial, we want a clear definition of the disease we are targeting while excluding patients for whom the likelihood of successfully completing the clinical trial is low. From a purely scientific basis, we would like to include patients for whom an improved treatment is desired, there is no contraindication to the use of the investigational treatment, and the investigational treatment might reasonably be expected to work. For clinical utility the definition of the target population should be based on information commonly available prior to start of treatment. More specifically, definitions for eligible patients that are based on diagnostic criteria available only after some delay should be avoided, since patients may be treated prior to the confirmation of eligibility. An example of this would be the use of bacterial culture that is often only available 24 hours after start of sepsis therapies. In addition, definitions based on diagnostic tests that are not routinely available in common clinical practice should be avoided, since it would be difficult to identify such patients in the general population if the treatment were actually approved for use.

Ideally, the study sample should look like a random sample from the subpopulation of all diseased patients who would ultimately be judged suitable for the intervention. Of course, we must also take into account safety considerations. To this end, we need
to consider whether at-risk patients should be exposed to unproven therapy. Common exceptions may be pregnant patients, children, patients with other comorbidities, and the elderly. These exceptions should be weighed against the potential generalizability of the trial results and the ultimate use of the therapy in the general population.

The precise definition of the study sample is formulated via the inclusion and exclusion criteria for the study. Simply put, inclusion criteria define the ultimate target population for the therapy, while exclusion criterion defines safety and ethical exceptions required for the clinical trial setting. To this end, common inclusion criteria will include an objective criteria of disease using common clinical definitions, possibly a measure of severity of disease, subgroups of interests (e.g., adults vs children or particular genetic subtypes), and any contraindications to treatment (e.g., liver or kidney disease). Exclusion criteria should consider contraindications to treatments in clinical trial setting (e.g., safety concerns with new drug that might lead to compliance issues or contraindications to the comparison treatment), requirements for evaluation of treatment outcome (e.g., inability to make clinical visits for evaluation), and requirements for ethical investigation (e.g., unwillingness or inability to provide informed consent).

**Defining the intervention**

A clinical trial will ultimately compare outcomes across populations receiving different treatments, but it is important to acknowledge that we never test a treatment. Instead, we test a treatment strategy. To see this, we need only note that we cannot ethically force participants to continue on therapy and in many cases it may not even be medically advisable to want a participant to continue. Instead, we prescribe an initial treatment that may be discontinued at any time. In addition, patients may also receive ancillary treatments that are not part of the assigned treatment. These may be precipitated by an experimental therapy or patients may progress to other therapies (e.g., first to second line therapies in cancer research).

Given that we are testing a treatment strategy, it is important that a full description of the treatment be given when planning a clinical trial. This entails stating the formulation of the treatment to be investigated, including the dose, administration, frequency, and duration of treatment. It also entails defining the rules for treatment of adverse events that may arise during the course of the trial, including dose reductions and discontinuations, the use of acceptable ancillary treatments.

**Defining the outcome**

The objective of a clinical trial is to find effective treatment indications, and the primary outcome is a crucial element of the indication. From a scientific perspective, a clinical trial is planned to detect the effect of a treatment on some outcome and hence statement of the outcome is a fundamental part of the scientific hypothesis being tested.
in the trial. From an ethical perspective, subjects participating in a clinical trial are hoping that they will benefit in some way from the trial, so that clinical endpoints (or endpoints that the patient can actually feel) are therefore of more interest than purely biological endpoints. When deciding upon a trial outcome, one should seek to choose (1) the most relevant clinical outcome (e.g., survival or quality of life), (2) the outcome the treatment is most likely to affect, and (3) the endpoint that can be assessed most accurately and precisely. Examples of clinical endpoints might be overall survival in cancer clinical trials, 6-minute walk distances in cardiovascular disease trials, or functional ability in dementia trials.

A common problem is that the clinical outcomes are rare or occur after a long time. This has an impact on trial design. Namely, larger sample sizes are required to detect treatment effects on rare events and/or long periods of follow-up may be needed to assess clinical endpoints. Because of this, many trial designers consider the use of surrogate outcomes. A surrogate outcome is a biological endpoint that ideally can be measured in a shorter time frame, can be measured precisely, and is predictive of the clinical outcome of interest. In this case, use of a surrogate may increase trial efficiency by reducing the number of patients or length of the trial in order to show an effect of treatment on the surrogate. Examples of surrogates include CD4 counts in HIV/AIDS trials (instead of overall survival) and arrhythmias in cardiovascular disease trials (instead of myocardial infarction (MI) or overall survival).

While substantial savings in time and patients may be realized by the use of a surrogate, this comes with some cost. Specifically, when a surrogate outcome is used, the basic assumption being made is that the treatment effect on the surrogate is a good indication of its effect on the clinical outcome. It is often believed that having a surrogate that is strongly associated with a clinical outcome will imply that the treatment effect on the surrogate will then be a good indication of its effect on the clinical outcome. However, this is not always the case and, in many situations, sensible, yet terribly misleading surrogates, have been used in clinical trials. A classic example is the use of arrhythmias as a surrogate endpoint in the Cardiac Arrhythmia Suppression Trial (CAST).\textsuperscript{5} CAST was designed to assess the efficacy of treatment with encainide or flecainide for reducing the incidence of sudden death among patients that had previously experienced a MI. Because arrhythmias are highly correlated with death and MI, and because encainide and flecainide had previously been shown to effectively reduce arrhythmias, there was a strong sentiment that the drugs would decrease the risk of sudden death in this patient population. While CAST showed a significant reduction in the rate of arrhythmias among patients receiving encainide and flecainide, relative to placebo, the study was stopped early by a data and safety monitoring board due to an increased risk of death associated with encainide and flecainide. This is a classic example of arrhythmias as a misleading and dangerous surrogate outcome. Many other examples of misleading surrogate outcomes have been pointed out in the literature.\textsuperscript{6,7}
Choosing a comparison group

Clinical trials can utilize no comparison group, a historical control, or concurrent comparison group(s). Having a comparison group is important when deciding whether a proposed treatment is effective and when deciding among the alternatives when treating a patient. Using no comparison group results in a single-arm clinical trial. This is only relevant when an absolute criterion for a treatment effect exists, but this is rare in practice. The use of historical controls also results in a single-arm trial but attempts to compare the experience of the single arm to either an absolute criterion derived from historical trials (e.g., the outcomes observed in the placebo arm of a past trial). While this sounds appealing at first (because only half of the patients would need to be collected if existing data on control patients is to be used), the assurance in making a “fair comparison” is heavily dependent upon the historical trial being comparable in every way to the patients enrolled in the current trial. This means that there can be no changes in the comparison treatment, no changes in the definition of study population, no changes in the ancillary treatments, no changes in the measurement of treatment outcomes, and no differences in important patient characteristics between patients in the historical trial data and those in the current trial. Because the previous criteria are difficult, if not impossible to guarantee, this has led to a lack of enthusiasm for the use of historical controls in clinical trials and most trialists avoid the use of historical controls whenever possible, opting instead to use concurrent control populations. In the case of concurrent control populations, two or more treatment arms are followed in the trial. A placebo arm should be used as a control if no other standard of therapy exists in practice. If a standard therapy does exist, it may be unethical to randomize participants to placebo. In this case an active treatment is used for comparison. In addition, there may be cases where multiple arms in the trial represent multiple levels of the same treatment. This strategy would be used when there is a desire to assess evidence of dose–response effect or to identify an “optimal dose.”

Statistical criteria for evidence

Last, but certainly not least, it is essential to a priori define the statistical criteria that will be used to conclude whether a treatment is deemed to be efficacious or not. By clearly specifying the trial outcome, treatment strategies, and comparison groups, one can fully specify the statistical test that will be used for comparing outcomes across treatment arms (or relative to a single criterion in the case of a single-arm study). It is fairly common to bound the type I error rate at 0.05 for a two-sided test (or 0.025 for a one-sided test). Further, in order to ensure adequate precision, the number of
subjects to be enrolled into the trial should be determined to ensure that sufficient statistical power is attained if a minimal clinically relevant effect exists. Further discussions of sample size and power can be found in Chapter 21, Sample size estimation.

22.3 REDUCING BIAS IN CLINICAL TRIALS: BLINDING AND RANDOMIZATION

The reason that controlled clinical trials are viewed as the gold standard for assessing cause-and-effect associated with medical treatments is due to our ability to reduce (or eliminate) bias by experimental design. Bias is a tendency of a statistical estimate to deviate in one direction from a true value. The EPIC study example presented in the “Why the need for clinical trials?” section is an example where estimates of the effect of vitamin C on CHD were biased in the observational study, indicating that vitamin C is protective. However, when compared to the results of the randomized clinical trial, we saw there was in fact no difference between patients treated with vitamin C and those treated with placebo. We saw that this bias was likely due to differences in socioeconomic factors between vitamin C users and nonusers in the EPIC studies, thus making an unfair comparison between the groups. Bias can also arise when patients or caregivers report outcome with the knowledge of what treatment the patient is receiving. This is because we as humans tend to believe that a new experimental treatment is likely to work (Why else would be investigating it?) and hence this can skew our thoughts about the occurrence, frequency, or severity of trial outcomes. This is most widely seen in the well-known “placebo effect” in cases where patients treated with a placebo report feeling better because they believe that they are being treated with an active therapy. The two most important clinical trial design strategies that can be used to reduce bias are blinding (or masking) and randomization.

Blinding

Blinding occurs when neither the study subject (single-blind) nor the study investigator (double-blind) has knowledge of the treatment being received or delivered. Blinding can serve to minimize the placebo effect, minimize investigator bias in assessing adverse events, and minimize bias due to missing data that arise when patients may be less likely to continue in the trial with knowledge that they are receiving placebo. Because of this, blinding should always be used whenever possible. Effective blinding requires the existence of a placebo that is identical in physical form and feel to the experimental treatment, has the same mode of delivery, and has a similar side-effect profile. Many times, a suitable placebo can be found; however, this may not always be possible. Sometimes, it is difficult or impossible to obtain a nonactive placebo that is identical in appearance, weight, and/or viscosity (if the active treatment is
CHAPTER 22 Clinical trials and group sequential testing

an injectable fluid). Many times, active treatments have common side-effect profiles that reveal them. One example of this is the use of methylene blue (LMTM), an experimental therapy that has been investigated for the treatment of Alzheimer’s disease and is known to make the urine of participants turn blue. Finally, for ethical reasons, it may not be possible to use a placebo. This is particularly the case when there exists an accepted standard of care for treatment or when the experimental intervention involves an invasive surgery or hospitalization.

Exercise 22.1
Consider a randomized clinical trial designed to assess a new topical hemostatic agent for the treatment of moderate bleeding during abdominal surgery. The new treatment will be compared to an existing treatment where patients will be randomized to one of the two arms at the first identification of a bleeding event during their surgery. Patients who agree to the trial provide informed consent to enter the study during their preoperative meeting. Patients are informed that they will not know which hemostatic agent they will be treated with. Surgeons in the study, however, will know which agent they are using because the two products are physically different, and the surgeon is the one who must apply the agents. What type of blinding, if any, is being implemented in the study?

Randomization
In order to establish cause-and-effect, it is necessary that when comparing groups differing in their treatment, the groups should be comparable in every other way (at the start of treatment). The best technique for ensuring this is to randomly assign trial participants to the trial’s treatment groups. When this is done, any imbalances between groups that may be observed in the trial will be strictly due to chance. The simplest, and most commonly employed, randomization scheme is known as a completely randomized trial design. In the case of a completely randomized design, treatment assignment is made by randomly allocating a subject to one of the treatment groups without considering previous treatment allocations or the subject’s covariates. Fig. 22.1A gives an example of a completely randomized design scheme between two groups, A and B. In the case of a 1:1 randomized trial such as Fig. 22.1A, where the intention is to have roughly an equal number of patients randomized to each of two treatment groups, each participant has equal probability of getting either of the treatments (such as flipping a fair coin). In some cases, it is desirable to allocate more patients to the active therapy, resulting in r:1 randomization. This may be the case when additional information is needed to assess the safety profile of the new experimental therapy. In this setting, we may wish to have 2:1 randomization in which approximately two-thirds of participants are randomized to treatment and one-third are randomized to control. Then a completely randomized design would assign patient to treatment with
22.3 Reducing Bias in Clinical Trials: Blinding and Randomization

Probability 2/3 and to control with probability 1/3. Note that with a completely randomized design, it is not guaranteed that under $r:1$ randomization the exact ratio of patients on two arms is actually $r:1$ at the end of the study. For example, in Fig. 22.1A, by chance eight patients are randomized to group B, while only four patients are randomized to group A.

Imbalances in important prognostic factors can still arise by chance between treatment groups when a completely randomized design is used. To avoid this, stratified randomization is often used. In this setting, important prognostic factors (e.g., age above or below 60, stage of disease, or geographic location) are a priori specified, then patients are randomized in a $r:1$ fashion within each stratum. The result of this is that balanced between treatment groups is forced within each stratum, eliminating the possibility of imbalances with respect to these important covariates. Fig. 22.1B gives an example of a stratified randomization scheme with three strata. Notice that within each stratum, six patients are randomized to group A and six patients are randomized to group B (by design). Thus balance between groups is observed in each of the three
strata. When stratified randomization is used, the analysis should be adjusted for the covariates used in the stratification otherwise a loss of precision can result.

Another concern with the completely randomized design is that large imbalances between treatment groups can occur over time. The third stratum in Fig. 22.1B provides an example of this. After six patients are randomized in the third stratum, five of the patients are randomized to group A, while only one patient is randomized to group B. While this will balance out in a large trial, it can have impacts on bias in some settings. One example is when there is a learning effect associated with the treatment intervention strategy. If this is the case, then it would not be surprising that outcomes for early enrolled patients would be worse than those for later enrolled patients. Thus having imbalances in randomization ratios over time would result in a biased estimate of the treatment effect. This is a special case of what is known as temporal confounding. To eliminate the possibility of temporal confounding, a common strategy is to use a blocked randomization scheme. In a blocked randomization scheme, one specifies the size of blocks, say \( k \), and then ensures that there is the desired ratio of patients randomized to each group after every \( k \) patients are randomized in the study. As a simple example, suppose we took \( k = 4 \) and desired 1:1 randomization in the study. This would mean that after every four patients randomized, we would force two patients to be randomized to each group. So we would flip a coin for patient 1. If they received Heads (H), they would be randomized to treatment. Then we would flip a coin for patient 2. If they also received H, they would be randomized to treatment, and patients 3 and 4 would be forced into the control group. However, if patient 2 received Tails (T), they would be randomized to the control, and patient 3 would receive a coin flip of either H or T. In any event, patient 4’s allocation would be completely determined by the first three patients, since the randomization ratio would need to be satisfied after each set of four patients are randomized. After the first four patients have been randomized, the process would then start over again for the next four patients. It is generally desirable to conceal the future group allocation of patients from investigators in order to avoid bias (e.g., a treating physician may seek to put sicker patients on an experimental arm if they believe all other options have been exhausted). If the block size becomes known to investigators, then the deterministic pattern at the end of each block eliminates concealment. To avoid this, it is common practice to implement randomly permuted blocks. In this case a discrete set of block sizes is defined (e.g., 2, 4, and 6) and then a string of these block sizes is randomly generated. After that, for each block, patient randomization then occurs as usual. Doing this means that an investigator cannot guess the block sizes because they are constantly (randomly) changing. Fig. 22.1C gives an example of a blocked randomization scheme where block sizes of 2, 4, and 6 are randomly permuted. We can see that at the end of each block, an equal number of patients are randomized to each arm, as desired.
Stratified randomization is effective for dealing with imbalances between groups with respect to important prognostic factors, while blocking is effective for dealing with imbalances between groups over time. A final and common strategy is to combine these two techniques into what is called a stratified blocked design. In this case, strata are defined as before, and blocking is used within each stratum. Fig. 22.1D provides an example of a stratified blocked randomization scheme. Notice that the string of block sizes differs in each stratum, however, after each block an equal number of patients is randomized to each arm.

**Exercise 22.2**

Again consider the randomized clinical trial designed to assess a new topical hemostatic agent for the treatment of moderate bleeding during abdominal surgery (see Exercise 22.1). The study sponsor believes that there may be a “learning effect” with the two agents to be used because surgeons will learn to apply them more adequately with increased practice. What randomization strategy would best address concerns of a potential unfair comparison due to these learning effects.

### 22.4 INTERIM ANALYSES IN CLINICAL TRIALS: GROUP SEQUENTIAL TESTING

Clinical trials represent experimentation in human volunteers. The objective of clinical trial design is therefore to find a procedure that ensures scientific credibility while at the same time protecting the safety of human subjects. The ethical constraints associated with protecting patients include both the individual ethics related to the subjects enrolled on the clinical trial and the group ethics related to the population of people who might benefit from the rapid adoption of a new treatment or preventive strategy.

During the conduct of a clinical trial, it is now common for the accruing data to be monitored repeatedly in order that patients on the study not be unnecessarily given a treatment known (or credibly demonstrated) to be inferior, and that new, beneficial treatments be adopted as rapidly as possible. The goal is often to allow the early termination of the clinical trial as soon as there is high confidence in a decision about whether to alter standard clinical practice. However, repeated analysis of accruing data can lead to multiplicity bias, thereby increasing the overall type I error rate of the trial to higher than the desired level. This is explored in the following subsection.

**Disadvantages of fixed-sample methods**

Consider a single-arm clinical trial designed to test whether a new cholesterol lowering drug is effective in reducing mean total cholesterol over 6 months. One way to carry out such a trial would be to measure the baseline total cholesterol value of each patient, prescribe them the treatment, then measure total cholesterol after 6 months of being
on treatment. If we let \( \mu_d = \mu_{\text{post}} - \mu_{\text{pre}} \) denote the true mean change in total cholesterol over 6 months, we would be interested in testing the null hypothesis \( H_0: \mu_d \geq 0 \) versus the alternative hypothesis \( H_A: \mu_d < 0 \). Suppose we wish to carry out a level 0.05 test of this hypothesis and that we were going to test the hypothesis after every 25 patients is observed (here we consider the unrealistic case where we only enroll a new cohort of 25 patients after the previous cohort of 25 patients has finished 6 months of follow-up). Further suppose that we will enroll 100 patients in total, so that we will test \( H_0 \) four times. Using the one-sample methods described in Chapter 11, Tests of location with continuous outcomes, we might carry out each of the abovementioned tests by comparing
\[
 z = \frac{m_{\text{post}} - m_{\text{pre}}}{SE(m_{\text{post}} - m_{\text{pre}})}
\]
to 1.96, the 0.95 quantile of the standard normal distribution. Implementing this procedure will not yield the desired type I error rate of 0.05 for the study. Instead, the true type I error rate is approximately 0.127. This means that if there really were no effect of treatment on cholesterol, we would erroneously reject \( H_0 \) 12.7% of the time if we repeated the experiment over and over again (a 2.5-fold increase in the desired type I error rate).

Why does the increase in the probability of falsely rejecting the null hypothesis occur in the situation above? It is because we have allowed ourselves to make a type I error (falsely rejecting the null hypothesis) 5% of the time at each of the four tests that were performed (recall that this is how we came up with the critical value of 1.96 in Chapter 11: Tests of location with continuous outcomes). Thus with each test we have a 5% error rate, so that over all four tests our rate must be higher than 5%. Table 22.3 provides that actual proportion of trials that would be erroneously rejected at each analysis and overall if we implement this “fixed-sample testing procedure.” These values are derived from 100,000 simulated clinical trials where \( \mu_d = 0 \) in truth (i.e., under the null hypothesis). Through simulation we can “generate” data on patients in each trial where

<table>
<thead>
<tr>
<th>Significant at</th>
<th>Proportion significant</th>
<th>Number significant</th>
<th>Proportion significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis 1</td>
<td>0.05075</td>
<td>Exactly 1</td>
<td>0.07753</td>
</tr>
<tr>
<td>Analysis 2</td>
<td>0.04978</td>
<td>Exactly 2</td>
<td>0.02975</td>
</tr>
<tr>
<td>Analysis 3</td>
<td>0.05029</td>
<td>Exactly 3</td>
<td>0.01439</td>
</tr>
<tr>
<td>Analysis 4</td>
<td>0.05154</td>
<td>All 4</td>
<td>0.00554</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Any</td>
<td>0.12721</td>
</tr>
</tbody>
</table>

At each analysis the proportion of trials falsely rejecting the null hypothesis is given on the left side of the table. The right side of the table provides the proportion of trials rejecting the null hypothesis at exactly 1, 2, 3, or all 4 of the analyses. The overall type I error rate is given as the sum of these values, representing the proportion of trials rejecting the null hypothesis for at least one analysis.
we know there is no treatment effect, then count the number of trials where we would have falsely concluded that the treatment works. As expected, for each analysis we have approximately a 5% chance of erroneously rejecting the null hypothesis and concluding that the treatment works for reducing cholesterol. Also presented in the table, however, is the proportion of trials where the null hypothesis was falsely rejected exactly at 1, 2, 3, or all 4 of the analyses. When added up, we obtain the proportion of trials where we falsely rejected the null hypothesis at least one time. This proportion is approximately 12.7% and is the overall type I error rate for the testing procedure.

Examples of group sequential stopping rules

We saw in the previous subsection that applying a standard fixed-sample testing procedure will not control the type I error rate for a study to the desired level. So how do we fix it? The answer lies in a domain of statistics known as group sequential testing. The idea behind group sequential testing is that we periodically test our data after groups of patients have been enrolled and observed for the outcome in our study. This differs from truly sequential testing where a test is performed after every observation. For logistical reasons, it is usually infeasible to implement truly sequential testing in clinical trials because of the amount of time it generally takes to gather and clean data for each patient.

One of the earliest group sequential tests was proposed by the statistician Pocock. Pocock asked the following natural question: If we cannot use a critical value of 1.96 to control the family-wise type I error rate at 0.05, what value should we use? Using results from Armitage et al., Pocock was able to compute that the correct value is 2.36 in the abovementioned example where we implement four tests at equally spaced numbers of patients. Thus Pocock looked for a single common value to compare the usual normalized z-statistic in order to yield the desired type I error rate. This value intuitively depends upon the number of tests being performed. This testing procedure that uses a common critical value at each analysis has become known the Pocock design. Table 22.4 provided the critical value needed for one-sided level 0.025 and two-sided

<table>
<thead>
<tr>
<th>Total number of interim analyses (J)</th>
<th>Pocock critical value (z-statistic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.96</td>
</tr>
<tr>
<td>2</td>
<td>2.18</td>
</tr>
<tr>
<td>3</td>
<td>2.29</td>
</tr>
<tr>
<td>4</td>
<td>2.36</td>
</tr>
<tr>
<td>5</td>
<td>2.41</td>
</tr>
<tr>
<td>10</td>
<td>2.56</td>
</tr>
</tbody>
</table>

The single critical value given is used at all analyses.
level 0.05 tests when \( J = 1, 2, 3, 4, 5, \) and 10 equally spaced analyses are performed. We can quickly see that the critical value increases with the number of analyses. This is because adding analyses gives us additional opportunities for rejecting the null hypothesis, and this must be accounted for in order to maintain the overall type I error rate of the test.

Not long after the Pocock design was created, O’Brien and Fleming noted that we may wish to be more conservative when testing a treatment effect early on in a clinical trial. The rationale for this is multifaceted. First, we will have highly variable estimates of treatment effects making us less confident in our result. Next, there are often “cohort effects” in clinical trials because the patient population enrolled into trials can change over time (e.g., sicker patients might be recruited first, then healthier patients might be recruited later). Also, some of the effects of a treatment might vary over time. For example, a new HIV drug may be effective at prolonging short-term progression to AIDS but after the virus mutates the drug may have no long-term effect (which patients generally care more about). This would result in a high difference in progression to AIDS at an early analysis, but little or no difference at a later analysis. Finally, we may wish to be conservative at early tests because we have not observed the long-term safety profile of the new investigative therapy for very long. Prolonging a clinical trial allows time to observe potentially rare but serious safety signals to arise. For the abovementioned reasons, O’Brien and Fleming proposed what is known as the O’Brien–Fleming design. While too technical for the current text book, the O’Brien–Fleming design is characterized by finding a fixed value to compare the sum of observations at each interim analysis in order to maintain the overall type I error rate (as opposed to a fixed value to compare the normalized \( z \)-statistic to in the Pocock design). The result of this is that it is much harder to reject the null hypothesis at early analyses when the O’Brien–Fleming design is used in comparison to the Pocock design. Table 22.5 provides the stopping boundaries on the normalized \( z \)-scale for the O’Brien–Fleming design (one-sided level 0.025 or two-sided level 0.05 tests) with \( J = 1, 2, 3, 4, \) and 5 equally spaced analyses. A comparison with the Pocock boundaries reveals the high early conservatism of this design, though both designs yield the same overall type I error rate. To illustrate the utility of the O’Brien–Fleming design for maintaining the overall type I error rate, we create the simulated example presented in the “Disadvantages of fixed-sample methods” section, but this time use the O’Brien–Fleming boundaries given in Table 22.5 (\( J = 4 \)) instead of the fixed-sample critical value of 1.96. In Table 22.6 we see that the probability of committing a type I error rate at the first analyses is quite low, indicating the conservatism of the test, but that the overall type I error (i.e., the proportion of trials falsely rejecting on any one of the four analyses) is approximately 5% as desired.

One additional benefit of this high conservatism at early analyses is that the O’Brien–Fleming design yields higher statistical power for the same maximal sample
size when compared to the Pocock design. This is intuitively true because there is less of a chance of making a type I error at early analyses with the O’Brien–Fleming design, and hence there is less of a “penalty” that must be paid to control for multiplicity. Because of the reasons stated previously regarding the desire for early conservatism in testing, the O’Brien–Fleming design is arguably the most commonly used
group sequential design in clinical trials. An infinite number of tests that maintain a desired family-wise type I error rate can be formulated and advanced reading on the topic can be found elsewhere, however, the O’Brien–Fleming and occasionally Pocock designs are most commonly implemented given their intuitiveness and historical positioning in group sequential testing methodology.

**Exercise 22.3**

Suppose you wish to design a clinical trial to compare the 6-month change in testosterone levels among hypogonadism patients treated with a new experimental testosterone replacement therapy versus placebo. In total, 240 patients will be randomized in a 1:1 fashion to the new therapy or placebo. To increase efficiency of the study, the trial investigators wish to conduct an O’Brien–Fleming group sequential test with interim analyses after 80, 160, and 240 patients have a 6-month measurement. A z-statistic is to be computed at each interim analysis and the investigators wish to test the difference between the new treatment and placebo using a level 0.025 one-sided test. What critical values should be used to compare the observed z-statistics to at each of the interim analyses?

**REFERENCES**

23.1 THE NATURE OF EPIDEMIOLOGY

Definitions

An epidemic is the occurrence in a community or region of cases of an illness, specific health-related behavior, or other health-related events clearly in excess of normal expectancy. Epidemiology is the study of the occurrence of illness in populations. To the epidemiologist comparing the rate of occurrence with expectancy, two cases of plague in a city of 6 million people may constitute an epidemic, whereas 10,000 cases of influenza may not. Epidemiology started with the analysis of mortality recordings, from which surprising insights into the patterns of public disease and health events immediately demonstrated its importance. It long concentrated on mortality from infectious diseases (e.g., smallpox), then evolved to include any disease (e.g., cardiovascular disease, cancer) and morbidity of diseases, and finally broadened to include any condition threatening public health (e.g., smoking, mercury in a water or food supply).

Epidemiology compared with other branches of medicine

Epidemiology, like all of medicine, seeks to understand and control illness. The main conceptual difference between epidemiology and other branches of medicine is its primary focus on health in the group rather than in the individual, from which arises the concept of public health. The main methodologic difference between epidemiology and other branches of medicine is its attention to rates and distributions. Basic to epidemiology is thinking probabilistically rather than in mutually exclusive categories. The epidemiologist is concerned not with a patient who is sick or well but rather with the proportion ill from a catchment of patients and the probability that the proportion will grow. To have a baseline for estimating these proportions and probabilities, the epidemiologist continually is concerned with denominator data.

Stages of scientific knowledge in epidemiology

The accrual of knowledge in epidemiology follows the same stages as in other fields of medicine, first met in the beginning words of this book: description, explanation, and prediction. Let us consider these stages in sequence. In description, the etiology of a public health disease or condition is documented. This stage generates basic epidemiologic data and generates the
scientific hypotheses to be tested. In explanation, the epidemiologic data are assessed to explain the outbreak of public disease. The scientific hypotheses are tested. The next stage is prediction, which integrates the test results and provides a model of the course of an epidemic, which can be used in prevention or control of the outbreak and even in assessment of the efficacy of potential treatments or measures of prevention and control.

**Epidemiology is an eclectic science**

Epidemiology must interact with many fields of study to isolate causal factors: human and veterinary clinical medicine in their various specialties; biologic, biochemical, and genetics investigations; social, economic, and demographic factors; geographic and historic patterns; and the personal habits of individuals. The epidemiologist must be a general health scientist.

*Exercise 23.1*

*How is epidemiology similar to clinical diagnosis? How is it different?*

### 23.2 SOME KEY STAGES IN THE HISTORY OF EPIDEMIOLOGY

The respective beginnings of the descriptive, explanatory, and predictive stages in epidemiology may be followed over three centuries. Quantified description in epidemiology began in Great Britain in 1662, when John Graunt published a tabular analysis of mortality records. In 1747, James Lind, also in Great Britain, pioneered isolating causes by experimentation, finding the nutritional rather than infectious basis of scurvy. Although epidemiology was to wait another century for formal predictive modeling, the theory of contagion allowed the predictive stage of epidemiologic science to begin on an informal basis in the 1800s. An example is the 1847 work of Ignaz Semmelweis in Austria, who traced the source of the highly fatal disease puerperal fever by epidemiologic methods and showed how to control it.

*Exercise 23.2*

*List the three major steps in the evolution of epidemiologic methodology.*

### 23.3 CONCEPT OF DISEASE TRANSMISSION

**Spectrum of disease**

The sequence of events from exposure to a disease to resolution, often death, is termed the *spectrum of disease*. It arises from the interaction among three main factors:
the host, the agent, and the environment. The *host* includes primary hosts in which a parasite develops, for example, intermediary hosts (the host is a component in the disease cycle) and carrier hosts (host only transports the disease in space or time). The host as a member of the epidemiologic catchment is characterized by the susceptibility to acquiring the disease, which includes behavior leading to exposure, immunization or resistance, and the like. The *agent* is not necessarily an infectious agent. It may simply consist of a deficiency of necessary nutrient (vitamin C in the case of scurvy), the excess of a deleterious substance (tobacco smoke), or a metabolic condition (metabolic acidosis). More dramatically, it may consist of microbes (the bacterium *Yersinia pestis* in the case of plague) or toxins (mercury poisoning, or *yusho*, the name that arose from the classic episode in Minamata, Japan). The *environment* as a component in the spectrum is characterized not only by the physical and biologic properties (in malaria: still water to breed mosquitos, access by the mosquitos to previously infected hosts, exposure of the catchment member to the infected mosquito) that affect disease transmission but also by the social properties (as in sexually transmitted diseases).

### Modes of transmission

The classic book by Lilienfeld and Stolley\(^2\) classifies two modes of transmission: common-vehicle epidemics, in which the vehicle is water, food, air, and so forth, and serial transfer epidemics, in which the vehicle is a host-to-host transfer, such as infectious transmission by touching, coughing, and so on.

### Herd immunity

Rates of infectious disease transmission are not constant. The pattern of an epidemic (depending considerably on incubation period) is usually an initial surge of infected patients, continuing until a substantial number of the epidemiologic catchment is immune. At this point, the rate slows, the return cycle to the infectious agent is interrupted, and the epidemic wanes. This high proportion of immunity within a demographic catchment has been termed *herd immunity*. When the herd immunity reaches a critical level, the rate of change of new infection changes from positive to negative and the epidemic starts to decline. The population’s susceptibility to epidemic may be reduced by increasing the herd immunity, for example, by vaccination. It should be noted that herd immunity is by no means the only mechanism to alter the rate of new infection. Many occurrences may alter it. For example, it was noted as early as the 1700s that a yellow fever epidemic often was terminated by a change in the weather.\(^3\)

**Exercise 23.3**

Name two diseases in which there can be no herd immunity.
23.4 DESCRIPTIVE MEASURES

Incidence and prevalence

Incidence and prevalence have been used informally, although not incorrectly, in previous chapters, assuming the student is familiar with the everyday connotations of the terms. In epidemiology, the terms are used with technical precision, as indeed they should be in all fields of medicine. The incidence rate of a disease is the rate at which new cases of the disease occur in the epidemiologic population. The prevalence rate of the disease is the proportion of the epidemiologic population with that disease at a point in time. Thus during an influenza epidemic in a certain city in November, the prevalence indicates how much of the population is sick and the incidence indicates how rapidly the epidemic is increasing. Incidence and prevalence rates should be scaled to make them meaningful. They are often given per 1000 members of the population, unless otherwise indicated. In the formulae and examples below we will express them per 1000, but in practice then scaling should be specific to the problem being addressed.

Mortality rate

Similar to incidence rate is mortality rate, the rate at which the population is dying rather than becoming ill. For a disease resulting in certain death, the mortality rate at the end of the duration period of the disease would be the same as the incidence rate at the beginning of the period.

FORMULAS

Incidence rate must be expressed in relation to an interval in time, for example, as “2000 new cases of illness per month.” This is because the number of new cases will be zero at a point in time, that is, when the interval goes to zero. Prevalence rate, on the other hand, can be measured at a point in time, although sometimes a short interval must be used to allow for the time to obtain the prevalence data. Let us denote the number of individuals in the epidemiologic population as $n$, the number of new cases in a specified interval as $n_{\text{new}}$, and the number of cases present at any one point in time by $n_{\text{present}}$. Then, incidence rate $I$ per 1000 is given by

$$I = 1000 \times \frac{n_{\text{new}}}{n},$$

(23.1)

and the prevalence rate $P$ is

$$P = 1000 \times \frac{n_{\text{present}}}{n}.$$  

(23.2)

If we denote by $n_{\text{dying}}$ the number of patients dying during the specified interval, the mortality rate $M$ becomes

$$M = 1000 \times \frac{n_{\text{dying}}}{n}.$$  

(23.3)
In the remainder of this chapter, the terms **incidence** and **prevalence** are used when incidence rate and prevalence rate are actually intended, because this is the general custom in medical articles, at least outside strict technical usage in epidemiologic articles.

**EXAMPLE: CERVICAL CANCER IN THE ACORNHOEK REGION OF THE TRANSVAAL**

The population served is given as 56,000. From diagrams of sex ratios and assumptions about the child/adult ratio, we believe the adult female population served to be 13,000. (A further subtraction should be made for adult women who never present to medical services, but this unknown number will be ignored for this illustration.) The study covered the 9 years from 1957 to 1966. Cervical cancer was seen in 53 cases, or about 6 per year. Substituting in Eq. (23.1) yields \( I = \frac{1000 \times 6}{13,000} = 0.462 \). The yearly incidence of cancer of the cervix is about 0.5. This type of cancer in women rarely goes into remission. To find prevalence, let us assume, quite arbitrarily, that a woman survives, on average, 2 years after diagnosis. Then there will be about 12 women with cancer of the cervix surviving at any one time, and Eq. (23.2) yields \( P = \frac{1000 \times 12}{13,000} = 0.923 \), or about 0.9. After the first 2 years, \( M \), from Eq. (23.3), is the same as \( I \).

**The odds ratio**

The odds ratio (OR), which was introduced in Section 10.1, also is used in epidemiology. Let us examine an epidemiologic study of the interrelation between occurrences of cervical cancer and schistosomiasis. (Schistosomiasis is a parasitic infection common in parts of the Third World that is contracted during immersion in river or stream water.) Table 23.1 reproduces Table 1 from a study by Riffenburgh et al. Some clinicians in Africa have suggested a protective effect against cancer of the cervix by schistosomiasis. Do the figures support this? Let us calculate the cervical cancer OR for schistosomiasis. Odds of cervical cancer given schistosomiasis, \( \frac{101}{165} = 0.6121 \), in ratio to odds of cervical cancer given not schistosomiasis, \( \frac{101}{165} = 0.6121 \), in ratio to odds of cervical cancer given not

<table>
<thead>
<tr>
<th>Observed frequencies</th>
<th>With schistosomiasis</th>
<th>Without schistosomiasis</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>With cervical cancer</td>
<td>101</td>
<td>5212</td>
<td>5313</td>
</tr>
<tr>
<td>Without cervical cancer</td>
<td>165</td>
<td>2604</td>
<td>2769</td>
</tr>
<tr>
<td>Totals</td>
<td>266</td>
<td>7816</td>
<td>8082</td>
</tr>
</tbody>
</table>

schistosomiasis, $5212/2604 = 2.0015$, provide OR = 0.3058. These figures indicate that the odds of having cervical cancer are more than three times as high as that for women who do not have schistosomiasis! The figures appear to support the claim of a protective effect. However, the epidemiologic OR has been derived from biased sampling. The rate of schistosomiasis without cervical cancer is markedly overreported. The often overwhelmed, undertrained African clinician who finds the presence of *Schistosoma haematobium* often ceases to look further; the patient’s complaints have been explained adequately, and the cervical cancer is not detected. Thus we could expect that many of the patients reported in the lower left cell of the table should be moved to the cell above it. Furthermore, the rate of cervical cancer with schistosomiasis is markedly underreported for an entirely different reason: schistosomiasis tends to be acquired at an earlier age (mean = 30 years), and these patients are less likely to live to the age at which cervical cancer tends to appear (mean = 45 years). If these two reports were corrected, the OR would be considerably changed. It should be clear to the student that high-quality data are crucial prerequisites to meaningful epidemiologic results.

**Exercise 23.4**

In a Canadian study including the effect of caffeine consumption on fecundability,\(^6\) 2355 women drank caffeine-containing drinks and 89 did not. Of those who consumed caffeine, 236 conceived within 6 months of beginning pregnancy attempts. Of those who did not consume caffeine, 13 became pregnant. Calculate the 6-month incidence of pregnancy for the two groups. Note carefully that the population being considered is women attempting pregnancy, not the general population. Among this group, prevalence might be thought of as the rate at which pregnancy occurs at all, whereas the population attempting it is equivalent to the population at risk in morbidity studies. A total of 1277 couples attempting pregnancy were followed for an extensive period of time and 575 were successful. Of these, 304 of the women consumed caffeine. Calculate the prevalence of caffeine consumption among women who became pregnant. Furthermore, calculate the OR (fecundity ratio) of a caffeine-consuming woman becoming pregnant.

**Exercise 23.5**

A 2003 article\(^7\) addresses the occurrence of acute gastroenteritis caused by Norwalk-like viruses among crew members on the US Navy vessels Peleliu (2800 crew members) and Constellation (4500 crew members) visiting ports in Southeast Asia. During a period of rapid increase in the disease, 28 cases appeared over 5 days on Peleliu and 31 cases appeared over 14 days on Constellation. Calculate the incidence during the initial increase for the two ships. During an ensuing stable period (before a reduction arising from herd immunity), an average of 16.56 cases were seen on Peleliu and 29.37 cases on Constellation. Calculate the prevalence during this stable period. The total number of cases for the outbreak was 162 on Peleliu and 425 on Constellation. Calculate the OR for Constellation as compared with Peleliu.
23.5 TYPES OF EPIDEMIOLOGIC STUDIES

Basic variables
All epidemiologic studies are framed in terms of exposures and outcomes. In simplest terms, the exposure is the putative cause under study, and the outcome is the disease or other event that may result from the exposure.

Experimental studies and intervention trials
The investigator assigns exposure (or nonexposure) according to a plan. The concept of an experimental control, introduced in Section 1.5, is used for unexposed subjects in experimental studies. In clinical trials, patients already with disease are subjects; an example would be James Lind’s study of scurvy. In field trials, patients without disease are subjects; an example would be a vaccine trial in a population. Community intervention trials are field trials that are conducted on a community-wide basis.

Nonexperimental or observational studies
In nonexperimental or observational studies, the investigator has no influence over who is exposed. Two types of such studies are cohort studies and case-control studies, both met in the context of clinical research in Section 1.6. Each study design has inherent strengths and weaknesses.

COHORT STUDIES
Cohort studies are also termed follow-up or incidence studies. In these studies, outcomes among two or more groups initially free of that outcome are compared. Subjects may be selected randomly or according to exposure. Indeed, if outcome comparison is the purpose, the study by definition is a cohort study. Cohort studies may be prospective (if exposed and unexposed subjects are enrolled before outcome is apparent) or retrospective (if subjects are assembled according to exposure after the outcome is known). Cohort studies are particularly well suited to evaluating a variety of outcomes from a single exposure (e.g., smoking). Population-based rates or proportions, or relative risk (see discussion in Section 10.1), may be computed.

CASE—CONTROL STUDIES
Case-control studies are also termed case-referent studies or, loosely and confusingly, sometimes retrospective studies. In case-control studies, the cases and noncases of the outcome in question are compared for their antecedent exposures. If exposure comparison is the purpose, the study by definition is usually a case-control study. The investigator generally has no influence over these antecedent exposures. The appropriate measure of risk in these studies is the exposure OR (see also Section 10.1 and below). The use of case-control designs is discussed further in the next section.
PREVALENCE OR CROSS-SECTIONAL STUDIES
Prevalence or cross-sectional studies are investigations of an entire population enrolled, regardless of exposure and outcome, with exposure and outcome ascertained at the same time. Effectively, these are “snapshots” of a population, where analysis may be performed as a cohort study or a case–control study.

Inferring causation
Identifying causal relationships in observational studies can be difficult if not impossible. If neither chance nor bias is determined to be a likely explanation of a study’s findings, a valid statistical association may be said to exist between an exposure and an outcome. Statistical association between two variables does not establish a cause-and-effect relationship, however. The next step of attempting to establish a cause follows a set of logical criteria by which associations could be judged for possible causality, which was first described by Sir Bradford Hill in 1965.

EVIDENCE SUPPORTING CAUSALITY
Seven criteria currently in widespread use facilitate logical analysis and interpretation of epidemiologic data:

1. Size of effect. The difference between outcomes given exposure and those not given exposure is termed effect. Conditional upon sample size, large effects are less likely to be the result of chance than small effects. Effect size is often summarized by the relative risk (RR), or the probability of having a disease when it is predicted in ratio to the probability of having the disease when the prediction is not having it; see Section 10.1 for further details. As a reference, an estimated RR > 2.0 in a well-designed study may be added to the accumulating evidence of causation.

2. Strength of association. Strength of association is based on the p-value, the probability of observing results as or more indicative of the alternative hypothesis if the null hypothesis were true. A weak association is more easily dismissed as resulting from random or systematic error. By convention, p-value < 0.05 is often accepted as an evidence of association.

3. Consistency of association. A particular effect should be reproducible in different settings and populations.

4. Specificity of association. Specificity indicates how exclusively a particular effect can be predicted by the occurrence of potential cause. Specificity is complete where one manifestation follows from only one cause.

5. Temporality. Putative cause must precede putative effect.

6. Biologic gradient. There should be evidence of a cause-to-outcome process, which frequently is expressed as a dose–response effect, the term being carried over from clinical usage.

7. Biologic plausibility. There should be a reasonable biologic model to explain the apparent association.
23.6 Retrospective Study Designs: The Case—Control Study Design

Epidemiologic studies are often focused on estimating the association between an outcome (e.g., disease) and some exposure of interest. Study designs for estimating such an association are a cross-sectional study design, where a random sample of individuals is gathered and disease status and exposure status are measured at the same time, or a prospective cohort study where a random sample of individuals without disease is sampled and exposure is measured, then the cohort is followed for the occurrence of disease. In the case of a rare outcome (e.g., childhood leukemia), these study designs would be nearly impossible to conduct since the required sample size would be prohibitively large to allow for any precision on the estimated probability of disease. For this reason, we might turn to a retrospective sampling design in which we first identify a sample of individuals known to have the outcome of interest (e.g., we begin by identifying a sample of children with leukemia). These are known as cases. We can then sample controls, i.e., individuals known to not have the disease of interest. From this sample, we can then measure exposure on cases and controls and compare the distribution between the groups. Specifically, if the distribution of exposure between cases and controls is the same, we would conclude that no association between exposure and disease exists.

Measure of association in the case—control study

In the case—control study design, the outcome is fixed by design. What is random is the exposure in cases and controls. If the exposure is binary (e.g., exposure to potential carcinogen or not), what we can estimate is the probability of being exposed given that one either has disease or does not:

\[ P(\text{Exposure} = 1|\text{Disease} = 1) \text{ and } P(\text{Exposure} = 1|\text{Disease} = 0). \]

A natural contrast would be to compare the probability of disease given exposure to the probability of disease given no exposure. The case—control design, however, does not allow us to do this. Fortunately, we can still estimate the OR for disease comparing
exposures just as we would have been able to if we had conducted a prospective or cross-sectional study. This can easily be seen by applying Bayes’ Theorem. Briefly, Bayes’ Theorem allows us to relate the conditional probabilities of random variables as follows:

\[ P(AB) = \frac{P(BA)P(A)}{P(B)} \]

Now, letting \( D \) denote disease and \( E \) denote exposure, the OR from a case–control study is given by

\[
\text{OR}_{\text{CC}} = \frac{P(E = 1|D = 1)/P(E = 0|D = 1)}{P(E = 1|D = 0)/P(E = 0|D = 0)}
\]

\[
= \frac{P(D = 1|E = 1)P(E = 1)/P(D = 1|E = 0)P(E = 0)}{P(D = 0|E = 1)P(E = 1)/P(D = 0|E = 0)P(E = 0)}
\]

\[
= \frac{P(D = 1|E = 1)}{P(D = 1|E = 0)}/P(D = 0|E = 1)
\]

The last line in the above is the OR for disease, comparing exposed to unexposed, that would be estimated from a prospective study design. Hence, the OR that would be computed from the retrospective case–control design is the same as the OR that would be computed from a prospective disease (where disease is random). The above result can be extended (in a far less trivial way) to continuous exposure measurements and to allow for adjustment for confounding factors via a logistic regression model (see Chapter 17: Logistic regression for binary outcomes).

**Number and selection of controls**

A natural question to ask when conducting a case–control design is the following: How many controls should I sample for each case? In the case of a rare outcome, controls are often far easier to obtain than cases. However, there may still be limitations in obtaining a large number of controls. Empirical studies have demonstrated that the precision for estimating associations in the case–control design tends to level off at about four controls per case. Sampling more than four controls per case tends to not provide much of an improvement in association estimates. The next question is as follows: How should we sample controls?

In a case–control design, we begin by obtaining a random sample of cases. We must then select controls. This can be done in an *unmatched* or a *matched* way. In an unmatched control design, the controls should be sampled from the same population as the cases and who were also at risk for the outcome of interest at the same time the
case became diseased. It is important that the controls are not sampled in such a way to influence the distribution exposure (apart from any relationship between diseases and exposure that might exist). In the case of unmatched controls, we can adjust for confounding via the use of logistic regression.

Many times, we would like to balance cases and controls on important confounding factors, by design. In this case, when selecting controls for each case, we would want to select controls that are from the same population as the cases and who were also at risk for the outcome of interest at the same time the case became diseased, but who also share the same confounding covariate value. For example, we may wish to match cases and controls on age. This forces the distribution of ages to be the same between cases and controls and removes any potential confounding that would be due to age. This is an effective way to adjust for confounding but does have drawbacks. First, if there are many matching criteria it may be difficult to find enough controls for each case. Also, matching on a variable means that we can no longer estimate the association between that variable and the outcome. In the above example we could not estimate the association between age and the outcome.

23.7 THE NESTED CASE—CONTROL STUDY DESIGN

In Chapter 19, Analysis of censored time-to-event data, we introduced survival analysis. The goal of survival analysis is to compare the time-to-event (e.g., death) between groups defined by their covariate values. A classic survival analysis is performed on a cohort of individuals that are all free of the event at the start of follow-up. It is further assumed that all covariate values have been measured on all subjects. If this is the case, the Cox proportional hazards model can be used to estimate the risk of the event that is associated with the covariates measured. In many cases, however, the covariate values for each subject are not all readily available. For example, we may have the time to death on subjects for whom stored serum has been collected, but not analyzed. In this case, it may be expensive to analyze the serum for all subjects. To save costs it would be advantageous to selectively analyze serum on only a subset of patients. This is the purpose of the nested case—control study design.

The nested case—control study design gets its name because it mimics a case—control study design in the survival analysis context. Specifically, the design is implemented by first sampling all patients that were observed to experience the event of interest. Then, at each event time, a random sample of controls (usually 1—4) are sampled from all patients that have not yet experienced the event and who have not been censored. Thus the nested case—control design is like a matched case—control design where matching is on the risk set (or pool of subjects still at risk to experience the event) for each case. A control at an early event can then also go on to become a case at a later event.
Analysis of a nested case-control study utilizes the Cox proportional hazards model as discussed in Chapter 19, Analysis of censored time-to-event data. The only difference between the nested case-control analysis and the usual Cox analysis is that we must stratify by each case’s matching set. This can easily be done with most modern statistical software packages. Nuño and Gillen provide further detail on the nested case-control design for interested readers.

### 23.8 THE CASE—COHORT STUDY DESIGN

One drawback of the case-control design and the nested case-control design is that sampling is performed for a specific outcome of interest. Hence, the same sample cannot be used to examine different outcomes (as this would require a new study that samples on that outcome). An alternative to this is called the case-cohort design. In this study design, we first randomly sample a subset of our total sample and measure the exposure of interest on all subjects in the subcohort. The subcohort size is often about 10%–20% of the total sample size, allowing substantial savings on measuring covariates for the whole sample. In order to increase efficiency, we then also obtain the covariate values for all other subjects that experience the event of interest who were not sampled into the original subcohort. This also means that if we wish to examine another type of outcome, we can simply supplement the original cohort with other individuals experiencing the new event.

Analysis of the case-cohort design also relies upon the Cox proportional hazards model. After the originally sampled subcohort is obtained and then augmented with all other patients that experience the event of interest, the Cox model is applied as usual (see Chapter 19: Analysis of censored time-to-event data); however, the use of a robust (or empirical) variance estimator is required. This accounts for the fact that the augmented sample is not truly a random sample. The option for a robust variance estimator is available in nearly all modern software packages. Again, Nuño and Gillen provide further detail on the case-cohort design for interested readers.

### 23.9 METHODS TO ANALYZE SURVIVAL AND CAUSAL FACTORS

#### Life tables list survival through time

Life tables, addressed in some detail in Section 19.3, give the proportion of a demographic group surviving to the end of each time interval. The life table is historic, having been used for survival analysis during the 1700s by Daniel Bernoulli in Switzerland. Survival can be used to see patterns of events other than life and death through time. For example, replacement of “time to death” by “time to onset of
illness” provides a window on morbidity rather than mortality. Replacement of “time to death” by “time to fail” provides a window on the reliability of medical instruments.

Life tables will be revisited in the context of epidemiology.

**EXAMPLE: LIFE TABLE FOR INFANT MALARIA**

Table 23.2 provides basic and calculated data for a life table on the malarial morbidity of 155 infants in the Cameroon (born of mothers without malarial infection of their placentas). Note that survival in this example is not thought of as remaining alive but as remaining disease free. $S$ for the first period ($> 0–13$ weeks) is $1.00 \times \left( \frac{141}{155} \right) = 0.9097$. In the second period ($>13–26$ weeks), 23 died and 3 were lost to follow-up, who were separated on two lines, with died first. $S$ for that period is $0.9097 \times \left( \frac{118}{141} \right) = 0.7613$. At the end of that period, we subtracted the 3 *lost* subjects, leaving 115. Because we assumed that they remained alive to the end of the period, they did not reduce the survival rate, but they are removed for calculating survival rate in the next period. For the third period ($>26–39$ weeks), *end* is divided by the value of *end* just above it, which has had the three *lost* subjects removed. $S$ for the third period is $0.7613 \times \left( \frac{98}{115} \right) = 0.6488$.

Calculations are continued in this fashion. At the end of the first year, 58% of the infants remain free of malaria. This also may be interpreted as an estimated 0.58 probability that an infant randomly chosen at the outset will remain disease free longer than 1 year.

<table>
<thead>
<tr>
<th>Interval (weeks)</th>
<th>Begin</th>
<th>Died</th>
<th>Lost</th>
<th>End</th>
<th>$S$ (survival rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (Outset)</td>
<td>155</td>
<td>0</td>
<td>0</td>
<td>155</td>
<td>1.0000</td>
</tr>
<tr>
<td>&gt;0–13</td>
<td>155</td>
<td>14</td>
<td>3</td>
<td>141</td>
<td>0.9097</td>
</tr>
<tr>
<td>&gt;13–26</td>
<td>141</td>
<td>23</td>
<td>0</td>
<td>118</td>
<td>0.7613</td>
</tr>
<tr>
<td>&gt;26–39</td>
<td>118</td>
<td>17</td>
<td>3</td>
<td>115</td>
<td>0.6488</td>
</tr>
<tr>
<td>&gt;39–52</td>
<td>98</td>
<td>11</td>
<td>4</td>
<td>87</td>
<td>0.5760</td>
</tr>
<tr>
<td>&gt;52–65</td>
<td>87</td>
<td>0</td>
<td>4</td>
<td>83</td>
<td>0.4788</td>
</tr>
<tr>
<td>&gt;65–78</td>
<td>83</td>
<td>14</td>
<td>0</td>
<td>69</td>
<td>0.4788</td>
</tr>
<tr>
<td>&gt;78–91</td>
<td>69</td>
<td>24</td>
<td>0</td>
<td>45</td>
<td>0.3123</td>
</tr>
<tr>
<td>&gt;91–104</td>
<td>45</td>
<td>0</td>
<td>2</td>
<td>43</td>
<td>0.1888</td>
</tr>
</tbody>
</table>

Table 23.2 Life table on malaria (*Plasmodium falciparum*) morbidity of 155 infants born in Cameroon 1993–95.
Graphing survival information

The graphical display of survival data, introduced in Section 19.3, is illustrated for the malarial morbidity example as Fig. 23.1. Note that the number lost to follow-up (censored) is shown as a small integer over the line for the period in which they were lost, distinguishing between those lost and those dying.

Confidence intervals for life table and Kaplan–Meier estimates of the survival function were also addressed in Section 19.3. The method for comparing two survival curves, the log-rank test, was addressed in Section 19.4. As an epidemiological application, consider examining the data of infant morbidity for cases in which the placentas had versus had not been infected with malaria. We ask if the two curves are the same in probability or different. If they are different, we would conclude that placental infection in the mother affects the morbidity of the child.

Serial correlation through time in epidemiology

Serial correlation, introduced in Chapter 20, Analysis of repeated continuous measurements over time, is another concept useful in epidemiology. Recall from Section 5.3 that a correlation coefficient is the adjusted covariance between two matching data sets. If the matching data sets are observations not taken at a point in time but taken sequentially through time, the correlation between them is termed serial correlation. A serial correlation in which the second of the matching sets is a repeat of the first, but lagged to start at a different point in time, is designated autocorrelation; if the second is a different variable, the serial correlation is a cross-correlation.

![Figure 23.1 A survival curve for the infant malaria morbidity data of Table 23.2.](image-url)
CROSS-CORRELATION

The correlation of infant malarial morbidity between malaria-free mothers and infected mothers through the 104 weeks is a cross-correlation. It tells us how closely related the two variables are through time. The correlation is based on the difference between them, not on their individual behavior through time. Thus, if they both increase and decrease together, the correlation is high even though the pattern may not be simple. Serial correlation can also be calculated with one of the sets lagged behind the other. For example, the appearance of symptoms of a disease having a 2-week incubation period can be correlated with exposure to the disease, where exposure observations are paired with the symptom observations that occurred 2 weeks later. By varying the lag, we may be able to find the incubation period. Alternatively, knowing the incubation period, finding lagged cross-correlations with several potential exposure candidates may help identify how the exposure occurred.

AUTOCORRELATION

Observations through time may be correlated with themselves. A set of observations through time is taken as the first set, and the same set is taken as the second, except lagged. If the autocorrelation coefficient retreats from 1.00 as the lag increases and then returns to nearly 1.00, we know that we have a periodically recurring disease. If that lag is 12 months, the disease is seasonal. A plot of autocorrelation coefficients on the vertical axis with different lags on the horizontal axis is termed a correlogram. Periodicities in a time series can be seen easily in a correlogram as time values at which the autocorrelation coefficient reapproaches 1.00.

Exercise 23.6

Data on the survival of women with untreated breast cancer give rise to a life table with format and essential data as in Table 23.3. Complete the table. What is the probability that a woman with untreated breast cancer will survive 5 years?

Exercise 23.7

Sketch a rough survival graph from the life table survival results of Exercise 23.6.

Exercise 23.8

Give one example each of an epidemiological phenomenon that could be detected using cross-correlation without lag, cross-correlation with lag, and (lagged) autocorrelation.

Exercise 23.9

Give another example each of an epidemiological phenomenon that could be detected using cross-correlation without lag, cross-correlation with lag, and (lagged) autocorrelation.
23.10 A HISTORICAL NOTE

Let us look at the steps Edward Jenner might have gone through in the 1790s in the discovery of a smallpox vaccine.

The first step would be to characterize a case definition: describing the disease with all the signs, symptoms, and other properties that characterize it in minute detail. We do not know which variable or combination of variables will relate to others. Smallpox is characterized by chills, high fever, backache, headache, and sometimes convulsions, vomiting, delirium, and rapid heart rate; after a few days, these symptoms retreat and papules erupt, becoming pustules that leave deep pockmarks. Complications include blindness, pneumonia, and kidney damage. The mortality rate approaches 250 per 1000 people. (We have no means to assess its viral origin; we do not even know viruses exist).

We investigate the incidence by geographic area. The disease never appears in the absence of an infected patient. Investigating in greater detail where the disease occurs, that is, at the household level, we find that no one contracts the disease without having been in proximity to an infected patient. It appears that the infection passes only from one person to the other. However, it seems to pass with any type of contact: airborne vaporized fluids, touch, or second-level touch (touch of things touched by an infected person).

We record the times an infected patient had contact with a previously infected patient. The period from contact to the origin of symptoms is identified as between 1 and 2 weeks, implying a 10 ± 3-day incubation period. Infectious contact may be from the very beginning to the very end of symptoms.

Only humans are affected. We gather demographic data, estimating the incidence of smallpox among all sorts of groups. We compare incidence by sex, age groupings, ethnic origin, socioeconomic level, and occupation. We gather biologic, physical, and genetic data and estimate incidences among groupings within these variables.
We gather personal data, estimating incidences by cleanliness habits, diet, food and drink types, food and water sources, and methods for disposal of bodily and food waste products. We find little difference among groupings. We are beginning to think that we must experiment with controlling food, water, waste, and so on, when a secondary occupational analysis turns up a surprising result: dairymaids do not contract smallpox!

We now further isolate the events and characteristics. We examine what is different about dairymaids that leads to immunity. We find that they differ little in any respect from the usual smallpox victim, except in socioeconomic level and disease history. Socioeconomic level has already been ruled out. In the disease history, we find that most dairymaids have undergone a course of cowpox, a mild form of pox that has some similarities. Exposure to cowpox conveys immunity! We are well on the way to a control for this heinous disease.

**Exercise 23.10**

Suppose you are working with James Lind in 1747, attempting to isolate the cause of scurvy. After a long period at sea, a number of crew members are suffering from scurvy, characterized by weakness, fatigue, listlessness, bleeding gums, loose teeth, and subcutaneous bleeding. You note that scurvy most often occurs aboard ships, never in a fertile countryside. Although the hold is stuffy, sailors get considerable fresh air working topside. Food and human waste are emptied over the side daily, not remaining to decay. Water and food are stored and are the same for all, except that the officers eat better. Bathing is available frequently, but in seawater, not fresh water. Scurvy never occurs at the beginning of a voyage, but always after a lengthy time without landfall. Once begun, scurvy only worsens aboard ship and improves only on return to shore. Anyone aboard ship is susceptible, but the incidence rate for officers is much lower. Crew members are in close contact over extended periods. Some contract it, but others do not. Therefore you conclude that it is not contagious. (You have no knowledge of other limits to contagion such as immunization.) How do you further your investigation?

**REFERENCES**

CHAPTER 23 Epidemiology

24.1 INTRODUCTION

The concept of meta-analysis

Meta-analysis is a pooling of data from various sources (journal articles, documented research sources, government statistics, etc.) to enlarge the sample size, thus reducing the sizes of the types I and II errors. It treats the different articles as replications of a single study, wherein lies its hazard. Medical studies are seldom if ever identical in experimental design and variable measurement. The crucial requirements to use meta-analysis are to decide which sources’ data are acceptable to pool and how to pool them.

Steps to conduct a meta-analysis

A meta-analysis should be developed as follows:

1. Define the inclusion/exclusion criteria for admitting articles. If one study admits all women older than 40 years and another admits women of any age but only with BMI > 30, the two studies cannot be pooled without introducing both an age bias and a weight bias. If the original data can be acquired (and very often it cannot), removing lighter women in the first and younger women in the second might allow pooling the remaining data, but the effect on other correlated criteria and upon the sample selection methods would have to be examined as well. This is a simple and obvious example. Too often the inclusion/exclusion criteria are not meticulously documented. As the number of candidate studies increases, the complication increases.

2. Search exhaustively and locate all articles addressing the issue. Searches must include more than checking published articles in commonly read medical journals. Relevant studies may exist in journals in other medical fields or even in nonmedical fields. Furthermore, many times study results yield only “negative results,” that is, they fail to find a putative effect and are therefore rejected by medical journals. These studies are notoriously underrepresented in meta-analyses but need to be included to avoid publication bias. The investigator should attempt to locate such studies by querying known workers in the specific subject matter and asking if they know of such studies.
3. Assess the articles on the basis of acceptable criteria. Some criteria are given in the next subsection. It may be useful to list the criteria in the left most column of a rectangular display, list the studies across the top, and fill in the cells with each criterion’s specification for each article. This is an easy way to assess the differences among the studies.

4. Quantify the admitted variables on common scales. Medical studies seldom document the scales in which their variables are addressed, deeming them to be obvious. However, some contain differences or discrepancies. An example is the number of decimal places reported; they may be quite accurate in one study but may not be carried farther than the least accurate recording in another. Another example is classifying a laboratory test result as normal or above normal, where the cut point between the states is different from one study to the next.

5. Aggregate or pool the admitted databases. The reader may see one of several aggregation methods referred to. The primary methods for pooling estimates across studies are discussed in more detail next.

Criteria for an acceptable meta-analysis

Because a meta-analysis depends a great deal on the investigator’s judgment, clear criteria are crucial. A minimum list of criteria follows, and the reader should verify, so far as it is possible, that each criterion has been met.

1. The study objectives were clearly identified.
2. Inclusion criteria of articles in general and specific data to be accepted were established before selection.
3. An active effort was made to find and include all relevant articles.
4. An assessment of publication bias was made.
5. Specific data used were identified.
6. Assessment of article comparability (e.g., controls and circumstances) was made.
7. The meta-analysis was reported in enough detail to allow replication.

Even after a careful meta-analysis, limitations may remain, for example, innate subjectivity, aggregation of data of uneven quality, and the forcing of results into a mold for which they were not intended.

24.2 PUBLICATION BIAS IN META-ANALYSES

Several biases may infiltrate an integrative literature review. A few are given here as follows:

1. **Data pruning.** Are original data given? Data that do not support the author’s agenda may be omitted, creating a data selection bias.
2. *Lack of scientific rigor.* Various biases can creep in when one article has been done less rigorously than another. Among other biases are varying enthusiasm in interpreting results, varying quality of randomization, and downplaying or omitting outliers.

3. *Inadequate reporting policy.* Authors tend to avoid submitting negative findings and journals tend to reject them when they are submitted, leading to overestimation of the success of an approach.

### 24.3 FIXED- AND RANDOM-EFFECTS ESTIMATES OF THE POOLED EFFECT

Once a number of studies have satisfied the inclusion criteria, and those that have not satisfied the criteria have been removed from consideration, results of the admitted studies must be combined. If the studies were total replications of one another, their data could be amalgamated into a single spreadsheet and the combined database analyzed in the form pursued in the studies. In fact, the admitted studies are seldom exact replications, and an appropriate aggregation method must be chosen. There are two standard approaches to conducting a meta-analysis, and they differ in how one interprets the studies being pooled together to estimate a common effect. The first method is known as a *fixed-effects* meta-analysis. For a fixed-effects meta-analysis, we begin with the assumption that all studies we wish to combine share one common true effect size that does not vary from one study to the next. The other approach to analyzing a meta-analysis data is termed a *random-effects* meta-analysis. For a random-effects meta-analysis, we assume that the study results that we are combining represent a random sample of studies that share a common mean effect size, but that each study has an individual true effect. The difference in the true effect between studies would then be due to, for example, differences in the patient populations from one study to the next, differences in how the intervention of interest was implemented, differences in the formulation of the treatment of interest, etc. While both methods ultimately seek to estimate one common effect by combining the estimates from individual studies via a weighted average, the two approaches differ in how the weights are assigned to each study. Namely, the fixed-effects analysis essentially assigns weights that are proportionate to the sample size in each study. The random-effects meta-analysis assigns weights that are increasingly shrunk toward one another so that the discrepancy between the weight given to the largest and smallest study is not as large as that in the fixed-effects meta-analysis.

**Fixed-effects estimate**

In the fixed-effects meta-analysis, we assume that all studies we wish to combine share one common true effect size that does not vary from one study to the next, and hence
any variation we see in the estimated effect sizes comes simply from measurement error. More formally, we assume that

\[ m_k = \mu + \epsilon_k, \]  

(24.1)

where \( m_k \) is the observed effect in study \( k \), \( \mu \) is the true common mean effect, and \( \epsilon_k \) is the within-study error observed for study \( k \). Our goal is to estimate \( \mu \).

In this case a natural way to pool the estimated effects to estimate a pooled common effect is to take an average of the estimated effects in each study, but to give more weight to those studies that have a low variance for the estimated effect. Note that this will naturally correspond to giving the highest weight to larger studies.

The fixed-effects estimate of \( \mu \) is given by

\[ m_{FE} = \frac{\sum_{k=1}^{K} w_k m_k}{\sum_{k=1}^{K} w_k}, \]  

(24.2)

where \( w_k = 1/v_k \), with \( v_k = \text{Var}[m_k] \). In addition, the variance of the \( m_{FE} \) is estimated as \( 1/\sum_{k=1}^{K} w_k \). In most cases, each of the estimates to be combined is approximately normally distributed, which implies that \( m_{FE} \) will also be approximately normally distributed. Thus a 100 \( (1 - \alpha) \% \) confidence interval for \( \mu \) is given by

\[ \left( m_{FE} - z_{1-\alpha/2} \sqrt{\frac{1}{\sum_{k=1}^{K} w_k}}, m_{FE} + z_{1-\alpha/2} \sqrt{\frac{1}{\sum_{k=1}^{K} w_k}} \right). \]  

(24.3)

**Random-effects estimate**

In the random-effects model, we assume that the study results that we are combining represent a random sample of studies that share a common mean effect size, but that each study has an individual true effect. This can be more formally written by letting

\[ m_k = \mu + \delta_k + \epsilon_k, \]  

(24.4)

where again \( m_k \) is the observed effect in study \( k \), \( \mu \) is the true common mean effect, \( \delta_k \) is the deviation of the true mean for study \( k \) from the common mean, and \( \epsilon_k \) is the within-study error observed for study \( k \). In the random-effects model, we typically assume that \( \delta_k \) and \( \epsilon_k \) are normally distributed with mean 0 and variance \( \tau^2 \) and \( \sigma^2 \), respectively. In addition, \( \delta_k \) and \( \epsilon_k \) are assumed to be independent of one another, and hence the variance of \( m_k \) is equal to \( \tau^2 + \sigma^2 \). Again, our goal in a meta-analysis is to estimate \( \mu \), the overall effect of the studies, and we will do so via a weighted average of the individual study means, \( m_k, k = 1, \ldots, K \).
In the random-effects model, we will again take the weight for each study to be the inverse of the variance for each mean, which will be $1/(\tau^2 + \sigma^2)$. Of course, we do not know $\tau^2$ and $\sigma^2$ and so we must estimate them. Noting that $\tau^2$ represents the variance between the sample means, an estimate of $\tau^2$ is given by

$$t^2 = \begin{cases} (Q - K + 1) \frac{\sum_{k=1}^{K} w_k}{(\sum_{k=1}^{K} w_k)^2 - (\sum_{k=1}^{K} w_k^2)}, & \text{if } Q > K, \\ 0, & \text{otherwise} \end{cases}, \quad (24.5)$$

where $w_k = 1/v_k$, with $v_k = \text{Var}[m_k]$ and

$$Q = \sum_{k=1}^{K} w_k (m_k - m_{FE})^2, \quad (24.6)$$

where $m_{FE}$ is the fixed-effects estimator defined previously.

Now, $\sigma^2$ represents the within-study variance for each study, so for study $k$ this would be estimated by $v_k$. Thus in the random-effects model the weight given to study $k$ is given by

$$w^*_k = \frac{1}{t^2 + v_k}, \quad (24.7)$$

and the random-effects estimate of $\mu$ is given by

$$m_{RE} = \frac{\sum_{k=1}^{K} w^*_k m_k}{\sum_{k=1}^{K} w^*_k}. \quad (24.8)$$

Similar to the fixed-effects estimator, the variance of the $m_{RE}$ is estimated as $1/\sum_{k=1}^{K} w^*_k$, and in most cases each of the estimates to be combined is approximately normally distributed which implies that $m_{RE}$ will also be approximately normally distributed. Thus a $100(1 - \alpha)\%$ confidence interval for $\mu$ is given by

$$\left( m_{RE} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{K} \frac{\sum_{k=1}^{K} w^*_k}{\sum_{k=1}^{K} w^*_k}}, m_{RE} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{K} \frac{\sum_{k=1}^{K} w^*_k}{\sum_{k=1}^{K} w^*_k}} \right). \quad (24.9)$$
24.4 TESTS FOR HETEROGENEITY OF ESTIMATED EFFECTS ACROSS STUDIES

We have discussed two common methods for pooling effects across studies to obtain an estimate of a common effect. A natural question to ask is which to use. Generally speaking, it is safest to use the random-effects estimate of the common effect because in most cases there will be some heterogeneity across the individual study effects due to differences in the study populations and designs being utilized. If, however, there is truly homogeneity across the studies, the fixed-effects estimator can provide a slightly more precise estimate of the aggregated effect.

In order to guide the decision on whether the fixed-effects estimator may be acceptable, some authors have advocated that one first test for heterogeneity of effects across the studies. This can be based upon the following statistic:

\[ Q = \sum_{k=1}^{K} w_k (m_k - m_{FE})^2. \]  

(24.10)

When each of the estimates to be combined is approximately normally distributed, \( Q \) follows a chi-squared distribution with \( K - 1 \) degrees of freedom under the null hypothesis that the fixed-effects model holds.

Some textbooks would advocate that if one conducts a test of heterogeneity and fails to reject the null hypothesis that the fixed-effects model holds, the fixed-effects estimator should be used. One should be cautious in taking this approach, however. Because the test begins by assuming that homogeneity holds across the studies, failure to reject the null hypothesis does not guarantee that the null hypothesis does indeed hold. Instead, failure to reject heterogeneity may simply stem from a lack of statistical power. Thus it is recommended that when deciding upon whether to use the fixed- or random-effects method, that the investigator carefully examines the differences in estimated effects across all available studies and also assesses how different the study populations and designs to be collected are. If differences are present, then the random-effects estimator is generally recommended.

24.5 REPORTING THE RESULTS OF A META-ANALYSIS

We previously discussed the importance of attempting to achieve homogeneity of the inclusion/exclusion criteria and general study design features among the studies to be utilized in a meta-analysis. This is obviously important because we seek to pool data across multiple studies in order to yield greater precision to estimate one common
effect. Of course, if the studies that we are pooling are inherently different in design, it is unlikely that one common effect can adequately describe the association of interest across the studies due to high variability across the studies.

Reporting a meta-analysis should include adequate information for a reader to judge whether the studies are roughly homogeneous in concept and design. Thus the general inclusion and exclusion criteria used to define which studies made it into a meta-analysis should be clearly stated. Just as important, the reason why any studies were not included in the meta-analysis should also be stated.

The results of a meta-analysis should also clearly reflect the precision afforded by each study being combined and differences in the estimated effect sizes across studies. To achieve this a forest plot is often a useful tool. Fig. 24.1 displays one example of a forest plot. Here we can see that, for each study included in the meta-analysis, the estimated effect size (odds ratio) and lower and upper limits of a 95% confidence interval for the true effect are given. This is further plotted on the right of the figure, where the box represents the estimated odds ratio and the horizontal line represents the 95% confidence interval. Finally, the estimated common effect size is also presented along with its uncertainty. Here the width of the diamond represents the length of the width of the 95% confidence interval for the combined estimate. The vertical line on the right of the figure illustrates a null effect (an odds ratio of 1 in this case), and the plot then depicts how consistent (or inconsistent) the effect is from one study to the next along with the common estimate of the effect by pooling information from all considered studies.

24.6 FURTHER REFERENCES

The historic book Statistical Methods for Meta-Analysis, by Hedges and Olkin,¹ provides a good reference for basics on meta-analysis. Section V of Medical Uses of Statistics, by Bailar and Hoaglin,² is a quite readable additional reference. Additional references are Borenstein et al.,³ Sutton et al.,⁴ and Cooper.⁵
REFERENCES

Bayesian statistics

25.1 WHAT IS BAYESIAN STATISTICS

Frequentist statistics

Historically, statisticians have considered sample frequency distributions and their theoretical counterparts, probability distributions, as containing the information required to make statistical inferences. (The dependence on frequency distributions has led to naming this approach to statistics as frequentist.) For example, a patient with elevated temperature could be suffering from either bacterial or viral influenza. From past data, we take the odds of bacterial to viral influenza as 1:4. An investigator suspects that the presence of elevated temperature ($temp$) can help distinguish between the two and obtains data from records of a sample of 50 patients having bacterial flu ($bac$) and another 50 having viral flu ($vir$). The rate of $temp$ occurrence for all patients is found to be 65%, for patients with $bac$ to be 90%, and for patients with $vir$ to be 40%. The frequencies of occurrences can be displayed as a contingency table, as in Table 25.1.

From these figures the following statistical results can be calculated. Fisher’s exact test (see Section 9.4) yields $p$-value < 0.001, showing the rates are highly significantly different. The odds ratio (see Section 10.1) is 13.5; the odds of $bac$ to $vir$ are more than 13-to-1 in the presence of an elevated temperature. The positive and negative predictive values (PPV and NPV) (also Section 10.1) are 0.69 and 0.86, respectively, strong predictions that an elevated temperature leads to a diagnosis of bacterial infection and its absence leads to one of viral infection. In particular, the PPV of 69% provides a prediction that, if we see a flu patient with an elevated temperature, this patient is more likely than not to have a bacterial infection. However, we started with equal numbers of bacterial and viral patients. Does this reflect reality? Hardly. We know that viral flu is much more common than bacterial flu. What really is the probability that an influenza patient with a temperature has a bacterial infection, or, in our symbolism, $P(bac \mid temp)$?

Bayesian statistics

If the occurrence rate of bacterial and viral infections was equal, $P(bac \mid temp)$ would be PPV. Usually, this is not the case. Suppose epidemiological records show viral flu to
CHAPTER 25 Bayesian statistics

Table 25.1 Frequencies of occurrence from the influenza study in which 50 patients each with bacterial and viral infection were sampled.

<table>
<thead>
<tr>
<th></th>
<th>Temp</th>
<th>No temp</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterial infection</td>
<td>45</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Viral infection</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Totals</td>
<td>65</td>
<td>35</td>
<td>100</td>
</tr>
</tbody>
</table>

The rates from the previous paragraph provide entries for the right side of Eq. (25.1). \( P(bac) = 0.20 \) (from 1 bac to 4 vir), \( P(temp) = 0.65 \), and \( P(temp|bac) = \) sensitivity = 0.90. By substitution, we find \( P(bac|temp) = 0.28 \). If we see elevated temperature in a flu patient, the chance is slightly over a quarter that the disease is bacterial. This is very different from the PPV of 0.69 arising from the equal occurrence of bac and vir in our sampling scheme of 50 patients of each type. If we had stopped with our PPV, we would have concluded that a bacterial infection is twice as likely as viral if we see an elevated temperature in the patient. By continuing the analysis using Bayes’ rule, we see that, in fact, the odds are still with a conclusion of viral.

Bayesian statistics, then, are methods that factor in the prior probability of occurrence of an event, in this case the prior probability (i.e., before seeing patients) that a randomly encountered flu patient will have a bacterial type of infection.

Revisiting the data and analytic approaches

Suppose we had thought about and obtained data giving the prior probabilities before conducting our study and had sampled 20% bac and 80% vir from the outset, obtaining the same outcome rates. Then we would have had outcome frequencies as shown in Table 25.2.

In this case, Fisher’s exact test still would have yielded \( p\)-value < 0.001 and the odds ratio would have remained 13.5. However, PPV would have been 0.36 and
NPV, 0.96. (The NPV would have implied that, given no fever, we would be quite sure the infection was viral.) Now using Bayes’ rule, we find $P(\text{bac}|\text{temp}) = 0.20 \times 0.90/0.50 = 0.36$, the same as the PPV. If the true rates had been used from the outset, the frequentist methods would have yielded the same results as the Bayesian methods.

### Some comments on frequentist versus Bayesian statistics

The frequentist says, I believe in my data and will consider inferential evidence based upon the theoretical variability of data I may have observed under specific hypotheses if my experiment were repeated many times. The Bayesian says, if you have information beyond your data, specifically a prior probability, it should be used and there is no need to condition upon unobserved hypothetical experiments. Both views can be argued convincingly. It was just shown previously that, if the prior probability arises from data, both schools of thought can incorporate it and arrive at the same conclusion.

One issue arises when the prior probability comes from personal belief rather than data, that is, is subjective rather than objective. A frequentist and two Bayesians may all believe, without data, that cause of illness A is much more frequent than cause of B. The frequentist cannot use the information for want of hard evidence. The first Bayesian believes A is twice as frequent as B. The second Bayesian believes A is five times as frequent as B. The Bayesians both incorporate their beliefs into their analysis and arrive at different results. Who is closer to the true answer? In the absence of data-based evidence, no one knows.

While there always is and should be controversy in scientific development, in the field of statistics it is usually well-meaning discussion about smaller issues, such as how close to a normal distribution a sample distribution must be to use a $t$ test, or if a mathematical relationship has been adequately proven. On the contrary, the frequentist versus Bayesian issue has arisen as a major rift and is sometimes even approached with rancor. Frequentist methods have been used successfully for generations and still represent the majority of applications. The use of Bayesian methods is growing and Bayesians claim that their methods are more complete and more accurate. Further, in

<table>
<thead>
<tr>
<th></th>
<th>Temp</th>
<th>No temp</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacterial infection</strong></td>
<td>18</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td><strong>Viral infection</strong></td>
<td>32</td>
<td>48</td>
<td>80</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 25.2 Frequencies of occurrence from the influenza study that would have occurred if the viral:bacterial sampling rate had been 4:1.
many settings of complicated statistical models, Bayesian statistics can offer more flexible model estimation procedures for obtaining estimates and predictions.

Most statisticians would insist on differentiating the two schools of thought using precise and detailed mathematical probability. However, this book is written not for statisticians but for health-care providers who want to conduct research. As such, this chapter approaches the distinction on a more conceptual level, introducing some Bayesian methods and showing how they may be applied. A reader who is mathematically capable and wants to pursue development and comparison more thoroughly than is found in the rather shallow glimpse of this chapter is referred to Frank Samaniego’s monograph, the chapter “Bayesian Methods in Public Health,” and the text by Christensen et al.

25.2 BAYESIAN CONCEPTS

Bayesian terminology

The components of Eq. (25.1) have been given the following names. $P(bac)$ is the prior probability. $P(temp \mid bac)$ is the likelihood of temp, given bac has occurred. $P(bac \mid temp)$ is the posterior probability. $P(temp) = 0.65$ is known as the marginal distribution, theoretically obtained using calculus by integrating $bac$ out of the product $P(temp \mid bac)P(bac)$.

Bayesian inference

In the classical statistical approach, we want to estimate an unknown parameter from a probability distribution, so we randomly sample data from that distribution and calculate the estimate from those data. Let us leave the individual patients of the influenza example and consider methods to deal with means. For example, suppose we want to estimate and draw inference on the mean glomerular filtration rate (GFR), say $\mu$, of a people from a culture with a different genetic, nutritional, and disease history pattern than those that have been established. We obtain a random sample of GFR levels, denoted $x_i$, from this population and use $\sum x_i / n = m$ to estimate $\mu$. We consider $\mu$ to be a constant and $x_i$ to have a sampling distribution, $f(m \mid \mu)$.

The Bayesian approach, in contrast, considers $\mu$ to be variable, with its own probability distribution, $P(\mu)$, denoted as the prior distribution. After specifying $P(\mu)$, the sample is obtained. The intent is to use Bayes’ rule to meld the sample result with the probability distribution of $\mu$ to form the posterior distribution, $P(\mu \mid m)$. To perform this, we still need to have the marginal distribution $P(m)$. $P(m)$ is obtained using calculus by integrating $f(m \mid \mu)$ over $\mu$ to remove it. Then we can use Bayes’ rule to obtain

(Continued)
One issue with Bayesian inference is the specification of the prior distribution. Since it is subjective, the user can choose the functional form, for example, if it is normal or some less frequently met form, as well as the parameter values, for example, rate or mean value. In practice, each commonly assumed likelihood distribution has a prior distribution that combines relatively simply with it, denoted a conjugate prior, so it is common to assume the conjugate prior for each likelihood form. Having made this assumption, the user has only to provide the prior's parameters from subjective belief.

On a conceptual level, for the GFR example, we might say that the frequentist estimates the value of the unknown mean GFR given the sample, while the Bayesian estimates the conditional probability distribution of the unknown mean GFR given the sample. Some medical investigators view this distinction as splitting theoretical hairs, but to statisticians, it is quite real.

What is the effect on the estimation and testing process? Specifically, how will the distinction affect our estimate of mean GFR? Both approaches provide a usable estimate and both types of estimates converge on the true \( \mu \) as the sample size grows large provided that the Bayesian prior probability places some mass on all values of \( \mu \). However, for smaller samples, the values of \( \mu \) differ.

For the level of this book, we will examine and compare methods only for a mean and a rate (specifically, a proportion).

### 25.3 Describing and Testing Means

**Example Posed: Baseline Glomerular Filtration Rate on a Humanitarian Mission**

Suppose we are on a humanitarian medical mission to a remote island that has experienced a disaster. Many residents suffer from dehydration and contaminated water. We want to estimate the mean of GFR for healthy middle-aged adults (40–50 years) and characterize the variability of our estimate to use as a baseline for comparison and learn if it is the same as that for developed nations—which we find recorded as mean 99 with standard deviation 14. We obtain (fictional) GFR readings on 8 healthy adults as 97, 84, 106, 69, 72, 108, 103, and 94.

What is the sample mean GFR using frequentist and Bayesian methods? How do these methods test the mean for difference from that listed for developing nations?
Mean estimation methods

Frequentist methods estimate the mean and variance as \( m = \frac{\sum x_i}{n} \) and \( s^2 = \frac{\sum x_i^2 - nm^2}{n-1} \) (Eqs. 5.1 and 5.3). Bayesian methods are a bit more involved.

Assume a sample of \( n \) observations is drawn from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), symbolized \( N(\mu, \sigma^2) \), and yields sample mean \( m \) (as in Eq. 5.1). We know from the central limit theorem (Section 4.9) that \( m \) follows the distribution \( N(\mu, \sigma^2/n) \). From this point in this section, we will be dealing only with means, so variances will refer to squared standard errors of the mean (SEMs). In Section 4.8, we referred to the SEM as \( \sigma_m \). For the Bayesian approach, we will assume that the prior distribution is also normal, in this case the likelihood’s conjugate, as \( N(\mu_0, \sigma_0^2) \); here \( \mu_0 \) represents a single value from the distribution and, as \( \sigma_0^2 \) is its variance, \( \sigma_0 \) is an SEM. The subscript 0 might be thought of as a time point prior to the sampling. Using Bayes’ rule, it can be shown that the distribution of the posterior distribution is

\[
N\left( \frac{m + \mu_0 \sigma_m^2/\sigma_0^2}{1 + \sigma_m^2/\sigma_0^2}, \frac{\sigma_m^2}{1 + \sigma_m^2/\sigma_0^2} \right). \tag{25.3}
\]

the first term in parentheses being the Bayesian mean estimate and the second, the Bayesian variance estimate.

EXAMPLE CONTINUED: GLOMERULAR FILTRATION RATE ESTIMATES ON A HUMANITARIAN MISSION

The frequentist estimates of the data are \( m = 91.6, s = 15.1, s_m^2 = 15.1^2/8 = 28.5 \), and standard error of the mean (SEM) = 5.3.

The first step in calculating the Bayesian mean, before obtaining our data, is to specify what we believe to be the parameters of the prior distribution. The investigator believes that the island residents will not be different from the historic record of GFR average and chooses \( \mu_0 = 99 \). He believes the standard deviation for that mean will be 4, so \( \sigma_0^2 = 16 \). Now he takes his sample, finding \( m = 91.6 \) and \( s_m^2 = 28.5 \). (Because these methods use the normal distribution rather than \( t \), we take \( s_m^2 \) as \( \sigma_m^2 \)). By substitution in the parameter entries of Eq. (25.3), we find that the distribution of \( \mu \) as a random variable is normal with mean 96.3 and standard deviation (of the mean) 3.2.

A frequentist would find average GFR 91.6, over 7 units below the historic average. This Bayesian finds that his belief in the islanders’ GFR being the same as the historic value leads to an island average only 2.7 units lower.

Another Bayesian that uses a different prior would achieve a different result. Suppose a colleague of the investigator believes the traditional shortage of water
during dry seasons has stressed the kidneys of islanders and their average GFR is more like 80. She assumes the same variability. Her Bayesian average is brought down by her subjective prior belief in a lower GFR to 84.2.

**Shrinkage**

It hardly surprises us that adding information (about the prior distribution) to data alters the parameter estimates. In most cases, the variability is reduced and the estimated mean moves toward the mean of the new information, that is, the prior’s mean, an effect termed *shrinkage*. Rarely, the information about the prior is so poor that its introduction has the opposite effect.

**Mean testing methods**

The frequentist approach is to test $H_0: \mu = 0$. If $H_0$ is rejected, $\mu \neq 0$ is concluded. The probability of a type I error is bounded by $\alpha$, but not the total probability of error. The other component of error, the estimate of $\beta$, depends on the true value of $\mu$ other than 0, and that is not known.

In contrast to the somewhat convoluted Bayesian estimation method, the Bayesian inferential method is simple and intuitive. The approach is to evaluate the probability that $\mu \neq 0$ as the area under the curve of the posterior probability, such as in Eq. (25.3) within the rejection areas.

**EXAMPLE CONCLUDED: GLOMERULAR FILTRATION RATE TESTING ON A HUMANITARIAN MISSION**

The frequentist does not know (may believe but has no evidence) whether the islander mean GFR is greater or less than the historic mean GFR, so $H_0$ and $H_1$ are $\mu = 99$ and $\mu \neq 99$, respectively. Using $s_m$ as $\sigma_m$, $z = (m - \mu)/\sigma_m = (91.6 - 99)/5.3 = -1.4$, yielding $p$-value $= 0.206$. The estimated probability of observing a sample mean as far, or farther, away from 99 if the true mean were actually 99 is approximately 20%, so he does not conclude a difference.

The first Bayesian uses his prior mean of 96.3 to find $z = (96.3 - 99)/3.2 = -0.844$, which gives the probability that the island GFR mean is less than the historic GFR mean as 0.199. He concludes that no difference has been shown.

His Bayesian colleague uses her prior mean of 84.2 to find $z = (84.2 - 99)/3.2 = -4.625$, which gives the probability that the island GFR mean is less than the historic GFR mean as $<0.001$. She concludes that a strong difference has been shown.

We note that, in commonly used methods, the frequentist cannot speak of the probability of no difference, just that no difference was found. Some would interpret this as an inconclusive test. If he had used an equivalence test as in Sections 12.2–12.5, he could have given a probability of no difference but then could not have spoken of the probability of a difference. By using a joint difference-and-
equivalence test as in 12.6, he can consider both, but this procedure is not yet widely used by frequentists. The Bayesians can speak of the probabilities of no difference and of a difference, but such probabilities depend on the subjective prior and we have seen that the Bayesians find conflicting outcomes from the same data. Neither school of thought is without limitations. In practice, since each individual is likely to have their own prior, it is generally advised to consider a sensitivity analysis in which the results of a Bayesian analysis are presented under a wide range of prior specifications.

**ADDITIONAL EXAMPLE: AN EMERGENCY DEPARTMENT TREATMENT FOR DYSPHAGIA**

Let us revisit the example in Section 11.2 using a Bayesian test. A “GI cocktail” of antacid plus viscous lidocaine was used to treat dyspepsia and the pre-to-post-treatment difference in pain on a 1-to-10 scale was recorded for \( n = 15 \) patients. Sample parameters were found to be \( m = 3.87 \) and \( s = 1.73 \). \( s_m^2 \) rounds to 0.2. For the Bayesian normal application, we take \( s_m^2 \) as \( \sigma_m^2 \). We further assume the prior distribution to be normal and ask the investigator for subjective estimates of the parameters, that is, what does she believe the mean and SEM to be before seeing any data. She has found the treatment to be extraordinarily successful in her recent patients and believes the average benefit (reduction in pain units) to be 5. She has no feel for the SEM and accepts the sample value of 0.2. This is done only for illustrative purposes and equates to given equal weight to the prior and the data. In practice, the variance associated with the prior is usually taken to be much larger than that observed in the data. She substitutes these values in Eq. (25.3). When \( \sigma_m^2/\sigma_0^2 = 1 \), Eq. (25.3) simplifies to \( N[(m + \mu_0)/2, \sigma_m^2/2] = N[4.44, 0.1] \). In this case, \( z = 4.44/0.1 = 44.4 \). The probability that the true mean \( \mu \leq 0 \) is the area under the normal curve to the left of 0, or \( < 0.001 \). She concludes that the treatment is effective, as she believed before the analysis.

Her colleague, on the other hand, is more pessimistic. He has not had a similar success with his patients and believes the treatment to be ineffective, that is, to have a mean reduction in pain of 0 units (\( \mu_0 = 0 \)). He also believes the participants in the study to have been influenced by the attention from the study itself and believes the variability to be greater than observed. He is willing to take the standard deviation as 2.5, yielding an SEM of 1.614 (\( \sigma_0^2 = 0.417 \)). He substitutes in Eq. (25.3) and obtains \( N[(3.87 + 0)/(1 + 0.2/0.417), 0.2/(1 + 0.2/0.417)] = N[2.616, 0.135] \). For him, \( z = 2.616/0.135 = 19.378 \). The probability that \( \mu \leq 0 \) is still < 0.001. The conclusions from frequentist methods, from an optimistic Bayesian, and from a pessimistic Bayesian, are all the same: the treatment is effective. We note, however, that both Bayesian values of \( z \) are larger than the frequentist \( z \), showing that the Bayesian testing, by lowering the variance through the introduction of the prior, resulted in higher estimated precision.
Exercise 25.1

Is early prostate-specific antigen (PSA) sampling the same as later sampling? In Section 11.2, we used frequentist methods to ask if the mean for the first 10 PSA values sampled (DB1) \( (m = 6.75) \) is different from that for the large number of patients sampled later \( (\mu = 8.85, \sigma_m = 1.01) \). The normal test produced \( z = (m - \mu) / \sigma_m = -2.08 \), yielding \( p = 0.037 \). There was sufficient evidence to conclude an answer of yes; the mean of the first 10 is less than that of the remaining 291. One Bayesian conjugate prior distribution might be a normal with parameters as from the entire 301 PSA values: \( \mu_0 = 8.79 \) and \( \sigma_0^2 = 0.98 \). Use Bayesian methods to estimate the parameters of the posterior distribution and find the Bayesian probability that the mean of the first 10 readings is less than that of the remainder.

25.4 ON PARAMETERS OTHER THAN MEANS

For means (as in the preceding section), distributed exactly (or even approximately) normal, the conjugate is normal. The reader will be well acquainted with the normal distribution and can calculate the Bayesian methods. For other parameters, the conjugate distributions have not been introduced in this book. They are familiar to workers in mathematical statistics but are seldom used in medical applications. For example, to perform a paired \( t \) test by Bayesian methods, the conjugate is an inverse gamma distribution; for a binomial test, a beta distribution; for a Poisson test, a gamma distribution; and for a multinomial test, a Dirichlet distribution. Software not requiring facility with these conjugates exists to perform many Bayesian analyses, but a more thorough understanding of the constraints and assumptions is required for practical use. Therefore treatment of remaining methods will be constrained to one case, that of estimating and testing a rate (proportion) and will be given only as an illustration. The reader will not be asked to participate actively in the calculations.

25.5 DESCRIBING AND TESTING A RATE (PROPORTION)

As said in Section 25.4, this section composes only an illustration; the reader is not expected to follow all aspects of the method.

EXAMPLE POSED: PREVENTING NAUSEA AFTER GALL BLADDER SURGERY

DB2 presents data on reducing postsurgical nausea by administering ondansetron hydrochloride prior to gall bladder removal. Let us use the patients' reporting of nausea or not, the third table under the heading "Data" in the DB2 entry. Of \( n = 40 \) patients given the drug, \( k = 9 \) experienced nausea. What is the rate of postdrug nausea?
Estimating the rate

As in Chapter 9, Tests on categorical data, let us denote the true but unknown rate by \( \pi \) and the estimated rate by \( p \). A frequentist would estimate \( \pi \) with \( p = k/n \).

A Bayesian would express a subjective prior probability and use it to influence the estimate of \( \pi \). In DB2 the 16 of 41, or 39%, of placebo patients became nauseous. In advance of the drug use data, the Bayesian believes the drug will diminish this rate to 25%. Remember that the Bayesian considers \( \pi \) not a fixed yet unknown value we want to estimate, but a random variable whose posterior probability distribution we want to identify. As we know from Section 4.8, the sum of successes is distributed as a binomial. The conjugate prior distribution of a binomial is what is known as a beta distribution. It is a function of the variable \( \pi \), having form

\[
  f(\pi) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha, \beta)},
\]

where \( \alpha \) and \( \beta \) are its parameters, \( B(\alpha, \beta) = (\alpha-1)!/(\beta-1)!/(\alpha+\beta-1)! \), and factorials [e.g., \( (\alpha-1)! \)] were defined in Section 3.1. The Bayesian accepts the conjugate beta distribution assumption and has only to subjectively choose the parameters \( \alpha \) and \( \beta \). He takes the peak, or mode, of the beta distribution as his subjective 25% postdrug nausea rate. He knows that the mode is \( (\alpha-1)/(\alpha+\beta-2) \), which reaches 0.25 when \( \alpha = 2 \) and \( \beta = 4 \). Furthermore, he knows from the Bayesian derivation that the posterior distribution is again a beta distribution, beta \( (k+\alpha, n-k+\beta) \), that estimates \( \pi \) by

\[
  p_B = \frac{k+\alpha}{n+\alpha+\beta}.
\]

**EXAMPLE CONTINUED: REDUCING POST—GALL BLADDER SURGERY NAUSEA**

The frequentist’s estimate would be \( p = k/n = 9/40 = 0.225 \); 22.5% of the treated patients became nauseous.

The Bayesian substitutes \( n = 40, k = 9, \alpha = 2, \) and \( \beta = 4 \) into Eq. (25.5) to obtain \( p_B = 0.239 \). The Bayesian estimate of the nausea rate among treated patients is 24%. Due to the “shrinkage,” the Bayesian estimate of the rate has moved from the frequentist estimate of 22.5% toward the Bayesian’s prior estimate of 25%. If the Bayesian’s impression of a larger nausea rate is correct, the Bayesian estimate will be closer to “truth.” If, on the other hand, the data are representative of the population, then the movement of the rate toward the subjective prior introduces a bias and represents a rate estimate that is farther from truth.
25.6 CONCLUSION

The reader who wants to pursue Bayesian statistics should seek reference material at the appropriate level. Many good books on a variety of levels of mathematical sophistication are available.

In the authors’ view, each school of thought has much to offer and each suffers from some deficiencies.

The authors subscribe to the following paradigm. If data-based evidence is available to estimate the parameters of a prior probability distribution, Bayesian methods should be considered. If the prior would have to be based on a subjective impression, not data-based evidence, and one that would differ more than minimally with each user, frequentist methods should be considered. In either case, one can (and usually should) consider the statistical properties of any procedure under the two paradigms. For example, if a Bayesian analysis is proposed, frequentist properties such as type I and type II errors can be assessed. Similarly, if a frequentist analysis is proposed and members of the scientific community have strong feelings about the prior information available on the problem then posterior probabilities utilizing a range of prior distributions can also be presented.

The reader will have to weigh the pros and cons and make a personal choice.

REFERENCES

1. Missing Sources. The Sources of a Few Examples Could Not Be Found Despite a Strong Effort to Locate Them. Such Data That Could Not Be Referenced Were Slightly Altered So as Not to Reflect on Any Investigator Later Appearing.
26.1 INTRODUCTION

In the medical statistics connotation, a survey is a systematic process of acquiring statistical information about a collection of subjects. It could refer to a chart review but is usually thought of as obtaining information directly from the subjects. Questionnaires—construed here to include interviews—are tools for obtaining such information, consisting of series of questions to be asked of the subjects. (Polls are a form of survey not typically used in medicine, so will not be pursued here.) This chapter, in the spirit of the rest of this book, is not intended to be a comprehensive guide to survey development but rather to provide the most important concepts and admonitions in survey and questionnaire design and start the user on the path.

26.2 SURVEYS

Design of survey

Design includes two parts: the design of the survey (the process of acquiring information), addressed in this section, and the design of the questionnaire (the tool of acquisition), addressed in Section 26.3.

As with any medical study, start with objectives. Clarify precisely what is wanted so that the sampling scheme and questionnaire can be designed to answer these objectives.

The next step is to identify the population to which you want to generalize and then plan sampling to obtain a representative sample from this population.

Specify in detail the logistics of the sampling plan, including how to reach the target sample. This will include, inter alia, funding, personnel acquisition and management, and the timing of the operation.

Design the questionnaire. Test the questionnaire with a “dry run” and revise it. Avoid whatever does not contribute directly to the objectives or to checking for possible bias.
Methods of collection

There are four major modes of data collection.

*Direct personal contact*, that is, “face-to-face” contact. This mode can accommodate either interview or written questionnaire and the questions can be either open or closed ended. Direct contact provides more complete responses and fewer erroneous responses in that the interviewer can insist on answers, clarify unclear answers, and explain meaning to confused respondents. Direct contact provides a higher response rate by denying respondents the ability to procrastinate or to ignore the request. Accuracy can be verified more easily because the interviewer has made the recording. However, direct contact is costly and time consuming, requiring trained interviewers.

*Direct contact over the telephone or Internet.* Telephone contact provides responses in a shorter time and with much reduced cost over face-to-face contact. Most of the advantages and disadvantages are retained, although somewhat muted: the interviewer’s perception of the respondent’s uncertainty or confusion is not as clear, many respondents will not feel as open as with face-to-face contact, and respondents find it easier to refuse response.

*Postal contact.* Mailing written questionnaires requires low expense and a low level of effort, but greatly decreased control over the response and answers. The respondent may trash the questionnaire or set it aside until later and never return to it; may ignore some questions; may misunderstand (or pretend to as a result of hostility) and provide erroneous answers; and may find that a desired answer is not an option, writing it in and causing deletion or questionable interpretation of the question. The response rate of postal contact is rather low. Various surveys of surveys have reported median response rates in the vicinity of 35%—40%. The greatest advantage of postal surveys is anonymity, found in no other form of contact, but that is not always a good characteristic. On the one hand, in most cases respondents can feel free to be honest by virtue of their anonymity. On the other hand, sometimes the relief from responsibility can lead to indifference or capriciousness.

*Internet contact.* Sending queries over the Internet is the least costly and fastest mode but exhibits the lowest response rate of all, with a median rate in the vicinity of 30%—35%. Of course, using this mode requires obtaining electronic addresses for all respondents and denies access to potential respondents who do not use the Internet. The greatest advantage of Internet contact is that it can accommodate branching questions, subordinate explanations, and explanatory graphics, to be accessed as the respondent may require.

Data and analysis

As with almost every medical study, after the data have been obtained they are transferred to a spreadsheet and analyzed statistically. Conclusions are made. Confidence
in the conclusions is assessed. The adequacy of the sample and accuracy of results are assessed. The results are generalized to the target population. Guidance on these steps may be found throughout this book.

**Bias in surveys**

Bias may occur in the design of the survey or of the questionnaire. It is not possible to be assured that a survey is free from any sort of bias, but the most severe and frequent forms of bias may be investigated.

The most frequent form of survey design bias is sampling bias: a difference exists between characteristics of the sample and of the target population. The bias may exist in the demographic character or in the nature of the subject being questioned, such as knowledge, belief, or attitude. If the demographic character of the population is known, demographic questions may be asked in the survey and the results compared with the character of the population. One must be careful not to sacrifice anonymity if that is intended in the design. There is no one satisfactory method to address bias in the nature of the subject matter being questioned. One approach is to randomly assign the respondents into two groups and compare the groups for consistency. This approach examines the consistency of response within the sample but does not relate the nature of the sample response to that of the population. Another approach is to investigate reasons for nonresponse. This involves following up and querying nonrespondents, a difficult and costly procedure. However, finding the reasons for nonresponse often uncovers bias.

Questionnaire bias is addressed in Section 26.3. Various causes of and approaches to identify bias may be found throughout this book.

**Survey response rate**

Response rate is calculated as the number of (legitimate or usable) responses received divided by the number of potential respondents contacted. It is usually expressed as a percentage.

Response rates are notoriously low. It has been noted that indirect contact rates typically run in the range of 30%—40%; direct contact rates are somewhat better. An investigator is usually satisfied to receive a response rate of 60%—70%. A low response rate can be increased by following up with additional contacts: a second visit, telephone call, letter, or e-mail. Indirect contact response rates can be markedly helped by employing direct contact for follow-up, although this mechanism removes anonymity. The difficulty of increasing response rate is reminiscent of the historically cited “20/80 rule”: the last 20% of responses will take 80% of the total time and effort.
However, the issue is not the rate at which the responses occur but rather the representativeness of the responses that do occur. A representative response with a 30% rate may be more accurate and informative than a biased response with a 70% rate.

A way to improve sampling accuracy of a written questionnaire is to revisit a random subsample of responses using direct contact (telephone or face-to-face) to relate the answers given to the respondent’s intention. Of course, such a procedure will uncover only sampling bias, not design bias. And, naturally, a direct contact follow-up consumes extensive time and effort and cannot be done in the presence of anonymity.

A comment on terminology

Most surveys present results in percentages. We often hear or read reports that quote results in “percentage points,” by which is meant percent, or, worse, just “points.” A point could be calculated in any number of ways and its meaning is not clear. A percent is precise and unequivocally understood. Using the terms points and percentage points to imply percent suggests a misunderstanding of the investigative process and should be scrupulously avoided.

26.3 QUESTIONNAIRES

The term “questionnaire” includes interviews (direct contact questionnaires) and written questionnaires. Each type may include the same questions in the same form, but each type has advantages or disadvantages relative to the other type.

Comparison of interviews versus written questionnaires

Advantages of interviews. Interviews offer the ability to clarify or reword questions to respondents who are uncertain as to what is wanted or make a first response that is facetious or hostile. Interviews offer a richness of response by allowing respondents to form their own answers or expand on answers. If an interviewer finds that a question does not provide the full or adequate information required, the interviewer can continue questioning until the issue is exhausted. The interviewer can often extract answers from a respondent who is hesitant to answer and would give no answer on a written form. The interviewer exercises some control over the order and emphasis of the questions. If the respondent’s thought process is seen to follow a certain track, the interviewer can reorder a sequence of questions to take advantage of that thought process. If the respondent is seen to inflate or deflate the importance of questions, the interviewer can adjust the respondent’s perception.

Advantages of written questionnaires. Written questionnaires are more economic of time and funding. Written questionnaires can be self-administered by the respondent.
At times (as in a meeting, classroom, or training session), written questionnaires can be administered to multiple respondents simultaneously. Written questionnaires may be anonymous, that is, submitted without the ability to identify the respondent, often permitting more sensitive questions to be answered than would be answered by identified respondents. Written questionnaires may be standardized, allowing for comparison between demographic groups or treatment groups and allowing for generalization of specific questions not possible if the question has been posed differently or answered differently by various respondents. Perhaps most important, written questionnaires, having a limited number of identical responses, may be quantified and thereby subjected to statistical analysis, leading to sounder conclusions.

Characteristics of written questionnaires and their questions

The length of the questionnaire and the difficulty of the questionnaire are both inversely proportional to both accuracy and response rate. The longer the questionnaire is, the poorer the accuracy will be and the poorer the response will be. The more difficult the questionnaire is for the respondent, the poorer the accuracy will be and the poorer the response will be. In this sense a questionnaire is a trade-off between the completeness with which an issue is explored and the accuracy of the information obtained.

Open-ended versus closed-ended questions

Open-ended questions are questions that allow the respondent to choose how to express the answer. Examples might be as follows:

- How much pain do you feel? __________________________
- The pain I feel is __________________________.

Closed-ended questions provide specifically worded possible answers and require the respondent to select one. Examples might be as follows:

- Circle your answer: I am feeling pain. YES NO
- Circle your level of pain from 0 to 10,
  where 0 is no pain and 10 is the worst pain you could imagine:
  0 1 2 3 4 5 6 7 8 9 10

Open-ended questions share most of the advantages and disadvantages of interviews. They can acquire greater subtlety, but at the cost of being different one from another, rendering them difficult if even possible to evaluate. Most importantly, they cannot be quantified and therefore cannot be objectively compared with other answers or from one respondent to another.
Closed-ended questions can be quantified and subjected to statistical analysis but are sensitive only to the subtleties the developer can anticipate. Respondents are required to select answers that may not agree with their true intentions. Closed-ended questions lead to more consistent conclusions because they are less subject to the interpretation of the investigator. Questionnaires with closed-ended questions are faster to administer, requiring less time and effort from the respondent, and are faster and easier to record and analyze. An extremely important characteristic of surveys with closed-ended questions is that they may be replicated to compare with other studies or to investigate issues raised by their first use.

Format of questions

In order to be effective, written questionnaires have aesthetic requirements not found in other forms of data acquisition and management. Many respondents are not engaged out of desire to participate and therefore easily become uncooperative. It is essential that the question-and-answer sheet used by the respondent be easy to read with pleasant spacing.

The investigator should position answer spaces in anticipation of the recording process. If the blanks can be positioned at the right edge of the page (without making it difficult for the respondent to record answers), then numerous pages can be lined up overlapping and the scores entered on the data spreadsheet more efficiently. One benefit is that scores for a single variable appearing as a row across subjects can be entered as a column on the spreadsheet using the 10-key pad on a computer keyboard by touch.

Topics are better amalgamated together rather than presented helter-skelter in order to take advantage of the respondent’s focus on a topic. Frequent “mental rebooting” is often annoying to a respondent.

The investigator can take advantage of branching to reduce the respondent’s time and effort. For example, if questions number 6–9 related to women only, then a statement may be entered prior to question 6 stating “If the respondent is male, skip to question #10.”

Types of answer

Questions can ask for answers in several forms. The most common are dichotomous, multiple choice, ranked, continuous, and rated. Dichotomous answers include Yes/No, Some/None, Satisfied/Dissatisfied, Male/Female, 0/1, 1/2, and the like. Multiple choice answers include a set of choices that is not necessarily orderable, such as ethnic origin. Ranked answers require the respondent to place a set of possible states in rank order. Continuous data answers ask the respondent to provide a number (e.g., age) or make a mark on a scale (e.g., visual analog scale). Rated answers ask the respondent to choose
a position in a succession of states. The design of the first four types is rather obvious; the design of ratings will be considered later.

**Instructions to respondent**

Clear and simple instructions are crucial. A respondent may be intellectually limited, distracted, indifferent, bored, or even hostile. Instructions should be so simple and so clear that they can be misunderstood only intentionally (which sometimes occurs). Plan the layout, topic and question sequence, wording, and even choice of questions in a way to facilitate simple and unmistakable instructions.

**Wording**

Clarity of expression is essential. A mechanism to improve wording is to assume that a respondent will misunderstand and choose your words accordingly. Avoid jargon, abbreviations, and unfamiliar acronyms. If any of these are necessary, be sure they are defined clearly and prominently.

Avoid ambiguity. Words such as “frequently” or “regularly” introduce ambiguity; be specific if a rate is wanted. A question such as “How many packs of cigarettes do you smoke?” leaves the respondent asking, “Per week? Per month? What?”.

Simplicity is essential. Remember that you have been thinking about this issue for a long time and are relatively well informed about it. In most cases the respondent is new to the ideas involved and relatively ignorant. What seems clear or even obvious to you may be obscure to the respondent. You are in danger of falling into “the technical manual syndrome” of assuming that items obvious to you are obvious to the respondent. (The author has been known to say, “If a technical person knows enough to write the manual, he knows too much to write the manual.”)

Avoid “double-barreled” questions. Examples might be, “How many cups of coffee or tea do you drink per day?” The respondent asks if that means the preferred or usual drink or a combination. “Do you carry health insurance and life insurance?” The respondent may want to give the two answers yes and no. If he carries one and not the other, you are receiving false information whether the single answer is yes or no. Also, identify and explain any possible misinterpretations.

Set the time frame for the question if the answer may change by past versus present or by season. For example, a respondent might want to answer the question, “Are you a smoker?” As, “I was, but I quit last month; I hope I can stay off it.” Should the answer be recorded yes or no? Better: Question 1, “Are you currently a smoker?” And Question 2, “Have you ever been a smoker?” As another example, “How many alcoholic drinks do you take per week?” might lead to an answer like, “This month, 15, because it is holiday time; the rest of the year, about 3.” Better: “What is your average weekly intake of alcoholic drinks over a year?”
CHAPTER 26 Questionnaires and surveys

Avoid questions the respondent is unlikely to be able to answer. A facetious example might be, “How many bowel movements did you have last year?” Avoid asking what the respondent intends to do in the future, for example, “Will you quit smoking in the next month?”

Coding
In preparing scoring for a questionnaire, anticipate how the scores will be recorded on a spreadsheet. In the spreadsheet, each column of data should be composed of only numbers. For example, sex might be coded 1 for male and 2 for female, not M or F. Male or Female on the questionnaire can easily be recorded in numbers in the spreadsheet.

Each column of data on the spreadsheet should have a single dimension, that is, contain no more than a single indicator. For example, instead of Cancer Stage = 0 for no cancer and 1, 2, 3, or 4 for stage, there should be two columns Cancer Present = 0 for no and 1 for yes plus Cancer Stage 1, 2, 3, or 4 wherever Cancer Present = 1 and blank where Cancer Present = 0. If data are entered in the questionnaire as more than one dimension, the person transferring data to the spreadsheet will have to make transformations while recording or will have to use an intermediate transform sheet.

Record original data whenever possible. For example, record age in years rather than decades (20–29, 30–39, . . .). We can always group original data later, but we cannot recover the original data if only groupings are recorded.

When data might be analyzed as ranks as regression predictors, the order listed will be germane. For example, answering a smoking question in the order 1: current smoker, 2: never smoked, 3: past smoker implies that the severity of habit is 1, 2, 3, when in fact it is 1, 3, 2. That is, a past smoker has a greater severity of habit than a subject who never smoked. The order presented should be 1: current smoker, 2: past smoker, 3: never smoked.

Rating scales
One popular source of respondent information is the rating scale. The respondent is asked to rate some perception, such as pain or satisfaction with medical care. There are two forms of scale: summative and cumulative scales.

Likert scales. Summative scales are usually called Likert scales after their 1932 founder, Rensis Likert. Such scales ask the respondent to select a position on an ordered set of positions, such as 1–5, where 1 may be the least and 5 the most of some value range, or the equivalent 0–4. A radiologist or pathologist may be asked to rate the severity of a pathological finding on a four-point scale as absent (0), mild (1), moderate (2), or severe (3).
Often several such scales are used and an overall score calculated as the sum of the individual scores. One must be careful to differentiate between scores on an individual item or element and a summative scale.

Investigators often ask how many choices should be used in a scale. The larger is the number of choices, the more is the information, but the more difficult it is for the respondent to differentiate. In a scale with too many choices, respondents confuse adjacent choices and accuracy is lost. Some psychologists claim that 7 is the maximum number of choices that can be distinguished clearly. In most cases, 5 is enough to provide the required information.

Investigators also often ask if the number of choices should be odd or even. An odd number allows for a central rating to be neutral. If a neutral position is wanted, use an odd number. If the respondent is to be guided into a forced choice on one side or the other of neutrality, use an even number.

Whether the investigator wants to list interpretive wording along with or in place of the rating numbers depends on whether the respondent is to be led into the interpretation or is to select his own interpretation. For example, a question on a patient’s satisfaction with level of care could be enumerated 1–5 versus could use interpretive wording Very Dissatisfied, Somewhat Dissatisfied, etc.

The internal consistency of the overall (summative) scale can be assessed using Cronbach’s $\alpha$ (see Chapter 14: Measuring association and agreement).

Cumulative scales. Cumulative, or Guttman, scales list a sequence of states, each of which includes all those previously listed. For example, a respondent might be asked to state if smoking can cause (1) undesirable symptoms, (2) illness, (3) severe illness, and (4) death. Guttman scales therefore allow multiple states to be selected as opposed to a single state on a Likert scale.

**Questionnaire bias**

Bias might be introduced into a survey by the form in which questions are asked. The investigator must be careful to maintain neutrality in each question as well as in the collection of questions together.

In each question, wording can introduce bias. (Intentional introduction of bias by wording is notorious in political fundraising “surveys.”) Reread your wording, asking yourself, “Can I detect what the investigator might want for an answer?” If so, reword.

Investigate the overall questionnaire for bias. The selection of topics and questions to be asked can introduce bias even when single questions are neutral. Ask colleagues or pilot respondents to complete the questionnaire and then ask if they can identify your preferences or intentions from it. If so, consider eliminating or adding questions to create a neutral balance.
Some mechanisms to improve questionnaires

When a draft questionnaire has been developed, test it on a pilot sample. Ask pilot respondents what they think you mean by the questions. Follow the steps in the process all the way through: record the data, transfer the data to a spreadsheet, analyze the spreadsheet, and form conclusions. Revise the questionnaire to correct the problems seen to arise. Try to shorten the questionnaire. (It can almost always be shortened to advantage.)

If the questionnaire is to be used in a large formal trial to assess some medical problem or condition, it may be necessary to standardize the questionnaire. This involves administering it to a large number of normal (in the medical, not statistical, sense) respondents and analyzing it in order to establish a baseline of normality to which putatively medically abnormal respondents may be compared. Normal statistical values are then available, such as the mean, median, standard deviation, and other properties of the data distribution.
Techniques to Aid Analysis

27.1 INTERPRETING RESULTS

If you have been careful in conducting your study, you asked specific questions, obtained data containing answers to those questions, and selected statistical methods that will tease out those answers. The interpretation is just those answers.

However, sometimes an investigator is not sure just what will emerge from data and, therefore, records measurements on many variables. Software provides a statistical result or two on each variable. What does one make of this mish-mash? Revisiting the steps in Section 7.8 will help sort it out.

As a start, list the goals, that is, what is being asked of the data, and see if the analysis results answer these questions. If they do not, there is a mismatch among question, data, and statistical analysis method.

EXAMPLE: QUESTION VERSUS DATA MISMATCH

An investigator wants to know if pre-injury to post-injury change in a certain short-term memory test, yielding a score between 0 and 50, is a marker of neurologic damage due to blast over-pressure (BOP) resulting from an explosive device. She obtains scores from 30 marines who have been injured in Afghanistan by BOP. She wants to compare these scores with the pre-BOP scores, but does not have them. She collects scores from marine recruits and performs a two-sample t test to see if the average is different between the two groups.

The question asks if a change in score occurs from pre-blast to post-blast. Her pre- and post-blast scores are not on the same people. The pre- to post-blast difference is confounded with individual differences. There is a question and data mismatch.

EXAMPLE: DATA VERSUS METHOD MISMATCH

Our investigator starts over, collecting memory test scores on a large number of marines in a pool from which those to be deployed to Afghanistan are drawn. A year later, she searches hospital records, identifying marines who suffered from BOP. She finds several for whom she had pre-deployment tests and she re-administers the memory test to them. Now she has a sample of marines with both pre- and post-injury scores. The average change in scores gives her the clinical information on the effect of BOP.
She then asks if this change is greater than might have occurred by chance. She calculates a correlation coefficient and finds it is significantly greater than chance.

Unfortunately, the correlation coefficient tells her if the high or low pattern of memory scores remains the same after BOP; it does not tell her if BOP damage changes the memory test scores. There is a data and statistical method mismatch. She should have performed a paired $t$ test on the pre- minus post-injury scores to see if there was a significant mean memory reduction.

**Imperfect Data and the Art of Statistics**

There are no perfect statistics because there are no perfect data. Statistical analysis might be thought of as opening a window into a room. You see one thing from one window and something a little different from the next. You try to find the window that allows you to see the contents of the room with scientific accuracy.

To an extent, the practice of statistical design and analysis contains elements of the art of statistics, in some ways like the art of medicine. If statistical design and analysis were purely technical with only one way to perform it properly, it could be completely automated. There is judgment involved in many aspects.

**27.2 SIGNIFICANCE IN INTERPRETATION**

**Definition of Significance**

The *significance level* of an event (such as a statistical test) is an estimate of the probability that the event could have occurred by chance. If the level is quite low, that is, the probability of occurring by chance is quite small, we say the event is *significant*.

Statistical significance does not mean that the event has any clinical meaning. It should not be confused with the common use of the term significant by society, meaning the event has some societal importance. Indeed, the confusion has led some statisticians to say we should not even use the word significant.

**The Purpose of Estimating the Significance Level**

The goal of estimating a significance level is to answer a question being asked of the data, such as, does this treatment work? Statistical tests are used in circumstances in which an unequivocal answer cannot be established. If cells always die when oxygen is cut off, we do not need a statistical test to establish that they require oxygen to live. Statistical significance says something like: “Odds are pretty good that this treatment is effective”. However, it will never give an unequivocal yes or no.

**Indicators of Significance**

Measures of significance approximate probabilities associated with outcomes. They usually are not exact probabilities, but only estimates. They may be thought of as markers or indicators, as temperature is a marker for infection.
The \( p \)-value is the most touted measure of significance. It gives crucial information, but certainly does not tell the whole story. The \( p \)-value estimates the probability of observing a result as or more indicative of the alternative hypothesis if the null hypothesis were, in fact, true. It does not address \( \beta \), the probability of a false negative. In addition, other aspects of the experimental design and analysis give other clues, some of which are estimated magnitude of effect, confidence intervals, correlation coefficients, indices, et al. Certainly a \( p \)-value should not be the only factor considered when assessing empirical evidence of association.

**What If Two Tests Give Different Significance Levels?**

A user often informally asks just if two samples are different, an oversimplification that causes confusion. A significance level is influenced by the form of analysis and underlying assumptions. For example, a two-sample \( t \) test and a rank-sum test comparing the same two samples will produce different significance levels. The difference occurs because the levels are calculated from different probability distributions. It should be so, for basically the precisely worded questions the two tests are answering are different. The \( t \) test asks if means of the two samples, drawn randomly from normal distributions with the same variances, are different, while the rank-sum test asks if the rank-order of a randomly drawn observation from one distribution precedes the rank-order of a randomly drawn observation from the other distribution.

If the user obtains different results from the \( t \) and the rank-sum tests, which should be believed? That answer is that the user should not be using both tests. The specific question answered by one is more appropriate than the other. The user should discern the difference and choose the one that best matches the data, assumptions, and situation. Ideally, this should have been done prior to conducting the analysis.

### 27.3 POST HOC CONFIDENCE AND POWER

**Post-Test Confidence in the One-Sample Mean**

We have done our test on a mean and reached a conclusion. What is our confidence interval on this mean? It does not affect the test result, but certainly provides additional understanding of the precision of our estimate and the plausible values of the parameter of interest. Consider testing \( H_0: \mu_0 = \mu \) versus \( H_0: \mu_0 \neq \mu \). For known standard error of the mean difference \( \sigma_m \), our test statistic \( z = (m-\mu)/\sigma_m \) (or its negative; the distribution is symmetric) is the number of standard errors apart are the theoretical and the observed means. Using the logic of Chapter 8, we may write this relationship as \( P[-z_{1-\alpha/2} < (m-\mu)/\sigma_m < z_{1-\alpha/2} | \mu_0 = \mu] = 1 - \alpha \). (For example,
\[ P \left[ -1.96 < (m-\mu)/\sigma_m < 1.96 \mid \mu_0 = \mu \right] = 0.95. \]

By adding \( m \) and then multiplying by \( \sigma_m \) throughout within the brackets, we achieve a form similar to Eq. (8.5):

\[
P \left[ m - z_{1-\frac{\alpha}{2}} \sigma_m < \mu < m + z_{1-\frac{\alpha}{2}} \sigma_m \mid \mu_0 = \mu \right] = 1 - \alpha
\]

(27.1)

the confidence interval on the unknown theoretical mean. If we estimate \( \sigma_m \) by \( s_m \), we can use similar logic to find the confidence interval on \( \mu \) for that case as

\[
P \left[ m - t_{1-\frac{\alpha}{2}} s_m < \mu < m + t_{1-\frac{\alpha}{2}} s_m \mid \mu_0 = \mu \right] = 1 - \alpha.
\]

(27.2)

Note that in the above probability statements, what is random are the confidence interval limits (we are considering \( m \) to be a random variable, not a fixed observed sample mean). This makes it clear that the confidence interval is an “inversion of a hypothesis test,” meaning that the confidence interval represents the set of all hypothesized values for \( \mu_0 \) that we would fail to reject H_0 given our sample. For this reason, a confidence interval can be interpreted as the set of all plausible values for the true population mean, where plausibility is defined by the failure to reject that parameter value with a level \( \alpha \) test.

**Power of the Test**

The upper bound on the probability of a false positive (rejecting the null hypothesis when the null hypothesis is, in fact, true), i.e. \( \alpha \), defines the critical value demarking rejection versus non-rejection regions for our test statistic. For example, a two-tailed test of a normally distributed variable yields critical values of \( \pm 1.96 \) for \( \alpha = 0.05 \). This selection occurs \textit{a priori}, i.e., before the test is made (and we hope before the data are gathered). At that point, the power of the test, that is, the probability of a true positive, or \( 1 - \beta \), is unknown, because \( \beta \) depends on the true value of the alternate mean, which is unknown. After completion of the test, many investigators ask themselves: Did the study have enough power? Some carry this further and want to perform a \textit{post hoc} power analysis. What does this tell the investigator? Is it even appropriate?

**Interpreting Post Hoc Power**

What does a \textit{post hoc} power analysis tell us? It adds nothing to the test, for that is complete. If our test result is significant, the observed mean has been estimated and may be taken as the alternate-hypothesis mean, allowing us to estimate \( 1 - \beta \). If our test result is significant, we did have enough power; most investigators will not see a need for a \textit{post hoc} power estimation.

If the test result is not significant, however, we have the problem of deciding whether (1) the null hypothesis is true or (2) a significant difference that exists in fact
failed to appear because of too small a sample size. Despite some investigators’ belief that a post hoc power analysis will distinguish between these states, it will not. In the face of insignificance, we find ourselves back in the same boat: we still do not know the value needed for the alternative hypothesis, upon which the power depends. A belief that a post hoc power estimation will add intuitive assurance about the study result is misguided. What it will add is estimation of the sample size needed if the study should be repeated. The assurance about the study result should come from the confidence interval.

The reader can find more detail on the inappropriateness of post hoc power analysis in interpreting a study result in the following three articles: Goodman and Berlin,\textsuperscript{1} Hoenig and Heisey,\textsuperscript{2} and Levine and Ensom.\textsuperscript{3}

### Calculating the Post Hoc Power

Because some users still want to find the post hoc power, the method of calculation is given here for the \( t \) test form. Recall that \( \alpha \) was obtained as the area under the probability curve defined by the null hypothesis to the right of the critical value. To calculate the post hoc power, we just substitute the observed mean for the null-hypothesized mean, standardize the distribution, and find the area under the curve to the right of the critical value. For example, continuing the asthma training test, our \( t \) statistic was calculated to be 4.05. To standardize the distribution, we move the center, or mean, 4.05 units to the left to make it 0. This, of course, also moves the critical value, previously at 2.04, 4.05 units to the left, so it becomes \(-2.01\). The power of the test is the area under a standard \( t \) with 31 df to the right of \(-2.01\). Recall that power \(= 1 - \beta\); it is easier to find \( \beta \), which is the area to the left of \(-2.01\), or, by symmetry, to the right of \(+2.01\). The closest right tail area in Table II for the right tail area (top row) is 0.025, associated with \( t \) value 2.04 for 30 df. A \( t \) of 2.01 for 31 df would be fairly close. Therefore, the power is close to \(1 - 0.025 = 0.975\). From a statistical software package, the post-hoc power \(= 0.973\).

**Exercise 27.1**

*What Was the Power of “Student’s” Original \( t \) Test?* Follow the steps to calculate the post hoc power for the result of Exercise 11.1.

### 27.4 Multiple Tests and Significance

#### The Issue

When making repeated tests on the same process, the \( p \)-values on each test can add up to produce an increasing overall type I error rate. For example, suppose we are testing
mean postoperative C-reactive protein (CRP) levels at 4 time points. For each time point, the probability of obtaining a spurious or false positive result is bounded by $\alpha$, taken by convention to be 0.05. If you have a spinner marked 1 to 100 and spin it 4 times, what is the chance that it will stop in the 1-to-5 region at least once? This is the same as performing four independent hypothesis tests. Using the methods from Chapter 3, we can see that it will be $1 - \alpha$, the probability of not stopping there at all, or $1 - 0.95^4 = 0.185$. We chose $\alpha = 0.05$ to bound the probability of a type I error result by 0.05, but now we have nearly 4-fold that chance.

As a more dramatic example, we record ESR levels at 6 time points (it responds more slowly than CRP), the points designated $t_1, t_2, \ldots, t_6$, and we want to identify the time at which mean ESR increased to significance; we test pairs of means, $t_1$ vs. $t_2$, $t_1$ vs. $t_3$, $t_2$ vs. $t_3$, $t_2$ vs. $t_4$, $t_2$ vs. $t_5$, $t_3$ vs. $t_6$. With 6 time points, the number of possible mean pairings to test is $\binom{6}{2} = 15$ [see Eq. (3.6)], so that the probability of a type I error being made on any of the tests (if they were assumed to be independent) has accumulated to 53.7% $[1 - (1 - 0.05)^{15}]$.

The problem of type I (and type II) errors has become quite serious with the advent of genomic testing, in which often thousands of tests of the same process are conducted.

**Some Approaches to Solving the Problem**

Approaches to testing a small number of means was addressed in Section 11.4 as multiple comparisons in interpreting a one-way ANOVA, and more completely in Section 16.9. Instead of repeated $t$ tests, an ANOVA provides an overall $p$-value and then post hoc tests on each pair of means are made, adjusting the risk for error so that the accumulated risk does not exceed $\alpha$.

For applications for which multiple comparisons methodology has not been well developed, such as tests of rank-order, a Bonferroni correction on the $p$-value may be made, in which the allowable risk of a false positive error is taken as $\alpha$ divided by the number of pairs to be tested. However, we should be aware that the Bonferroni approach is overly conservative. It requires a more stringent level of evidence than most investigators believe is fair and seriously inflates the number of false negatives.

These and related methods are collectively called family-wise error rate (FWER) controls.

**The False Discovery Rate: An Approach to Large Scale Testing**

If the objective in large scale testing is exploration, that is, statistical tests suggest rather than define results, the false discovery rate (FDR) may be calculated. The FDR identifies a set of potential, or “candidate”, positive test results, more likely than other
results, to be further investigated. \(FDR\) controls are less conservative than \(FWER\) controls, leading to higher power (lower false negative rate) at the cost of a higher false positive rate.

The \(FDR\) method depends somewhat on whether or not the statistics being tested are mutually correlated. It will be given first for the case in which they are independent or positively correlated; the case for negative correlation is similar with an additional term included. The method comes from Benjamini and Hochberg\(^4\) and Benjamini and Yekutieli.\(^5\)

Let \(m\) denote the number of test results. We place their \(p\)-values in increasing rank order, \(p(1), \ldots, p(k), \ldots, p(m)\). Find the value of \(k\), say \(k'\), such that

\[
p(k') \leq \frac{\alpha}{m}.
\] (27.3)

Accept these first \(k\) tests as those with rejected null hypotheses, that is, with significant results.

For example, suppose we have 6 test results and their \(p\)-values in increasing order are 0.001, 0.005, 0.009, 0.020, 0.080, and 0.100. With \(\alpha = 0.05\) and \(m = 6\), the values on the right side of Eq. (27.3) corresponding to \(k = 1, 2, 3, 4, 5, 6\) are 0.008, 0.017, 0.025, 0.033, 0.042, and 0.050. The first 4 \(p\)-values are less than the corresponding \(\alpha k/m\) values, that is 0.001 < 0.008, \ldots, 0.020 < 0.033. The remaining \(p\)-values are larger than the corresponding \(\alpha k/m\) values: 0.080 > 0.042 and 0.100 > 0.050. Thus, the critical \(k'\) takes on the value 4. The first 4 null hypotheses are considered rejected.

If the statistics being tested are negatively correlated, the term

\[
c(m) = \sum_{i=1}^{m} \frac{1}{i}.
\] (27.4)

is included altering the form of Eq. (27.3) to

\[
p(k') \leq \frac{\alpha}{m \times c(m)} k.
\] (27.5)

Pursuing the same example of 6 test results, the right side values are 0.003, 0.007, 0.010, 0.014, 0.017, and 0.020. Now \(k' = 3\); only the first 3 test results are taken as significant.

In interpretation, it might be useful to have an approximate mean false positive risk for the overall set of \(m\) tests. Such a rough \(FDR\) (\(RFDR\)) is given by

\[
RFDR = \frac{\alpha(m + 1)}{2m}.
\] (27.6)

In the example, with \(\alpha = 0.05\) and \(m = 6\), \(RFDR = 0.029\).
**EXAMPLE: DOES ANESTHETIC TYPE AFFECT RECOVERY TIME?**

In a variety of surgery types, 87 patients were given 4 different types of anesthetic, scaled to the same level of intended anesthesia, and time (minutes) to recovery from the anesthesia recorded. (The data appear in database DB34, provided by internet access.) The anesthetics (sample size followed by mean time in parentheses) were: (1) ketamine IV (26; 55.8), (2) ketamine IM (21; 61.8), (3) a narcotic (e.g. versed, fentanyl, morphine) (27; 67.5), and (4) propofol (13; 37.3). Is mean recovery time different for different anesthetic types? We need to compare type 1 with type 2, type 1 with type 3, ..., type 3 with type 4, yielding 6 pairings.

Performing two-sample $t$ tests on mean recovery times yields the following table of $p$-values:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.517</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.366</td>
<td>0.694</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.050</td>
<td>0.026</td>
<td>0.083</td>
</tr>
</tbody>
</table>

Of the 6 $p$-values, we see that one is significant at the 0.05 level and one is on the border. The probability of finding one or more $p$-values < 0.05 under the null hypothesis could be as high as $1 - 0.95^6 = 0.265$ if the tests were all independent. In this case we would expect a quarter of our results, that is, 1 or 2, to be false positives. Thus, we have no confidence that any result is a true positive.

A better approach would have been to perform a one-way ANOVA, followed by multiple comparisons. Such an ANOVA yields overall $p = 0.180$. We do not have sufficient evidence to conclude a difference among the population means. Performing Scheffé’s multiple comparisons yields the following table of six pair-wise $p$-values:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.969</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.783</td>
<td>0.973</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.626</td>
<td>0.419</td>
<td>0.200</td>
</tr>
</tbody>
</table>

We can see that no pair yields a significant test result. The 1 or 2 significant results from independent testing were likely spurious.

As illustration, the Bonferroni procedure would have compared the 6 test results to a critical value of $0.05/6 = 0.0083$; a test’s $p$-value would have to be less than 0.0083 to be considered significant. As the smallest $p$-value is 0.026 ($> 0.0083$), none of the mean pairs of recovery times was significant.

While we would not pursue the FDR approach in this case, it may be instructive to perform the calculations for illustration. We have $m = 6$ test results. The $p_{(i)}$-values with which to compare the rank-ordered test result $p$-values are given by $(\alpha/m)$
The rank-ordered test $p$-values and their corresponding $p_i$-values are (0.026, 0.0083), (0.050, 0.0166), (0.083, 0.0249), (0.366, 0.0332), (0.517, 0.0415), and (0.694, 0.0500). We can see that no test $p$-value is less than its corresponding $p_i$, so that $k' = 0$. No significant differences in mean recovery times are found. RFDR = $\alpha (m + 1)/2 = 0.029$. On average, we would expect results less than this value to be significant.

Exercise 27.2

Database DB24 (access on line) gives data from pediatric snakebite records. We ask if age makes a difference between being bitten in the hand/arm area versus the foot/leg area by season (seasons 1, 2, 3, 4). (Snakes and children are more active in warm seasons.) Using 4 seasons, $m = 4$. We could perform 4 $t$ tests on age for hand v. foot, one for each season. (a) Calculate the accumulated $p$-value over the 4 tests. The $p$-values for the individual tests are, respective for season, 0.100, 0.279, 0.034, and 0.695. However, an ANOVA on age by the interaction of season-by-bite-location yields a non-significant $p = 0.069$. (b) Comment on the probability that the significant 0.034 result represents a true positive. (c) Calculate the Bonferroni critical value and indicate what significance the Bonferroni procedure provides. (d) Assuming these tests were exploratory, find and explain $k'$ and the RFDR.

27.5 BOOTSTRAPPING, RESAMPLING, AND SIMULATION

What Bootstrapping Is

We are interested in a parameter (perhaps the mean, for example) of a distribution. We take an initial sample, from which we can make a single estimate of the parameter. We would like additional samples in order to achieve a better grasp on the distribution of the estimator for the population parameter but we cannot obtain additional samples. Bootstrapping is a scheme of resampling, i.e. taking additional samples repeatedly from the initial sample, to provide estimates of its variability. In a case where the distribution of our estimator is unknown via theory, bootstrapping is of especial help, in that it provides information about this distribution.

How Bootstrapping Works

Suppose we have a sample of readings from a distribution, perhaps 20 readings on estimated blood loss (EBL) in liters during open prostatectomies. We do not know the distribution from which this sample was drawn, but suspect it is not normal; EBL cannot be less than zero on the left tail of the distribution, but may extend on occasion to large losses on the right tail, creating a right skew to the distribution. We estimate the median EBL. How can we put a confidence interval on this median taking into account the skewed distribution? We resample using the bootstrap.
To form our first resample, we randomly select a reading, return that reading to the sample, randomly select a reading again (which has a 1 in 20 chance of being the same reading as on the first draw), and repeat this process until we have 20 readings. This composes our first resample. We find the median of this resample and proceed to a second resample. Resampled median estimates are generated many times, perhaps 1000. Obviously, this must be done by computer. To put a 95% confidence interval on the median, we put the resampled median estimates in order and find the EBL medians corresponding to the 2.5th and 97.5th estimates. This is a nonparametric outcome, that is, not dependent on assumptions about the nature of the initial distribution, such as normality (which includes symmetry) as would be required to use the mean $\pm 2SD$ form of a confidence interval. The bootstrap can be used to generate other descriptors of the initial distribution, such as the mean, the interquartile range, regression estimates, or nearly any parameter of the distribution of interest.

The Fundamental Assumption of the Bootstrap

The fundamental assumption of bootstrapping is that the initial sample from which all resampling occurs is representative of the target distribution. If the initial sample is biased, all the resamples will be biased.

Another assumption is that the observations in the sampling are identically and independently distributed, that is, they are all drawn from the same distribution with equal probability of occurring and are uncorrelated with each other. An example of the violation of this assumption is an ophthalmological sample of measurements on 30 eyes from 20 patients; some of the eyes arise from the same patient and therefore the readings may be correlated with each other while the readings on single eyes per patient are not.

The Jackknife

Related to the bootstrap is an earlier and less general development, the jackknife. In the jackknife, a randomly chosen observation is removed from the initial sample and the estimate (e.g. the sample mean) is recalculated. The observation is returned and the process repeated over and over. Though not used often today, the jackknife is still useful in certain circumstances.

EXAMPLE: CONFIDENCE INTERVAL ON THE MEDIAN SIZE OF A TUMOR I

We have measurements on the size (ml) of 115 tumors of the liver. In this example, we will use the first five readings as an unrealistic sample size to illustrate the bootstrap process; the full data set is used in the additional example below. The readings are 1.1, 2.6, 2.7, 2.1, and 5.1. If we put these tumor sizes in order, we see that the middle or median value, i.e. the 3rd, is 2.6 ml. We are going to resample 10 times, finding the
median of each. Then we will put the medians in order. The value between the first and second medians and the value between the ninth and tenth medians will enclose 80% of the medians, producing an 80% confidence interval on the original median of 2.6. To perform resampling, we choose one of the five sizes at random. We get the first, 1.1. Then we choose again from the five, getting the fourth, 2.1. Repeating, we get 1.1, 2.1, and 2.7 on the next three draws. Note that the first and fourth values occur twice, the third value once, and the second and fifth values not at all. That composes our first resample. The median is 2.1. Then we repeat this process 9 more times. The 10 medians, in increasing order, that arose were 2.1, 2.1, 2.1, 2.1, 2.6, 2.6, 2.6, 2.6, 2.6, 2.7. The 80% C.I. is 2.1 to 2.65.

If the process in this example were repeated, the result would not be exactly the same, because the tumor sizes chosen randomly would be different each time. This is why a large number of replications is needed; the accuracy increases with the number of replications.

**ADDITIONAL EXAMPLE: CONFIDENCE INTERVAL ON THE MEDIAN SIZE OF A TUMOR II**

In the actual study, there were 115 tumor sizes (ml), with mean = 2.77, median = 2.6, and range = 0.7 to 6.0. Because the distribution of sizes is limited on the left of the distribution and not on the right, it tends to be right skewed, so the median is a better representation of the central location of the distribution than the mean. Using statistical software, we produce 1000 resamples of 115 random drawings each (with replacement, i.e., the randomly drawn value is retained in the sample for the next random drawing) and find the median in each resampling. 95% of the medians thus produced lie between the 2.5th and 97.5th of the ordered list, producing the 95% C.I. We find that this C.I. is 2.3 to 2.9.

**Exercise 27.3**

One question in the liver tumor study was whether or not patients’ BMIs were related to tumor size. As part of the process, we wanted to describe the BMI distribution. Like tumor size, BMI is limited on the smaller size but may have some quite large values, rendering it right skewed; we use the median and its C.I. for description. To form an exercise that can be done by hand, use the first four BMI values: 29, 24, 27, 38, which has median 27. Produce 5 resamples of 4 numbers each, finding the median of each. Order the five medians and calculate the value between the 1st and 2nd and between the 4th and 5th as giving a 60% confidence interval.

**Simulation**

We have seen sampling as drawing a sample randomly from a parent distribution, usually the true probability distribution. Bootstrapping is repeatedly drawing a sample
randomly from the initial sample. More generally, *resampling* is repeatedly drawing a sample randomly from the same distribution, not necessarily the initial sample.

If we have reason to believe that a variable of interest arises from a known distribution, say a normal distribution with mean $\mu$ and standard deviation $\sigma$, we could sample and then resample from this parent distribution to obtain parameter estimates $m$ and $s$ that converge on $\mu$ and $\sigma$ as the number of resamples grows large. But we already know $\mu$ and $\sigma$, so why would we do this?

If we have a model in which several variables interact in ways not exactly known, the effect of sampling in one variable (mimicking a “real-life” occurrence) on the outcomes of the other variables simulates a *system* effect. Thus, investigating the interaction of several variables in a model by resampling in one or a subset of variables is called a system simulation, or more usually just a *simulation*.

As an example, suppose we know that, for a fixed exercise regimen, body mass index (BMI) has a known regression equation on calorie intake. We can control calorie intake directly, but not BMI. We record calorie intake and identify its frequency distribution. We resample this distribution, recording with each sample the corresponding BMI value from the regression. We then have a distribution of BMI values from which we can calculate mean and confidence interval for any calorie intake level.

Most simulations are much more involved, some requiring teams of investigators working over many months to accomplish. Most are beyond the scope of this book and infrequently undertaken by the clinical medicine investigator. One simulation model being used with increasing frequency is Markov chain Monte Carlo (MCMC). More on that topic is briefly described in Sections 28.5 to 28.8 in the next chapter and treated more thoroughly in Chapter 20 of the 3rd edition of this book.

### 27.6 BLAND-ALTMAN PLOT: A DIAGNOSTIC TOOL

A Bland-Altman plot is a useful display of the relationship between two paired variables using the same scale. It allows you to perceive a phenomenon but does not test it, that is, does not give a probability of error on a decision about the variables as would a test.

A Bland-Altman plot consists of a plot of the difference between paired readings of two variables [e.g., central venous pressure (CVP) and peripheral venous pressure (PVP)] over the average of these readings, with $\pm 2SD$ lines (Confidence Interval or CI) parallel to the mean difference line.

(Continued)
The Bland-Altman plot displays four types of data misbehavior: (1) systematic error (mean offset), (2) proportional error (trend), (3) inconsistent variability, and (4) excessive or erratic variability.

EXAMPLE: COMPARING BALANCE WITH VERSUS WITHOUT HEARING PROTECTION

To illustrate how a Bland-Altman plot shows these types of data anomaly, let us examine five possible pairings in 20 subjects of balance with and without combat arms ear plugs (CAEPs). It is asked if CAEPs affect the balance of subjects performing balance-affected tasks, such as piloting an aircraft, making terrain judgments, or making analog mechanical or computer assignments (e.g., aiming a weapon). The subject stands on a balance-sensitive plate and performs a standardized task, while the plate records the mean variability in degrees from vertical. The data appear in Table 27.1.

Table 27.1 Imbalance (degrees) of subjects while performing a task with ear protection (CAEP3) and without (No CAEP) along with similar data columns fabricated for illustration.

<table>
<thead>
<tr>
<th>Pt. No.</th>
<th>No CAEP</th>
<th>CAEP 0</th>
<th>CAEP 1</th>
<th>CAEP 2</th>
<th>CAEP 3</th>
<th>CAEP 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.4</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>2</td>
<td>.5</td>
<td>.3</td>
<td>.5</td>
<td>0</td>
<td>.4</td>
<td>.4</td>
</tr>
<tr>
<td>3</td>
<td>.6</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>4</td>
<td>.4</td>
<td>.4</td>
<td>.6</td>
<td>.1</td>
<td>.4</td>
<td>.4</td>
</tr>
<tr>
<td>5</td>
<td>.6</td>
<td>.8</td>
<td>1.0</td>
<td>.8</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>6</td>
<td>.7</td>
<td>1.0</td>
<td>1.2</td>
<td>1.2</td>
<td>1.0</td>
<td>.8</td>
</tr>
<tr>
<td>7</td>
<td>1.2</td>
<td>1.0</td>
<td>1.2</td>
<td>1.3</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>8</td>
<td>.5</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>.9</td>
<td>.8</td>
<td>1.0</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td>10</td>
<td>.6</td>
<td>.8</td>
<td>1.0</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td>11</td>
<td>.5</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>12</td>
<td>.7</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>13</td>
<td>.6</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>14</td>
<td>.8</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.4</td>
</tr>
<tr>
<td>15</td>
<td>.8</td>
<td>.8</td>
<td>1.0</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
</tr>
<tr>
<td>16</td>
<td>.5</td>
<td>.5</td>
<td>.7</td>
<td>.3</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>17</td>
<td>.7</td>
<td>.5</td>
<td>.7</td>
<td>.3</td>
<td>.5</td>
<td>.6</td>
</tr>
<tr>
<td>18</td>
<td>.6</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>19</td>
<td>.5</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>20</td>
<td>.8</td>
<td>.9</td>
<td>1.1</td>
<td>1.0</td>
<td>.9</td>
<td>.9</td>
</tr>
<tr>
<td>21</td>
<td>.6</td>
<td>.6</td>
<td>.8</td>
<td>.5</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>22</td>
<td>.4</td>
<td>.4</td>
<td>.6</td>
<td>.1</td>
<td>.4</td>
<td>.4</td>
</tr>
</tbody>
</table>

(CAEP: Combat arms ear plugs).
Figure 27.1 a. A B-A plot for No CAEP paired with CAEP0. The points center about difference = 0, provide a reasonable confidence interval, and remain in the same general pattern for all horizontal axis values. There is no remarkable data behavior. b. A B-A plot for No CAEP paired with CAEP1. In this type, the cluster of points may lie above or below the mean, indicating an offset, a bias, a systematic error. It might be wise to test means. In Fig. 27.1b, we can see that the mean of the data lies considerably above 0, indicating that the CAEP wearers’ degrees of sway (imbalance) are consistently greater than with no CAEP. c. A B-A plot for No CAEP paired with CAEP2. The cluster of points go from below the mean on the left to above as you move to the right, showing a trend or error proportional to size of measure. d. A B-A plot for No CAEP paired with CAEP3. The cluster of points surround the mean tightly on the left and spread out to greater variability upon moving right, sort of cone-shaped, showing variability dependent on the magnitude of the measure. This effect appearing in Fig. 27.1d is not very strong. e. A B-A plot for No CAEP paired with CAEP4. Only about 5% of the data should lie outside the plus-minus CI and these points should not be dramatically far outside the confidence interval. If too many points lie outside, then the data are erratically variable. This effect needs a larger sample size (>100 data points) to see dependably if 5% of the data lie outside the 95% confidence interval. In Figs. 27.1e, 3 of the 22 points, nearly 14%, lie outside the confidence interval, a suspicious proportion.
The Bland-Altman plots pair No CAEP with CAEP0 through CAEP4, representing no anomaly and anomaly types 1 through 4, respectively, appearing in Figs. 27.1a through 27.1e, respectively. These levels are credible but columns labeled 0, 1, 2, and 4 are fictional in order to show the plot types using the same type of measurement. The variables No CAEP and CAEP 3 are actual recordings.
ADDITIONAL EXAMPLE: CLINIC VERSUS LABORATORY DIFFERENCES IN INR VALUES

Fig. 27.2 shows a Bland-Altman plot of the 104 clinic versus laboratory data on warfarin INR values in DB13. Three of the anomalies identified in the Fig. 27.1 plots can be seen. The mean is offset, lying above zero, suggesting a mean bias. The data cluster
more greatly in the upper left and the cluster moves downward as we pass from left to right, suggesting that the mean bias is greater for smaller INR values. No data lie above the upper confidence bound (2 or 3, about 2.5%, would be expected), but more than 7.5% lie below the lower bound, suggesting a skew in the data.

**Exercise 27.4**

*Shown in Fig. 27.3 are the preoperative versus postoperative breast implant plasma silicon levels from DB5. Interpret the results of the Bland-Altman plot.*

---

### 27.7 Cost Effectiveness

*Cost per unit of benefit* was addressed in Section 22.17 in the context of number needed to treat. More generally, the cost per unit of benefit is a frequent issue arising in all types of modeling. *Cost* is not restricted to monetary units but may be evaluated as the loss of anything we do not want to lose: money, lives, time, materials, quality of life, etc. *Benefit* may be evaluated as the gain of anything we want to gain: a similar list. The word *evaluated* was chosen intentionally; both cost and benefit must be quantitatively measurable. In medical screening of patients for disease, we might speak of dollars spent (cost) per disease detected (benefit). Suppose we have a new drug that cures diabetes but has a side effect of (a low rate of) myocardial infarction (MI). We might speak of dollars spent (cost) per cure (benefit) or of MIs (cost) per cure (benefit).
The foregoing criteria are easily quantifiable: count dollars or detections or cures or MIs. Other criteria are more difficult to quantify. Quality of life (QoL) is often approached as a questionnaire: several questions (e.g., ability to function, independence, level of pain) are asked as Likert scales to be rated on a, perhaps, 1-to-5 scale and then the ratings are added to form a QoL score. Risk of heart disease from obesity may require adding “apples to oranges”, as in combining the effects of high cholesterol, body mass index, frequency of exercise, and quality of diet. Measures of benefit may involve trading off costs and benefits. A very controversial example is the rationing of health care.

As the cost of health care rises world-wide, we are destined to see more and more use of cost effectiveness to allocate funds to treatment categories and to trade off high cost per patient for a few against low cost per patient for many. For a fixed available amount of money, how does an organ transplant for an aged person compare with anti-malarial vaccination for 1000 healthy children? Is it fair to let the person with a failing organ die to reduce the risk of malaria in children? Such difficult questions will inevitably depend on relative cost effectiveness evaluation.

In summary, the concept of cost per unit of benefit is simple and straightforward. However, measuring costs and/or benefits may be difficult and controversial. In the difficult issue of allocating limited health care resources, cost effectiveness will play an ever increasing role.

**EXAMPLE: COST EFFECTIVENESS OF PNEUMOCOCCAL VACCINE IN THE U.S. NAVY**

Vaccination for *streptococcus pneumoniae* has long been recognized as cost effective for elderly and immunocompromised populations. Is it cost effective for healthy military personnel? Costs include the expense and allocation of medical resources to conduct a vaccination program and to treat side effects. Benefits include savings in expense and allocation of medical resources currently used to treat the many types of pneumococcal infection, along with saving manpower currently lost to such illnesses, and quality of life improvement due to the avoidance of such illnesses.

Markov chain analysis (noted in Chapter 28 and addressed in some detail in the 3rd Edition of this book) is an apt statistical method for evaluating cost versus benefit tradeoffs. One study used it for the purpose of this example, providing an occurrence diagram and drawing transition probabilities from a number of medical sources over a 24-year period. Costs and benefits were standardized in dollar values, except for quality of life which was standardized in life-years. Vaccination was given practically simultaneously Navy/Marine Corps-wide. Pre- and post-vaccination costs and benefits were compared by a Markov chain analysis.

It was shown that the incidence of illness due to *streptococcus pneumoniae* was reduced 70%. Net gain was found to be $5,681,000 (1997 U.S. dollars) and 54 quality of life years across the service (at the time 575,000 personnel).
REFERENCES

28 Methods you might meet, but not every day

28.1 OVERVIEW

Medical investigators often encounter statistical methods that they have seldom or even never seen before in journal articles, references from mentors and reviewers, and the like. In previous chapters, this book has addressed the most commonly used statistical methods. What of those not commonly met? This chapter lists and provides definitions for a few of the less commonly met statistical methods. In cases in which a definition is not adequately descriptive, an example is given of a situation in which it might be applied. This chapter does not intend to present methodology. The purpose is solely orientation.

28.2 ANALYSIS OF VARIANCE ISSUES

Additional pairwise comparison methods following analysis of variance

Section 11.4 addresses the interpretation of a significant analysis of variance (ANOVA) with three or more categories. Which pair is, or which pairs are, responsible for the ANOVA’s significance? Subordinate testing of each pair may be done by multiple comparisons (or by multiple-range tests) that provide the same sort of interpretation. Table 11.11 lists five multiple comparisons tests. The reader might also meet those of Dunnett, Gabriel, or Hochberg, or the multiple-range tests of Duncan, Student—Newman—Keuls, or Tukey.

Equality of variances assumption

Levene’s test is a test of the equality of variances assumption in ANOVA.

The Brown—Forsythe and Welch tests are calculations of ANOVA significance by adjusting results for unequal variances.

The Box—Cox transform is designed to make the distributions of residuals closer to normal and their variances more similar.
Nonadditivity

Additivity is a property often assumed in a two-way ANOVA in which the contribution of the two variables to a cell value is the sum of individual contributions. In some cases, too small a sample leads to a lack of $df$ to test for nonadditivity. Tukey\(^1\) provided a revised model for the ANOVA, leaving $1\ df$ for the test, and provided the test. The test indicates the requirement for a transformation, suggests what transformation is appropriate, and then tells how well the transformation did its job.

Nonlinear models

The general linear model is not restricted to linear algebraic models; it may include nonlinear forms, for example, log, exponential, and Gompertz (traditional biologic growth model), among others, as well as combinations of algebraic and/or nonlinear forms.

28.3 REGRESSION ISSUES

Bias in regression variables

The *Heckman selection* model provides an adjustment for biased regression variables.

Analysis of residuals

A great deal can be determined from the set of points given by the data minus the fit, that is, by the *residuals*. Residuals have been mentioned, but not treated thoroughly. The pattern of residuals shows where the model does not fit and how serious this discrepancy is. Outliers are more clearly identified by residuals. Characteristics of the residuals can assess how well the model meets the assumptions, such as independence of errors, normality of distribution, and so forth.

As an example, consider the plot of Fig. 16.9, survival of malarial rats given a placebo treatment. The data were fitted with linear and quadratic (second-degree) models and the fitted values subtracted from the data. These residuals are shown in Fig. 28.1. In the plot of residuals from a linear fit (A), the leftmost data lie above the 0-line, the central data lie below 0, and the rightmost data are again above 0, demonstrating a curved (quadratic) pattern. In the plot of residuals from a quadratic fit (B), the data lie alternately (by pairs) above and below 0, a more nearly linear pattern with no meaningful departures from the fit. Thus if we had started an analysis with a putative linear fit, the residuals would have shown us a residual (quadratic) pattern remaining after the fit, indicating that the fit could be improved. Upon sophisticating to a quadratic fit the residual would have shown no meaning pattern, indicating an improved fit.

An analysis of residuals involves testing and fitting procedures much the same as with the original data, revealing not only the quality of the fit but also in some cases relevant variables and relationships not apparent in the original data.
If groups or categories of a phenomenon being measured are far from linear but non-linear models are not appropriate, ridit analysis may assist the analysis. For example, consider the World Health Organization’s severity of bleeding grading scale. Grades 0–2 require no intervention, Grade 3 requires transfusion, and Grade 4 leads to death. The difference from Grade 1 (petechiae appear) to Grade 2 (clinically notable loss), for example, is not nearly as serious a difference as from Grades 3 to 4. In ridit analysis a reference group is defined, equivalent measures are obtained from it, and ridits are calculated as the proportion of the reference group falling into a function of proportions in the affected group. Mean ridits are then used in calculating probabilities and odds and testing associated factors now adjusted for the severity discrepancy. Ridit analysis has received some criticism and is not widely used.

**28.5 MULTIVARIATE METHODS**

**Canonical correlation**

Multiple regression, met in Chapter 16, Multiple linear and curvilinear regression and multifactor analysis of variance, and following, is a form of multivariate analysis. In this case, one dependent variable is predicted by several independent variables. A coefficient of determination $R^2$ is calculated and may be considered as a multiple correlation coefficient, that is, the correlation between the dependent variable and the set of independent variables. If this design is generalized to multiple dependent variables, a correlation relationship between the two sets is of interest.

*Canonical correlation* is a term for an analysis of correlation among items in two lists (vectors of variables). For example, does a list of lab test results correlate with a list of
clinical observations on a patient? The goal is to find two linear combinations, one for each list of variables that maximize the correlation between them. The coefficients (multipliers of the variables) act as weights on the variables providing information on the interrelationships.

Suppose an investigator is interested in differentiating forms of meningitis. Blood test results (C-reactive protein, blood counts, etc.) along with MRI and lumbar puncture (LP) findings are taken from samples of meningitis patients having different forms. Canonical correlations are generalizations of simple correlations between individual variables to correlations between groups. In this case, canonical correlations are found between blood test results as a group and MRI/LP results as a group for each form of meningitis and may then be compared with one another.

**Hotelling’s $T^2$**

A two-sample $t$ test compares means on a single continuous measure between two groups. The $T^2$ test of Harold Hotelling compares means of two or more continuous measures simultaneously for the two groups. For example, it has been suggested that because increased levels of vitamin D have been shown to reduce tooth loss, it may increase the rate of correction of pediatric scoliosis. An investigator measures the difference between left and right thoracic height, Cobb angle, kyphosis, and lordosis on two groups of children with scoliosis. One group was given large vitamin D doses and the other group was not. The investigator uses $T^2$ to test the difference between the two groups on the basis of the four spinal measurements simultaneously. $T^2$ will also test a single or paired sample against a hypothesis of zero, as a generalization of the paired $t$ test.

**Discriminant analysis**

Discriminant analysis classifies sets of patients or measures into groups on the basis of multiple measures simultaneously. A line (or plane or hyperplane, depending on a number of classifying variables) is constructed between the two groups in a way that minimizes misclassifications. Patients whose results appear on one side of this separator are classified as most likely to have arisen from one group and those with results on the other side are classified as likely to be from the other group. An otolaryngologist wants to classify snoring patients into two groups: one group of patients requires surgery, whereas the other group requires less aggressive treatment. He uses apnea—hypopnea index, maximum decibels (dB) of snoring, and percentage of snoring time at greater than 50 dB as classifying variables. The method produces a plane in three dimensions separating the groups, and the otolaryngologist classifies patients according to their position relative to the plane in the three dimensions. More advanced forms of discriminant analysis address classification into more than two groups.
Mahalanobis' distance

Mahalanobis' distance is a measure of distance $D$ between the multiple means (centroids) of the two groups used in discriminant analysis. $D^2$ may be used to test the significance of this distance, as does $T^2$.

Multivariate analysis of variance and covariance

Multivariate ANOVA (MANOVA) and analysis of covariance (MANCOVA) extend those methods to situations having more than one dependent variable. In DB10, we investigate the effects of surgery on hamstrings or quadriceps by (1) time to perform hops and (2) centimeters covered in hops for the operated leg compared with the non-operated leg. We could analyze either (1) or (2) by the methods of this book. If we wanted to contrast legs by both (1) and (2) simultaneously, we could use Hotelling's $T^2$ (see earlier). If we had not only legs but also two types of surgery to compare, we would have a $2 \times 2$ ANOVA for either (1) or (2) or a MANOVA for both simultaneously. If we added the age of the patient (continuous) to leg and surgery type (categorical) and analyzed both (1) and (2) simultaneously, we would have a MANCOVA.

Principal components analysis

Let us start with an example. We have a sample of patients at risk for heart disease with the following measures: cholesterol level; age; socioeconomic level; dietary intakes of saturated fats, carbohydrates, and protein; and mean daily burn off of calories through exercise. Several of the variables are correlated, but cause and effect is not clear. Is there a component of several variables that accounts for their correlation? We do not know what this component might be, but if we find it, we might be able to divine a name or meaning for it by noting the relative contributions from each variable. The search for such a component, and for a second after the first has been removed, and so on, is termed principal component analysis. The matrix (array) of variances and covariances is subjected to mathematical procedures that produce "eigenvalues," a measure of "strength" of each component, and "eigenvectors," a list of contributions to the eigenvalue of each variable (but that may not have been adjusted for scale differences). The production of eigenvalues and eigenvectors is mathematical and unambiguous. The trick is to make sense of their meaning.

Factor analysis

Factor analysis is similar in method to principal component analysis, but it uses correlations in the matrix that produce eigenvalues and eigenvectors. The goal is to identify one or a few factors existing in a set of many variables that will explain most of the outcomes. Again, the mathematical results are unambiguous. The challenge is to name and characterize the emerging factors.
Cluster analysis

Suppose you have clinical and laboratory test results on a large sample of patients. If we should group similar sets of results together, perhaps we could see a pattern of disease similarity in a cluster. We might find that diseases we had believed to be quite different have similarities. We might find that we can form a list of potential diseases to be considered when we are given a pattern of test results. Clustering is obtained mathematically by statistics that compare numerical similarities or distances in multivariate space.

Multivariate data displays

Every day, we see two-dimensional displays, particularly graphs, and sometimes we see three-dimensional graphs. (We speak of graphs that display values from three variables, not two-dimensional graphs with a third dimension added to provide drama; these latter sorts provide no additional information and often prevent readers’ ability to perceive values accurately.) However, displaying more than three dimensions is a challenge. A fourth dimension has sometimes been displayed as varying symbols for data points, as varying colors, or as a sequence of patterns through time.

One very creative technique that can display several variables simultaneously was representing them as facial characteristics, devised by Chernoff\(^2\) of Harvard University. Professor Chernoff told the author that people are particularly sensitive to subtle differences in facial structure or expression and therefore should be able to perceive differences in data characteristics displayed in this manner. One variable’s values might be displayed as the length of the eyebrows, another as the tilt of the eyebrows. Height, length, and tilt of the eyes could display values of three more variables, and so forth.

The drawback is that different facial characteristics carry different weights in interpretation and, further, people weigh facial characteristics differently, one person to another, so that an unimportant variable might show the most striking effect while an important variable might become unnoticed due to their assignment to facial characteristics. Nevertheless, Chernoff’s faces can be very helpful in the right circumstances.

28.6 MARKOV CHAINS: FOLLOWING MULTIPLE STATES THROUGH TIME

Markov chains might be placed under multivariate techniques because they involve multiple variables and use much of multivariate methodology, such as matrix algebra. However, they also involve some rather different concepts, such as progression through time and Monte Carlo methods, so are juxtaposed to multivariate methods but listed as separate sections.
Markov chain modeling

Markov chains are sets of occurrences followed through steps in time. As an example, think of flu passing among the crew of a ship at sea across time steps of a week. A ship’s hand can be healthy and stay so for the week or pass to sick. The next week this hand can be still sick or have regained health or have died. The probabilities of a random hand passing from one state to another during this week are arrayed in a rectangular array called a matrix. Similar to ordinary arithmetic, matrices can be added, subtracted, multiplied, and raised to powers, although these procedures are limited to matrices having conforming dimensions. The definitions, mechanism, and chain name were devised by Markov in 1906.3

A shipboard flu epidemic

We can start with a ship’s complement healthy and follow the passage of flu through the crew through the chain of probability matrices. If we have beginning and ending numbers, we can estimate probabilities. With these probabilities, we can predict the health patterns of crews on future voyages, learning how much lost manning will require adjustment or additional crew. We can alter the probability of illness by, say, giving flu vaccine and learn what effect the vaccine has on crew health.

Where This method can be found

The third edition of this book goes into Markov chain modeling in some detail, but it is too infrequently used to compete with other needed methods and was therefore dropped for the fourth edition. An investigator who suspects it would be useful is directed toward that edition.

Further examples

Other examples included there are the pattern of traumatic brain injury (TBI), post-traumatic stress disorder (PTSD), and suicide among troops in insurgency warfare combat; the pattern of actinic keratosis and skin cancers across periods of sun exposure; and the pattern of an H1N1 pandemic in the catchment of a large hospital. These examples are begun with Markov chain modeling and followed through the more advanced methods next.

28.7 Markov Chain Monte Carlo: Evolving Models

A real situation would have some variability in the outcomes along with the ability to winnow out real from chance outcomes. In short, the Markov chain’s probabilities would have distributions as would occur in real life rather than be fixed values. That
situation and others of a more advanced nature fall under the heading of Markov chain Monte Carlo (MCMC).

Some history
Historically, effective system simulations, such as Markov chain modeling and MCMC, had to await the development of computers due to their complicated nature. During the Manhattan Project in World War II, the development of early computers enabled the first extensive use of resampling in simulations, named Monte Carlo sampling by early users after the famous gambling casino in the small nation by that name.

We see two classes of MCMC. The first, that we refer to as evolving models, uses resampling from probability distributions to generate confidence intervals (or other distribution descriptors) on the mean outcome probabilities of Markov chains. It is called evolving because the time periods used are clinically relevant, and there are typically only a few such time periods included. The goal is to follow—and conduct virtual experiments on—the evolution of a clinical process, such as an epidemic. The second, stationary models, is addressed in the next section.

Markov chain Monte Carlo as Bayesian statistics
MCMC is often thought of as a Bayesian method because a matrix of prior probabilities is postulated, and then a statistical procedure is engaged to develop a matrix of posterior probabilities, a process basic to the Bayesian approach.

Processing Markov chain Monte Carlo
We know that each of the probability estimates in the transition matrix has a probability distribution and so together they have a joint probability distribution. Computer-based sampling and processing are necessary because the simulation is very computationally intensive. The Gibbs sampler, named for physicist J. W. Gibbs but developed by the Geman brothers, provides a mechanism to pass quickly through thousands of loops but cannot be used in all applications. More general is a set of algorithms, developed earlier by Metropolis et al. and extended by Hastings, known as the Metropolis-Hastings algorithms, that allow faster processing of the set of probability states.

Examples of Markov chain Monte Carlo’s evolving models
The examples of Markov chain modeling—shipborne flu, the pattern of TBI, PTSD, and suicide among troops in insurgency warfare combat; the pattern of actinic keratosis and skin cancers across periods of sun exposure; and the pattern of an H1N1 pandemic in the catchment of a large hospital—can all be pursued by
MCMC with more complete and more accurate outcomes. These were all examined in this book’s third edition.

28.8 MARKOV CHAIN MONTE CARLO: STATIONARY MODELS

The second class of MCMC, usually referred to as stationary models, more informally described as “stable” or “fixed,” seeks to determine the actual probabilities of a clinical process that is in equilibrium, that is, is stationary. Extensive repeated Monte Carlo sampling is conducted from frequency distributions of all the probabilities involved simultaneously, that is, joint frequency distributions. If the outcome values cease to change after a large number of repetitions, they may be taken as the actual probabilities. The purpose is not to influence a developing process as in the previous section but to identify and describe aspects of processes that are fixed in time; it deals more with prevalence than incidence. It is a tool that is proving very useful in a number of fields of medicine, including medical genetics and epidemiology.

Requirements of stationary models are that the process eventually converges to a nonchanging state and that correlations among variables “damp out” after a large number of iterations. If that correlation does not disappear, one solution is to “thin” the process by recording only every, say, 1000th outcome, skipping the intervening iterations. A more recent solution is parallel sampling, in which several Monte Carlo simulations are developed simultaneously. The parallel outcomes are then combined to achieve convergence.

Examples from ophthalmology

The third edition of this book contains two examples: Providing early detection of the progression of glaucoma affecting a visual field and, from the literature (Levine and Ensom), showing that short-wavelength automated perimetry was not a successful predictor of glaucoma onset.

28.9 FURTHER NONPARAMETRIC TESTS

The Runs Test (Wald–Wolfowitz) is a simple procedure that tests if there are too many or too few binary observations of the same type in a sequence to be randomly occurring. For example, when allocating patients to two arms of a study, it tests if there are too many of one characteristic, for example, sex or advanced age, in sequence for the sample to be unbiased.

The sign test is another name for a binomial test of proportion (see Section 9.6).

A number of tests not seen frequently were included in this book because they provide tests for designs an investigator might need to use but which appear in few if
any commonly used books on biostatistics. In Section 11.11, we found that ranked outcomes of three or more treatments given to each patient of a sample could be tested by Friedman’s test. Cochran’s Q, met in Section 9.9, is a version of Friedman’s test in which the outcomes are not ranks but rather are binary outcomes (healed or not, reaction or not, etc.). Testing for trends (consistent change across the sequence of x-axis values) are the ptrend (two categories) and nptrend (three or more categories) tests in Sections 9.10 and 11.12, respectively. Comparing matched ranked samples with ranked outcomes is done by Page’s L test in Section 11.13.

28.10 IMPUTATION OF MISSING DATA

Imputation is the term applied to estimating the values of missing data. Missing data may imply bias. For example, the Connecticut Tumor Registry started in 1941 contained data from records back to 1936, from which patients who had died were omitted, leading to an underestimation of mortality. Imputation may not rectify such bias, because imputed data are estimated only from data that are present. For unbiased data, missing data cause little to huge effects, depending on the analysis being used. Many statistical procedures may be conducted without concern for missing data. If it is a concern, the computer user should investigate the software to see how it handles missing data. Some software drops every case (patient) in which there is a single missing datum. A study starting with 1000 patients may result in analysis conducted on 100 because of missing data deletions, one variable missing in some patients, another variable missing in other patients. Often the user is not alerted to this reduction in sample size and proceeds with the analysis of 100, thinking it is of 1000. Usually, one or a few missing data may be imputed without devastating the results, but it is not advisable to impute a large number of missing data. Imputation may be done by a variety of methods. The simplest, but not the best, method is to insert the average of that variable for other cases. A more sophisticated method is to use all cases except that with the missing datum to predict the missing-datum variable in a multiple regression of all variables other than that with the missing datum and then use that regression to predict the missing datum. Other methods exist as well.

28.11 FRAILTY MODELS IN SURVIVAL ANALYSIS

Frailty models are extensions of Cox proportional hazards regression. The Cox model assumes a homogeneous database, that is, all subjects have the same hazard rate. Do some patients have shorter time to survive, that is, are more failure-prone or “frail” than others? If so, the database is not homogeneous. Frailty models, having a conceptual history in accident proneness, were formed to account for deviations from this
homogeneity due to unmeasured variables. Other variables that were not used in the initial Cox regression are taken into account to explain the frailty occurring in the data. For example, the disease-free survival state following gallbladder treatment among women increases with age for European Americans and decreases with age among African Americans. If a Cox proportional hazards regression was conducted using health characteristics and age as variables, the effect of age as a hazard would be muted. Frailty modeling would take ethnic origin into account.

**28.12 BONFERRONI “CORRECTION”**

Bonferroni correction extends the Bonferroni approach from multiple comparisons discussed in Section 11.4 to multiple test results. If $k$ significance tests, each with error rate $\alpha$, are conducted on the same set of data answering aspects of the same data question, the overall error will increase to $1 - (1 - \alpha)^k$. The critical value for concluding significance is adjusted to $\alpha/k$.

**28.13 LOGIT AND PROBIT**

Suppose we recorded the survival or death of a pathogen as depending on increasing dosages of an antibiotic 5 days after administration to a sample of ill patients. Logit (for “logistic unit”) and probit (probability unit) are mechanisms to transform the data into a form suitable for regression analysis. Logit analysis is almost the same as logistic regression (see Chapter 17: Logistic regression for binary outcomes), yielding similar results. Probit analysis is a little different but again yields similar results in the end. In probit analysis the residuals (errors) are assumed to follow a standard normal distribution; in logit analysis the residuals follow a logistic distribution. Both methods give coefficients; logit’s coefficients are the natural logs of odds ratios, whereas probit’s coefficients are more difficult to interpret.

**28.14 ADJUSTING FOR OUTLIERS**

Outliers are extreme values in a sample that are impossible or highly improbable to have arisen from a “well-behaved” population distribution. They may arise as a natural but rare occurrence, as a fluke occurrence, or as a data collection or recording error. The first should be retained and the last should be deleted, but there is no statistical method to identify the cause. Cook’s distance, met in Section 17.5, measures the effect on regression outcomes of removing each case (e.g., patient) in turn so that large Cook’s distances signal outliers. Two approaches to addressing outliers are followed next.
Sample (or mean) trimming

Sample (or mean) trimming is using an established policy for removing (deleting) outliers. A sample mean lying closer to the population mean usually results from the trimming. However, rank-order methods, which are not affected by outliers and do not lose sample size, have taken precedence. The current use of sample trimming is rare.

Winsorized sample

Instead of deleting outliers, and thereby losing degrees of freedom, Charles P. Winsor devised the strategy of replacing extreme data by duplicating values of less extreme data in the sample. Winsor’s approach was to replace an outlier with the next datum closer to the mean. If several outliers must be treated, their values are replaced by the values associated with a selected percentile range, such as 5%—95% for 90% Winsorization. In this case, outliers below the 5th percentile sample value are replaced by that value, and above the 95th percentile sample value by that value. Winsorized means and standard deviations are usually closer to the corresponding population parameters than those from the original data, but, of course, the data have been manipulated. Again, rank-order methods are more commonly utilized than Winsorization.

28.15 CURVE FITTING TO DATA

The ideal fitting procedure occurs when the theoretical (e.g., physiologic) cause of data generation is known and specified mathematically. An example is the probability distribution of a sample mean, which is known to be normal (see Section 4.8). When the theoretical curve is unknown, curves may be fitted to data by various means (e.g., the spline fit).

Theoretical fits

In the theoretical fits orientation the reader need only appreciate the general shape of various theoretical distributions. They may be classified into the following six patterns, but keep in mind that a certain choice of parameters might make them appear different: (1) bell-shaped curves, typified by the normal and $t$ distributions introduced in Chapter 4, Distributions, to which we can add the logistic and Cauchy distributions; (2) skewed bell-shaped curves, typified by the chi-square and $F$ distributions introduced in Chapter 4, Distributions, to which we can add the beta, gamma, Rayleigh, and Weibull distributions; (3) curves declining with decreasing slope, tailing toward horizontal in the extreme, characterizing the negative exponential, log-normal, and Pareto distributions; (4) curves increasing with increasing slope, becoming indefinitely large in the extreme, characterizing the exponential distribution; (5) a cusp, pointing upward, characterizing the Laplace distribution; and (6) curves increasing at first, then
decreasing, often used to represent growth (of an infecting pathogen, of a cancer, of a population, or of the height of a human), characterizing the Gompertz distribution.

**Spline fit**

A *spline* curve is often fitted by cubic equation solutions for each segment in a sequence of data segments. It fits the data well but gives little insight into why it fits.

**28.16 SEQUENTIAL ANALYSIS**

In sequential analysis, analysis is conducted during data acquisition. A statistical test and resulting decision are made after acquiring each datum as to whether (A) to stop in favor of the null hypothesis, (B) to stop in favor of the alternate hypothesis, or (C) to continue sampling. Fig. 28.2 illustrates the concept. Envision moving along the horizontal axis, taking a datum, testing, and deciding (A), (B), or (C). If it is (A) or (B), you are done. If it is (C), you take another datum and repeat. In the figure, sampling is continued through the first 6 data, $H_0$ is rejected on the 7th, and sampling is terminated. This method has been available since Wald’s pioneering 1947 book but is used more frequently in engineering than medicine. Chapter 22, Clinical trials and group sequential testing, covers group sequential testing, a more commonly used version of sequential testing where tests are done after groups of observations are sampled. A more complete coverage of truly sequential analysis may be found in Chapter 24, Sequential analysis and time series, in the third edition of this book.

![Figure 28.2](image_url)

**Figure 28.2** A graph of a decision statistic plotted against $k$, the increasing sample size. As long as the statistic falls between the lines, sampling continues. If the statistic falls above the upper line, $H_0$ is rejected with risk $\alpha$ of making the wrong decision. If the statistic falls below the lower line, $H_0$ is accepted with risk $\beta$ of making the wrong decision.
AN EXAMPLE
Suppose, this method had been available to Ignaz Semmelweis in Vienna in 1847
for his experiment to show the effect on patient mortality of handwashing between
patients. He plans to randomize 1000 patients into two arms, one treated by physicians
who wash between patients and the other not, and to compare the mortality rates. After
50 patients in each arm, he finds only one death from sepsis in the wash arm, but nine
deaths in the nonwash arm. The probability that such a difference would occur by chance
is 0.033, and he concludes that handwashing prevents mortality and stops sampling. If he
had completed the full sample at these rates, 77 additional patients in the study would
have died.

Advantages of sequential analysis
The ethical benefit is obvious. In addition to mortality, terminating a study at an early
decision point may prevent unnecessary morbidity, inconvenience to patients,
increased study costs, demand on facilities, and delay in reporting the useful informa-
tion. In the long run, sample sizes from sequential sampling techniques are smaller
than those from fixed sample size techniques.

Disadvantages of sequential analysis
If the sample size generally will be smaller with sequential analysis, why is it not used
more frequently?
One issue is that the alternate hypothesis value of the decision statistic is seldom
known. Why would Semmelweis have believed that handwashing would have
reduced the mortality rate to 2% as opposed to 5% or 10%?
Another issue is that the method is demanding computationally and has not found
its way into many statistical software packages.
In addition, there is the issue of accumulating risk in repeated testing. We have
noted before that testing on the same decision statistic using repeated samples leads
to an accumulating probability of error. For example, if five tests, each with proba-
bility $\alpha = 0.05$ of a spurious rejection of $H_0$, is conducted, the probability that at
least one will be spurious becomes $1 - (0.95)^5 = 0.226$, nearly 1 in 4, no longer an
acceptable risk. Sequential sampling does not involve testing different samples in
this way, as the same sample augmented by new observations is used, but the prob-
lem persists. To solve this problem, Wald developed the test using a theoretical
basis different from the least squares we have seen on many tests. Since $k$ could be
expressed as depending on the error risks, the solution was developed for a general
$k$, whatever value it might take on. Thus the sequential test is based on a fixed $\alpha$
and $\beta$ for any $k$. 

CHAPTER 28 Methods you might meet, but not every day
Finally, there is the possibility that an atypical sequence will occur. A complication $C$ on a form of surgery occurs with some probability $P$. The occurrence of no complication $N$ is, of course, $1 - P$. For four patients, we might find a sequence such as $NCNN$ or by chance the sequence $CCCC$. We could just by luck fall into a decision region by a sequence of patients with more Cs (or more Ns) than represents typical and thereby make an erroneous conclusion.

Nevertheless, in the circumstance of advantages exceeding disadvantages, sequential analysis has much to offer.

### 28.17 ANOTHER TEST OF NORMALITY

The Lilliefors test of normality\textsuperscript{11} may be added to the Shapiro–Wilk, one-sample Kolmogorov–Smirnov, and chi-square goodness–of–fit tests addressed in Section 13.6. The Lilliefors test is a variation of the Kolmogorov–Smirnov test, having a more incisive outcome but much more difficult to work with.

### 28.18 DATA MINING

Until the middle of the 20th century, a major problem in scientific research was obtaining a sufficient amount of data to reach a valid conclusion. The science of statistics concentrated on drawing maximum information from sparse databases. Since that time scientists in many fields have developed data sources that produce prodigious amounts of data, and considerable attention has shifted to large sample summary and analysis and modes of valid selection of analyzable database subsets. By the start of the 21st century, many databases had become so voluminous that often information was obscured by the very magnitude of data, or “buried,” so completely that scientists could not even conjecture about relationships enough to form testable hypotheses. The “mining” of such information buried in data began.

To mine information from massive data sets, the data are first put into a format acceptable to the processing software and the database is adjusted for corrupted and missing data. The database is divided into a search portion and a validation portion. Search methods include clustering and association methods, multiple regression, and classification procedures (decision trees, discriminant analysis, neural networks, nearest neighbor methods, etc.). Computer technology is used to optimize search methods and minimize search time and effort. After information is obtained from searching, it is tested by analyzing its presence and relationships in the validation subdatabase to reduce false-positive errors.

Partially automated data-mining programs are beginning to appear in statistical software packages.
28.19 DATA SCIENCE AND THE RELATIONSHIP AMONG STATISTICS, MACHINE LEARNING, AND ARTIFICIAL INTELLIGENCE

Data science is a term that is increasingly used in society. The foundations of science rest on the theory and application of statistics and computer science. These disciplines form the fundamental tool that data scientists use to collect, organize, maintain, analyze, and interpret empirical data in the quest to answer scientific questions. Falling under the umbrella of data science are artificial intelligence (AI) and machine learning. Unfortunately, there is much confusion regarding these two terms and their relationship with statistics. Some seem to think all three are basically the same thing while others think they are completely separate. Neither is true.

As can be gleaned from this book, statistics is the scientific discipline primarily aimed at using empirical data to estimate the association between covariates and to draw inference regarding associations using a sample of data. Of course, this also incorporates many other aspects as we have discussed in this book. These include study design, the quantification of uncertainty or variability, and the ability to predict new observations from a statistical model, to name a few.

Machine learning is a discipline devoted to learning from empirical data to predict future observations. Given this, it should be clear that much of the methodology utilized in machine learning either stems directly from statistics or has its roots in statistics. Machine learning, however, is not generally concerned with population inference regarding specific associations. As such, interpretability of prediction models is often of little or no concern in the field. Indeed, the primary goal is generally to find a suitable statistical model or algorithm to generate predicted values that are close to those that would be observed in new settings. This is typically done by training a model (or estimating the parameters within the model) on one set of data, validating the model on another set of data to obtain a good estimate of the true predictive performance of the model, then testing the model on a new independent set of data to best evaluate the model in the setting it is meant to be used in. While statistical models can also form predictions, the field of machine learning is interested in increased computational speed and prediction performance, and hence the field increasingly relies on algorithmic procedures.

AI is the use of machines to make intelligent decisions. An example of the use of AI that is becoming more common is self-driving automobiles. In this case, visual and spatial sensors are placed all around a vehicle. These sensors collect vast amounts of data in real time, and these data are then analyzed to make (hopefully) rational decisions with respect to the speed and direction of the vehicle. Another example of AI is the use of computers to identify and classify colon polyps in real time via video obtained during a colonoscopy. Thus AI has a very broad definition.
So why does AI often get confused with machine learning and statistics? The answer is that AI relies heavily on machine learning, since predictive models must generally be used to classify objects and make decisions. For example, a self-driving automobile must be able to distinguish a manhole cover in the road from an animal, using data from the visual sensors. This is a prediction problem. Using that prediction, the vehicle must then decide what action should be taken (e.g., stopping, slowing, evading, or continuing as usual). Thus machine learning is a key tool in the development of AI systems. Further, as machine learning techniques rely heavily on the foundations and methods of statistics, statistics is also fundamental to most AI processes.

REFERENCES

Answers to exercises: Final

CHAPTER 1

Exercise 1.1 There is no unique answer for this exercise.

Exercise 1.2 (a) Randomized controlled trial. Drug versus placebo is the independent variable. Nausea scores are the dependent variables. (b) Cohort study. EIB is the independent variable. eNO differences are the dependent variables.

Exercise 1.3 No. Among the reasons why not are the following. Rather than being a prospective study, the sample is a “convenience sample,” that is, drawn from patients presenting already tattooed and who have decided on removal; there is no control group for comparison; and the investigator is not masked from the decision about “response” to the treatment.

Exercise 1.4 (a) Yes. Nothing can be done; age and sex were not recorded. (b) Yes. Age and sex can be tested for bias.

Exercise 1.5 One-sided. Although the postoperative minus preoperative difference could be either positive or negative, adding the silicone to the body is believed to be unable to decrease the plasma silicone level; thus any decrease would be an anomalous chance outcome. Furthermore, it would not be harmful to the patient. The test should be to detect an increase.

CHAPTER 2

Exercise 2.1 (a) Name, (b) name, (c) operator, (d) operator, (e) relationship, (f) indicator.

Exercise 2.2 Mean: accuracy; SD: precision.

Exercise 2.3 (a) Rank-order data, (b) categorical data, (c) continuous data.

Exercise 2.4 (a) Mean, (b) SD, (c) rate, (d) median, (e) Pearson correlation, (f) pie chart.

Exercise 2.5 Independent variable: EIB, two samples, for all answers. (a) Dependent variable: Continuous. Table position 1,4: t test. (b) Dependent variable: two categories. Table position 1,1: FET. (c) Dependent variable: Continuous. Table position 1,4: t test. However—too advanced for readers who have not read later chapters—the 5-minute difference has outliers in the non-EIB distribution and the EIB distribution is
asymmetric, violating \( t \) test assumptions. Go to rank methods in table position 1,3: the rank-sum test.

CHAPTER 3

Exercise 3.1 (a) By Eq. (3.1), \( P(M \text{ or } S) = P(M) + P(S) = \frac{5}{25} + \frac{10}{25} = 0.6 \). (b) By Eq. (3.3), \( P(\text{not} Y \text{ and not} Y) = P[(M \text{ or } S) \text{ and } (M \text{ or } S)] = P(M \text{ or } S) \times P(M \text{ or } S) = 0.6^2 = 0.36 \).

Exercise 3.2 (a) By Eq. (3.2), \( P(H_b \text{ or } C) = P(H_b) + P(C) - P(H_b \text{ and } C) = \frac{10}{25} + \frac{10}{25} - \frac{6}{25} = 0.56 \). (b) By Eq. 3.4, \( P(H_b \text{ and } C) = P(H_b) \times P(C|H_b) = \left(\frac{10}{25}\right) \times \left(\frac{6}{10}\right) = 0.24 \).

Exercise 3.3 From Eq. (3.5), \( P(H_b|J) = P(H_b) \times P(J|H_b)/P(J) = \left(\frac{10}{25}\right) \times \left(\frac{0.90}{12/25}\right) = 0.75 \).

Exercise 3.4 By Eq. (3.8), the combination of 5 things 2 at a time is \( \frac{5(5 - 1)}{2} = 10 \) possible pairings.

Exercise 3.5 The relative frequency, or rate, of cirrhosis of our sample is \( \frac{2}{6} = 0.333 \ldots \). The probability estimated by the relative frequency is \( \frac{10}{25} = 0.4 \). As we sampled more and more charts, we would expect the rate to move closer and closer to 0.4.

Exercise 3.6 (a) \( 1 < 1.5, 1.5 < 2, 2 < 2.5, \ldots, 5 < 5.5, 5.5 < 6 \). (b) Tally can be verified by comparing frequencies with plot in (d). (c) Median = 2.5. (d)
(e) Distribution is skewed to the right and does not grow negligible in the tails, but it is not dramatically abnormal.

(f) Mean appears to be about 3. Mode is 2.25. Mode < median < mean, as expected in a right-skewed distribution.

**Exercise 3.7**


(b) Tally can be verified by comparing frequencies with plot in (d).

(c) Yes, median of accumulating data converges to final median of 51.

(e) Distribution is a little skewed to the left, a little too “bulky” in the center, but is not dramatically abnormal.
(f) The mean appears to lie in the 46–50 intervals. The mode is 62.5. The mean < median < mode, as expected in a left-skewed distribution.

Exercise 3.8  
(a) 1 − < 1.25, 1.25 − < 1.50, . . . , 4 − 4.25.  
(b) Tally can be verified by comparing frequencies with plot in (d).  
(c) Median = 2.38.  
(d)  

(e) Mean of 2.40 is almost the same as median. Although this does not imply symmetry in all respects, as can be seen from inspection of the distribution in (d), it does imply that the distribution is not badly skewed.

Exercise 3.9  
(a) 1 − < 1.25, 1.25 − < 1.50, . . . , 3.50 − 3.75.  
(b) Tally can be verified by comparing frequencies with plot in (d).  
(c) Median = 2.15.  
(d)
(e) Mean of 2.28 is somewhat to right of median, implying that the
data extend farther to the right of the median than to the left, indicat-
ing a slight skew.

Exercise 3.10 Chi-square. The reason is because chi-square is based on squared data.
Thus variances, for example, are distributed as chi-square.

Exercise 3.11 The distribution is more similar to a normal.

CHAPTER 4

Exercise 4.1 (a) 0.4. (b) 0.35.
Exercise 4.2 (a) 0.425. (b) 0.45.
Exercise 4.3 (a) 0.0467. (b) 0.2160.
Exercise 4.4 (a) 0.0092. (b) 0.0959.

Exercise 4.5 The sample estimate we would use is the mean (averaged over patients)
of changes in serum theophylline levels from baseline to 5 days. We
assume this mean is drawn from a normal distribution. We choose the
upper bound for the probability of a false positive (concluding there is
a real change when there is not), usually (but not always) 0.05. We use
the \( t \) probability table (for this small sample of size 16) to establish an
interval about the theoretical mean of 0 (no difference) within which
there is probably no difference. We reject the hypothesis of no differ-
ence if the calculated \( t \) falls outside this interval or conclude that no dif-
ference was shown if \( t \) falls in the interval.

Exercise 4.6 The sample estimate we would use is the mean (averaged over patients)
of changes in eNO from before exercise to 20 minutes after. We
assume this mean is drawn from a normal distribution. We choose the
upper bound for the probability of a false positive (concluding there is
a real change when there is not), usually (but not always) 0.05. We use
the \( t \) probability table (for this small sample of size 39) to establish an
interval about the theoretical mean of 0 (no change) within which
there is probably no difference. We reject the hypothesis of no change
if the calculated \( t \) falls outside this interval or conclude that no change
was shown if \( t \) falls in the interval.

Exercise 4.7 \( df = n - 1 = 29 \). Subtract 0, the theoretical mean difference if pre-op and
post-op arise from identical distributions; divide by the standard devia-
tion of the sample mean (SEM). The cut point is named the \textit{critical
value}.

Exercise 4.8 \( df = n - 1 = 103 \).

Exercise 4.9 Proportion is not close to 0: Binomial.
Exercise 4.10 Proportion is very close to 0: Poisson.
Exercise 4.11 Covariance.
Exercise 4.12 SEM=0.2160/√4=0.1080.
Exercise 4.13 SEM=0.2868/√8=0.1014.

CHAPTER 5

Exercise 5.1 (a) 154.2, (b) 154.1, (c) 576.4, (d) 24.0, (e) 140.1, (f) 164.9, (g) 5.7
Exercise 5.2 (a) 452.75, (b) 462.5, (c) 8416.2140, (d) 91.7399, (e) 372.5, (f) 532.0, (g) 32.4350

Exercise 5.3

<table>
<thead>
<tr>
<th>Treatment</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(f)</th>
<th>(g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.97</td>
<td>31.75</td>
<td>16.77</td>
<td>4.10</td>
<td>28.68</td>
<td>33.20</td>
<td>0.84</td>
</tr>
<tr>
<td>2</td>
<td>15.06</td>
<td>9.20</td>
<td>165.25</td>
<td>12.86</td>
<td>5.40</td>
<td>24.75</td>
<td>2.63</td>
</tr>
</tbody>
</table>

Exercise 5.4 62
Exercise 5.5 39.04
Exercise 5.6 (a) 5539.75, (b) 0.9823
Exercise 5.7 (a) 0.20007, (b) 0.8398
Exercise 5.8 0.963
Exercise 5.9 0.976
Exercise 5.10 $C=(8 \times 30/5 \times 7)^{π/4}=4.5363$. $r_I=3.5363/5.5363=0.64$.
Exercise 5.11
Exercise 5.12

Exercise 5.13  (a)

(b)
Exercise 5.14

### Frequency Distribution of 20-Min Change in eNO

- Frequency on the y-axis
- 20-Min change in eNO on the x-axis
- Bars represent the distribution of change values

### Pie Chart

- Age 1: 44%
- Age 2: 33%
- Age 3: 11%
- Age 4: 11%
Exercise 5.15

Exercise 5.16
APPENDIX 1 Answers to exercises: Final

Exercise 5.17

Exercise 5.18
Exercise 5.19 (a)

Exercise 5.19 (b)
Exercise 5.20  (a)

[Comment on 5.20: Note how the one patient with outlying values distorts the means and magnifies the standard deviations so that 5.20(a) gives little of the information about the data pattern seen in 5.20(b) and renders the result misleading.]
Exercise 5.21

Exercise 5.22
CHAPTER 6

Exercise 6.1  \( z = \frac{(126 - 120)}{5} = 1.20 \). Your patient is 1.2 \( \sigma \) above the mean. From Table 13.1, the area in the right tail is 0.115. Therefore 11.5% of the population has greater DBP.

Exercise 6.2  For 7 \( df \), a two-tailed \( \alpha = 0.05 \) is obtained by \( m \pm 2.365 \times s \), or the interval from 235.9 to 669.7.

Exercise 6.3  \( \chi^2 = 17 \times 576.48/259.85 = 37.71 \). From the row for 17 \( df \) in Table 13.3, we see that a \( \chi^2 \) value of 37.71 lies between 35.72, associated with an area under the curve in the right tail of 0.5%, and 40.78, associated with an area of 0.1%. By interpolation, which is approximate, the area associated with 37.71 will be about 0.4 of the way from 0.5% to 0.1%, or about 0.3%. If we sampled standard deviations repeatedly from a normal patient population, we would find a standard deviation of 24.01 or larger only about three times out of 1000.

Exercise 6.4  For samples so small, numerator and denominator \( df \)s both 3, the \( F \) ratio as shown in Table 13.4 would have to be larger than 9.28. For these types of assay, \( F = 0.0586/0.0120 = 4.88 \). As 4.88 < 9.28, we do not have sufficient evidence to conclude a difference in variability.

Exercise 6.5  In Table 6.5, which displays the \( n = 6 \) block of Table VI, the probability 0.114 lies at the intersection of the column \( \pi = 0.10 \) and the row \( n_o = 2 \). The probability that there would be two patients with unhealed ulcers is more than 11%.

Exercise 6.6  \( \pi = 0.01 \) and \( n = 70 \); therefore the expected number of AIDS cases is \( \lambda = 0.7 \). \( n_o = 5 \) cases occurred. In Table 13.6 the intersection of the column \( \lambda = 0.7 \) and \( n_o = 5 \) shows the probability to be 0.001.
CHAPTER 7

Exercise 7.1 It is a clinical hypothesis, stating what the investigator suspects is happening. A statistical hypothesis would be the following: Protease inhibitors do not change the rate of pulmonary admissions. By stating no difference, the theoretical probability distribution can be used in the test. (If there is a difference, the amount of difference is unknown, so that the associated distribution is unknown.)

Exercise 7.2 $H_0: \mu_w = \mu_{w/o}$; $H_1: \mu_w \neq \mu_{w/o}$.

Exercise 7.3 (a) The $t$ distribution is associated with small data sample hypotheses about means. (b) Assumptions include: The data are independent one from another. The data samples are drawn from normal populations. The standard deviations at baseline and at 5 days are equal. (c) The type I error would be concluding that the baseline mean and the 5-day mean are different when in fact they are not. The Type II Error would be concluding that the two means are the same when in fact they are different. (d) The a priori upper bound placed on the risk of a type I error is designated $\alpha$. The risk of a Type II Error for an assumed true difference in means is designated $\beta$.

Exercise 7.4 (a) The $\chi^2$ distribution is associated with a hypothesis about the variance. (b) Assumptions include: The data are independent one from another. The data sample is drawn from a normal population. (c) The type I error would be concluding the platelet standard deviation is different from 60,000 when in fact it is 60,000. The Type II Error would be concluding that the platelet standard deviation is 60,000 when in fact it is not. (d) The a priori upper bound placed on the risk of a type I error is designated $\alpha$. The risk of a Type II Error for an assumed true value of the platelet standard deviation (different from 60,000) is designated $\beta$.

Exercise 7.5 (a) The $F$ distribution is associated with a hypothesis about the ratio of two variances. (b) Assumptions include: The data are independent one from another. The data samples are drawn from normal populations. (c) The type I error would be concluding that the variance (or standard deviation) of serum silicon before the implant removal is different from the variance (or standard deviation) after when in fact they are the same. The Type II Error would be concluding that the before and after variances (or standard deviations) are the same when in fact they are different. (d) The a priori upper bound placed on the risk of a type I error is designated $\alpha$. The risk of a type II error for an assumed difference in variances is designated $\beta$. 
Exercise 7.6  The “above versus below” view would interpret this decrease as not significant, end of story. Exercise has no effect on the eNO of normal subjects. The “level of p” view would say that, there is an estimated 2.15 ppb decrease in eNO associated with exercise and the probability of observing an estimate equal to or higher than this if there were no association between smoking and eNO is about 8%. The magnitude of the estimated effect should be interpreted clinically and the precision of the estimate should be evaluated. Perhaps the effect of exercise on eNO in normal subjects should be further investigated.

Exercise 7.7  (a) A categorical variable. (b) A rating, which might be any type, depending on circumstance. In this case, treating it as a ranked variable is recommended, as it avoids the weaker methods of categorical variables and a five-choice is rather small for use as a continuous variable. (c) A continuous variable.

Exercise 7.8  1, 6, 4, 7, 5, 8, 2, 3.

Exercise 7.9  (a)

<table>
<thead>
<tr>
<th>Sex</th>
<th>1 (male)</th>
<th>2 (female)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIB</td>
<td>0</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Totals</td>
<td>28</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>5-min difference</th>
<th>5-min ranks</th>
<th>Sex</th>
<th>Ranks for males (1)</th>
<th>Ranks for females (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>-1.0</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>-3.9</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2.0</td>
<td>2.5</td>
<td>1</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>-2.0</td>
<td>2.5</td>
<td>1</td>
<td>2.5</td>
<td></td>
</tr>
</tbody>
</table>

(c)

<table>
<thead>
<tr>
<th>5-min difference</th>
<th>Sex</th>
<th>Difference for males (1)</th>
<th>Difference for females (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>-1.0</td>
<td>1</td>
<td>-1.0</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>1</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>-3.9</td>
<td>2</td>
<td></td>
<td>-3.9</td>
</tr>
<tr>
<td>-2.0</td>
<td>1</td>
<td>-2.0</td>
<td></td>
</tr>
<tr>
<td>-2.0</td>
<td>1</td>
<td>-2.0</td>
<td>m = -3.9, no s</td>
</tr>
</tbody>
</table>

\(m = 1.62, \ s = 4.54\)
Exercise 7.10  (1) Does the drug reduce nausea score following gall bladder removal? (2 and 3) Drug/no drug against nausea/no nausea. (4) $H_0$: nausea score is independent of drug use; $H_1$: nausea score is influenced by drug use. (5) The population of people having laparoscopic gall bladder removals who are treated for nausea with Zofran. The population of people having laparoscopic gall bladder removals who are treated for nausea with a placebo. (6) My samples of treated and untreated patients are randomly selected from patients who present for laparoscopic gall bladder removal. (7) A search of the literature failed to indicate any proclivity to nausea by particular subpopulations. (8) These steps seem to be consistent. (9) A chi-square test of the contingency table is appropriate; let us use $\alpha=0.05$. (10) (Methodology for step 10 has not been addressed yet; the student who wishes to pursue this can see Chapter 21: Sample size estimation.)

Exercise 7.11  (1) Are our clinic’s INR readings different from the laboratory’s? (2) Difference between clinic and laboratory readings. (3) Mean of difference. (4) $H_0$: mean difference=0; $H_1$: mean difference$\neq 0$. (5) Population: All patients, past and future, are subject to the current INR evaluation methods in our Coumadin clinic. (6) Sample: The 104 consecutive patients taken in this collection. (7) Biases: The readings in this time period might not be representative. We can examine records to search for any cause for nonrepresentativeness. Also, we could test a small sample from a different time and test it for equivalence. (8) Recycle: These steps seem consistent. (9) Statistical test and $\alpha$: Paired $t$ test of mean difference against 0 with $\alpha=0.05$.

CHAPTER 8

Exercise 8.1  Mean is 2, standard deviation 3. $z=(x - 2)/3$.

Exercise 8.2  $2 \pm 1.96 \times 3$, or $-3.88, 7.88$.

Exercise 8.3  The 95% tolerance limits on individuals are 18.8–76.8. Your patient’s age lies 2.77 standard deviations below the mean; only 0.28% of patients would be younger (using Table I). Your patient is improbably young.

Exercise 8.4  The tolerance interval on an individual patient would be $1.96 \times \sigma$ or 1.23 units below and above the mean of 2.28 U. Thus we would expect that 95% of patients would lie in the interval 1.05–3.51 INR units. This data-based tolerance interval is much larger than the in range interval.
Exercise 8.5 To substitute in Eq. (8.5), the \( z \) values must be found from Table 8.1 (or Table I). Under the column for “Two-tailed \( 1 - \alpha \)”, 0.900 (90\%) yields \( z = 1.645 \) and 0.990 (99\%) yields 2.576. Substituting \( m = 47.84 \) and \( \sigma_m = 1.36 \), we find the 90\% CI to be \((47.84 - 1.645 \times 1.36, 47.84 + 1.645 \times 1.36) = (45.60, 50.08)\) and the 99\% CI to be \((47.84 - 2.576 \times 1.36, 47.84 + 2.576 \times 1.36) = (44.34, 51.34)\).

Exercise 8.6 Using Table I, (a) \((8.59, 9.73)\); (b) \((8.42, 9.90)\); (c) \((8.69, 9.63)\).

Exercise 8.7 Confidence limits for mean INR, based on 104 patients, are \(2 = 28.61\), \(1 = 9.63\), \(3 = 63\) leading to the interval \((2.16, 2.40)\). The confidence interval on the mean is much smaller than the individual patient in range interval, as would be expected, but, interestingly, lies close to the lower In range limit rather than lying in the center.

Exercise 8.8 The surgeon calculates \( s_m = s/\sqrt{n} = 4.57/\sqrt{20} = 1.02\). From Table II, \( t \) for 95\% “two-tailed \( 1 - \alpha \) (except both tails)” for 19 \( df \) is 2.093. The surgeon substitutes these values in Eq. (8.9) to obtain a 95\% CI of \((6.56 - 2.093 \times 1.02, 6.56 + 2.093 \times 1.02) = (4.43, 8.69)\). Plausible values for the true mean time to full feeding lie between 4.4 and 8.7 hours.

Exercise 8.9 The mean difference is \( m = 0.0073 \) with standard deviation \( s = 0.1222 \). The \( SEM = 0.1222/\sqrt{30} = 0.0223 \). In Table II the intersection of the column for two-tailed \( 1 - \alpha \) = 0.95 with the row for \( n - 1 = 29 \) \( df \) gives \( t_{1 - \alpha/2} = 2.045 \), which tells us that the critical values will be the sample mean \pm \) a bit over two SEMs. After substitution in Eq. (8.9), a 95\% CI is given by \((0.0073 - 2.045 \times 0.0223, 0.0073 + 2.045 \times 0.0223) = (-0.038, 0.053)\).

Exercise 8.10 Using Table II with 7 \( df \), \((60.19, 75.31)\).

Exercise 8.11 Confidence limits for treatment 1 are \(34.40 \pm 2.201 \times 3.25/\sqrt{12} \) leading to a CI of \((32.34, 36.46)\), and for treatment 2 are \(36.12 \pm 2.201 \times 1.17/\sqrt{12} \) leading to a CI of \((35.38, 36.86)\). The confidence interval on the mean for treatment 2 is notably smaller.

Exercise 8.12 \( p \) is not near 0 or 1, so \( \sigma = \sqrt{p(1-p)/n} = \sqrt{0.261 \times 0.739/226} = 0.0292 \). Substitute in Eq. (8.10) to obtain a 95\% CI given by \((0.261 - 1.96 \times 0.029 - 1/452, 0.261 + 1.96 \times 0.029 + 1/452) = (0.202, 0.320)\).

Exercise 8.13 \( s^2 = 36,000,000 \) and \( df = 18 \). From Table 8.3 the intersection of the 0.025 column with 18 \( df \) is 31.53 for the right tail and 8.23 for the left tail. Substitution in Eq. (8.12) yields a 95\% CI for \( \sigma^2 \) of \((20,551.855, 78,736,330)\) and taking square roots of the limits yields a 95\% CI for \( \sigma \) of \((4533, 8873)\)=. Because \( \sigma \) for a healthy population would be about 1200–1300, the variability does indeed appear to be abnormally large.
Exercise 8.14 $s^2 = 19,039.7^2 = 362,510,180$. $df = n - 1 = 19$. By looking in Table 8.3 under right tail area=0.025 for 19 $df$, we find 32.85. Similarly, under left tail area=0.025 for 19 $df$, we find 8.91. By substituting these values in Eq. (8.12), we find the interval given by

$$(s^2 / \chi^2_{R}, s^2 / \chi^2_{L}) = (362, 510, 180 \times 19 / 32.85, 362, 510, 180 \times 19 / 8.91)$$

$$= (209, 671, 030, 773, 029, 560).$$

Taking square roots of both limits we obtain a 95% CI for $\sigma$ given by (14,480.0, 27,803.4).

Exercise 8.15 Using Table III in calculating the lower bound and Table IV, the upper bound, with 19 $df$, (a) (12.08, 44.54) and (b) (3.48, 6.67).

Exercise 8.16 Using Table III in calculating the lower bound and Table IV, the upper bound, with 7 $df$, (a) (35.73, 338.49) and (b) (5.98, 18.40).

Exercise 8.17 For treatment 1, we have a CI for the variance of $(10.56 \times 11 / 21.92, 10.56 \times 11 / 3.82) = (5.30, 30.41)$, leading to a 95% CI for the standard deviation of (2.30, 5.51). For treatment 2, we have a CI for the variance of $(1.37 \times 11 / 21.92, 1.37 \times 11 / 3.82) = (0.69, 3.95)$, leading to a 95% CI for the standard deviation of (0.83, 1.99). The confidence interval on the standard deviation for treatment 2 is notably smaller.

Exercise 8.18 $1 + r = 1.267$, $1 - r = 0.733$, and the exponential term is 0.5718. Substitution in Eq. (8.15) yields $P[-0.0123 < \rho < 0.5076] = 0.95$. The confidence interval crosses zero, indicating that the population correlation coefficient could plausibly be zero or negative and the 0.267 sample coefficient could have occurred by chance.

CHAPTER 9

Exercise 9.1 Critical value is 3.84. $\chi^2 = 5.44$ that is greater than 3.84. Reject $H_0$ at level 0.05 and conclude that protease inhibitors are effective. (Actual $p$-value=0.020.)

Exercise 9.2 Critical value is 3.84. $\chi^2 = 8.32$ that is greater than 3.84. Reject $H_0$ at level 0.05 and conclude that titanium-containing ink is harder to remove. (Actual $p$-value=0.0039.)
Exercise 9.3 Critical value is 3.84. \( \chi^2 = 9.32 \) that is greater than 3.84. Reject \( H_0 \) at level 0.05 and conclude that nausea score is reduced by the drug. (Actual \( p \)-value = 0.0023.)

Exercise 9.4 The \( 2 \times 2 \) table is:

<table>
<thead>
<tr>
<th></th>
<th>Out of range</th>
<th>In range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth (Laboratory)</td>
<td>30</td>
<td>27</td>
</tr>
<tr>
<td>Out of range</td>
<td>3</td>
<td>44</td>
</tr>
<tr>
<td>In range</td>
<td>33</td>
<td>71</td>
</tr>
</tbody>
</table>

From Eq. (6.2), \( \chi^2 = 25.44 \), which is greater than 3.84, the 5% critical value of \( \chi^2 \). Clinic results were discrepant from laboratory results.

Exercise 9.5 The FET’s \( p \)-value = 0.041, indicating a significant difference between the success of the two treatments with a risk of error less than 5%; MB is better than CS. \( \chi^2 = 5.132 \), yielding \( p \)-value = 0.023, also showing a significant difference. A cell entry of less than 5 indicates that the chi-square approximation may not be adequate. FET is the appropriate test.

Exercise 9.6 The FET’s \( p \)-value = 0.085, not significant. To calculate chi-square, \( e_{11} = 19 \times 100/195 = 9.7436 \), etc. \( \chi^2 = (8 - 9.7436)^2/9.7436 + \ldots = 4.92 \). \( df = 2 \). From Table III, 4.92 for 2 \( df \) lies between 0.05 and 0.10. A computer evaluates the \( p \)-value as 0.085, the same as that for FET. As \( p \)-value > 0.05, there is insufficient evidence to reject the null hypothesis and conclude a difference in systolic and diastolic behavior. [However, the rank-sum test (see Chapter 11: Tests of location with continuous outcomes) yields \( p \)-value = 0.033, significant, leading to the reverse conclusion.]

Exercise 9.7 \( n = 16 \), \( n_b = 11 \), and \( \pi = 0.50 \). From Table VI the probability of 11 or more of 16 occurring if \( \pi = 0.50 \) is 0.105.

Exercise 9.8 If \( \pi_b \) is the true population proportion of boys presenting, \( H_0: \pi_b = \pi \) and \( H_1: \pi_b \neq \pi \). Then, \( p_b = 0.68 \). Using the normal approximation, \( \sigma = 0.0707 \). From Eq. (9.5), \( z = (|p_b - \pi|)/\sigma = 0.18/0.0707 = 2.55 \). From Table I, \( z = 2.55 \) yields two-tailed \( p \)-value = 0.010. We reject \( H_0 \) and conclude that there is an association between gender and limb fractures.

Exercise 9.9 We judge that introducing better testing can only reduce the rate; therefore the test is one-sided. From Table I a critical value of \( z \) at \( \alpha = 0.05 \) will be \( -1.645 \). We note that \( \lambda = n \pi = 5 \), so we may use Table VII. We want the probability of three or fewer cases [\( P(3 \ or \ fewer) \)], which is \( 1 - P(4 \ or \ more) \), occurring by chance alone. Using \( n_0 = 4 \), \( P(3 \ or \ fewer) = 1 - 0.735 = 0.265 \). There is inadequate evidence to
infer an improvement due to the additional screening tests. For illustration, let us calculate the normal approximation, although we know the result will be erroneously different from the result of Table VII because of the small $\lambda$. From Eq. (9.10),

$$z = (p_s - \pi) \sqrt{\frac{n}{\pi}} = (0.000200 - 0.000333) \times \sqrt{\frac{15,000}{0.000333}} = -0.893,$$

which is less distant from zero than the critical value; therefore there is inadequate evidence to infer an improvement. From a statistical software package, $p$-value=0.186.

**Exercise 9.10** The data tally appears in Table 9.22. $\chi^2_{1 df} = (|b-c| - 1)^2/(b+c) = (4 - 1)^2/8=1.125$. From Table III the critical value of $\chi^2$ for $\alpha=0.05$ with 1 $df$ is 3.84. Because $1.125 < 3.84$, you conclude that there is insufficient evidence that boys have a greater accident rate. The actual $p$-value is 0.289.

<table>
<thead>
<tr>
<th>Female member of pair had accident(s)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male member of pair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Had accident(s)</td>
<td>Yes</td>
<td>$a=2$</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>$c=2$</td>
</tr>
</tbody>
</table>

**Exercise 9.11** $c = 3$, $r = 10$, $n = 16$, $SS_r = 28$, and $SS_c = 90$. $Q = (3 \times 2 \times 90 - 2 \times 256)/(2 \times 16 - 28) = 28/4 = 7$, larger than the critical 5.99, so we have evidence of a discrepancy among abnormal cholesterol rates. Using computer software, a $\chi^2$ value of 7 with 2 $df$ gives $p$-value=0.030. We note that this result does not agree with the anticipation of lower cholesterol with fasting. In the true data the cholesterol after fasting showed more cases exceeding normal than after eating. Using McNemar’s test on the true data gives $\chi^2=0$, no evidence of difference. The significance resulting from Cochran’s $Q$ was due to the fabricated data.

**Exercise 9.12** The correlation coefficient is 0.369. $\chi^2 = 36 \times 0.369^2 = 4.91$, less than the 5% critical value 3.84. A computer calculates $p$-value=0.027. We conclude that the additional information is associated patients’ decision on the surgery. The FET $p$-value=0.109. FET tests the information options as unordered categories, whereas $ptrend$ tests them as increasing ranks. The ranking adds additional information to the test.
CHAPTER 10

**Exercise 10.1** Sensitivity=0.981, the chance of detecting appendicitis using MRI. Specificity=0.979, the chance of correctly diagnosing nonappendicitis. Accuracy=0.980, the overall chance of making a correct diagnosis. PPV=0.981, the chance that a diagnosis of appendicitis is correct. NPV=0.979, the chance that a diagnosis of nonappendicitis is correct. That these values are the same as sensitivity and specificity is an unusual coincidence due to the equality of the marginal totals. RR=46.11, the chance of appendicitis when predicted in ratio the chance of appendicitis when not predicted. OR=2392, the odds of correctly diagnosing appendicitis in ratio the odds of incorrectly diagnosing appendicitis. LR=46.11, the chance of correctly diagnosing appendicitis in ratio to the chance of incorrectly diagnosing appendicitis, which is coincidentally the same as RR due to identical marginal totals. AR=0.96, the proportion of appendicitis cases discovered by the clinical diagnostic procedure. Some of these figures, lacking in clinical meaning or usefulness, were calculated for the exercise.

**Exercise 10.2** LRR=1.1527, se(LRR)=0.5751, $\chi^2=4.0174$, $p$-value for testing $H_0: \theta = 0$ is 0.045; Reject $H_0$ at level 0.05 and conclude that there is a difference between the Modified Bostrom and Chrisman–Snook methods of repair.

**Exercise 10.3** LOR=1.5506; se(LOR)=0.7327; $\chi^2=(1.5506/0.7327)^2=4.4787$. From Table III the $p$-value for this result is between 0.025 and 0.05, a significant result. The actual $p$-value is 0.034, which compares to the 0.041 FET result in Exercise 15.1. We have shown that quality of result is significantly associated with method of ligament repair.

**Exercise 10.4** AR=0.266, telling us that the disease incidence is increased more than 25% by exposure to putatively contaminated dust. $r_t=0.575$, a substantial level of correlation between suspected dust and disease occurrence.

**Exercise 10.5** LOR=1.6685, se(LOR)=0.3271. 95% CI for OR is (2.79, 10.07).

**Exercise 10.6** Clearly a high value of sensitivity leads to a very low specificity. In the vicinity where the errors of misclassification are not very different, the point most distant from the 45 degrees line of indifference shows sensitivity=71% and specificity=50%. This leads to a WBC of 11.5.

**Exercise 10.7** Duration of snoring lies along the line of no discrimination, indeed falling a bit below (AUC < 0.50), while loudness provides a potentially useful discriminator. The difference is highly significant. We need not differentiate patients’ BMI when using duration but must separate patients for BMI when using loudness.
Exercise 11.1 Substitution in Eq. (11.2) (with renamed elements) yields \( t = (d - 0)/s_d = 1.58/(1.23/\sqrt{10}) = 4.06 \). From Table II the critical value for a 5% two-tailed \( t \) with 9 \( df \) is 2.26. As 4.06 is much larger, we reject \( H_0 \) and conclude that the mean of Laevo is higher. A statistical software package gives the \( p \)-value as 0.003.

Exercise 11.2 (Equal variances) Using Table 11.3, we note that the sample sizes are small, \( \sigma \)'s are only estimated, and the standard deviations are not significantly different (refer to Exercise 13.3). Here we show the two-sample \( t \) test with equal standard deviations result. \( df=n_1+n_2-2=14 \). From Eq. (11.6),

\[
s_d = \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right) \left[ \frac{(m_1 - 1)s_1^2 + (m_1 - 1)s_2^2}{n_1 + n_2 - 2} \right]} = \sqrt{\left( \frac{1}{8} + \frac{1}{8} \right) \left[ \frac{7 \times 0.88^2 + 7 \times 1.12^2}{8 + 8 - 2} \right]} = 0.5036.
\]

From Eq. (11.5),

\[
t = \frac{m_1 - m_2}{s_d} = \frac{96.41 - 97.86}{0.5036} = -2.88.
\]

From Table II the \( p \)-value for \( t=2.88 \) standard deviations from the mean for 14 \( df \) is between 0.01 and 0.02. We have evidence to infer a difference in mean readings between ears. From statistical software, \( p=0.012 \).

Exercise 11.3 (a) (Subscript \( s \) for survivors, \( d \) for died.) \( n_s=117, \ n_d=17, \ m_s=48.05, \ m_d=46.41, \ \sigma=15.78, \ \text{SEM}=4.10, \ z=0.4. \ z < 1.96, \) the normal’s critical value for two-tailed \( \alpha=0.05 \). (Actual \( p \)-value=0.345.) The age difference between patients who survived and who died is not significant in probability.

(b)\( n_s=117, \ n_d=17, \ m_s=2.82, \ m_d=3.96, \ \sigma=1.54, \ \text{SEM}=0.32, \ z=-3.56. \ z < -1.96, \) the normal’s critical value for two-tailed \( \alpha=0.05 \). (Actual \( p \)-value <0.001.) The extent of resection is significantly greater in patients who died.

Exercise 11.4 \( n_w=6, \ n_m=10, \ m_w=3.33, \ m_m=-0.41, \ s_w=4.268, \ s_m=2.067, \ s_d=1.571, \ t_{14} \ df=2.381, \ t > 2.145, \) the critical value for two-tailed \( \alpha=0.05 \).
Serum theophylline levels drop significantly more in women than in men; indeed, levels tend to increase in men.

**Exercise 11.5**

\[
m_w=3, \quad n_m=15, \quad m_w=148.97, \quad m_m=155.27, \quad s_w=15.79, \quad s_m=25.64, \quad s_d=15.57, \quad t_{16, df}=-0.405, \quad t > -2.120, \quad \text{the critical value for two-tailed } \alpha=0.05. \quad (\text{Actual } p\text{-value}=0.691.)
\]

There is inadequate evidence to say that there is a difference in bone density between men and women.

**Exercise 11.6** (Unequal variances) By substituting in Eq. (11.8) and then Eq. (11.5), we find

\[
s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{29,451.98^2}{24} + \frac{50,116.50^2}{22}} = 12,260.06
\]

and

\[
t = \frac{21,457.3 - 32174.5}{12,260.06} = -0.874.
\]

To find the critical value of \(t\), we need \(df\). Substituting in Eq. (11.9), we find

\[
\text{approx}(df) = \frac{((s_1^2/n_1) + (s_2^2/n_2))^2}{((s_1^2/n_1)^2/(n_1 - 1)) + ((s_2^2/n_2)^2/(n_2 - 1)) = 33.35.}
\]

From Table II the critical value of \(t\) for a two-tailed 5% \(\alpha\) for 33 \(df\) is \(\pm 2.03\). \(-0.874\) is not outside the \(\pm 2.03\) critical values, so \(H_0\) is not rejected. From statistical software, \(p\)-value=0.388.

**Exercise 11.7**

From the formulae in Table 11.4, calculate \(SST=(n-1)s^2=19 \times 4799.358=91,187.805, \quad SSM=5[(291.4 - 314.9)^2+(323.2 - 314.9)^2+(274.8 - 314.9)^2+(371.2 - 314.9)^2]=27,234.20, \quad \text{and } \quad SSE=SST-SSM=91,187.80 - 27,234.20=63,953.60. \quad \text{The table below shows the ANOVA results. From Table V the critical } F_{3,16, df}=3.24. \quad \text{The calculated } F=2.27 \text{ is smaller, so } H_0 \text{ is not rejected; there is insufficient evidence to conclude that mean parasite infestation is different from batch to batch. (With a nonsignificant } F\text{, we are not}
concerned that a subordinate pair will be significant, so we do not pursue a multiple comparison procedure.) From a statistical software package, the calculated $p$-value=0.120.

**ANOVA table**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS (sum of squares)</th>
<th>$df$</th>
<th>MS (mean square)</th>
<th>Calculated $F$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SSM=27,234.20</td>
<td>$k-1=3$</td>
<td>MSM=27,234.20/3=9078.07</td>
<td>MSM/ MSE=2.27</td>
<td>0.120</td>
</tr>
<tr>
<td>Error</td>
<td>SSE=63,953.60</td>
<td>$n-k=16$</td>
<td>MSE=63,953.60/16=3997.10 ($\sigma^2=4799.36$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SST=91,187.80</td>
<td>$n-1=19$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Exercise 11.8** The data differences (98.6 minus each datum) are 0.5, 2.8, 1.1, 1.4, 0.9, −0.7, −0.6, 0.5. The unsigned ranks are 1.5, 8, 6, 7, 5, 4, 3, 1.5, and the signed ranks are 1.5, 8, 6, 7, 5, −4, −3, 1.5. The smaller rank sum is 7. From Table VIII the probability that a signed rank sum is 7, given $n=8$, is 0.149. This $p$-value is too large to infer a difference; the patient’s median temperature is not shown to be different from 98.6°F.

**Exercise 11.9** Rank the combined data, keeping track of which ear gave rise to each datum. The ranks are 3.5, 2, 1, 5, 6, 9.5, 9.5, 7, 13.5, 3.5, 11, 8, 12, 16, 15, 13.5. The first eight ranks for the left ear sum to 43.5; the remainder for the right sum to 92.5. Since the $ns$ are the same, either $n$ could be designated $n_1$. The associated $U$-values are $U=n_1n_2+n_1(n_1+1)/2=43.5=56.5$ and $U=n_1n_2+n_1(n_1+1)/2=92.5=7.5$. Since 56.5 is off the table, we choose $U=7.5$. (The test is symmetric.) In Table IX, for $n_1=n_2=8$, $U=7.5$ designates a $p$-value lying halfway between 0.006 and 0.010, or 0.008. The data provide strong evidence that the temperature readings for the two ears tend to be different.

**Exercise 11.10** Follow the answer to Exercise 11.8 until the smaller rank sum $T=7$ is found. Calculate $\mu=n(n+1)/4=8 \times 9/4=18$, $\sigma^2=(2n+1)\mu/6=17 \times 18/6=51$, and then $z=(T-\mu)/\sigma=(7-18)/\sqrt{51}=-1.54$. From Table I, the two-tailed $\alpha$ lies between 0.134 and 0.110, or 0.124 by interpolation. We do not have enough evidence to reject the null hypothesis of no difference. Note that the use of Table VIII yielded $p$-value=0.149. The discrepancy is due to too small a sample for a good approximation.

**Exercise 11.11** Follow the answer to Exercise 11.10 above until the rank sums $T=43.5$ or 92.5 are found. Calculate $\mu=n_1(n_1+n_2+1)/2=8 \times 17/
\[ \sigma^2 = n_1 n_2 (n_1 + n_2 + 1)/12 = 8 \times 8 \times 17/12 = 90.67, \quad \sigma = 9.52, \] and then \[ z = (T - \mu)/\sigma = (43.5 - 68)/9.52 = -2.57. \] (The other \( T = 92.5 \) yields +2.57; note symmetry.) From Table I, \( p \)-value = 0.010. The result differs little from Table IX’s 0.008.

**Exercise 11.12** Frequency plots per group are as follows:
Sed rates, separated by treatment from which each datum arose, and their corresponding ranks are as follows:

<table>
<thead>
<tr>
<th>Sed 1</th>
<th>Ranks 1</th>
<th>Sed 2</th>
<th>Ranks 2</th>
<th>Sed 3</th>
<th>Ranks 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6.5</td>
<td>9, 9, 9</td>
<td>4 × 3</td>
<td>10</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15, 15</td>
<td>9.5 × 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>22</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>15.5</td>
<td>25</td>
<td>15.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>30, 30, 30</td>
<td>22.5 × 4</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>27</td>
<td>64</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( T = )</td>
<td>155</td>
<td></td>
<td>153.5</td>
<td>108</td>
<td>30</td>
</tr>
<tr>
<td>( T^2 = )</td>
<td>24,025</td>
<td></td>
<td>23,562.25</td>
<td>24,492.25</td>
<td></td>
</tr>
</tbody>
</table>
\[
H = \frac{12}{n(n+1)} \left( \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} \right) - 3(n+1)
\]

\[
= \frac{12}{30 \times 31} \left( \frac{24,025}{10} + \frac{23,562.25}{10} + \frac{24,492.25}{10} \right) - 3 \times 31 = 0.0058
\]

From Table III the 5% critical value for 2 \(df\) is 5.99, which is much larger than \(H\); we have insufficient evidence to conclude that \(Sed\) is different for the three treatments. From a statistical software package, \(p\)-value=0.997.

**Exercise 11.13** The ranks for each patient of his PSA levels, with rank sums at the bottom, are as follows:

<table>
<thead>
<tr>
<th>Patient</th>
<th>Rank of first PSA</th>
<th>Rank of second PSA</th>
<th>Rank of third PSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>(T)</td>
<td>15</td>
<td>19.5</td>
<td>19.5</td>
</tr>
<tr>
<td>(T^2)</td>
<td>225</td>
<td>380.25</td>
<td>380.25</td>
</tr>
</tbody>
</table>

\[
F_r = \frac{12}{nk(k+1)} (T_1^2 + T_2^2 + \cdots + T_k^2) - 3n(k+1)
\]

\[
= \frac{12}{9(3)(4)} (225 + 380.25 + 380.25) - 3(9)(4) = 1.50
\]

From Table III the critical chi-square for \(k-1=2\) degrees of freedom for \(\alpha=0.05\) is 5.99. Because 1.50 < 5.99, there is insufficient evidence to conclude that PSA level is increasing over time in this population. The \(p\)-value is 0.472.

**Exercise 11.14** \(z=18.89\) represents a normal curve with tail area starting 18 standard deviations out from the mean! Obviously \(p\)-value < 0.001. We conclude that snoring loudness is positively associated with BMI.
**Exercise 11.15** \( L = 16.5 + 44 + 64.5 = 125 \). \( \chi^2 = 1.25 \). \( \chi^2 \) is less than 2.71, so that \( p \)-value > 0.10. \( L \) is not significant at the 0.05 level. A computer calculates \( p \)-value = 0.264.

**CHAPTER 12**

**Exercise 12.1** Mean nonoperated time is 2.54 seconds, so we take \( \Delta = 0.25 \) second. The mean of differences between leg times is \( d = 0.1600 \), and the standard deviations of differences is \( s = 0.2868 \). From the second row of Table 12.1, \( s_d = s / \sqrt{\bar{n}} = 0.2868 / \sqrt{8} = 0.1014 \). The standard deviation is estimated from the sample, so we use the \( t \) distribution. For 7 \( df \), \( t_{0.95} = 1.895 \). If the test statistic of Eq. (12.1) \( \geq 1.895 \), there is evidence to reject \( H_0 \). The calculated \( t = (\Delta - d) / s_d = (0.25 - 0.16) / 0.1014 = 0.8876 \) does not exceed 1.895; therefore we have no evidence to reject the null hypothesis of difference. We have not established noninferiority.

**Exercise 12.2** Follow the steps in equivalence testing. (1) \( \Delta = 2 \). (2) \( H_{0L} \): \( \delta = -2 \); \( H_{0R} \): \( \delta = 2 \); and \( H_1: -2 < \delta < 2 \). (3) \( \alpha = 0.05 \). (4) Row six of Table 12.1 indicates Table II to find the critical value. From Table II, in the column “0.95” for one-tailed \( 1 - \alpha \), for \( n_1 + n_2 - 2 = 14 \ df \), the critical value of 1.761 is to be used twice, once for each \( H_0 \). (5) We calculate \( d = m_1 - m_2 = 2.99 \), and \( s_d = 1.8504 \). The test statistics are \( t_L = (2.99 + 2) / 1.8684 = 2.6967 \), exceeding 1.761, but \( t_R = (2 - 2.99) / 1.8504 = -0.5350 \), less than 1.761. The data fail to establish that baseline theophylline level is unbiased by sex because not both \( t \)-values exceed the critical value. To perform the test using a confidence interval, we put the calculated values in the form \( (d - t_{1-0.05} \cdot s_d, d + t_{1-0.05} \cdot s_d) \), specifically \( (2.99 - 1.761 \times 1.8504, 2.99 + 1.761 \times 1.8504) \), or \( (-0.269, 6.249) \). \( \Delta = 2 \) falls within the interval. The data do not reject the null hypotheses.

**CHAPTER 13**

**Exercise 13.1** \( H_0: \sigma_o^2 = \sigma^2 \) and \( H_1: \sigma_o^2 > \sigma^2 \). From Table III the critical \( \chi^2 \) value for \( \alpha = 0.05 \) with 15 \( df \) is 25.00.

Substituting in Eq. (13.1), we find

\[
\chi^2 = \frac{df \times s^2}{\sigma^2} = \frac{15 \times 1.230^2}{0.5^2} = 90.77.
\]
As the calculated \( \chi^2 \) is much larger than the critical value, we reject \( H_0 \); the tympanic thermometer is too variable for clinical use. From a computer package the \( p \)-value is 0 to many decimal places; we would report \( p \)-value, 0.001.

Exercise 13.2 \( H_0: \sigma^2_L = \sigma^2_R \) and \( H_1: \sigma^2_L \neq \sigma^2_R \). From Table V the critical \( F \)-value for \( \alpha=0.05 \) with 7,7 \( df \) is 3.79. By substituting in Eq. (13.2), we find \( F = \frac{s^2_L}{s^2_R} = \frac{1.12^2/0.88^3}{1.62} = 1.620 \). As 1.62 < 3.79, we have insufficient evidence to reject the null hypothesis. From a statistical software package, the actual \( p \)-value=0.27.

Exercise 13.3 From Eq. (13.4) and then Eq. (13.5), calculate

\[
\hat{s}^2 = \frac{\sum (n_i - 1) \hat{s}_i^2}{n - k} = \frac{4 \times 3248.30 + \cdots + 4 \times 3623.70}{16} = 3997.10
\]

and

\[
M = \frac{(n - k) \ln(\hat{s}^2) - \sum (n_i - 1) \ln(\hat{s}_i^2)}{1 + (1/(3(k - 1))) \left( \sum (1/(n_i - 1)) - (1/(n - k)) \right)}
\]

\[
= \frac{16 \ln(3997.10) - 4 \ln(3248.30) - \cdots - 4 \ln(3623.70)}{1 + (1/9) \left( (1/4) + \cdots + (1/4) - (1/16) \right)} = 0.1557.
\]

From Table III the critical \( \chi^2 \) \( df=7.81 \). The calculated \( M=0.1557 \) is very much smaller, so we do not have evidence to reject \( H_0 \); variability of parasite infestation does not differ significantly from batch to batch. From a statistical software package,

Exercise 13.4 (KS) For \( n=16 \), our 5\% critical value is \( 1.36/4 - 1/(4.5 \times 16)=0.326 \). The data were given in ascending order. We prepare Table 13.20 in the format of Table 13.9. \( x \) is Hct each time the value is different from the preceding value. \( k \) is the number of values preceding \( x \). From the Hct readings, we calculate \( m=37.25, s=6.9937 \). We calculate \( z=(x-m)/s \). \( F_e \) is the area under the normal curve to the left of \( z \). And finally, the last column is the absolute value of the difference between the \( F \)'s. From inspection of the last column, the largest is \( L=0.1164 \). As 0.1164 < 0.326, we do not reject the hypothesis of normality.

Exercise 13.5 \( \chi^2 \). With nine intervals, \( df=8 \). From Table 13.12 (or Table III), the critical value of \( \chi^2 \) is 15.51. The data and calculations are set up in Table 13.21 in the format of Table 13.11. By using Eq. (13.7), we find
\[ \chi^2 = \sum \frac{n_i^2}{e_i} - n = \frac{2^2}{3.0832} + \frac{9^2}{9.1840} + \cdots - 164 = 9.2112, \]

which is less than the critical value. (From statistical software, \( p\)-value=0.325.) We conclude that the distribution is not significantly different from normal and that our initial judgment was correct.

| \( x \) | \( k \) | \( F_n(x) \) | \( z \) | \( F_n(x) \) | \( |F_n - F_e| \) | \( x \) | \( k \) | \( F_n(x) \) | \( z \) | \( F_n(x) \) | \( |F_n - F_e| \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 25.6 | 1 | 0.0625 | -1.0795 | 0.1402 | 0.0777 | 38.3 | 8 | 0.5000 | 0.1501 | 0.5597 | 0.0597 |
| 29.7 | 2 | 0.1250 | -0.7507 | 0.2264 | 0.1014 | 39.0 | 10 | 0.6250 | 0.2502 | 0.5988 | 0.0262 |
| 32.0 | 2 | 0.1250 | -0.7507 | 0.2264 | 0.1014 | 39.0 | 10 | 0.6250 | 0.2502 | 0.5988 | 0.0262 |
| 32.0 | 2 | 0.1250 | -0.7507 | 0.2264 | 0.1014 | 39.0 | 10 | 0.6250 | 0.2502 | 0.5988 | 0.0262 |
| 32.1 | 4 | 0.3125 | -0.6506 | 0.2577 | 0.0548 | 43.3 | 12 | 0.7500 | 0.8651 | 0.8065 | 0.0565 |
| 32.7 | 5 | 0.3125 | -0.6506 | 0.2577 | 0.0548 | 43.3 | 12 | 0.7500 | 0.8651 | 0.8065 | 0.0565 |
| 33.9 | 6 | 0.3750 | -0.4790 | 0.3160 | 0.0590 | 46.7 | 14 | 0.8750 | 1.3512 | 0.9117 | 0.0367 |
| 34.0 | 7 | 0.4375 | -0.4647 | 0.3211 | 0.1164 | 52.0 | 15 | 0.9375 | 2.1090 | 0.9825 | 0.0450 |

**Exercise 13.6** The 5% critical value from Eq. (20.3b) is

\[ 1.36 \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.36 \sqrt{\frac{9 + 10}{9 \times 10}} = 0.6249. \]

He ranks orders of the values of percent functionality (% func.) in Table 13.22 in the format of Table 13.16. The numbers in the respective sample preceding each current observation are listed for leave in place \( (k_1) \) and remove \( (k_2) \). \( F_1=k_1/10 \) and \( F_2=k_2/9 \) are entered, and finally \( |F_1 - F_2| \). Inspection of \( |F_1 - F_2| \) shows the largest value as \( L=0.5778 \), denoted by an asterisk. \( L \) is a little smaller than critical; there is not enough evidence to conclude that the distributions are different.


APPENDIX 1 Answers to exercises: Final

| % func. | $k_1$ | $k_2$ | $F_1$ | $F_2$ | $|F_1 - F_2|$ | % func. | $k_1$ | $k_2$ | $F_1$ | $F_2$ | $|F_1 - F_2|$ |
|---------|-------|-------|-------|-------|--------------|---------|-------|-------|-------|-------|--------------|
| 35      | 0     | 0     | 0     | 0     | 0            | 90      | 0     | 0.2   | 0.7778 | 0.5778 |             |
| 45      | 0     | 1     | 0     | 0.1111| 0.1111      | 90      | 3     | 0.3   | 0.7778 | 0.4778 |             |
| 63      | 1     | 2     | 0.2222| 0.2222| 95           | 4       | 0     | 0.4   | 0.7778 | 0.3778 |             |
| 70      | 3     | 0     | 0.3333| 0.3333| 95           | 6       | 0.5   | 0.7778| 0.2778 |        |             |
| 72      | 1     | 4     | 0.1   | 0.3333| 98           | 7       | 0.6   | 0.7778| 0.1778 |        |             |
| 75      | 4     | 5     | 0.1   | 0.4444| 99           | 8       | 0.7   | 0.7778| 0.0778 |        |             |
| 78      | 5     | 6     | 0.1   | 0.5556| 100          | 9       | 0.8   | 0.8889| 0.0889 |        |             |
| 80      | 6     | 0.1   | 0.5667| 0.5667| 100          |         | 9     | 0.9   | 0.8889| 0.0111 |             |
| 85      | 2     | 0.2   | 0.4667| 0.4667|             |         |       |       |       |        |             |

CHAPTER 14

Exercise 14.1 Eq. (14.1) yields $\Phi=0.3392$. By using $\chi^2$ calculated as 9.3205 and $n=81$, Eq. (14.2) yields $\Phi=0.3392$ This level of $\Phi$ shows a perceptible but not substantial association between nausea relief and the drug.

Exercise 14.2 $\chi^2=6.6702$, $n=185$, and $k=3$. $V=0.13$, showing a weak association between race and HSV2 occurrence.

Exercise 14.3 $C=8(30+54)+20(54)=1752$. $D=20(11)+72(11+30)=3172$. $\gamma=-0.29$. There is a slightly greater tendency for BP to fall in SBP than in DBP.

Exercise 14.4 Using the $2 \times 5$ table in DB2, we find $C=31(6+5+2+3)+7(5+2+3)=566$. $D=7(25)+2(25+6)=237$. $\gamma=0.4097$. There is a moderate tendency for nausea scores to be greater with placebo.

Exercise 14.5 $s_{\text{op}}=0.5279$, $s_{\text{non}}=0.4513$, $s_{\text{op,non}}=0.2001$, $r=0.840$. $r_s=1-6(6)/504=0.929$. $C=25$, $D=3$, $\tau=(25-3)/28=0.786$. OR=25/3=8.3. Correlation coefficients are all high, although not as high as for distance, but are all somewhat different. $\tau$ is closer to $r$ than is $r_s$. Odds are about 8–1 that an increase in hop time on the operated leg will be paired with an increase in hop time on the nonoperated leg.

Exercise 14.6 While the correlation is high and the variability is not significantly different between clinic and laboratory, the mean difference is significant, raising the question of adequate agreement. However, this mean difference is 0.125 INR, not of clinical concern. Therefore we can conclude adequate agreement.

Exercise 14.7 Although actual readings agree adequately, clinical decisions are made on readings falling in or outside of INR range. Kappa=0.51 is rather low for clinical decision use. 6% of the laboratory in readings were
labeled *out* by the clinic (false negatives) and 38% of the laboratory *out* readings were labeled *in* by the clinic (false positives). We must conclude that too many clinical decision errors occur for the agreement to be considered adequate.

**Exercise 14.8**

*W*: \( n=8, \ k=5 \). Rank sums are found as the sum of products of rank number \( \times \) row entry; the rank sum for the first row is \( 3 \times 4 \times 0 \times 12 + 0 = 19 \). The rank sums for the four remaining therapies are, respectively, 21, 19, 27, and 34. Mean rank sum is \( n(k+1)/2 = 24 \). \( S = (19 - 24)^2 + 3^2 + 5^2 + 3^2 + 10^2 = 168 \). Maximum sum of squares of deviations = \( n^2(k^3 - k)/12 = 640 \). \( W = 168/640 = 0.2625 \). The overall agreement is rather low.

\( A \): In the first row the sequence of ranks is 1, 1, 1, 2, 2, 4, 4, 4. The median of these is 2. The medians of the remaining four therapies, respectively, are 3, 2.5, 4, and 5. For the first therapy, \( d_{\text{obs}} = 3 \times |2 - 1| + 0 + 0 + 3 \times |2 - 4| + 0 = 9 \). \( d_{\exp} = (8/5) \times (|3 - 1| + |3 - 2| + |3 - 3| + |3 - 4| + |3 - 5|) = 9.6 \). \( A_{\text{SurgXRT}} = 1 - 9/9.6 = 0.06 \). The values of \( A \) for the remaining therapies, respectively, are 0.48, 0.27, 0.06, and 0.38.

**Interpretation**: A *W* of 0.26 and *A*’s for the therapies ranging from 0.48 down to 0.06 indicate a discouragingly low level of agreement. This is unfortunately often the case with such a complex clinical picture. The forced ranking is useful in demonstrating the lack of agreement among tumor board members. Further discussion is required with perhaps consideration of aspects not yet included.

**Exercise 14.9** \( m_r = 0.80093 \) and \( \alpha = 0.95 \). The esophageal thermometer is reliable.

**CHAPTER 15**

**Exercise 15.1** Respiration rate is the dependent variable. Age is the independent variable. The curve goes slightly downward from left to right. No clear pattern is discernible; it would be sensible to start with a simple straight-line model, as in Eq. (15.1).

**Exercise 15.2** \( r = 39.0 = -0.34(a - 10.9) \).
**Exercise 15.3**  $b_1=94.3817$ and $b_0=8.9388$. The slope-mean form is $y = 25.6538 + 94.3817(x - 0.1771)$ and the slope-intercept form is $y = 8.9388 + 94.3817x$. If SVmR = 0.3, the change in SBP is predicted to be 37.2533. The axes and regression are shown on the accompanying graph along with the data.

**Exercise 15.4** Preexercise eNO increases with 20-minute eNO. It appears that a straight line will express the relationship. (If the outlying point in the upper right was absent, this may have suggested a curved relationship. When that point is included, we cannot suggest any pattern other
than linear.) Using the mean-and-slope form, \( \text{eNO}_{20} = 28.1 + 0.95 (\text{eNO}_0 - 29.3) \).

Exercise 15.5  \( R^2 = 0.038 \). Age exerts only about 4% of the causal influence on respiration rate. (For the student’s information, a \( t \) test on the slope of the regression line, which is addressed in Section 15.4, yields a \( p \)-value of 0.003, significant at level 0.05. We can say that the influence of age on respiration rate is real, but small in comparison to other influences.)

Exercise 15.6 The coefficient of determination \( R^2 = (0.96)^2 = 0.92 \). It may be concluded that baseline eNO is quite a good predictor of 20-minute eNO.

Exercise 15.7 \( s_e = 14.01 \) and \( s_b = 18.04 \). The two-tailed 95% critical \( t \) for 24 \( df \) is 2.064. Calculated \( t \) to test the slope is 5.23, which is significantly greater than the critical \( t \); the slope of the regression line is greater than chance. \( R^2 = 0.5323 \). SVmR is a major predictor; 47% of the total variation in the response is not accounted for by the model. \( t \) for testing \( R^2 = 5.23 \), which is far greater than 2.064; \( R^2 \) is highly significant. A 95% CI for \( \beta_1 \) is (57, 132). \( s_{m|0.3} = 3.53 \) and \( s_{y|0.3} = 14.45 \). A 95% CI for the mean at 0.3 is (30, 45). A 95% prediction interval for a subject with covariate value 0.3 is (7.4, 67.1).

Exercise 15.8 Corr(\( r, a \)) = −0.195. The form is a downward slope. Some correlation is apparent, but it is not strong.

Exercise 15.9 The correlation coefficient \( r = 0.96 \). This is quite a high correlation coefficient. As baseline eNO rises, 20-minute eNO rises proportionally.
Exercise 15.10  \( r = \frac{2.2746}{0.1552 \times 20.0757} = 0.7300 \). A correlation coefficient of 0.73 is rather high; there is clearly an association between the variables. \( r_s = 0.7050 \), which is close to \( r \).

Exercise 15.11  \( r^2 = 0.5322 \). The two-tailed 95% critical \( t \) for 24 df is 2.064. The calculated \( t \) is \( \sqrt{24 \times 0.5322/0.4678} = 5.2253 \), which is much larger than the critical \( t \). The population correlation coefficient is in probability greater than 0. (Slight differences appearing in prior calculations based on these data are due to rounding.) Denote \( \rho_s \) as the correlation coefficient of the population from which the sample was drawn. To test the null hypothesis \( H_0: \rho = 0.5 \), substitute the appropriate values in Eqs. (15.25) through (15.28). \( m = 0.9287 \), \( \mu = 0.5493 \), \( \sigma = 0.2085 \), and \( z = 1.82 \). The critical \( z \) is 1.96. The calculated \( z \) is less; therefore we must conclude that there is inadequate evidence to reject \( H_0 \). The correlation has not been shown to be different from the theoretical value.

CHAPTER 16

Exercise 16.1  Model: \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \). Data. Enter data for the triple hop distance using the operated leg into the \( y \) position, the same using the unoperated leg into the \( x_1 \) position, and the time to hop using the unoperated leg into the \( x_2 \) position. Interpretation. The predictive ability using both \( x \)'s is real and clinically useful. Dropping \( x_2 \) should be considered. The reduced model also is significant and very nearly as clinically useful. The predictive equation is \( y = -300.868 + 1.466 x_1 \). For a nonoperated leg hop distance of 504, the predicted distance on the operated leg is 438 cm, which agrees very well with the observed 436 cm.

Exercise 16.2  Model: \( SBP = b_0 + b_1 (SVmR) + b_2 (SVmR)^2 \). Data: The square of SVmR would be calculated. SBP, SVmR, and \( SVmR^2 \) would be entered. Interpretation: The final predictive equation is \( SBP = -2.92 + 289.00 (SVmR) - 414.42 (SVmR)^2 \). The fit is shown in the figure. The \( p \)-values for tests of both models are <.001, which is significant. The modeled prediction accounts for the majority of variability in both models; SVmR is a major predictor in either model. Moving from the simple model to the parabolic model increases the \( R^2 \) from about 53% to about 68%, indicating that the curved model is the better fit.
Exercise 16.3  Means were as follows:

<table>
<thead>
<tr>
<th>Effect</th>
<th>Others</th>
<th>Stones</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>20.71</td>
<td>-0.14</td>
<td>10.29</td>
</tr>
<tr>
<td>Drug</td>
<td>6.71</td>
<td>10.14</td>
<td>8.43</td>
</tr>
<tr>
<td>Combined</td>
<td>13.71</td>
<td>5.00</td>
<td>9.36</td>
</tr>
</tbody>
</table>

The ANOVA table was as follows:

<table>
<thead>
<tr>
<th>Effect</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>24.14286</td>
<td>1</td>
<td>24.14286</td>
<td>0.22</td>
<td>0.643</td>
</tr>
<tr>
<td>Stones</td>
<td>531.57143</td>
<td>1</td>
<td>531.57143</td>
<td>4.85</td>
<td>0.038</td>
</tr>
<tr>
<td>Drug × stones interaction</td>
<td>1032.14286</td>
<td>1</td>
<td>1032.14286</td>
<td>9.42</td>
<td>0.005</td>
</tr>
<tr>
<td>Residual (or error) component</td>
<td>2630.57143</td>
<td>24</td>
<td>109.60714</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4218.42857</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We note that the drug is not shown to be effective ($p$-value=0.643) overall. However, there is a significant difference in pain change between the stones and others patients ($p$-value=0.038) and the interaction is highly significant ($p$-value=0.005), indicating that pain change has a different pattern between stones and others patients. Examining the means table, we see that, while combined groups no-drug-to-drug means are similar, mean pain is reduced by placebo (!) for the other group but is reduced by the drug for the stones group. This reversal has obscured the effect of the drug for the two groups and seems to account for the mixed result pattern seen in prior medical articles.
APPENDIX 1 Answers to exercises: Final

CHAPTER 17

Exercise 17.1 The estimated probability of CHD among males is $p_{\text{males}} = (823/2049) = 0.402$. The estimated probability of CHD among females is $p_{\text{females}} = (650/2650) = 0.245$. The estimated odds of CHD among males is $\text{odds}_{\text{males}} = (p_{\text{males}}/(1 - p_{\text{males}})) = 0.671$. The estimated odds of CHD among females is $\text{odds}_{\text{females}} = (p_{\text{females}}/(1 - p_{\text{females}})) = 0.325$. The estimated odds ratio for CHD comparing females to males is $(\text{odds}_{\text{females}}/\text{odds}_{\text{males}}) = (0.325/0.671) = 0.484$, which matches the resulting estimate from the logistic regression model.

Exercise 17.2 The estimated odds ratio comparing females to males of a similar age is $e^{-0.761} = 0.467$. The estimated odds ratio comparing individuals differing in age by 1 year who have the same sex is $e^{0.040} = 1.04$.

Exercise 17.3 The estimated odds ratio comparing females to males of a similar age and BMI is $e^{-0.740} = 0.477$.

Exercise 17.4 67-year-old female: $p = (e^{-2.22 -0.761+0.040 \times 67})/(1+ e^{-2.22 -0.761+0.040 \times 67}) = 0.425$; 83-year-old male: $p = (e^{-2.22+0.040 \times 83})/(1+ e^{-2.22+0.040 \times 83}) = 0.750$.

Exercise 17.5 67-year-old female with BMI of 27.2: $p = (e^{-3.80-0.740+0.035 \times 67+0.069 \times 27.2})/(1+ e^{-3.80+0.035 \times 83+0.069 \times 22.4}) = 0.421$; 83-year-old male with BMI of 22.4: $p = (e^{-3.80+0.035 \times 83+0.069 \times 22.4})/(1+ e^{-3.80+0.035 \times 83+0.069 \times 22.4}) = 0.657$.

Exercise 17.6 $z = (0.06937/0.00801) = 8.66$. The $p$-value of $Z$ for testing the null hypothesis that BMI is not associated with CHD after adjustment for both sex and age is less than 0.001, as seen in Table I.

Exercise 17.7 \( (e^{0.06937-1.96 \times 0.00801}, e^{0.06937+1.96 \times 0.00801}) = (1.06, 1.09) \).

CHAPTER 18

Exercise 18.1 The observed rate of seizures over 1 day: 3/14=0.214; over 1 year: 365.25 × (3/14)=78.27; over 5 years: 5 × 365.25 × (3/14)=391.34.

Exercise 18.2 The estimated rate of repeats among normal participants is 187/3887=0.048 per word, or 2.4 per 50 words. The estimated rate of repeats among demented participants is 104/852=0.122 per word, or 6.1 per 50 words. The rate ratio comparing demented to normal participants is 0.122/0.048=2.54.

Exercise 18.3 (a) $e^{0.30} = 1.35$; (b) $e^{0.86} = 2.36$; (c) $e^{0.86-0.30} = 1.75$; (d) $e^{0.02} = 1.02$

Exercise 18.4 67-year-old normal participant with 16 years education: $50 \times e^{-3.48+0.02 \times 67-0.05 \times 16} = 2.64$ repeats per 50 words; 72-year-old demented participant with 14 years education: $50 \times e^{-3.48+0.86+0.02 \times 72-0.05 \times 14} = 7.63$ repeats per 50 words.
Exercise 18.5  The value of $z$ for testing that the rate of repeats comparing demented participants to cognitively normal participants is 4.58 and the resulting $p$-value is less than 0.001, as seen in Table 1. If conducting a level .05 test, we would reject the null hypothesis and conclude that demented participants do have a higher rate of repeats when compared to normal participants.

Exercise 18.6  \( \left( e^{0.86-1.96 \times 0.18727}, e^{0.86+1.96 \times 0.18727} \right) = (1.64\ 3.41). \)

CHAPTER 19

Exercise 19.1  See Table 19.5. The estimated probability that a diabetic woman survives more than 10 years is 0.715. See Figure 19.5.

Exercise 19.2  From Table III the critical $\chi^2$ value for $\alpha=0.05$ for 1 df is 3.84. The calculated 0.63 is much less than 3.84. Conclusion: No difference between the survival patterns of diabetic women for the two decades has been shown. (Actual $p$-value=0.427.)

Exercise 19.3  Because mean survival time is less in the salvage group, the hazard ratio would say that the probability of survival of a nonsalvage patient is nearly twice that of a salvage patient. However, the confidence interval extends well into either side of 0, so we are not confident that the hazard ratio is really greater than 1 (equal chances for both groups). The $p$-value of 0.261 supports the conclusion that the survival time
difference between the groups is not significant. Clearly, we have no evidence that salvage therapy benefits the patient’s survival.

### CHAPTER 20

**Exercise 20.1** The ANOVA table is as follows:

<table>
<thead>
<tr>
<th>Effect</th>
<th>Sums of squares</th>
<th>df</th>
<th>Mean squares</th>
<th>F</th>
<th>Crit. F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug treatment</td>
<td>10,945.2460 (SSR)</td>
<td>1</td>
<td>10,945.2460</td>
<td>7.166</td>
<td>4.12</td>
<td>0.011</td>
</tr>
<tr>
<td>Between-drug error</td>
<td>54,980.5270</td>
<td>36</td>
<td>1527.2369</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>27,196.9370 (SSC)</td>
<td>2</td>
<td>13598.4690</td>
<td>29.664</td>
<td>3.13</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Drug × time interaction</td>
<td>10,042.4940 (SSI)</td>
<td>2</td>
<td>5021.7456</td>
<td>10.955</td>
<td>3.13</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Within-time error</td>
<td>33,005.8360</td>
<td>72</td>
<td>458.4144</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>136,171.0400 (SST)</td>
<td>113</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A significant difference appears between drugs. From the means table, we can see that prochlorperazine reduces pain more than diazepam does. A highly significant difference appears in the repeated effects. Both drugs reduce the headache over time. However, their pattern through time is different. From the means table, we can see that prochlorperazine reduces pain through time more effectively than diazepam.

**Exercise 20.2**

(a) $0.20815 - 0.44781 = -0.23966$; (b) We estimate that the cognitive scores for AD subjects decrease by 0.24 per year. As this is negative, there does not appear to be a learning effect in this group though it cannot be proven with these data; (c) The Z statistic for testing if the slope of AD participants differs from that of cognitively normal participants is $-0.44781/0.0155522 = -28.79$. The corresponding p-value is less than 0.001, as seen in Table I. If conducting a level 0.205 test, we would reject the null hypothesis and conclude that the slope of cognitive performance differs between AD and normal participants.

**Exercise 20.3**

(a) $0.2085 - 0.1186 = 0.0899$; (b) We estimate that the cognitive scores for MCI subjects increase by 0.0899 per year. As this is positive and we would not expect cognition to increase with aging, there does appear to be a learning effect in this group though variation in the estimate must also be considered; (c) The Z statistic for testing if the
slope of MCI participants differs from that of cognitively normal participants is \(-0.1186/0.012620 = -9.40\). The corresponding p-value is less than 0.001, as seen in Table I. If conducting a level 0.05 test, we would reject the null hypothesis and conclude that the slope of cognitive performance differs between MCI and normal participants;

(d) We can see that the estimates are nearly identical. This is true in general as the two models differ only slightly in how the covariance structure parameters are estimated.

**Exercise 20.4** It appears that the heart exhibits greater efficiency when the HR is in the normal range of about 70–80 bpm, with inferior efficiency below 65 bpm and efficiency diminishing gradually as the HR increases above 80 bpm. This pattern revealed by time-series analysis is unlikely to be detected with regression or other curve-fitting techniques.

**Exercise 20.5** The critical \(F_{2,3}(0.95) = 9.55\). See Figure 20.12.

![Figure 20.12](image-url)  
*Figure 20.12* Moving F for Dr. Shipman’s excess number of death certificates with critical \(F_{2,3} = 9.55\). The significant \(F_{2,3} = 9.55\) was breeched in 1984–86 and again 1994–98.

Moving F first become significant in 1984. Had he been stopped then, 15 years of patient murders would have been prevented, representing 178 patients.
CHAPTER 21

Exercise 21.1  (a). One-sided. Ondansetron hydrochloride could decrease nausea score but does not seem to increase it. Also, the clinical decision to not use the drug is the same whether nausea remains the same or is increased.

(b). Two-sided. There is no reason to believe that the spectrophotometric readings will be higher in one assay than the other.

Exercise 21.2  (a). A false positive implies that the patient takes the drug when it will not help nausea. A false negative implies the patient suffers nausea when it could have been reduced. The assessment depends on the side effects and cost of the drug. If these should be minimal, a false negative is worse than a false positive, so that $\alpha$ should be larger than $\beta$. If these are important clinically, the commonly used 5% $\alpha$ and 20% $\beta$ trade-off should be rethought.

(b). A false positive, that is, implant appears to raise plasma silicone when it does not, which may lead to unnecessary removal surgery. A false negative, that is, implant appears to be of no risk when in fact it is, which implies allowing possible side effects. Both errors have undesirable clinical implications, but a $\beta$ larger than $\alpha$ is reasonable.

Exercise 21.3  The posttrauma heart rate can be either greater or less than the healthy rate, so the test is two-sided; $z_{1-\alpha/2}=1.96$ and $z_{1-\beta}=0.84$. From Eq. (21.2),

$$n_1 = n_2 = (z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2 / \delta^2 = (1.96 + 0.84)^2 (9.1)^2 / 6^2 = 18.0.$$ 

He needs at least 18 patients. For reasons set forth in Section 22.2, he should choose a few more than 18.

Exercise 21.4  Either trauma type could evoke a higher heart rate, so the test is two-tailed; $z_{1-\alpha/2}=1.96$, $z_{1-\beta}=0.84$. From Eq. (21.3),

$$n_1 = n_2 = (z_{1-\alpha/2} + z_{1-\beta})^2 (6.13^2 + 6.34^2) / 6^2 = 16.9.$$ 

He needs at least 17 in each group and, by Section 18.2, is advised to choose a few more than 17.

Exercise 21.5  $n_1=n_2=71$.

Exercise 21.6  $n_1=n_2=142$.

Exercise 21.7  $n_1=n_2=(1.96+1.28)^2 (0.54^2+0.63^2) / 0.25^2=115.6$. We require a minimum of 116 readings in each group for a total of 232.

Exercise 21.8  $n_1=n_2=31$.

Exercise 21.9  The minimum sample size estimates provide a sample size that will just barely reach the 5% $\alpha$ on average. It is not surprising that
repeated sampling yields $p$-values on both sides of $\alpha=0.05$. A slightly
larger sample size would yield higher precision for estimating the
parameter of interest.

**Exercise 21.10** By substituting in Eq. (21.6), we find

$$n = \frac{z_{1-\alpha/2}^2 \sigma^2}{\omega^2} = \frac{1.96^2 \times 24^2}{10^2} = 22.1.$$ 

We will need a sample of at least 23 patients and should take a
few more, as per Section 18.2.

**Exercise 21.11** $\pi=0.18$. As $\pi$ is not near 0 or 1, the binomial form (22.8) will be
used. The test is one-sided (the training will not increase the number
of enhancements), so $z_{1-\alpha}=1.645$.

$$n = \frac{[1.645\sqrt{0.18 \times 0.82} + 0.84\sqrt{0.12 \times 0.88}]^2}{(0.12 - 0.18)^2} = 227.5.$$ 

At least 228 residents would be required, too large a number to
evaluate the training device in one medical center in a reasonable
period of time; the plan for this study must be either dropped or
arranged as a multicenter study.

**Exercise 21.12** The one-sided test uses $z_{1-\alpha}=1.645$. $p_m=(0.23+0.13)/2=0.18$, not
near 0 or 1, so the binomial form Eq. (22.11) is used. Substitution
yields

$$n_1 = n_2 = \frac{[1.645\sqrt{2 \times 0.18 \times 0.82}}{0.12 - 0.18} = 181.25.$$ 

He would need a minimum of 182 eyes per method.

**Exercise 21.13** $\pi$, is central, so he uses the form of Eq. (21.14). $p=0.7$, $w=0.1$, and, for 95% confidence, he retains the 1.96 given in Eq. (21.14). Substitution yields

$$n = \frac{1.96^2 \pi(1-\pi)}{w^2} = \frac{3.8416 \times 0.7(1-0.7)}{0.1^2} = 80.7.$$ 

**Exercise 21.14** The predicted required sample size will be the same as if the $r$ were
positive, due to the symmetry property. In Table 22.1 $r=0.23$ corre-
sponds to $n=74$.

**Exercise 21.15** Solve Eq. (21.17) for $n_d$. $n_d=p_d \times n_s=0.00767 \times 4172=31.9992$, rounding to 32 cases expected to be discovered by screening. $p_m=n_m/
\[ n_r = 25/4172 = 0.00599. \] Substituting in Eq. (21.18) yields \[ \text{NNT}_{\text{addit}} = 1/(0.00767^2 - 0.00599) = 595.2. \] 596 inmates must be screened for each detection not found otherwise.

**CHAPTER 22**

**Exercise 22.1** This is a single-blind study. The patients are blinded because they are not aware of which agent they will or did receive. The surgeons, however, are not blinded to the agent being used. In this case, bias could arise if surgeons are the individuals who subjectively determine if and when bleeding has stopped (given their knowledge of which agent they are using).

**Exercise 22.2** The concern about a learning effect is that if randomization is unbalanced toward the beginning of the trial, this would imply that surgeons are learning more on one agent than another. To avoid any bias this could cause, blocking would be advised. For example, if blocks of size 6 were utilized then after every 6 randomized patients 3 would be on each arm implying that there would be balance over time in the allocation of patients and hence the learning effects.

**Exercise 22.3** From Table 22.5 the correct critical values are 3.47 at interim analysis 1, 2.45 at interim analysis 2, and 2.00 at the final analysis.

**CHAPTER 23**

**Exercise 23.1** In both a cause of illness is being sought by trying to isolate a cause from among many potential causes. They are different in that clinical diagnosis seeking to treat an individual, whereas epidemiology seeks to treat a population.

**Exercise 23.2** (1) Quantified description, 1662, by means of mortality records. (2) Verified explanation, 1747, by means of isolating causes by experimentation. (3) Prediction of cause-and-effect, 1847, by means of demonstrating the theory of contagion. These steps chosen are somewhat arbitrary. The student’s attention to epidemiological history evoked by the effort serves the learning purpose.

**Exercise 23.3** Cancer caused by tobacco smoke or occupational exposure. Vitamin or mineral deficiency. Airway obstruction caused by allergic reaction.

**Exercise 23.4** \[ I_{\text{caff}} = 1000 \times \text{newly pregnant caffeine consumers/total caffeine consumers} = 100 \text{ per thousand}. \] \[ I_{\text{nocaff}} = 1000 \times \text{newly pregnant not caffeine consumers/total not caffeine consumers} = 100 \text{ per thousand}. \]
consumers = $1000 \times 13/89 = 146$ per thousand. \( P_{\text{caff}} = 1000 \times \) pregnant women who used caffeine/all pregnant women = $1000 \times 304/575 = 529$ per thousand. OR = number pregnant who use caffeine/number pregnant who do not use caffeine = $304/271 = 1.12$.

**Exercise 23.5**

\( I_{\text{Pel}} = 1000 \times 28/2800 \) divided by 5 days = 2.0 per day per thousand. 
\( I_{\text{Con}} = 1000 \times 31/4500 \) divided by 14 days = 0.49 per day per thousand. 
\( P_{\text{Pel}} = 1000 \times 16.56/2800 = 5.9 \) cases per thousand. 
\( P_{\text{Con}} = 1000 \times 29.37/4500 = 6.5 \) cases per thousand. Odds of gastroenteritis on *Peleliu* were 162/2638 = 0.0614 and on *Constellation* were 425/4075 = 0.1043. OR of *Constellation* versus *Peleliu* was $0.1043/0.0614 = 1.6987$. A randomly chosen sailor was about 1.7 times as likely to acquire gastroenteritis on *Peleliu* as on *Constellation*.

**Exercise 23.6**

<table>
<thead>
<tr>
<th>Interval (years)</th>
<th>Begin</th>
<th>Died</th>
<th>Lost</th>
<th>End</th>
<th>( S ) (survival rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (outset)</td>
<td>1824</td>
<td>0</td>
<td>0</td>
<td>1824</td>
<td>1.000</td>
</tr>
<tr>
<td>1</td>
<td>1824</td>
<td>545</td>
<td>0</td>
<td>1279</td>
<td>0.701</td>
</tr>
<tr>
<td>2</td>
<td>1279</td>
<td>816</td>
<td>0</td>
<td>463</td>
<td>0.254</td>
</tr>
<tr>
<td>3</td>
<td>463</td>
<td>274</td>
<td>0</td>
<td>189</td>
<td>0.104</td>
</tr>
<tr>
<td>4</td>
<td>189</td>
<td>88</td>
<td>0</td>
<td>101</td>
<td>0.055</td>
</tr>
<tr>
<td>5</td>
<td>101</td>
<td>44</td>
<td>0</td>
<td>57</td>
<td>0.031</td>
</tr>
<tr>
<td>6</td>
<td>57</td>
<td>24</td>
<td>0</td>
<td>33</td>
<td>0.018</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
<td>0</td>
<td>2</td>
<td>31</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>5</td>
<td>0</td>
<td>26</td>
<td>0.014</td>
</tr>
</tbody>
</table>

The probability that a woman with untreated breast cancer will survive 5 years is about 3%.

**Exercise 23.7**

![Graph showing survival probability over years since diagnosis]
Exercise 23.8 Many examples are possible. One each is given here for illustration. 

*Cross correlation without lag:* Cases of yellow fever appearing in a certain tropical nation, suspected to arise from a mutated virus, seem to exhibit longer lasting thrombocytopenia. From records, proportions of patients with below normal platelet counts from before and after the appearance of the new strain are correlated over the course of the disease. The low correlation verifies the mutated behavior. *Cross correlation with lag:* A group of people who recently traveled in a tropical nation contract a new type of viral disease. The dates of visit and of symptom onset are correlated with various lags. The lag that provides the largest cross correlation coefficient estimates the average incubation period.

*Autocorrelation:* Incidence of aspergillosis is recorded for several years. The autocorrelation coefficient is calculated for a lag of 1 year and is seen to be high. It is clear that aspergillosis is a seasonal disease.

Exercise 23.9 Many examples are possible. One each is given here for illustration. 

*Cross correlation without lag:* A measure of pollen level in air correlates in time with prevalence of allergic reaction. *Cross correlation with lag:* Days in hospital correlates with incidence of a certain nosocomial disease with incubation period equal to the lag. *Autocorrelation:* Malarial symptoms in a patient correlate with themselves when the lag equals the reproduction cycle time of the parasite.

Exercise 23.10 Because all the sailors are exposed to the same air, food, water, and personal habits, and no outside influences of any type occur, it appears that the cause is not something present, but something absent. You suspect diet. You choose a sample of patients (Lind chose 12), dividing them into experimental groups (Lind chose 2 per group for 6 treatments). You assign them six different diets, each including a component not regularly eaten aboard ship. It would be sensible to choose foods eaten regularly ashore but not at sea. (Lind did some of this but some of his treatments were bizarre, such as drinking seawater.) Fortunately, one of your choices is fresh citrus fruits, which quickly cures that group while no other group improves. You have found a cure for scurvy, despite not knowing that vitamins exist or that their deficiency can cause disease.

CHAPTER 24

(No exercises in this chapter.)
CHAPTER 25

Exercise 25.1 Substitute \( m, \sigma_m^2, \mu_0, \) and \( \sigma_0^2 \) into Eq. (25.3) to obtain the posterior distribution as \( N(7.82, 0.50) \). The probability that \( \mu < 8.85 \) uses \( z = (6.75 - 7.82)/0.50 = -2.14 \) to find the probability as 0.016.

CHAPTER 26

(No exercises in this chapter.)

CHAPTER 27

Exercise 27.1 The critical value of 9 \( df \) was 2.26 and the \( t \) statistic was 4.06. Moving values to the left 4.06 units to standardize the curve resets the critical value at \(-1.80\). \( \beta \) is the area left of \(-1.80\), or, by symmetry, right of 1.80. The closest entry in the 9 \( df \) row is 1.833, associated with a right tail area of 0.05. The power is close to 0.95. From a software package, the power is 0.947.

Exercise 27.2 (a) The accumulated type I error assuming independent tests is \( 1 - 0.95^4 = 0.185 \). (b) From the accumulative type I error, we would expect about 1 in 6 results to be a false positive. That we obtained 1 in 4 is not too surprising. Since the ANOVA result indicates no overall significance, we may conclude that the significant result is most likely not a true positive. (c) The Bonferroni critical value for selecting significance is \( 0.05/4 = 0.0125 \). None of the test result \( p \)-values are less than this critical value, so we conclude from the Bonferroni procedure that no test results are significant. The average age for hand versus foot snakebites is not different from season to season. (d) The four ordered \( p \)-values and their corresponding \( p_0 \)-values are \( (0.034, 0.0125), (0.100, 0.0250), (0.279, 0.0375), \) and \( (0.695, 0.0500) \). None of the \( p \)-values are below the calculated critical values, so \( k' = 0 \); no test shows a significant result. The RFDR = \( 0.05(5)/8 = 0.031 \).

Exercise 27.3 The user’s result will likely differ from the following solution, due to the very small sample. The resamples found were 1st: 29, 27, 24, 27, . . . , 5th: 38, 29, 27, 24. The five medians were 27, 28, 31, 29, 28. Ordering gives 27, 28, 28, 29, 31 for a 60% CI of 27.5 to 30.
Exercise 27.4  While there is no overall mean bias, the bulk of the data lie below 0 on the left and move to above 0 on the right, suggesting a mean bias in one direction for smaller plasma silicon levels and the other direction for larger levels. Lying near or outside the confidence intervals are four points (three outside), about 13%, while the remainder lies much closer to the mean, suggesting an unusual number of anomalies.

CHAPTER 28

(No exercises in this chapter.)
Databases

Databases are abbreviated “DBs” followed by the database number. Superscript number at end of each title keys to Reference and Data Source list. Databases may be downloaded in MS Excel format from the Elsevier website (access to information is available via the Internet at http://www.elsevierdirect.com/companions/9780123848642).

PURPOSE FOR REPEATED USE OF THESE DATA SETS

In this text, these 17 databases are used for primary illustration of many of the statistical methods presented. Once familiar with a data set, the reader may concentrate on the statistical aspects being discussed rather than focusing on the medical aspects of the data. These data sets were selected as much for data reasons as for medical reasons: A mix of data types illustrates different methods, adequate sample size, and other features.

ADDITIONAL DATABASES

Also available through the Elsevier website are 20 additional databases used at times throughout the book. References for these databases are the followings: DB18, DB19, DB20, DB21, DB22, DB23, DB24, DB25, DB26, DB27, DB28, DB29, DB30, DB31, DB32, DB33, BD34, DB35, DB36, DB37.

DB1 INDICATORS OF PROSTATE BIOSPY RESULTS

Background

These data were taken from 301 male patients examined in the Urology Department at the Naval Medical Center San Diego who exhibited one of several reasons to suspect the health of their prostate glands. They were recorded in the order they presented. These data have the advantage that they include biopsy outcomes regardless of the results of preliminary examinations. Thus, false negatives (negative biopsy outcomes when a tumor is present) as well as false positives (positive biopsy outcomes when no tumor is present) can be examined.
Data

For each patient, recorded data (units indicated in parentheses) are as follows:

- age (years, >0, integers)
- digital rectal examination result (DRE) (0 = negative, 1 = positive)
- transurethral ultrasound result (TRU) (0 = negative, 1 = positive)
- prostate specific antigen level (PSA) (ng/ml, >0, to one decimal place)
- volume of prostate (VOL) (ml, >0, to one decimal place)
- PSA density level (PSAD) (PSA/VOL, >0, to two decimal places)
- biopsy result (BIOP) (0 = negative, 1 = positive).

There are too many data for the reader to do calculations without a computer. Some summary statistics are given in the following table for the full set of 301 patients, for 296 patients who are at risk of prostate cancer (data excluded for 5 patients with severe benign prostatic hypertrophy), and for all patients except the first 10:

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Age</th>
<th>PSA</th>
<th>Volume</th>
<th>PSAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>301 patients: mean</td>
<td>66.76</td>
<td>8.79</td>
<td>36.47</td>
<td>0.269</td>
</tr>
<tr>
<td>standard deviation</td>
<td>8.10</td>
<td>16.91</td>
<td>18.04</td>
<td>0.488</td>
</tr>
<tr>
<td>296 patients: mean</td>
<td>66.75</td>
<td>8.79</td>
<td>35.46</td>
<td>0.272</td>
</tr>
<tr>
<td>standard deviation</td>
<td>8.13</td>
<td>17.05</td>
<td>16.35</td>
<td>0.491</td>
</tr>
<tr>
<td>Patients nos. 11-301: mean</td>
<td>66.82</td>
<td>8.85</td>
<td>36.60</td>
<td>0.270</td>
</tr>
<tr>
<td>standard deviation</td>
<td>8.14</td>
<td>17.19</td>
<td>18.12</td>
<td>0.496</td>
</tr>
</tbody>
</table>

Data for the first 10 cases are given in Table DB1.1 for use in calculations. Note that all data are numeric. Most analysis software packages require this, and it is much easier to transfer data from one package to another (e.g. from a spreadsheet to a statistical package) with data in numeric form.

Table DB1.1 Prostate Test Data from the First 10 of the 301 Patients in the Data Set.

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Age</th>
<th>Digital rectal exam</th>
<th>Transurethral ultrasound</th>
<th>PSA</th>
<th>Volume</th>
<th>PSAD</th>
<th>Biopsy result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75</td>
<td>0</td>
<td>1</td>
<td>7.6</td>
<td>32.3</td>
<td>0.24</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>1</td>
<td>0</td>
<td>4.1</td>
<td>27.0</td>
<td>0.15</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>5.9</td>
<td>16.2</td>
<td>0.36</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>1</td>
<td>1</td>
<td>9.0</td>
<td>33.0</td>
<td>0.27</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>6.8</td>
<td>30.9</td>
<td>0.22</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>61</td>
<td>0</td>
<td>1</td>
<td>8.0</td>
<td>73.7</td>
<td>0.11</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>7.7</td>
<td>30.5</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>61</td>
<td>1</td>
<td>1</td>
<td>4.4</td>
<td>30.5</td>
<td>0.14</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>73</td>
<td>1</td>
<td>1</td>
<td>6.1</td>
<td>36.8</td>
<td>0.17</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>74</td>
<td>1</td>
<td>1</td>
<td>7.9</td>
<td>16.4</td>
<td>0.48</td>
<td>0</td>
</tr>
</tbody>
</table>
**DB2 EFFECTIVENESS OF A DRUG IN REDUCING NAUSEA FOLLOWING GALLBLADDER REMOVAL**

**Background**

A sample of 81 gall bladder removals by laparoscope were randomized into two groups to receive preoperative ondansetron hydrochloride or a placebo. (Data were edited slightly for the exercise and do not reflect exact original results.) Patients rated their nausea on a 1 (no nausea) to 5 (unbearable nausea) 2 hours after the end of surgery.

**Data**

<table>
<thead>
<tr>
<th>Nausea Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>Drug</td>
</tr>
<tr>
<td>Placebo</td>
</tr>
<tr>
<td>Totals</td>
</tr>
</tbody>
</table>

To illustrate some analytic techniques, rating scores 2—5 may be combined to form a 2 × 2 contingency table of nausea or no nausea:

<table>
<thead>
<tr>
<th>No nausea (score 1)</th>
<th>Nausea (score 2–5)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>34</td>
<td>6</td>
</tr>
<tr>
<td>Placebo</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>Totals</td>
<td>56</td>
<td>25</td>
</tr>
</tbody>
</table>

Separately from a rating scale, patients were asked to choose no nausea versus some nausea, generating the following table:

<table>
<thead>
<tr>
<th>No nausea</th>
<th>Nausea</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>31</td>
<td>9</td>
</tr>
<tr>
<td>Placebo</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>Totals</td>
<td>56</td>
<td>25</td>
</tr>
</tbody>
</table>

**DB3 EFFECT OF AZITHROMYCIN ON SERUM THEOPHYLLINE LEVELS OF PATIENTS WITH EMPHYSEMA**

**Background**

Theophylline dilates airways in emphysema patients. When azithromycin is indicated for an infection in such a patient, does it alter the serum theophylline level? Serum theophylline levels (mg/dl) were measured in 16 emphysema patients at baseline (just before the start of azithromycin), 5 days (at the end of the course of antibiotics), and at 10 days. Clinically, it is anticipated that the antibiotic will raise the theophylline level.
**Data**

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Age</th>
<th>Sex</th>
<th>Baseline</th>
<th>5 days</th>
<th>10 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61</td>
<td>F</td>
<td>14.1</td>
<td>2.3</td>
<td>10.3</td>
</tr>
<tr>
<td>2</td>
<td>70</td>
<td>F</td>
<td>7.2</td>
<td>5.4</td>
<td>7.3</td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>M</td>
<td>14.2</td>
<td>11.9</td>
<td>11.3</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>M</td>
<td>10.3</td>
<td>10.7</td>
<td>13.8</td>
</tr>
<tr>
<td>5</td>
<td>64</td>
<td>F</td>
<td>15.4</td>
<td>15.2</td>
<td>13.6</td>
</tr>
<tr>
<td>6</td>
<td>76</td>
<td>M</td>
<td>5.2</td>
<td>6.8</td>
<td>4.2</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
<td>M</td>
<td>10.4</td>
<td>14.6</td>
<td>14.1</td>
</tr>
<tr>
<td>8</td>
<td>69</td>
<td>F</td>
<td>10.5</td>
<td>7.2</td>
<td>5.4</td>
</tr>
<tr>
<td>9</td>
<td>66</td>
<td>M</td>
<td>5.0</td>
<td>5.0</td>
<td>5.1</td>
</tr>
<tr>
<td>10</td>
<td>62</td>
<td>M</td>
<td>8.6</td>
<td>8.1</td>
<td>7.4</td>
</tr>
<tr>
<td>11</td>
<td>65</td>
<td>F</td>
<td>16.6</td>
<td>14.9</td>
<td>13.0</td>
</tr>
<tr>
<td>12</td>
<td>71</td>
<td>M</td>
<td>16.4</td>
<td>18.6</td>
<td>17.1</td>
</tr>
<tr>
<td>13</td>
<td>51</td>
<td>F</td>
<td>12.2</td>
<td>11.0</td>
<td>12.3</td>
</tr>
<tr>
<td>14</td>
<td>71</td>
<td>M</td>
<td>6.6</td>
<td>3.7</td>
<td>4.5</td>
</tr>
<tr>
<td>15</td>
<td>_b</td>
<td>M</td>
<td>9.9</td>
<td>10.7</td>
<td>11.7</td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>M</td>
<td>10.2</td>
<td>10.8</td>
<td>11.2</td>
</tr>
</tbody>
</table>

aAge for patient 15 is not available. F, female; M male.

aData are serum theophylline levels (mg/dl) in 16 patients with emphysema.

**DB4 EFFECT OF PROTEASE INHIBITORS ON PULMONARY ADMISSIONS**

**Background**

Protease inhibitors (PIs) are used in the treatment of infection by human immunodeficiency virus (HIV). A military hospital provides unlimited access to a full spectrum of HIV medications. Has this access reduced admissions due to secondary infections? In particular, has unlimited access to PIs reduced the annual admission rate of patients with pulmonary complications? The number of admissions for HIV, separated into patients with and those without pulmonary complications, were obtained for the four years (1992–1995) prior to access to PIs and for the two years (1996–1997) after PIs became available.

**Data**

<table>
<thead>
<tr>
<th></th>
<th>Patients with Pulmonary Complications</th>
<th>Patients without Pulmonary Complications</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before PIs were</td>
<td>194</td>
<td>291</td>
<td>485</td>
</tr>
<tr>
<td>available</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After PIs were</td>
<td>25</td>
<td>67</td>
<td>92</td>
</tr>
<tr>
<td>available</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>219</td>
<td>358</td>
<td>577</td>
</tr>
</tbody>
</table>
**DB5 EFFECT OF SILICONE IMPLANTS ON PLASMA SILICON**

**Background**

Silicone implants in women’s breasts occasionally rupture, releasing silicone into the body. A large number of women have received silicone breast implants, and some investigators believe that the presence of the implants raises plasma silicon levels, leading to side effects. A study was begun to test this belief. A method was developed for accurately measuring the silicon level in blood. For each of 30 women, plasma silicon levels (microg per gram dry weight) were taken prior to surgical placement of the implants. A post-surgery washout period was allowed, and plasma silicon levels were retaken.

**Data**

<table>
<thead>
<tr>
<th>Patient Number</th>
<th>Preoperative Level</th>
<th>Postoperative Level</th>
<th>Patient Number</th>
<th>Preoperative Level</th>
<th>Postoperative Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.15</td>
<td>0.21</td>
<td>16</td>
<td>0.28</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>0.13</td>
<td>0.24</td>
<td>17</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>0.39</td>
<td>0.10</td>
<td>18</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
<td>0.12</td>
<td>19</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>0.39</td>
<td>0.28</td>
<td>20</td>
<td>0.18</td>
<td>0.14</td>
</tr>
<tr>
<td>6</td>
<td>0.42</td>
<td>0.25</td>
<td>21</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>7</td>
<td>0.24</td>
<td>0.22</td>
<td>22</td>
<td>0.48</td>
<td>0.20</td>
</tr>
<tr>
<td>8</td>
<td>0.18</td>
<td>0.21</td>
<td>23</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td>9</td>
<td>0.26</td>
<td>0.22</td>
<td>24</td>
<td>0.22</td>
<td>0.18</td>
</tr>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.23</td>
<td>25</td>
<td>0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>11</td>
<td>0.10</td>
<td>0.22</td>
<td>26</td>
<td>0.19</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>0.11</td>
<td>0.24</td>
<td>27</td>
<td>0.32</td>
<td>0.26</td>
</tr>
<tr>
<td>13</td>
<td>0.19</td>
<td>0.45</td>
<td>28</td>
<td>0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>14</td>
<td>0.15</td>
<td>0.38</td>
<td>29</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>15</td>
<td>0.27</td>
<td>0.23</td>
<td>30</td>
<td>0.21</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**DB6 LASER REMOVAL OF TATTOOS AS RELATED TO TYPE OF INK USED**

**Background**

A frequently used component of tattoo ink is titanium. A dermatologist suspected that tattoos applied with titanium ink are more difficult to remove than those applied with other inks. Fifty patients wanting tattoo removal were tested for ink type, and laser removal was attempted. The numbers responding and not responding to the removal attempt are given here according to the type of ink used.
DB7 RELATION OF BONE DENSITY TO INCIDENCE OF FEMORAL NECK STRESS FRACTURES

Background

Femoral neck stress fractures occur in young people engaged in sports, combat training, etc. Such fractures have a relatively high rate of poor healing, which in many cases leaves the patient somewhat impaired. Many such events could be avoided if risk factors, i.e., those characteristics that predict a higher probability of fracture, could be identified. One potential risk factor is bone density, measured in mg/ml. A large database (several thousand) of bone density measures of normal healthy people has been compiled at the University of California – San Francisco (UCSF). Norms were established for age and sex. Do people with femoral neck stress fractures tend to have lower bone density measures than healthy individuals? Bone density was measured for 18 young people with femoral neck stress fractures.

Data

The following data include age, sex, bone density measure, and UCSF norm-value for that age and sex.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Sex</th>
<th>Age (years)</th>
<th>Bone density (mg/ml)</th>
<th>UCSF norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>22</td>
<td>139.6</td>
<td>185</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>24</td>
<td>167.2</td>
<td>188</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>39</td>
<td>140.1</td>
<td>181</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>18</td>
<td>163.6</td>
<td>225</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>18</td>
<td>140.0</td>
<td>225</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>18</td>
<td>177.4</td>
<td>225</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>18</td>
<td>147.0</td>
<td>225</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>19</td>
<td>164.9</td>
<td>223</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>19</td>
<td>155.8</td>
<td>223</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>19</td>
<td>126.0</td>
<td>223</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>19</td>
<td>163.4</td>
<td>223</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>21</td>
<td>194.9</td>
<td>219</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>21</td>
<td>199.0</td>
<td>219</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>21</td>
<td>164.6</td>
<td>219</td>
</tr>
<tr>
<td>15</td>
<td>M</td>
<td>22</td>
<td>140.8</td>
<td>218</td>
</tr>
</tbody>
</table>
DB8 COMPARING TWO TYPES OF ASSAY ON THE EFFECT OF GAG ON THE BLADDER SURFACE\textsuperscript{25}

**Background**

Glycosaminoglycans (GAGs) alters the adherence properties of the bladder surface. Four spectrophotometric readings (millimoles/kg) on the GAG levels in tissues of normal bladders from cadavers exposed to GAGs were taken by each of two types of assay. (Mean instrument error has been established as 0.002.)

**Data**

| Type 1 assay: | 0.74, 0.37, 0.26, 0.20 |
| Type 2 assay: | 0.51, 0.53, 0.42, 0.29 |

DB9 PREDICTION OF GROWTH FACTORS BY PLATELET COUNTS\textsuperscript{26}

**Background**

There are two types of platelet gel growth factors: platelet-derived growth factor (PDGF), used primarily in cutaneous wound healing, and transforming growth factor (TGF), used primarily to promote bone healing. Users wish to know the level of a growth factor in platelet gel, but the assay is costly and time consuming. Can a simple platelet count predict the level of a growth factor in platelet gel? Platelet counts and levels of PDGF and TGF in platelet gel were measured for 20 patients. Because the large numbers in the data are a little tedious to calculate, correlation coefficients are given to assist the student.

**Data**

<table>
<thead>
<tr>
<th>Variable Pair</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platelet Count with gel PDGF</td>
<td>0.926</td>
</tr>
<tr>
<td>Platelet Count with gel TGF</td>
<td>0.979</td>
</tr>
<tr>
<td>Gel PDGF with gel TGF</td>
<td>0.944</td>
</tr>
</tbody>
</table>
DB10 TESTS OF RECOVERY AFTER SURGERY ON HAMSTRINGS OR QUADRICEPS

Background

Tests of strength and control following surgery on hamstring or quadriceps muscles or tendons have had varying success in indicating the quality of recovery. Two proposed tests are the time to perform and the distance covered in a triple hop along a marked line, contrasted between the healthy leg and the operated leg. Strength of the operated muscle in the two legs has not been shown to be different for 8 post-operative patients. The question is whether or not the proposed tests will detect a difference.

Data

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Time to perform (sec)</th>
<th>Distance covered (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Operated leg</td>
<td>Non-operated leg</td>
</tr>
<tr>
<td>1</td>
<td>2.00</td>
<td>1.89</td>
</tr>
<tr>
<td>2</td>
<td>2.82</td>
<td>2.49</td>
</tr>
<tr>
<td>3</td>
<td>2.62</td>
<td>2.39</td>
</tr>
<tr>
<td>4</td>
<td>2.97</td>
<td>3.13</td>
</tr>
<tr>
<td>5</td>
<td>2.64</td>
<td>2.86</td>
</tr>
</tbody>
</table>
DB11 SURVIVAL OF MALARIAL RATS TREATED WITH HEMOGLOBIN, RBCS, OR A PLACEBO

Background

The malaria parasite affects the transmission of oxygen to the brain. The infusion of red blood cells (RBCs) in a malaria-infected body slows deterioration, giving the body more time to receive the benefit of antimalarial medication. Does the infusion of hemoglobin provide the same benefit? Three groups of 100 rats each were infected with a rodent malaria. One group was given hemoglobin, a second was given RBCs, and the third was given hetastarch (an IV fluid with protein). Survival numbers (and, coincidentally, percent) at the outset (day 0) and over 10 succeeding days were recorded.

Data

<table>
<thead>
<tr>
<th>Day Number</th>
<th>Hemoglobin</th>
<th>Red Blood Cells</th>
<th>Hetastarch</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>92</td>
<td>83</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>92</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
<td>83</td>
<td>58</td>
</tr>
<tr>
<td>5</td>
<td>69</td>
<td>83</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>83</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>83</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>38</td>
<td>75</td>
<td>33</td>
</tr>
<tr>
<td>9</td>
<td>31</td>
<td>58</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>58</td>
<td>33</td>
</tr>
</tbody>
</table>

DB12 IDENTIFICATION OF RISK FACTORS FOR DEATH FOLLOWING CARINAL RESECTION

Background

Resection of the tracheal carina is a dangerous procedure. From 134 cases, the following data subset is given from the variables that were recorded: age at surgery (years), prior tracheal surgery (no = 0, yes = 1), extent of the resection (cm), intubation...
required at end of surgery (no = 0, yes = 1), and patient death (no = 0, yes = 1).
Because the data set is large for an exercise, some descriptive statistics are given to assist the student.

**Data**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>47.84</td>
<td>15.78</td>
<td>Extent (cm)</td>
<td>2.96</td>
<td>1.24</td>
</tr>
<tr>
<td>Age, died = 0</td>
<td>48.05</td>
<td>16.01</td>
<td>Extent, died = 0</td>
<td>2.82</td>
<td>1.21</td>
</tr>
<tr>
<td>Age, died = 1</td>
<td>46.41</td>
<td>14.46</td>
<td>Extent, died = 1</td>
<td>3.96</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Age at surgery</th>
<th>Prior surgery</th>
<th>Extent of resection</th>
<th>Intubated</th>
<th>Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>66</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>41</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>44</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>67</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>54</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>36</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>63</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>46</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>80</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>47</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>41</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>62</td>
<td>1</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>53</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>62</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>22</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>65</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>73</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>38</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>55</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>56</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>66</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>36</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>48</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>49</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>29</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>52</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>39</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>33</td>
<td>27</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>55</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>62</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>36</td>
<td>37</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>37</td>
<td>26</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>57</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>60</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>76</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>28</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>25</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>62</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>63</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>34</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>56</td>
<td>0</td>
<td>2.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>47</td>
<td>51</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>42</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>54</td>
<td>1</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>62</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51</td>
<td>47</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>52</td>
<td>34</td>
<td>1</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>53</td>
<td>52</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>49</td>
<td>0</td>
<td>2.5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>69</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>56</td>
<td>32</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>57</td>
<td>19</td>
<td>1</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>58</td>
<td>38</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>59</td>
<td>55</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>40</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>61</td>
<td>63</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>62</td>
<td>51</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>63</td>
<td>8</td>
<td>1</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>54</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>58</td>
<td>1</td>
<td>2.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>66</td>
<td>48</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>67</td>
<td>19</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>68</td>
<td>25</td>
<td>0</td>
<td>2.5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>69</td>
<td>49</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>27</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>47</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>72</td>
<td>58</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>73</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>74</td>
<td>41</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>75</td>
<td>18</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>76</td>
<td>31</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>77</td>
<td>52</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>78</td>
<td>72</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Row</td>
<td>First Column</td>
<td>Second Column</td>
<td>Third Column</td>
<td>Fourth Column</td>
<td>Fifth Column</td>
</tr>
<tr>
<td>-----</td>
<td>--------------</td>
<td>---------------</td>
<td>--------------</td>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>79</td>
<td>62</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>38</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>59</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>82</td>
<td>62</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>83</td>
<td>33</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>84</td>
<td>62</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>85</td>
<td>25</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>86</td>
<td>46</td>
<td>0</td>
<td>3.2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>87</td>
<td>52</td>
<td>1</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>88</td>
<td>43</td>
<td>0</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>89</td>
<td>55</td>
<td>0</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90</td>
<td>66</td>
<td>1</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>91</td>
<td>30</td>
<td>0</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>92</td>
<td>41</td>
<td>0</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>93</td>
<td>53</td>
<td>0</td>
<td>3.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>94</td>
<td>70</td>
<td>0</td>
<td>3.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>95</td>
<td>61</td>
<td>1</td>
<td>3.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>96</td>
<td>28</td>
<td>1</td>
<td>3.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>97</td>
<td>60</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>98</td>
<td>46</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>99</td>
<td>52</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>17</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>61</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>102</td>
<td>46</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>103</td>
<td>51</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>104</td>
<td>71</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>105</td>
<td>24</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>106</td>
<td>67</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>107</td>
<td>68</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>108</td>
<td>56</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>109</td>
<td>39</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>110</td>
<td>60</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>111</td>
<td>62</td>
<td>1</td>
<td>4.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>112</td>
<td>57</td>
<td>1</td>
<td>4.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>113</td>
<td>60</td>
<td>0</td>
<td>4.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>114</td>
<td>32</td>
<td>0</td>
<td>4.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>115</td>
<td>34</td>
<td>0</td>
<td>4.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>116</td>
<td>48</td>
<td>1</td>
<td>4.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>117</td>
<td>69</td>
<td>0</td>
<td>4.5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>118</td>
<td>27</td>
<td>1</td>
<td>4.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>119</td>
<td>66</td>
<td>0</td>
<td>4.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>120</td>
<td>43</td>
<td>0</td>
<td>4.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>121</td>
<td>60</td>
<td>0</td>
<td>4.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>122</td>
<td>66</td>
<td>0</td>
<td>4.8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>123</td>
<td>57</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>124</td>
<td>71</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
DB13 QUALITY TEST ON WARFARIN INTERNATIONAL NORMALIZED RATIO VALUES

Background

Warfarin is a blood anticoagulant used to treat certain cardiac cases. It is measured in International Normalized Ratio units (INRs). As small doses are ineffective and large are quite dangerous (in very strong doses, it is used as a rodenticide), maintaining the INR level within range is crucial. In-clinic readings are immediately available, while the slower laboratory readings are the gold standard. To test the quality of its clinic tests, a Coumadin Clinic obtained both readings on all 104 patients treated during a certain period. We should like to compare them. A patient is “In Range” if $2 < \text{INR} \leq 3$.

Data

<table>
<thead>
<tr>
<th>Patient</th>
<th>INR Clinic</th>
<th>INR Laboratory</th>
<th>Difference</th>
<th>INR Clinic</th>
<th>INR Laboratory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.46</td>
<td>2.30</td>
<td>0.16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2.86</td>
<td>3.20</td>
<td>-0.34</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2.63</td>
<td>2.40</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2.26</td>
<td>2.00</td>
<td>0.26</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2.30</td>
<td>2.20</td>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2.16</td>
<td>1.80</td>
<td>0.36</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>2.24</td>
<td>2.40</td>
<td>-0.16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2.38</td>
<td>1.90</td>
<td>0.48</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>1.50</td>
<td>2.00</td>
<td>-0.50</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1.60</td>
<td>1.40</td>
<td>0.20</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>11</td>
<td>1.30</td>
<td>1.80</td>
<td>-0.50</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>12</td>
<td>3.18</td>
<td>2.90</td>
<td>0.28</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>2.20</td>
<td>2.00</td>
<td>0.20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1.60</td>
<td>1.50</td>
<td>0.10</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>15</td>
<td>2.40</td>
<td>2.20</td>
<td>0.20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>2.30</td>
<td>1.90</td>
<td>0.40</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>17</td>
<td>2.60</td>
<td>2.30</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>1.70</td>
<td>1.30</td>
<td>0.40</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>19</td>
<td>3.10</td>
<td>3.00</td>
<td>0.10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>2.40</td>
<td>2.00</td>
<td>0.40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>2.33</td>
<td>1.90</td>
<td>0.43</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>22</td>
<td>4.01</td>
<td>3.70</td>
<td>0.31</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>2.45</td>
<td>2.30</td>
<td>0.15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>2.10</td>
<td>1.80</td>
<td>0.30</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>25</td>
<td>2.86</td>
<td>2.70</td>
<td>0.16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>2.93</td>
<td>3.40</td>
<td>-0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>2.00</td>
<td>1.70</td>
<td>0.30</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>28</td>
<td>1.92</td>
<td>1.60</td>
<td>0.32</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>29</td>
<td>2.34</td>
<td>2.10</td>
<td>0.24</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>2.40</td>
<td>2.20</td>
<td>0.20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>2.35</td>
<td>1.80</td>
<td>0.55</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>32</td>
<td>2.47</td>
<td>2.20</td>
<td>0.27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>2.44</td>
<td>2.50</td>
<td>-0.06</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>1.29</td>
<td>1.40</td>
<td>-0.11</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>35</td>
<td>1.71</td>
<td>1.50</td>
<td>0.21</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>36</td>
<td>1.80</td>
<td>1.70</td>
<td>0.10</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>37</td>
<td>2.20</td>
<td>2.10</td>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>38</td>
<td>1.91</td>
<td>1.90</td>
<td>0.01</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>39</td>
<td>1.29</td>
<td>1.10</td>
<td>0.19</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>40</td>
<td>2.10</td>
<td>2.60</td>
<td>-0.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>2.68</td>
<td>2.40</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>3.41</td>
<td>2.80</td>
<td>0.61</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>1.92</td>
<td>1.50</td>
<td>0.42</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>44</td>
<td>2.85</td>
<td>3.30</td>
<td>-0.45</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>3.41</td>
<td>3.50</td>
<td>-0.09</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>2.70</td>
<td>2.60</td>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>47</td>
<td>2.43</td>
<td>2.10</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>1.21</td>
<td>1.10</td>
<td>0.11</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>49</td>
<td>3.14</td>
<td>3.30</td>
<td>-0.16</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>2.50</td>
<td>2.50</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51</td>
<td>3.00</td>
<td>3.10</td>
<td>-0.10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>52</td>
<td>2.50</td>
<td>2.50</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>53</td>
<td>1.90</td>
<td>1.70</td>
<td>0.20</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>54</td>
<td>2.40</td>
<td>2.00</td>
<td>0.40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>3.00</td>
<td>3.70</td>
<td>-0.70</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>56</td>
<td>2.08</td>
<td>1.90</td>
<td>0.18</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>57</td>
<td>2.66</td>
<td>2.50</td>
<td>0.16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>58</td>
<td>2.00</td>
<td>1.70</td>
<td>0.30</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>59</td>
<td>2.38</td>
<td>2.10</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>2.69</td>
<td>2.40</td>
<td>0.29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>61</td>
<td>2.28</td>
<td>1.90</td>
<td>0.38</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>62</td>
<td>3.23</td>
<td>3.30</td>
<td>-0.07</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>63</td>
<td>2.47</td>
<td>2.20</td>
<td>0.27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>2.14</td>
<td>2.10</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>1.93</td>
<td>1.90</td>
<td>0.03</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>66</td>
<td>2.71</td>
<td>2.50</td>
<td>0.21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>67</td>
<td>2.14</td>
<td>1.90</td>
<td>0.24</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>68</td>
<td>1.99</td>
<td>1.90</td>
<td>0.09</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>69</td>
<td>2.95</td>
<td>2.40</td>
<td>0.55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>2.15</td>
<td>2.00</td>
<td>0.15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>2.69</td>
<td>3.20</td>
<td>-0.51</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>72</td>
<td>2.98</td>
<td>3.40</td>
<td>-0.42</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>2.01</td>
<td>1.60</td>
<td>0.41</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>74</td>
<td>3.35</td>
<td>3.40</td>
<td>-0.05</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>2.29</td>
<td>2.00</td>
<td>0.29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>76</td>
<td>3.31</td>
<td>3.10</td>
<td>0.21</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>77</td>
<td>3.00</td>
<td>3.40</td>
<td>-0.40</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>78</td>
<td>2.20</td>
<td>2.00</td>
<td>0.20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>79</td>
<td>2.15</td>
<td>1.84</td>
<td>0.31</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>80</td>
<td>2.19</td>
<td>2.20</td>
<td>-0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>2.07</td>
<td>1.70</td>
<td>0.37</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>82</td>
<td>2.09</td>
<td>1.90</td>
<td>0.19</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>83</td>
<td>2.75</td>
<td>2.80</td>
<td>-0.05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>84</td>
<td>3.50</td>
<td>3.50</td>
<td>0.00</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>85</td>
<td>2.50</td>
<td>3.10</td>
<td>-0.60</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>86</td>
<td>2.23</td>
<td>1.90</td>
<td>0.33</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>87</td>
<td>2.15</td>
<td>1.90</td>
<td>0.25</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>88</td>
<td>2.26</td>
<td>1.90</td>
<td>0.36</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>89</td>
<td>2.94</td>
<td>2.60</td>
<td>0.34</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>90</td>
<td>2.92</td>
<td>2.30</td>
<td>0.62</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>91</td>
<td>2.86</td>
<td>2.80</td>
<td>0.06</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>92</td>
<td>2.28</td>
<td>1.90</td>
<td>0.38</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>93</td>
<td>2.65</td>
<td>2.60</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>94</td>
<td>2.62</td>
<td>2.50</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>95</td>
<td>2.10</td>
<td>2.00</td>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>96</td>
<td>2.90</td>
<td>3.10</td>
<td>-0.20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>97</td>
<td>1.50</td>
<td>1.30</td>
<td>0.20</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>98</td>
<td>3.50</td>
<td>3.40</td>
<td>0.10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>99</td>
<td>1.50</td>
<td>1.30</td>
<td>0.20</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>100</td>
<td>2.50</td>
<td>2.90</td>
<td>-0.40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>2.43</td>
<td>2.20</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>102</td>
<td>2.68</td>
<td>2.40</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>103</td>
<td>2.05</td>
<td>1.70</td>
<td>0.35</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>104</td>
<td>1.65</td>
<td>1.40</td>
<td>0.25</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>
DB14 EXHALED NITRIC OXIDE AS AN INDICATOR OF EXERCISE-INDUCED BRONCHOCONSTRICTION

Background

Exercise-induced bronchoconstriction (EIB) is an indicator of asthma. If exhaled nitric oxide (eNO) can flag EIB, then it may be used in the diagnosis of incipient asthma in a subject prior to engaging in vigorous exercise. In this experiment, eNO was measured on 38 subjects, who then underwent vigorous exercise. Six were found through a decrease in forced expiratory volume in 1 second (FEV₁) to have EIB (EIB = 1), whereas 32 did not have EIB (EIB = 0). After 5 and 20 minutes, eNO was again measured. Does the change in eNO at 5 minutes or at 20 minutes flag the EIB? (Rel hum is relative humidity in %. Age is in years. Sex = 1 is male, 2 is female.)

Data

<table>
<thead>
<tr>
<th>Patient Number</th>
<th>EIB</th>
<th>0 Before</th>
<th>5 Minutes</th>
<th>20 Minutes</th>
<th>5 Minutes</th>
<th>20 Minutes</th>
<th>Rel Hum (%)</th>
<th>Age (yrs)</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>57.7</td>
<td>36.9</td>
<td>38.5</td>
<td>20.8</td>
<td>19.2</td>
<td>38</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>51.3</td>
<td>53.0</td>
<td>44.8</td>
<td>−1.7</td>
<td>6.5</td>
<td>42</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>21.8</td>
<td>17.8</td>
<td>20.6</td>
<td>4.0</td>
<td>1.2</td>
<td>36</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>43.5</td>
<td>48.9</td>
<td>51.4</td>
<td>−5.4</td>
<td>−7.9</td>
<td>43</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>10.5</td>
<td>10.1</td>
<td>11.8</td>
<td>0.4</td>
<td>−1.3</td>
<td>51</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>11.6</td>
<td>9.2</td>
<td>10.0</td>
<td>2.4</td>
<td>1.6</td>
<td>39</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>15.8</td>
<td>13.4</td>
<td>15.9</td>
<td>2.4</td>
<td>−0.1</td>
<td>28</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>25.0</td>
<td>24.0</td>
<td>37.5</td>
<td>1.0</td>
<td>−12.5</td>
<td>33</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>31.6</td>
<td>28.3</td>
<td>37.4</td>
<td>3.3</td>
<td>−5.8</td>
<td>34</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>68.0</td>
<td>50.2</td>
<td>53.9</td>
<td>17.8</td>
<td>14.1</td>
<td>44</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>41.2</td>
<td>30.1</td>
<td>38.7</td>
<td>11.1</td>
<td>2.5</td>
<td>51</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>15.8</td>
<td>13.4</td>
<td>15.9</td>
<td>2.4</td>
<td>−0.1</td>
<td>28</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>4.8</td>
<td>4.7</td>
<td>4.4</td>
<td>0.1</td>
<td>0.4</td>
<td>49</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>71.0</td>
<td>49.9</td>
<td>55.9</td>
<td>21.1</td>
<td>15.1</td>
<td>33</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>36.4</td>
<td>28.7</td>
<td>42.1</td>
<td>7.7</td>
<td>−5.7</td>
<td>33</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>35.8</td>
<td>28.6</td>
<td>31.9</td>
<td>7.2</td>
<td>3.9</td>
<td>32</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>65.2</td>
<td>49.0</td>
<td>52.9</td>
<td>16.2</td>
<td>12.3</td>
<td>34</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>47.1</td>
<td>30.5</td>
<td>39.0</td>
<td>16.6</td>
<td>8.1</td>
<td>28</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>21.2</td>
<td>20.9</td>
<td>19.6</td>
<td>0.3</td>
<td>1.6</td>
<td>33</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>20.5</td>
<td>24.4</td>
<td>23.8</td>
<td>−3.9</td>
<td>−3.3</td>
<td>37</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>19.4</td>
<td>14.1</td>
<td>16.5</td>
<td>5.3</td>
<td>2.9</td>
<td>33</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>9.0</td>
<td>7.0</td>
<td>4.0</td>
<td>2.0</td>
<td>5.0</td>
<td>34</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>7.9</td>
<td>8.9</td>
<td>9.7</td>
<td>−1.0</td>
<td>−1.8</td>
<td>34</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>21.0</td>
<td>22.0</td>
<td>25.0</td>
<td>−1.0</td>
<td>−4.0</td>
<td>44</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>19.0</td>
<td>11.9</td>
<td>16.2</td>
<td>7.1</td>
<td>2.8</td>
<td>42</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>12.0</td>
<td>13.2</td>
<td>12.0</td>
<td>−1.2</td>
<td>0.0</td>
<td>51</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>10.8</td>
<td>9.8</td>
<td>15.2</td>
<td>1.0</td>
<td>−4.4</td>
<td>49</td>
<td>33</td>
<td>2</td>
</tr>
</tbody>
</table>
DB15 COMPARISON OF KIDNEY COOLING METHODS USED TO PROLONG SURGICAL TIME WINDOW

Background

In kidney surgery, the length of operation often causes tissue necrosis unless the kidney is quickly cooled to below 15°C. The traditional method is to pack the kidney in crushed ice as soon as it is exposed. The question arose: Could the kidney be cooled as well or better by infusing it with cold saline solution? Six anesthetized pigs were opened on both sides. One randomly chosen kidney was cooled internally with cold saline (treatment 1), the other, externally with ice (treatment 2). The temperature, measured at both surface (depth 1) and depth (depth 2) was recorded prior to cooling (baseline, time 0) and at 5, 10, and 15 minutes after start of cooling.

Data

<table>
<thead>
<tr>
<th>Observation Number</th>
<th>Pig Number</th>
<th>Time (minutes)</th>
<th>Treatment</th>
<th>Depth</th>
<th>Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>36.7</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>33.2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>24.5</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>26.6</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>35.9</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>12.7</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>6.9</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>3.9</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>36.8</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>33.4</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>24.8</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>36.1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>21.1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>19.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>8.7</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>34.6</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>31.5</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>30.7</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>37.4</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>10.6</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>8.5</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>3.7</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>34.7</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>31.7</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>28</td>
<td>2</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>27.8</td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>37.6</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>24.6</td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>14.5</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>12.9</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>32.6</td>
</tr>
<tr>
<td>34</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>32.3</td>
</tr>
<tr>
<td>35</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>30.8</td>
</tr>
<tr>
<td>36</td>
<td>3</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>29.9</td>
</tr>
<tr>
<td>37</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>35.8</td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>9.6</td>
</tr>
<tr>
<td>39</td>
<td>3</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>41</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>33.7</td>
</tr>
<tr>
<td>42</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>43</td>
<td>3</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>28.7</td>
</tr>
<tr>
<td>44</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>26.7</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>46</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>17.4</td>
</tr>
<tr>
<td>47</td>
<td>3</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>14.4</td>
</tr>
<tr>
<td>48</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>12.2</td>
</tr>
<tr>
<td>49</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>37.1</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>51</td>
<td>4</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>33.1</td>
</tr>
<tr>
<td>52</td>
<td>4</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>33.2</td>
</tr>
<tr>
<td>53</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>35.7</td>
</tr>
<tr>
<td>54</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>14.3</td>
</tr>
<tr>
<td>55</td>
<td>4</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>4.7</td>
</tr>
<tr>
<td>56</td>
<td>4</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>4.7</td>
</tr>
<tr>
<td>57</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>37.3</td>
</tr>
<tr>
<td>58</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>34.9</td>
</tr>
<tr>
<td>59</td>
<td>4</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>33.1</td>
</tr>
<tr>
<td>60</td>
<td>4</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>32.6</td>
</tr>
<tr>
<td>61</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>35.9</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>21.1</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>------</td>
</tr>
<tr>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td></td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>7.2</td>
</tr>
<tr>
<td>64</td>
<td></td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>10.5</td>
</tr>
<tr>
<td>65</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>66</td>
<td></td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>28.7</td>
</tr>
<tr>
<td>67</td>
<td></td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>68</td>
<td></td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>22.7</td>
</tr>
<tr>
<td>69</td>
<td></td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>37.3</td>
</tr>
<tr>
<td>70</td>
<td></td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>14.5</td>
</tr>
<tr>
<td>71</td>
<td></td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>6.7</td>
</tr>
<tr>
<td>72</td>
<td></td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>4.1</td>
</tr>
<tr>
<td>73</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>28.4</td>
</tr>
<tr>
<td>74</td>
<td></td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>27.5</td>
</tr>
<tr>
<td>75</td>
<td></td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>24.1</td>
</tr>
<tr>
<td>76</td>
<td></td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>20.2</td>
</tr>
<tr>
<td>77</td>
<td></td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>37.4</td>
</tr>
<tr>
<td>78</td>
<td></td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>25.5</td>
</tr>
<tr>
<td>79</td>
<td></td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>10.2</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>10.6</td>
</tr>
<tr>
<td>81</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>36.4</td>
</tr>
<tr>
<td>82</td>
<td></td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>33.1</td>
</tr>
<tr>
<td>83</td>
<td></td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>84</td>
<td></td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>28.6</td>
</tr>
<tr>
<td>85</td>
<td></td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>34.1</td>
</tr>
<tr>
<td>86</td>
<td></td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>15.4</td>
</tr>
<tr>
<td>87</td>
<td></td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>8.8</td>
</tr>
<tr>
<td>88</td>
<td></td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>6.1</td>
</tr>
<tr>
<td>89</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>36.5</td>
</tr>
<tr>
<td>90</td>
<td></td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>91</td>
<td></td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>29.3</td>
</tr>
<tr>
<td>92</td>
<td></td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>28.5</td>
</tr>
<tr>
<td>93</td>
<td></td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>34.2</td>
</tr>
<tr>
<td>94</td>
<td></td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>16.6</td>
</tr>
<tr>
<td>95</td>
<td></td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>16.6</td>
</tr>
<tr>
<td>96</td>
<td></td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>9.6</td>
</tr>
</tbody>
</table>

**DB16 WORD REPETITIONS AMONG COGNITIVELY NORMAL AND DEMENTED PATIENTS**

A clinical diagnosis of dementia is generally performed by a trained neuropsychologists and is based upon a collection of standardized neuropsychological tests as well as subjective complaints from the patient as well as family and caregivers. One easy to administer neuropsychological test that is routinely given to subjects is the Controlled
Oral Word Association Test (COWAT)-FAS test in which patients are asked to list as many different words starting with the letters F, A, or S in a 60 second time period.

The Alzheimer's Disease Research Center (ADRC) at the University of California Irvine (UCI) campus routinely follows a cohort of volunteers over time and performs annual neuropsychological testing on the participants with the hopes of better understanding the early predictors and consequences of cognitive impairment, dementia, and Alzheimer's disease. Participants range from cognitively normal to clinically diagnosed dementia.

One such test that is performed on participants in the UCI ADRC is the COWAT-FAS. Data are available for one visit on each of N = 182 participants, comprised of 80 cognitively normal participants, 71 participants diagnosed with mild cognitive impairment (MCI), and 31 participants diagnosed with dementia. Variables available in the data set are: each participant's unique id (id), total number of repeated words and words stated on the COWAT-FAS (fas.tr and fas.ts, respectively), the dementia diagnosis category for the participant (demcat, recorded as 0 = normal, 1 = MCI, and 2 = dementia), total years of education (educ), age in years (age), and the sex of the participant (sex, recorded as 0 for males and 1 for females). The first six rows of the data set are given below.

<table>
<thead>
<tr>
<th>id</th>
<th>fas.tr</th>
<th>fas.ts</th>
<th>demcat</th>
<th>educ</th>
<th>age</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>37</td>
<td>0</td>
<td>16</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>48</td>
<td>0</td>
<td>12</td>
<td>86</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>62</td>
<td>0</td>
<td>14</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>61</td>
<td>0</td>
<td>18</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>42</td>
<td>0</td>
<td>16</td>
<td>77</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>47</td>
<td>0</td>
<td>18</td>
<td>85</td>
<td>1</td>
</tr>
</tbody>
</table>

A clinical diagnosis of Alzheimer's disease relies upon neuropsychological testing, imaging, and subjective clinical judgment including established guidelines for diagnosis decided by large interdisciplinary research groups and professional organizations. Built into these guidelines is a pattern of deficits that include dysfunction in certain cognitive areas which must be documented by accepted neuropsychological testing as well as documented decline over time. However, a typical problem of longitudinal cognitive testing is that repeated testing may yield increases in score upon subsequent testing times. These increases over time due to repeated testing are known as retest effects and may mask the very effect that is supposed to provide evidence of illness.
Longitudinal data have been collected on \( N = 10,900 \) unique subjects, with each subject having at least two cognitive tests and a maximum of 12 visits. Cognitive testing visits were scheduled to occur approximately once a year, however there is variation in the times at which subjects returned for visits. These data contain information on the Logical Memory test at each visit. The Logical Memory test is a subset of the Wechsler Memory Scale, with total scores ranging from 0 to 25. For this test, the items/questions do not change from one visit to the next. Because of this it is of interest to determine if there exists retest effects in which patients scores may actually increase with increased number of visits (despite increasing age between visits).

Available data are diagnosis (normal, MCI, AD) visit number, and the Logical Memory test score at each visit. The table below summarizes the mean and standard deviation of the logical memory test score at the first visit for each diagnosis group. This file is for educational use only. It is a limited excerpt from the National Alzheimer’s Coordinating Center (NACC) database. Further distribution or reproduction of this file, research analysis for publication, or any other purpose beyond the specific examples provided in [book title] are strictly prohibited. Persons wishing to access the NACC database for scientific research must visit the NACC website, www.alz.washington.edu to apply for access.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Normal (( N = 6,583 ))</th>
<th>MCI (( N = 2,723 ))</th>
<th>AD (( N = 1,594 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Score (Logical Memory)</td>
<td>13.62 (3.85)</td>
<td>9.66 (4.2)</td>
<td>6.21 (3.92)</td>
</tr>
</tbody>
</table>

**REFERENCES**


11. Missing sources. The sources of a few examples could not be found despite a strong effort to locate them. Such data that could not be referenced were slightly altered so as not to reflect on any investigator later appearing.


33. La Ferla, F. *Clinical Data, Alzheimer's Disease Research Center;* University of California: Irvine, 2019.

### Table I Normal distribution.

<table>
<thead>
<tr>
<th>z (no. std. deviations to right of mean)</th>
<th>One-tailed applications</th>
<th>Two-tailed applications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One-tailed $\alpha$ (area in right tail)</td>
<td>$1 - \alpha$ (area except right tail)</td>
</tr>
<tr>
<td>0</td>
<td>.500</td>
<td>.500</td>
</tr>
<tr>
<td>.10</td>
<td>.460</td>
<td>.540</td>
</tr>
<tr>
<td>.20</td>
<td>.421</td>
<td>.379</td>
</tr>
<tr>
<td>.30</td>
<td>.382</td>
<td>.618</td>
</tr>
<tr>
<td>.40</td>
<td>.345</td>
<td>.655</td>
</tr>
<tr>
<td>.50</td>
<td>.308</td>
<td>.692</td>
</tr>
<tr>
<td>.60</td>
<td>.274</td>
<td>.726</td>
</tr>
<tr>
<td>.70</td>
<td>.242</td>
<td>.758</td>
</tr>
<tr>
<td>.80</td>
<td>.212</td>
<td>.788</td>
</tr>
<tr>
<td>.90</td>
<td>.184</td>
<td>.816</td>
</tr>
<tr>
<td>1.00</td>
<td>.159</td>
<td>.841</td>
</tr>
<tr>
<td>1.10</td>
<td>.136</td>
<td>.864</td>
</tr>
</tbody>
</table>

(Continued)
Table I (Continued)

<table>
<thead>
<tr>
<th>$z$ (no. std. deviations to right of mean)</th>
<th>One-tailed applications</th>
<th>Two-tailed applications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$ (area in right tail)</td>
<td>$1 - \alpha$ (area except right tail)</td>
</tr>
<tr>
<td>1.20</td>
<td>.115</td>
<td>.885</td>
</tr>
<tr>
<td>1.281</td>
<td>.100</td>
<td>.900</td>
</tr>
<tr>
<td>1.30</td>
<td>.097</td>
<td>.903</td>
</tr>
<tr>
<td>1.40</td>
<td>.081</td>
<td>.919</td>
</tr>
<tr>
<td>1.50</td>
<td>.067</td>
<td>.933</td>
</tr>
<tr>
<td>1.60</td>
<td>.055</td>
<td>.945</td>
</tr>
<tr>
<td>1.645</td>
<td>.050</td>
<td>.950</td>
</tr>
<tr>
<td>1.70</td>
<td>.045</td>
<td>.955</td>
</tr>
<tr>
<td>1.80</td>
<td>.036</td>
<td>.964</td>
</tr>
<tr>
<td>1.90</td>
<td>.029</td>
<td>.971</td>
</tr>
<tr>
<td>1.960</td>
<td>.025</td>
<td>.975</td>
</tr>
<tr>
<td>2.00</td>
<td>.023</td>
<td>.977</td>
</tr>
<tr>
<td>2.10</td>
<td>.018</td>
<td>.982</td>
</tr>
<tr>
<td>2.20</td>
<td>.014</td>
<td>.986</td>
</tr>
<tr>
<td>2.30</td>
<td>.011</td>
<td>.989</td>
</tr>
<tr>
<td>2.326</td>
<td>.010</td>
<td>.990</td>
</tr>
<tr>
<td>2.40</td>
<td>.008</td>
<td>.992</td>
</tr>
<tr>
<td>2.50</td>
<td>.006</td>
<td>.994</td>
</tr>
<tr>
<td>2.576</td>
<td>.005</td>
<td>.995</td>
</tr>
<tr>
<td>2.60</td>
<td>.0047</td>
<td>.9953</td>
</tr>
<tr>
<td>2.70</td>
<td>.0035</td>
<td>.9965</td>
</tr>
<tr>
<td>2.80</td>
<td>.0026</td>
<td>.9974</td>
</tr>
<tr>
<td>2.90</td>
<td>.0019</td>
<td>.9981</td>
</tr>
<tr>
<td>3.00</td>
<td>.0013</td>
<td>.9987</td>
</tr>
<tr>
<td>3.100</td>
<td>.0010</td>
<td>.9990</td>
</tr>
<tr>
<td>3.20</td>
<td>.0007</td>
<td>.9996</td>
</tr>
<tr>
<td>3.300</td>
<td>.0005</td>
<td>.9995</td>
</tr>
<tr>
<td>3.40</td>
<td>.0003</td>
<td>.9997</td>
</tr>
<tr>
<td>3.50</td>
<td>.0002</td>
<td>.9998</td>
</tr>
</tbody>
</table>

*For selected distances ($z$) to the right of the mean are given (1) one-tailed $\alpha$, the area under the curve in the positive tail; (2) one-tailed $1 - \alpha$, the area under all except the tail; (3) two-tailed $\alpha$, the areas combined for both positive and negative tails; and (4) two-tailed $1 - \alpha$, the area under all except the two tails. Entries for the most commonly used areas are in italics.*
### Table II: \( t \) Distribution.\(^a\)

<table>
<thead>
<tr>
<th>( df = 2 )</th>
<th>( 1.886 )</th>
<th>( 2.920 )</th>
<th>( 4.303 )</th>
<th>( 6.965 )</th>
<th>( 9.225 )</th>
<th>( 22.327 )</th>
<th>( 31.598 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 3 )</td>
<td>( 1.638 )</td>
<td>( 2.353 )</td>
<td>( 3.182 )</td>
<td>( 4.541 )</td>
<td>( 5.841 )</td>
<td>( 10.215 )</td>
<td>( 12.941 )</td>
</tr>
<tr>
<td>( 4 )</td>
<td>( 1.533 )</td>
<td>( 2.132 )</td>
<td>( 2.776 )</td>
<td>( 3.747 )</td>
<td>( 4.604 )</td>
<td>( 7.173 )</td>
<td>( 8.610 )</td>
</tr>
<tr>
<td>( 5 )</td>
<td>( 1.476 )</td>
<td>( 2.015 )</td>
<td>( 2.571 )</td>
<td>( 3.365 )</td>
<td>( 4.032 )</td>
<td>( 5.893 )</td>
<td>( 6.859 )</td>
</tr>
<tr>
<td>( 6 )</td>
<td>( 1.440 )</td>
<td>( 1.943 )</td>
<td>( 2.447 )</td>
<td>( 3.143 )</td>
<td>( 3.707 )</td>
<td>( 5.208 )</td>
<td>( 5.959 )</td>
</tr>
<tr>
<td>( 7 )</td>
<td>( 1.415 )</td>
<td>( 1.895 )</td>
<td>( 2.365 )</td>
<td>( 2.998 )</td>
<td>( 3.499 )</td>
<td>( 4.785 )</td>
<td>( 5.405 )</td>
</tr>
<tr>
<td>( 8 )</td>
<td>( 1.397 )</td>
<td>( 1.860 )</td>
<td>( 2.306 )</td>
<td>( 2.896 )</td>
<td>( 3.355 )</td>
<td>( 4.501 )</td>
<td>( 5.041 )</td>
</tr>
<tr>
<td>( 9 )</td>
<td>( 1.383 )</td>
<td>( 1.833 )</td>
<td>( 2.262 )</td>
<td>( 2.821 )</td>
<td>( 3.250 )</td>
<td>( 4.297 )</td>
<td>( 4.781 )</td>
</tr>
<tr>
<td>( 10 )</td>
<td>( 1.372 )</td>
<td>( 1.812 )</td>
<td>( 2.228 )</td>
<td>( 2.764 )</td>
<td>( 3.169 )</td>
<td>( 4.144 )</td>
<td>( 4.587 )</td>
</tr>
<tr>
<td>( 11 )</td>
<td>( 1.363 )</td>
<td>( 1.796 )</td>
<td>( 2.201 )</td>
<td>( 2.718 )</td>
<td>( 3.106 )</td>
<td>( 4.025 )</td>
<td>( 4.437 )</td>
</tr>
<tr>
<td>( 12 )</td>
<td>( 1.356 )</td>
<td>( 1.782 )</td>
<td>( 2.179 )</td>
<td>( 2.681 )</td>
<td>( 3.055 )</td>
<td>( 3.930 )</td>
<td>( 4.318 )</td>
</tr>
<tr>
<td>( 13 )</td>
<td>( 1.350 )</td>
<td>( 1.771 )</td>
<td>( 2.160 )</td>
<td>( 2.650 )</td>
<td>( 3.012 )</td>
<td>( 3.852 )</td>
<td>( 4.221 )</td>
</tr>
<tr>
<td>( 14 )</td>
<td>( 1.345 )</td>
<td>( 1.761 )</td>
<td>( 2.145 )</td>
<td>( 2.624 )</td>
<td>( 2.977 )</td>
<td>( 3.787 )</td>
<td>( 4.140 )</td>
</tr>
<tr>
<td>( 15 )</td>
<td>( 1.341 )</td>
<td>( 1.753 )</td>
<td>( 2.131 )</td>
<td>( 2.602 )</td>
<td>( 2.947 )</td>
<td>( 3.733 )</td>
<td>( 4.073 )</td>
</tr>
</tbody>
</table>

\(^a\) (Continued)
**Table II (Continued)**

<table>
<thead>
<tr>
<th></th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.01</th>
<th>.005</th>
<th>.001</th>
<th>.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-tailed ( \alpha ) (right tail area)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One-tailed ( 1 - \alpha ) (except right tail)</td>
<td>.90</td>
<td>.95</td>
<td>.975</td>
<td>.99</td>
<td>.995</td>
<td>.999</td>
<td>.9995</td>
</tr>
<tr>
<td>Two-tailed ( \alpha ) (area both tails)</td>
<td>.20</td>
<td>.10</td>
<td>.05</td>
<td>.02</td>
<td>.01</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>Two-tailed ( 1 - \alpha ) (except both tails)</td>
<td>.80</td>
<td>.90</td>
<td>.95</td>
<td>.98</td>
<td>.99</td>
<td>.998</td>
<td>.999</td>
</tr>
<tr>
<td>16</td>
<td>1.337</td>
<td>1.746</td>
<td>2.120</td>
<td>2.583</td>
<td>2.921</td>
<td>3.686</td>
<td>4.015</td>
</tr>
<tr>
<td>17</td>
<td>1.333</td>
<td>1.740</td>
<td>2.110</td>
<td>2.567</td>
<td>2.898</td>
<td>3.646</td>
<td>3.965</td>
</tr>
<tr>
<td>18</td>
<td>1.330</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
<td>3.610</td>
<td>3.922</td>
</tr>
<tr>
<td>19</td>
<td>1.328</td>
<td>1.729</td>
<td>2.093</td>
<td>2.539</td>
<td>2.861</td>
<td>3.579</td>
<td>3.883</td>
</tr>
<tr>
<td>20</td>
<td>1.325</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.552</td>
<td>3.850</td>
</tr>
<tr>
<td>21</td>
<td>1.323</td>
<td>1.721</td>
<td>2.080</td>
<td>2.518</td>
<td>2.831</td>
<td>3.527</td>
<td>3.819</td>
</tr>
<tr>
<td>22</td>
<td>1.321</td>
<td>1.717</td>
<td>2.074</td>
<td>2.508</td>
<td>2.819</td>
<td>3.505</td>
<td>3.792</td>
</tr>
<tr>
<td>23</td>
<td>1.319</td>
<td>1.714</td>
<td>2.069</td>
<td>2.500</td>
<td>2.807</td>
<td>3.485</td>
<td>3.767</td>
</tr>
<tr>
<td>24</td>
<td>1.318</td>
<td>1.711</td>
<td>2.064</td>
<td>2.492</td>
<td>2.797</td>
<td>3.467</td>
<td>3.745</td>
</tr>
<tr>
<td>25</td>
<td>1.316</td>
<td>1.708</td>
<td>2.060</td>
<td>2.485</td>
<td>2.787</td>
<td>3.450</td>
<td>3.725</td>
</tr>
<tr>
<td>26</td>
<td>1.315</td>
<td>1.706</td>
<td>2.056</td>
<td>2.479</td>
<td>2.779</td>
<td>3.435</td>
<td>3.707</td>
</tr>
<tr>
<td>27</td>
<td>1.314</td>
<td>1.703</td>
<td>2.052</td>
<td>2.473</td>
<td>2.771</td>
<td>3.421</td>
<td>3.690</td>
</tr>
<tr>
<td>28</td>
<td>1.313</td>
<td>1.701</td>
<td>2.048</td>
<td>2.467</td>
<td>2.763</td>
<td>3.408</td>
<td>3.674</td>
</tr>
<tr>
<td>29</td>
<td>1.311</td>
<td>1.699</td>
<td>2.045</td>
<td>2.462</td>
<td>2.756</td>
<td>3.396</td>
<td>3.659</td>
</tr>
<tr>
<td>30</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.385</td>
<td>3.646</td>
</tr>
<tr>
<td>40</td>
<td>1.303</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>3.307</td>
<td>3.551</td>
</tr>
<tr>
<td>60</td>
<td>1.296</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>3.232</td>
<td>3.460</td>
</tr>
<tr>
<td>100</td>
<td>1.290</td>
<td>1.660</td>
<td>1.984</td>
<td>2.364</td>
<td>2.626</td>
<td>3.174</td>
<td>3.390</td>
</tr>
<tr>
<td>( \infty )</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.090</td>
<td>3.291</td>
</tr>
</tbody>
</table>

\(^a\)Selected distances (\( t \)) to the right of the mean are given for various degrees of freedom (\( df \)) and for (1) one-tailed \( \alpha \), area under the curve in the positive tail; (2) one-tailed \( 1 - \alpha \), area under all except the tail; (3) two-tailed \( \alpha \), areas combined for both positive and negative tails; and (4) two-tailed \( 1 - \alpha \), area under all except the two tails.
Table III  Chi-square distribution, right tail.a

<table>
<thead>
<tr>
<th>( \alpha ) (area in right tail)</th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.01</th>
<th>.005</th>
<th>.001</th>
<th>.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - ( \alpha ) (except right tail)</td>
<td>.90</td>
<td>.95</td>
<td>.975</td>
<td>.99</td>
<td>.995</td>
<td>.999</td>
<td>.9995</td>
</tr>
<tr>
<td>( df=1 )</td>
<td>2.71</td>
<td>3.84</td>
<td>5.02</td>
<td>6.63</td>
<td>7.88</td>
<td>10.81</td>
<td>12.13</td>
</tr>
<tr>
<td>2</td>
<td>4.61</td>
<td>5.99</td>
<td>7.38</td>
<td>9.21</td>
<td>10.60</td>
<td>13.80</td>
<td>15.21</td>
</tr>
<tr>
<td>3</td>
<td>6.25</td>
<td>7.81</td>
<td>9.35</td>
<td>11.34</td>
<td>12.84</td>
<td>16.26</td>
<td>17.75</td>
</tr>
<tr>
<td>5</td>
<td>9.24</td>
<td>11.07</td>
<td>12.83</td>
<td>15.08</td>
<td>16.75</td>
<td>20.52</td>
<td>22.15</td>
</tr>
<tr>
<td>6</td>
<td>10.64</td>
<td>12.59</td>
<td>14.45</td>
<td>16.81</td>
<td>18.54</td>
<td>22.46</td>
<td>24.08</td>
</tr>
<tr>
<td>7</td>
<td>12.02</td>
<td>14.07</td>
<td>16.01</td>
<td>18.47</td>
<td>20.28</td>
<td>24.35</td>
<td>26.02</td>
</tr>
<tr>
<td>8</td>
<td>13.36</td>
<td>15.51</td>
<td>17.53</td>
<td>20.09</td>
<td>21.95</td>
<td>26.10</td>
<td>27.86</td>
</tr>
<tr>
<td>9</td>
<td>14.68</td>
<td>16.92</td>
<td>19.02</td>
<td>21.67</td>
<td>23.59</td>
<td>27.86</td>
<td>29.71</td>
</tr>
<tr>
<td>10</td>
<td>15.99</td>
<td>18.31</td>
<td>20.48</td>
<td>23.21</td>
<td>25.19</td>
<td>29.58</td>
<td>31.46</td>
</tr>
<tr>
<td>11</td>
<td>17.28</td>
<td>19.68</td>
<td>21.92</td>
<td>24.72</td>
<td>26.75</td>
<td>31.29</td>
<td>33.13</td>
</tr>
<tr>
<td>12</td>
<td>18.55</td>
<td>21.03</td>
<td>23.34</td>
<td>26.22</td>
<td>28.30</td>
<td>32.92</td>
<td>34.80</td>
</tr>
<tr>
<td>13</td>
<td>19.81</td>
<td>22.36</td>
<td>24.74</td>
<td>27.69</td>
<td>29.82</td>
<td>34.54</td>
<td>36.47</td>
</tr>
<tr>
<td>14</td>
<td>21.06</td>
<td>23.69</td>
<td>26.12</td>
<td>29.14</td>
<td>31.32</td>
<td>36.12</td>
<td>38.14</td>
</tr>
<tr>
<td>15</td>
<td>22.31</td>
<td>25.00</td>
<td>27.49</td>
<td>30.57</td>
<td>32.81</td>
<td>37.71</td>
<td>39.73</td>
</tr>
<tr>
<td>16</td>
<td>23.54</td>
<td>26.30</td>
<td>28.84</td>
<td>32.00</td>
<td>34.27</td>
<td>39.24</td>
<td>41.31</td>
</tr>
<tr>
<td>17</td>
<td>24.77</td>
<td>27.59</td>
<td>30.19</td>
<td>33.41</td>
<td>35.72</td>
<td>40.78</td>
<td>42.89</td>
</tr>
<tr>
<td>18</td>
<td>25.99</td>
<td>28.87</td>
<td>31.53</td>
<td>34.80</td>
<td>37.16</td>
<td>42.32</td>
<td>44.47</td>
</tr>
<tr>
<td>19</td>
<td>27.20</td>
<td>30.14</td>
<td>32.85</td>
<td>36.19</td>
<td>38.58</td>
<td>43.81</td>
<td>45.97</td>
</tr>
<tr>
<td>20</td>
<td>28.41</td>
<td>31.41</td>
<td>34.17</td>
<td>37.57</td>
<td>39.99</td>
<td>45.31</td>
<td>47.46</td>
</tr>
</tbody>
</table>

(Continued)
**APPENDIX 3 Tables of probability distributions**

**Table III (Continued)**

<table>
<thead>
<tr>
<th>$\alpha$ (area in right tail)</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
<th>0.005</th>
<th>0.001</th>
<th>0.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 - \alpha$ (except right tail)</td>
<td>0.90</td>
<td>0.95</td>
<td>0.975</td>
<td>0.99</td>
<td>0.995</td>
<td>0.999</td>
<td>0.9995</td>
</tr>
<tr>
<td>21</td>
<td>29.62</td>
<td>32.67</td>
<td>35.48</td>
<td>38.94</td>
<td>41.40</td>
<td>46.80</td>
<td>49.04</td>
</tr>
<tr>
<td>22</td>
<td>30.81</td>
<td>33.92</td>
<td>36.78</td>
<td>40.29</td>
<td>42.80</td>
<td>48.25</td>
<td>50.45</td>
</tr>
<tr>
<td>23</td>
<td>32.01</td>
<td>35.17</td>
<td>38.08</td>
<td>41.64</td>
<td>44.19</td>
<td>49.75</td>
<td>52.03</td>
</tr>
<tr>
<td>24</td>
<td>33.20</td>
<td>36.41</td>
<td>39.36</td>
<td>42.97</td>
<td>45.56</td>
<td>51.15</td>
<td>53.44</td>
</tr>
<tr>
<td>25</td>
<td>34.38</td>
<td>37.65</td>
<td>40.65</td>
<td>44.31</td>
<td>46.93</td>
<td>52.65</td>
<td>51.93</td>
</tr>
<tr>
<td>26</td>
<td>35.56</td>
<td>38.88</td>
<td>41.92</td>
<td>45.64</td>
<td>48.30</td>
<td>54.05</td>
<td>56.43</td>
</tr>
<tr>
<td>27</td>
<td>36.74</td>
<td>40.11</td>
<td>43.20</td>
<td>46.97</td>
<td>49.65</td>
<td>55.46</td>
<td>57.83</td>
</tr>
<tr>
<td>28</td>
<td>37.92</td>
<td>41.34</td>
<td>44.46</td>
<td>48.28</td>
<td>51.00</td>
<td>56.57</td>
<td>59.24</td>
</tr>
<tr>
<td>29</td>
<td>39.09</td>
<td>42.56</td>
<td>45.72</td>
<td>49.59</td>
<td>52.34</td>
<td>58.27</td>
<td>60.73</td>
</tr>
<tr>
<td>30</td>
<td>40.26</td>
<td>43.77</td>
<td>46.98</td>
<td>50.89</td>
<td>53.68</td>
<td>59.68</td>
<td>62.23</td>
</tr>
<tr>
<td>35</td>
<td>46.06</td>
<td>49.80</td>
<td>53.20</td>
<td>57.34</td>
<td>60.27</td>
<td>66.62</td>
<td>69.26</td>
</tr>
<tr>
<td>40</td>
<td>51.80</td>
<td>55.76</td>
<td>59.34</td>
<td>63.69</td>
<td>66.76</td>
<td>73.39</td>
<td>76.11</td>
</tr>
<tr>
<td>50</td>
<td>63.17</td>
<td>67.51</td>
<td>71.42</td>
<td>76.16</td>
<td>79.50</td>
<td>86.66</td>
<td>89.56</td>
</tr>
<tr>
<td>60</td>
<td>74.40</td>
<td>79.08</td>
<td>83.30</td>
<td>88.38</td>
<td>91.96</td>
<td>99.58</td>
<td>102.66</td>
</tr>
<tr>
<td>70</td>
<td>85.53</td>
<td>90.53</td>
<td>95.02</td>
<td>100.43</td>
<td>104.22</td>
<td>112.32</td>
<td>115.66</td>
</tr>
<tr>
<td>80</td>
<td>96.58</td>
<td>101.88</td>
<td>106.63</td>
<td>112.32</td>
<td>116.32</td>
<td>124.80</td>
<td>128.32</td>
</tr>
<tr>
<td>100</td>
<td>118.50</td>
<td>124.34</td>
<td>129.56</td>
<td>135.81</td>
<td>140.16</td>
<td>149.41</td>
<td>153.11</td>
</tr>
</tbody>
</table>

*Selected chi-square values (distances above zero) are given for various degrees of freedom (df) and for (1) $\alpha$, the area under the curve in the right tail and (2) $1 - \alpha$, the area under all except the right tail.*
Table IV  Chi-square distribution, left tail.\(^a\)

<table>
<thead>
<tr>
<th>(\alpha) (area in left tail)</th>
<th>0.0005</th>
<th>0.001</th>
<th>0.005</th>
<th>0.01</th>
<th>0.025</th>
<th>0.05</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - (\alpha) (area except left tail)</td>
<td>0.9995</td>
<td>0.999</td>
<td>0.995</td>
<td>0.99</td>
<td>0.975</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>(df=1)</td>
<td>0.000999</td>
<td>0.00020</td>
<td>0.00100</td>
<td>0.0020</td>
<td>0.0051</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>2</td>
<td>0.015</td>
<td>0.024</td>
<td>0.072</td>
<td>0.12</td>
<td>0.22</td>
<td>0.35</td>
<td>0.58</td>
</tr>
<tr>
<td>3</td>
<td>0.065</td>
<td>0.091</td>
<td>0.210</td>
<td>0.30</td>
<td>0.48</td>
<td>0.71</td>
<td>1.06</td>
</tr>
<tr>
<td>4</td>
<td>0.163</td>
<td>0.210</td>
<td>0.410</td>
<td>0.55</td>
<td>0.83</td>
<td>1.15</td>
<td>1.61</td>
</tr>
<tr>
<td>5</td>
<td>0.308</td>
<td>0.380</td>
<td>0.680</td>
<td>0.87</td>
<td>1.24</td>
<td>1.64</td>
<td>2.20</td>
</tr>
<tr>
<td>6</td>
<td>0.480</td>
<td>0.600</td>
<td>0.990</td>
<td>1.24</td>
<td>1.69</td>
<td>2.17</td>
<td>2.83</td>
</tr>
<tr>
<td>7</td>
<td>0.710</td>
<td>0.860</td>
<td>1.340</td>
<td>1.65</td>
<td>2.18</td>
<td>2.73</td>
<td>3.49</td>
</tr>
<tr>
<td>8</td>
<td>0.970</td>
<td>1.150</td>
<td>1.730</td>
<td>2.09</td>
<td>2.70</td>
<td>3.33</td>
<td>4.17</td>
</tr>
<tr>
<td>9</td>
<td>1.260</td>
<td>1.480</td>
<td>2.160</td>
<td>2.56</td>
<td>3.25</td>
<td>3.94</td>
<td>4.87</td>
</tr>
<tr>
<td>10</td>
<td>1.580</td>
<td>1.830</td>
<td>2.600</td>
<td>3.05</td>
<td>3.82</td>
<td>4.57</td>
<td>5.58</td>
</tr>
<tr>
<td>11</td>
<td>1.930</td>
<td>2.210</td>
<td>3.070</td>
<td>3.57</td>
<td>4.40</td>
<td>5.23</td>
<td>6.30</td>
</tr>
<tr>
<td>12</td>
<td>2.310</td>
<td>2.610</td>
<td>3.570</td>
<td>4.11</td>
<td>5.01</td>
<td>5.89</td>
<td>7.04</td>
</tr>
<tr>
<td>13</td>
<td>2.700</td>
<td>3.040</td>
<td>4.070</td>
<td>4.66</td>
<td>5.63</td>
<td>6.57</td>
<td>7.79</td>
</tr>
<tr>
<td>14</td>
<td>3.100</td>
<td>3.480</td>
<td>4.600</td>
<td>5.23</td>
<td>6.26</td>
<td>7.26</td>
<td>8.55</td>
</tr>
<tr>
<td>15</td>
<td>3.540</td>
<td>3.940</td>
<td>5.140</td>
<td>5.81</td>
<td>6.91</td>
<td>7.96</td>
<td>9.31</td>
</tr>
<tr>
<td>16</td>
<td>3.980</td>
<td>4.420</td>
<td>5.700</td>
<td>6.41</td>
<td>7.56</td>
<td>8.67</td>
<td>10.09</td>
</tr>
<tr>
<td>18</td>
<td>4.920</td>
<td>5.410</td>
<td>6.840</td>
<td>7.63</td>
<td>8.91</td>
<td>10.12</td>
<td>11.65</td>
</tr>
<tr>
<td>19</td>
<td>5.410</td>
<td>5.920</td>
<td>7.430</td>
<td>8.26</td>
<td>9.59</td>
<td>10.85</td>
<td>12.44</td>
</tr>
<tr>
<td>28</td>
<td>10.240</td>
<td>10.990</td>
<td>13.120</td>
<td>14.25</td>
<td>16.05</td>
<td>17.71</td>
<td>19.77</td>
</tr>
<tr>
<td>30</td>
<td>13.800</td>
<td>14.680</td>
<td>17.190</td>
<td>18.51</td>
<td>20.57</td>
<td>22.46</td>
<td>24.80</td>
</tr>
<tr>
<td>40</td>
<td>16.920</td>
<td>17.930</td>
<td>20.710</td>
<td>22.16</td>
<td>24.43</td>
<td>26.51</td>
<td>29.05</td>
</tr>
<tr>
<td>50</td>
<td>23.470</td>
<td>24.680</td>
<td>27.990</td>
<td>29.71</td>
<td>32.36</td>
<td>34.76</td>
<td>37.69</td>
</tr>
<tr>
<td>60</td>
<td>30.320</td>
<td>31.730</td>
<td>35.530</td>
<td>37.49</td>
<td>40.48</td>
<td>43.19</td>
<td>46.46</td>
</tr>
<tr>
<td>70</td>
<td>37.440</td>
<td>39.020</td>
<td>43.280</td>
<td>45.44</td>
<td>48.76</td>
<td>51.74</td>
<td>55.33</td>
</tr>
<tr>
<td>80</td>
<td>44.820</td>
<td>46.490</td>
<td>51.170</td>
<td>53.54</td>
<td>57.15</td>
<td>60.39</td>
<td>64.28</td>
</tr>
<tr>
<td>100</td>
<td>59.940</td>
<td>61.920</td>
<td>67.320</td>
<td>70.07</td>
<td>74.22</td>
<td>77.93</td>
<td>82.36</td>
</tr>
</tbody>
</table>

\(^a\)Selected chi-square values (distances above zero) are given for various degrees of freedom (\(df\)) and for (1) \(\alpha\), the area under the curve in the left tail and (2) \(1 - \alpha\), the area under all except the left tail.
Table V  $F$ Distribution.$^a$

<table>
<thead>
<tr>
<th>Numerator df</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>12</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>60</th>
<th>100</th>
<th>∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denominator df</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7.71</td>
<td>6.94</td>
<td>6.59</td>
<td>6.39</td>
<td>6.26</td>
<td>6.16</td>
<td>6.09</td>
<td>6.04</td>
<td>6.00</td>
<td>5.96</td>
<td>5.91</td>
<td>5.86</td>
<td>5.80</td>
<td>5.77</td>
<td>5.75</td>
<td>5.72</td>
<td>5.69</td>
<td>5.66</td>
<td>5.63</td>
</tr>
<tr>
<td>5</td>
<td>6.61</td>
<td>5.79</td>
<td>5.41</td>
<td>5.19</td>
<td>5.05</td>
<td>4.95</td>
<td>4.88</td>
<td>4.82</td>
<td>4.77</td>
<td>4.74</td>
<td>4.68</td>
<td>4.62</td>
<td>4.56</td>
<td>4.52</td>
<td>4.50</td>
<td>4.46</td>
<td>4.43</td>
<td>4.41</td>
<td>4.36</td>
</tr>
<tr>
<td>6</td>
<td>5.99</td>
<td>5.14</td>
<td>4.76</td>
<td>4.53</td>
<td>4.39</td>
<td>4.28</td>
<td>4.21</td>
<td>4.15</td>
<td>4.10</td>
<td>4.06</td>
<td>4.00</td>
<td>3.94</td>
<td>3.87</td>
<td>3.83</td>
<td>3.81</td>
<td>3.77</td>
<td>3.74</td>
<td>3.71</td>
<td>3.67</td>
</tr>
<tr>
<td>7</td>
<td>5.59</td>
<td>4.74</td>
<td>4.35</td>
<td>4.12</td>
<td>3.97</td>
<td>3.87</td>
<td>3.79</td>
<td>3.73</td>
<td>3.68</td>
<td>3.64</td>
<td>3.57</td>
<td>3.51</td>
<td>3.44</td>
<td>3.40</td>
<td>3.38</td>
<td>3.34</td>
<td>3.30</td>
<td>3.27</td>
<td>3.23</td>
</tr>
<tr>
<td>8</td>
<td>5.32</td>
<td>4.46</td>
<td>4.07</td>
<td>3.84</td>
<td>3.69</td>
<td>3.58</td>
<td>3.50</td>
<td>3.44</td>
<td>3.39</td>
<td>3.35</td>
<td>3.28</td>
<td>3.22</td>
<td>3.15</td>
<td>3.11</td>
<td>3.08</td>
<td>3.04</td>
<td>3.01</td>
<td>2.97</td>
<td>2.93</td>
</tr>
<tr>
<td>9</td>
<td>5.12</td>
<td>4.26</td>
<td>3.86</td>
<td>3.63</td>
<td>3.48</td>
<td>3.37</td>
<td>3.29</td>
<td>3.23</td>
<td>3.18</td>
<td>3.14</td>
<td>3.07</td>
<td>3.02</td>
<td>2.98</td>
<td>2.91</td>
<td>2.85</td>
<td>2.77</td>
<td>2.73</td>
<td>2.66</td>
<td>2.59</td>
</tr>
<tr>
<td>10</td>
<td>4.96</td>
<td>4.10</td>
<td>3.71</td>
<td>3.48</td>
<td>3.33</td>
<td>3.22</td>
<td>3.14</td>
<td>3.07</td>
<td>3.02</td>
<td>2.98</td>
<td>2.91</td>
<td>2.85</td>
<td>2.77</td>
<td>2.73</td>
<td>2.70</td>
<td>2.66</td>
<td>2.62</td>
<td>2.59</td>
<td>2.54</td>
</tr>
<tr>
<td>11</td>
<td>4.84</td>
<td>3.98</td>
<td>3.59</td>
<td>3.36</td>
<td>3.20</td>
<td>3.09</td>
<td>3.01</td>
<td>2.95</td>
<td>2.90</td>
<td>2.85</td>
<td>2.79</td>
<td>2.72</td>
<td>2.65</td>
<td>2.60</td>
<td>2.57</td>
<td>2.53</td>
<td>2.49</td>
<td>2.46</td>
<td>2.40</td>
</tr>
<tr>
<td>12</td>
<td>4.75</td>
<td>3.89</td>
<td>3.49</td>
<td>3.26</td>
<td>3.11</td>
<td>3.00</td>
<td>2.91</td>
<td>2.85</td>
<td>2.80</td>
<td>2.75</td>
<td>2.69</td>
<td>2.62</td>
<td>2.54</td>
<td>2.50</td>
<td>2.47</td>
<td>2.43</td>
<td>2.38</td>
<td>2.35</td>
<td>2.30</td>
</tr>
<tr>
<td>13</td>
<td>4.67</td>
<td>3.81</td>
<td>3.41</td>
<td>3.18</td>
<td>3.03</td>
<td>2.92</td>
<td>2.83</td>
<td>2.77</td>
<td>2.71</td>
<td>2.67</td>
<td>2.60</td>
<td>2.53</td>
<td>2.46</td>
<td>2.41</td>
<td>2.38</td>
<td>2.34</td>
<td>2.30</td>
<td>2.26</td>
<td>2.21</td>
</tr>
<tr>
<td>14</td>
<td>4.60</td>
<td>3.74</td>
<td>3.34</td>
<td>3.11</td>
<td>2.96</td>
<td>2.85</td>
<td>2.76</td>
<td>2.70</td>
<td>2.65</td>
<td>2.60</td>
<td>2.53</td>
<td>2.46</td>
<td>2.39</td>
<td>2.34</td>
<td>2.31</td>
<td>2.27</td>
<td>2.22</td>
<td>2.19</td>
<td>2.13</td>
</tr>
<tr>
<td>15</td>
<td>4.54</td>
<td>3.68</td>
<td>3.29</td>
<td>3.06</td>
<td>2.90</td>
<td>2.79</td>
<td>2.71</td>
<td>2.64</td>
<td>2.59</td>
<td>2.54</td>
<td>2.48</td>
<td>2.40</td>
<td>2.33</td>
<td>2.28</td>
<td>2.25</td>
<td>2.20</td>
<td>2.16</td>
<td>2.12</td>
<td>2.07</td>
</tr>
<tr>
<td>16</td>
<td>4.49</td>
<td>3.63</td>
<td>3.24</td>
<td>3.01</td>
<td>2.85</td>
<td>2.74</td>
<td>2.66</td>
<td>2.59</td>
<td>2.54</td>
<td>2.49</td>
<td>2.42</td>
<td>2.35</td>
<td>2.28</td>
<td>2.23</td>
<td>2.19</td>
<td>2.15</td>
<td>2.11</td>
<td>2.07</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>22</td>
<td>25</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>80</td>
<td>100</td>
<td>150</td>
<td>$\infty$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.45</td>
<td>4.41</td>
<td>4.38</td>
<td>4.35</td>
<td>4.30</td>
<td>4.24</td>
<td>4.17</td>
<td>4.08</td>
<td>4.03</td>
<td>4.00</td>
<td>3.96</td>
<td>3.94</td>
<td>3.90</td>
<td>3.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.59</td>
<td>3.55</td>
<td>3.52</td>
<td>3.49</td>
<td>3.44</td>
<td>3.39</td>
<td>3.32</td>
<td>3.23</td>
<td>3.18</td>
<td>3.15</td>
<td>3.11</td>
<td>3.09</td>
<td>3.06</td>
<td>3.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.20</td>
<td>3.16</td>
<td>3.13</td>
<td>3.10</td>
<td>3.05</td>
<td>2.99</td>
<td>2.92</td>
<td>2.84</td>
<td>2.79</td>
<td>2.76</td>
<td>2.72</td>
<td>2.70</td>
<td>2.66</td>
<td>2.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.96</td>
<td>2.93</td>
<td>2.90</td>
<td>2.87</td>
<td>2.82</td>
<td>2.76</td>
<td>2.69</td>
<td>2.61</td>
<td>2.56</td>
<td>2.53</td>
<td>2.49</td>
<td>2.46</td>
<td>2.43</td>
<td>2.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.81</td>
<td>2.77</td>
<td>2.74</td>
<td>2.71</td>
<td>2.66</td>
<td>2.60</td>
<td>2.53</td>
<td>2.45</td>
<td>2.40</td>
<td>2.37</td>
<td>2.33</td>
<td>2.31</td>
<td>2.27</td>
<td>2.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.70</td>
<td>2.66</td>
<td>2.63</td>
<td>2.60</td>
<td>2.55</td>
<td>2.49</td>
<td>2.42</td>
<td>2.45</td>
<td>2.40</td>
<td>2.37</td>
<td>2.33</td>
<td>2.31</td>
<td>2.27</td>
<td>2.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.61</td>
<td>2.58</td>
<td>2.54</td>
<td>2.51</td>
<td>2.46</td>
<td>2.40</td>
<td>2.34</td>
<td>2.39</td>
<td>2.36</td>
<td>2.32</td>
<td>2.28</td>
<td>2.25</td>
<td>2.22</td>
<td>2.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.55</td>
<td>2.51</td>
<td>2.48</td>
<td>2.45</td>
<td>2.40</td>
<td>2.34</td>
<td>2.33</td>
<td>2.39</td>
<td>2.38</td>
<td>2.35</td>
<td>2.32</td>
<td>2.28</td>
<td>2.25</td>
<td>2.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.49</td>
<td>2.46</td>
<td>2.42</td>
<td>2.47</td>
<td>2.40</td>
<td>2.34</td>
<td>2.33</td>
<td>2.39</td>
<td>2.38</td>
<td>2.35</td>
<td>2.32</td>
<td>2.28</td>
<td>2.25</td>
<td>2.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.38</td>
<td>2.34</td>
<td>2.32</td>
<td>2.31</td>
<td>2.28</td>
<td>2.23</td>
<td>2.22</td>
<td>2.28</td>
<td>2.28</td>
<td>2.28</td>
<td>2.26</td>
<td>2.23</td>
<td>2.20</td>
<td>2.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.31</td>
<td>2.27</td>
<td>2.23</td>
<td>2.22</td>
<td>2.19</td>
<td>2.16</td>
<td>2.16</td>
<td>2.21</td>
<td>2.21</td>
<td>2.21</td>
<td>2.19</td>
<td>2.16</td>
<td>2.14</td>
<td>2.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.23</td>
<td>2.19</td>
<td>2.16</td>
<td>2.16</td>
<td>2.14</td>
<td>2.11</td>
<td>2.11</td>
<td>2.16</td>
<td>2.16</td>
<td>2.16</td>
<td>2.14</td>
<td>2.11</td>
<td>2.10</td>
<td>2.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.18</td>
<td>2.14</td>
<td>2.11</td>
<td>2.11</td>
<td>2.08</td>
<td>2.04</td>
<td>2.04</td>
<td>2.08</td>
<td>2.08</td>
<td>2.08</td>
<td>2.06</td>
<td>2.03</td>
<td>2.02</td>
<td>2.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.15</td>
<td>2.11</td>
<td>2.06</td>
<td>2.06</td>
<td>2.02</td>
<td>1.98</td>
<td>1.98</td>
<td>2.02</td>
<td>2.02</td>
<td>2.02</td>
<td>2.01</td>
<td>1.97</td>
<td>1.97</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.10</td>
<td>2.06</td>
<td>2.02</td>
<td>2.02</td>
<td>1.98</td>
<td>1.91</td>
<td>1.91</td>
<td>2.02</td>
<td>2.02</td>
<td>2.02</td>
<td>2.01</td>
<td>1.96</td>
<td>1.95</td>
<td>1.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.06</td>
<td>2.01</td>
<td>1.98</td>
<td>1.98</td>
<td>1.95</td>
<td>1.91</td>
<td>1.91</td>
<td>2.02</td>
<td>2.02</td>
<td>2.02</td>
<td>2.01</td>
<td>1.96</td>
<td>1.95</td>
<td>1.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.02</td>
<td>1.98</td>
<td>1.94</td>
<td>1.94</td>
<td>1.91</td>
<td>1.89</td>
<td>1.89</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>1.98</td>
<td>1.95</td>
<td>1.94</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.96</td>
<td>1.92</td>
<td>1.89</td>
<td>1.89</td>
<td>1.87</td>
<td>1.84</td>
<td>1.84</td>
<td>1.95</td>
<td>1.95</td>
<td>1.95</td>
<td>1.93</td>
<td>1.92</td>
<td>1.91</td>
<td>1.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.85</td>
<td>1.78</td>
<td>1.77</td>
<td>1.77</td>
<td>1.76</td>
<td>1.74</td>
<td>1.74</td>
<td>1.86</td>
<td>1.86</td>
<td>1.86</td>
<td>1.84</td>
<td>1.83</td>
<td>1.82</td>
<td>1.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.78</td>
<td>1.71</td>
<td>1.67</td>
<td>1.67</td>
<td>1.65</td>
<td>1.64</td>
<td>1.64</td>
<td>1.76</td>
<td>1.76</td>
<td>1.76</td>
<td>1.74</td>
<td>1.73</td>
<td>1.72</td>
<td>1.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.71</td>
<td>1.62</td>
<td>1.59</td>
<td>1.59</td>
<td>1.57</td>
<td>1.56</td>
<td>1.56</td>
<td>1.68</td>
<td>1.68</td>
<td>1.68</td>
<td>1.66</td>
<td>1.65</td>
<td>1.64</td>
<td>1.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.62</td>
<td>1.51</td>
<td>1.52</td>
<td>1.52</td>
<td>1.51</td>
<td>1.50</td>
<td>1.50</td>
<td>1.63</td>
<td>1.63</td>
<td>1.63</td>
<td>1.61</td>
<td>1.60</td>
<td>1.59</td>
<td>1.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.51</td>
<td>1.44</td>
<td>1.48</td>
<td>1.48</td>
<td>1.47</td>
<td>1.46</td>
<td>1.46</td>
<td>1.59</td>
<td>1.59</td>
<td>1.59</td>
<td>1.58</td>
<td>1.57</td>
<td>1.56</td>
<td>1.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.44</td>
<td>1.32</td>
<td>1.43</td>
<td>1.43</td>
<td>1.42</td>
<td>1.41</td>
<td>1.41</td>
<td>1.54</td>
<td>1.54</td>
<td>1.54</td>
<td>1.53</td>
<td>1.52</td>
<td>1.51</td>
<td>1.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.32</td>
<td>1.28</td>
<td>1.39</td>
<td>1.39</td>
<td>1.38</td>
<td>1.37</td>
<td>1.37</td>
<td>1.52</td>
<td>1.52</td>
<td>1.52</td>
<td>1.51</td>
<td>1.50</td>
<td>1.49</td>
<td>1.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.28</td>
<td>1.22</td>
<td>1.34</td>
<td>1.34</td>
<td>1.33</td>
<td>1.32</td>
<td>1.32</td>
<td>1.48</td>
<td>1.48</td>
<td>1.48</td>
<td>1.47</td>
<td>1.46</td>
<td>1.45</td>
<td>1.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.22</td>
<td>1.15</td>
<td>1.18</td>
<td>1.18</td>
<td>1.17</td>
<td>1.16</td>
<td>1.16</td>
<td>1.31</td>
<td>1.31</td>
<td>1.31</td>
<td>1.30</td>
<td>1.29</td>
<td>1.28</td>
<td>1.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.18</td>
<td>0.80</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
<td>0.82</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> Selected distances ($F$) are given for various degrees of freedom ($df$) for $\alpha=5\%$, the area under the curve in the positive tail, and $1 - \alpha=95\%$, the area under all except the tail. Numerator $df$ appears in column headings, denominator $df$ in row headings, and $F$ at the intersection of row and column in the table body.
Table VI  Binomial distribution.\textsuperscript{a}

\begin{tabular}{cccccccccc}
\hline
n & n_o & .05 & .10 & .15 & .20 & .25 & .30 & .35 & .40 & .45 & .50 \\
\hline
2 & 1 & .098 & .190 & .278 & .360 & .438 & .510 & .578 & .640 & .698 & .750 \\
 & 2 & .003 & .010 & .023 & .040 & .063 & .090 & .123 & .160 & .203 & .250 \\
3 & 1 & .143 & .271 & .386 & .488 & .578 & .657 & .725 & .784 & .834 & .875 \\
 & 3 & .000 & .001 & .003 & .008 & .016 & .027 & .043 & .064 & .091 & .125 \\
 & 2 & .014 & .052 & .110 & .181 & .262 & .348 & .437 & .525 & .609 & .688 \\
 & 3 & .001 & .004 & .012 & .027 & .051 & .084 & .127 & .179 & .242 & .313 \\
 & 4 & .000 & .000 & .000 & .002 & .004 & .008 & .015 & .026 & .041 & .063 \\
5 & 1 & .226 & .410 & .556 & .672 & .763 & .832 & .884 & .922 & .950 & .969 \\
 & 2 & .023 & .082 & .165 & .263 & .367 & .472 & .572 & .663 & .744 & .813 \\
 & 4 & .000 & .001 & .002 & .007 & .016 & .031 & .054 & .087 & .131 & .188 \\
6 & 1 & .265 & .469 & .629 & .738 & .822 & .882 & .925 & .953 & .972 & .984 \\
 & 2 & .033 & .114 & .224 & .345 & .466 & .580 & .681 & .767 & .836 & .891 \\
 & 3 & .002 & .016 & .047 & .099 & .169 & .256 & .353 & .456 & .559 & .656 \\
 & 4 & .000 & .001 & .006 & .017 & .038 & .071 & .117 & .179 & .255 & .344 \\
 & 5 & .000 & .000 & .000 & .002 & .005 & .011 & .022 & .041 & .069 & .109 \\
7 & 1 & .302 & .522 & .679 & .790 & .867 & .918 & .951 & .972 & .985 & .992 \\
 & 2 & .044 & .150 & .283 & .423 & .555 & .671 & .766 & .841 & .898 & .938 \\
 & 3 & .004 & .026 & .074 & .148 & .244 & .353 & .468 & .580 & .684 & .773 \\
 & 4 & .000 & .003 & .012 & .033 & .071 & .126 & .200 & .290 & .392 & .500 \\
 & 5 & .000 & .000 & .001 & .005 & .013 & .029 & .056 & .096 & .153 & .227 \\
 & 6 & .000 & .000 & .000 & .001 & .004 & .009 & .019 & .036 & .063 & .096 \\
 & 7 & .000 & .000 & .000 & .001 & .004 & .009 & .019 & .036 & .063 & .096 \\
 & 3 & .006 & .038 & .105 & .203 & .322 & .448 & .572 & .685 & .780 & .856 \\
 & 4 & .000 & .005 & .021 & .056 & .114 & .194 & .294 & .406 & .523 & .637 \\
 & 5 & .000 & .000 & .003 & .010 & .027 & .058 & .106 & .174 & .260 & .363 \\
 & 6 & .000 & .000 & .001 & .004 & .011 & .025 & .050 & .089 & .145 & .205 \\
 & 7 & .000 & .000 & .000 & .001 & .004 & .009 & .018 & .035 & .063 & .096 \\
 & 8 & .000 & .000 & .000 & .001 & .004 & .009 & .018 & .035 & .063 & .096 \\
9 & 1 & .370 & .613 & .768 & .866 & .925 & .960 & .979 & .990 & .995 & .998 \\
 & 2 & .071 & .225 & .401 & .564 & .700 & .804 & .879 & .930 & .962 & .981 \\
 & 3 & .008 & .053 & .141 & .262 & .399 & .537 & .663 & .768 & .851 & .910 \\
 & 4 & .001 & .008 & .034 & .086 & .166 & .270 & .391 & .517 & .639 & .746 \\
 & 5 & .000 & .001 & .006 & .020 & .045 & .099 & .172 & .267 & .379 & .500 \\
 & 6 & .000 & .000 & .001 & .003 & .010 & .025 & .054 & .099 & .166 & .254 \\
 & 7 & .000 & .000 & .000 & .001 & .004 & .011 & .025 & .050 & .090 & .135 \\
 & 8 & .000 & .000 & .000 & .001 & .004 & .009 & .018 & .035 & .063 & .096 \\
 & 9 & .000 & .000 & .000 & .001 & .004 & .009 & .018 & .035 & .063 & .096 \\
\hline
\end{tabular}

(Continued)
### Table VI (Continued)

<table>
<thead>
<tr>
<th>( n )</th>
<th>( n_0 )</th>
<th>.05</th>
<th>.10</th>
<th>.15</th>
<th>.20</th>
<th>.25</th>
<th>.30</th>
<th>.35</th>
<th>.40</th>
<th>.45</th>
<th>.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>.401</td>
<td>.651</td>
<td>.803</td>
<td>.893</td>
<td>.944</td>
<td>.972</td>
<td>.987</td>
<td>.994</td>
<td>.998</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.086</td>
<td>.264</td>
<td>.456</td>
<td>.624</td>
<td>.756</td>
<td>.851</td>
<td>.914</td>
<td>.954</td>
<td>.977</td>
<td>.989</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.012</td>
<td>.070</td>
<td>.180</td>
<td>.322</td>
<td>.474</td>
<td>.617</td>
<td>.738</td>
<td>.833</td>
<td>.900</td>
<td>.945</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.001</td>
<td>.013</td>
<td>.050</td>
<td>.121</td>
<td>.224</td>
<td>.350</td>
<td>.486</td>
<td>.618</td>
<td>.734</td>
<td>.828</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.000</td>
<td>.002</td>
<td>.010</td>
<td>.033</td>
<td>.078</td>
<td>.150</td>
<td>.249</td>
<td>.367</td>
<td>.496</td>
<td>.623</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>.000</td>
<td>.001</td>
<td>.006</td>
<td>.020</td>
<td>.047</td>
<td>.095</td>
<td>.166</td>
<td>.262</td>
<td>.377</td>
<td>.493</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.000</td>
<td>.001</td>
<td>.004</td>
<td>.011</td>
<td>.026</td>
<td>.055</td>
<td>.102</td>
<td>.172</td>
<td>.271</td>
<td>.372</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>.000</td>
<td>.000</td>
<td>.002</td>
<td>.005</td>
<td>.012</td>
<td>.027</td>
<td>.055</td>
<td>.113</td>
<td>.192</td>
<td>.277</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>.000</td>
<td>.001</td>
<td>.002</td>
<td>.005</td>
<td>.011</td>
<td>.025</td>
<td>.054</td>
<td>.106</td>
<td>.181</td>
<td>.266</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.002</td>
<td>.005</td>
<td>.012</td>
<td>.027</td>
<td>.055</td>
<td>.113</td>
<td>.192</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.002</td>
<td>.005</td>
<td>.012</td>
<td>.027</td>
<td>.055</td>
<td>.113</td>
<td>.192</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.002</td>
<td>.005</td>
<td>.012</td>
<td>.027</td>
<td>.055</td>
<td>.113</td>
<td>.192</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.002</td>
<td>.005</td>
<td>.012</td>
<td>.027</td>
<td>.055</td>
<td>.113</td>
<td>.192</td>
</tr>
</tbody>
</table>

(Continued)
Table VI (Continued)

<table>
<thead>
<tr>
<th>n</th>
<th>(n_o)</th>
<th>.05</th>
<th>.10</th>
<th>.15</th>
<th>.20</th>
<th>.25</th>
<th>.30</th>
<th>.35</th>
<th>.40</th>
<th>.45</th>
<th>.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>1</td>
<td>.512</td>
<td>.771</td>
<td>.897</td>
<td>.956</td>
<td>.982</td>
<td>.993</td>
<td>.998</td>
<td>.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.153</td>
<td>.415</td>
<td>.643</td>
<td>.802</td>
<td>.899</td>
<td>.953</td>
<td>.980</td>
<td>.992</td>
<td>.997</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.030</td>
<td>.158</td>
<td>.352</td>
<td>.552</td>
<td>.719</td>
<td>.839</td>
<td>.916</td>
<td>.960</td>
<td>.983</td>
<td>.993</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>.004</td>
<td>.044</td>
<td>.147</td>
<td>.302</td>
<td>.479</td>
<td>.645</td>
<td>.780</td>
<td>.876</td>
<td>.937</td>
<td>.971</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>.000</td>
<td>.009</td>
<td>.047</td>
<td>.130</td>
<td>.259</td>
<td>.416</td>
<td>.577</td>
<td>.721</td>
<td>.833</td>
<td>.910</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>.002</td>
<td>.012</td>
<td>.044</td>
<td>.112</td>
<td>.220</td>
<td>.360</td>
<td>.514</td>
<td>.663</td>
<td>.788</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.000</td>
<td>.002</td>
<td>.012</td>
<td>.038</td>
<td>.093</td>
<td>.184</td>
<td>.308</td>
<td>.454</td>
<td>.605</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Values of cumulative binomial distribution, depending on \(\pi\) (theoretical proportion of occurrences in a random trial), \(n\) (sample size), and \(n_o\) (number of occurrences observed). Given \(\pi\), \(n\), and \(n_o\), the corresponding entry in the table body represents the probability that \(n_o\) or more occurrences (or alternatively that \(n_o/n\) proportion observed occurrences) would have been observed by chance alone.
Table VII  Poisson distribution.a

<table>
<thead>
<tr>
<th>$n_0$</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
<th>1.4</th>
<th>1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.095</td>
<td>.181</td>
<td>.259</td>
<td>.330</td>
<td>.394</td>
</tr>
<tr>
<td>2</td>
<td>.005</td>
<td>.018</td>
<td>.037</td>
<td>.062</td>
<td>.090</td>
</tr>
<tr>
<td>3</td>
<td>.000</td>
<td>.001</td>
<td>.004</td>
<td>.008</td>
<td>.014</td>
</tr>
<tr>
<td>4</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.002</td>
</tr>
<tr>
<td>5</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>6</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n_0$</th>
<th>1.6</th>
<th>1.7</th>
<th>1.8</th>
<th>1.9</th>
<th>2.0</th>
<th>2.1</th>
<th>2.2</th>
<th>2.3</th>
<th>2.4</th>
<th>2.5</th>
<th>2.6</th>
<th>2.7</th>
<th>2.8</th>
<th>2.9</th>
<th>3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.798</td>
<td>.817</td>
<td>.835</td>
<td>.850</td>
<td>.865</td>
<td>.878</td>
<td>.889</td>
<td>.900</td>
<td>.909</td>
<td>.918</td>
<td>.926</td>
<td>.933</td>
<td>.939</td>
<td>.945</td>
<td>.950</td>
</tr>
<tr>
<td>2</td>
<td>.475</td>
<td>.507</td>
<td>.537</td>
<td>.566</td>
<td>.594</td>
<td>.620</td>
<td>.645</td>
<td>.669</td>
<td>.692</td>
<td>.713</td>
<td>.733</td>
<td>.751</td>
<td>.769</td>
<td>.785</td>
<td>.801</td>
</tr>
<tr>
<td>3</td>
<td>.217</td>
<td>.243</td>
<td>.269</td>
<td>.296</td>
<td>.323</td>
<td>.350</td>
<td>.377</td>
<td>.404</td>
<td>.430</td>
<td>.456</td>
<td>.482</td>
<td>.506</td>
<td>.531</td>
<td>.554</td>
<td>.577</td>
</tr>
<tr>
<td>5</td>
<td>.024</td>
<td>.030</td>
<td>.036</td>
<td>.044</td>
<td>.053</td>
<td>.062</td>
<td>.073</td>
<td>.084</td>
<td>.096</td>
<td>.109</td>
<td>.123</td>
<td>.137</td>
<td>.152</td>
<td>.168</td>
<td>.185</td>
</tr>
<tr>
<td>6</td>
<td>.006</td>
<td>.008</td>
<td>.010</td>
<td>.013</td>
<td>.017</td>
<td>.020</td>
<td>.025</td>
<td>.030</td>
<td>.036</td>
<td>.042</td>
<td>.049</td>
<td>.057</td>
<td>.065</td>
<td>.074</td>
<td>.084</td>
</tr>
<tr>
<td>7</td>
<td>.001</td>
<td>.002</td>
<td>.003</td>
<td>.003</td>
<td>.005</td>
<td>.006</td>
<td>.008</td>
<td>.009</td>
<td>.012</td>
<td>.014</td>
<td>.017</td>
<td>.021</td>
<td>.024</td>
<td>.029</td>
<td>.034</td>
</tr>
<tr>
<td>8</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.002</td>
<td>.003</td>
<td>.003</td>
<td>.004</td>
<td>.005</td>
<td>.007</td>
<td>.008</td>
<td>.010</td>
<td>.012</td>
</tr>
<tr>
<td>9</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.002</td>
<td>.003</td>
<td>.004</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n_0$</th>
<th>3.1</th>
<th>3.2</th>
<th>3.3</th>
<th>3.4</th>
<th>3.5</th>
<th>3.6</th>
<th>3.7</th>
<th>3.8</th>
<th>3.9</th>
<th>4.0</th>
<th>4.1</th>
<th>4.2</th>
<th>4.3</th>
<th>4.4</th>
<th>4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.955</td>
<td>.959</td>
<td>.963</td>
<td>.967</td>
<td>.970</td>
<td>.973</td>
<td>.975</td>
<td>.978</td>
<td>.980</td>
<td>.982</td>
<td>.983</td>
<td>.985</td>
<td>.986</td>
<td>.988</td>
<td>.989</td>
</tr>
<tr>
<td>2</td>
<td>.815</td>
<td>.829</td>
<td>.841</td>
<td>.853</td>
<td>.864</td>
<td>.874</td>
<td>.884</td>
<td>.893</td>
<td>.901</td>
<td>.908</td>
<td>.916</td>
<td>.922</td>
<td>.928</td>
<td>.934</td>
<td>.939</td>
</tr>
<tr>
<td>$n_o$</td>
<td>3.1</td>
<td>3.2</td>
<td>3.3</td>
<td>3.4</td>
<td>3.5</td>
<td>3.6</td>
<td>3.7</td>
<td>3.8</td>
<td>3.9</td>
<td>4.0</td>
<td>4.1</td>
<td>4.2</td>
<td>4.3</td>
<td>4.4</td>
<td>4.5</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>5</td>
<td>0.202</td>
<td>0.219</td>
<td>0.237</td>
<td>0.256</td>
<td>0.275</td>
<td>0.294</td>
<td>0.313</td>
<td>0.332</td>
<td>0.352</td>
<td>0.371</td>
<td>0.391</td>
<td>0.410</td>
<td>0.430</td>
<td>0.449</td>
<td>0.468</td>
</tr>
<tr>
<td>6</td>
<td>0.094</td>
<td>0.105</td>
<td>0.117</td>
<td>0.130</td>
<td>0.142</td>
<td>0.156</td>
<td>0.170</td>
<td>0.184</td>
<td>0.199</td>
<td>0.215</td>
<td>0.231</td>
<td>0.247</td>
<td>0.263</td>
<td>0.280</td>
<td>0.297</td>
</tr>
<tr>
<td>7</td>
<td>0.039</td>
<td>0.045</td>
<td>0.051</td>
<td>0.058</td>
<td>0.065</td>
<td>0.073</td>
<td>0.082</td>
<td>0.091</td>
<td>0.101</td>
<td>0.111</td>
<td>0.121</td>
<td>0.133</td>
<td>0.144</td>
<td>0.156</td>
<td>0.169</td>
</tr>
<tr>
<td>8</td>
<td>0.014</td>
<td>0.017</td>
<td>0.020</td>
<td>0.023</td>
<td>0.027</td>
<td>0.031</td>
<td>0.035</td>
<td>0.040</td>
<td>0.045</td>
<td>0.051</td>
<td>0.057</td>
<td>0.064</td>
<td>0.071</td>
<td>0.079</td>
<td>0.087</td>
</tr>
<tr>
<td>9</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.010</td>
<td>0.012</td>
<td>0.014</td>
<td>0.016</td>
<td>0.019</td>
<td>0.021</td>
<td>0.025</td>
<td>0.028</td>
<td>0.032</td>
<td>0.036</td>
<td>0.040</td>
</tr>
<tr>
<td>10</td>
<td>0.001</td>
<td>0.002</td>
<td>0.003</td>
<td>0.004</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.010</td>
<td>0.011</td>
<td>0.013</td>
<td>0.015</td>
<td>0.017</td>
<td>0.015</td>
<td>0.017</td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.004</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>12</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>13</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n_o$</th>
<th>4.6</th>
<th>4.7</th>
<th>4.8</th>
<th>4.9</th>
<th>5.0</th>
<th>5.1</th>
<th>5.2</th>
<th>5.3</th>
<th>5.4</th>
<th>5.5</th>
<th>5.6</th>
<th>5.7</th>
<th>5.8</th>
<th>5.9</th>
<th>6.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.990</td>
<td>0.991</td>
<td>0.992</td>
<td>0.993</td>
<td>0.993</td>
<td>0.994</td>
<td>0.994</td>
<td>0.995</td>
<td>0.995</td>
<td>0.996</td>
<td>0.996</td>
<td>0.996</td>
<td>0.997</td>
<td>0.997</td>
<td>0.998</td>
</tr>
<tr>
<td>2</td>
<td>0.944</td>
<td>0.948</td>
<td>0.952</td>
<td>0.956</td>
<td>0.960</td>
<td>0.963</td>
<td>0.966</td>
<td>0.969</td>
<td>0.971</td>
<td>0.973</td>
<td>0.976</td>
<td>0.978</td>
<td>0.979</td>
<td>0.981</td>
<td>0.983</td>
</tr>
<tr>
<td>3</td>
<td>0.837</td>
<td>0.848</td>
<td>0.858</td>
<td>0.867</td>
<td>0.875</td>
<td>0.884</td>
<td>0.891</td>
<td>0.898</td>
<td>0.905</td>
<td>0.912</td>
<td>0.918</td>
<td>0.923</td>
<td>0.929</td>
<td>0.933</td>
<td>0.938</td>
</tr>
<tr>
<td>4</td>
<td>0.674</td>
<td>0.690</td>
<td>0.706</td>
<td>0.721</td>
<td>0.735</td>
<td>0.749</td>
<td>0.762</td>
<td>0.775</td>
<td>0.787</td>
<td>0.798</td>
<td>0.809</td>
<td>0.820</td>
<td>0.830</td>
<td>0.840</td>
<td>0.857</td>
</tr>
<tr>
<td>5</td>
<td>0.487</td>
<td>0.505</td>
<td>0.524</td>
<td>0.542</td>
<td>0.560</td>
<td>0.577</td>
<td>0.594</td>
<td>0.611</td>
<td>0.627</td>
<td>0.643</td>
<td>0.658</td>
<td>0.673</td>
<td>0.687</td>
<td>0.701</td>
<td>0.715</td>
</tr>
<tr>
<td>6</td>
<td>0.314</td>
<td>0.332</td>
<td>0.349</td>
<td>0.367</td>
<td>0.384</td>
<td>0.402</td>
<td>0.419</td>
<td>0.437</td>
<td>0.454</td>
<td>0.471</td>
<td>0.488</td>
<td>0.505</td>
<td>0.522</td>
<td>0.538</td>
<td>0.554</td>
</tr>
<tr>
<td>7</td>
<td>0.182</td>
<td>0.195</td>
<td>0.209</td>
<td>0.223</td>
<td>0.238</td>
<td>0.253</td>
<td>0.268</td>
<td>0.283</td>
<td>0.298</td>
<td>0.314</td>
<td>0.330</td>
<td>0.346</td>
<td>0.362</td>
<td>0.378</td>
<td>0.394</td>
</tr>
<tr>
<td>8</td>
<td>0.095</td>
<td>0.104</td>
<td>0.113</td>
<td>0.123</td>
<td>0.133</td>
<td>0.144</td>
<td>0.155</td>
<td>0.167</td>
<td>0.178</td>
<td>0.191</td>
<td>0.203</td>
<td>0.216</td>
<td>0.229</td>
<td>0.242</td>
<td>0.256</td>
</tr>
<tr>
<td>9</td>
<td>0.045</td>
<td>0.050</td>
<td>0.056</td>
<td>0.062</td>
<td>0.068</td>
<td>0.075</td>
<td>0.082</td>
<td>0.089</td>
<td>0.097</td>
<td>0.106</td>
<td>0.114</td>
<td>0.123</td>
<td>0.133</td>
<td>0.143</td>
<td>0.153</td>
</tr>
<tr>
<td>10</td>
<td>0.020</td>
<td>0.022</td>
<td>0.025</td>
<td>0.028</td>
<td>0.032</td>
<td>0.036</td>
<td>0.040</td>
<td>0.044</td>
<td>0.049</td>
<td>0.054</td>
<td>0.059</td>
<td>0.065</td>
<td>0.071</td>
<td>0.077</td>
<td>0.084</td>
</tr>
<tr>
<td>11</td>
<td>0.008</td>
<td>0.009</td>
<td>0.010</td>
<td>0.012</td>
<td>0.014</td>
<td>0.016</td>
<td>0.018</td>
<td>0.020</td>
<td>0.023</td>
<td>0.025</td>
<td>0.028</td>
<td>0.031</td>
<td>0.035</td>
<td>0.039</td>
<td>0.043</td>
</tr>
<tr>
<td>12</td>
<td>0.003</td>
<td>0.003</td>
<td>0.004</td>
<td>0.005</td>
<td>0.006</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.010</td>
<td>0.011</td>
<td>0.013</td>
<td>0.014</td>
<td>0.016</td>
<td>0.018</td>
<td>0.020</td>
</tr>
<tr>
<td>13</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.004</td>
<td>0.005</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td>14</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>(n_o)</td>
<td>6.1</td>
<td>6.2</td>
<td>6.3</td>
<td>6.4</td>
<td>6.5</td>
<td>6.6</td>
<td>6.7</td>
<td>6.8</td>
<td>6.9</td>
<td>7.0</td>
<td>7.1</td>
<td>7.2</td>
<td>7.3</td>
<td>7.4</td>
<td>7.5</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>.998</td>
<td>.998</td>
<td>.998</td>
<td>.998</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>2</td>
<td>.984</td>
<td>.985</td>
<td>.987</td>
<td>.988</td>
<td>.989</td>
<td>.990</td>
<td>.991</td>
<td>.991</td>
<td>.992</td>
<td>.993</td>
<td>.993</td>
<td>.994</td>
<td>.994</td>
<td>.995</td>
<td>.995</td>
</tr>
<tr>
<td>3</td>
<td>.942</td>
<td>.946</td>
<td>.950</td>
<td>.954</td>
<td>.957</td>
<td>.960</td>
<td>.963</td>
<td>.966</td>
<td>.968</td>
<td>.970</td>
<td>.973</td>
<td>.975</td>
<td>.976</td>
<td>.978</td>
<td>.980</td>
</tr>
<tr>
<td>4</td>
<td>.858</td>
<td>.866</td>
<td>.874</td>
<td>.881</td>
<td>.888</td>
<td>.895</td>
<td>.901</td>
<td>.907</td>
<td>.913</td>
<td>.918</td>
<td>.923</td>
<td>.928</td>
<td>.933</td>
<td>.937</td>
<td>.941</td>
</tr>
<tr>
<td>5</td>
<td>.728</td>
<td>.741</td>
<td>.753</td>
<td>.765</td>
<td>.776</td>
<td>.787</td>
<td>.798</td>
<td>.808</td>
<td>.818</td>
<td>.827</td>
<td>.836</td>
<td>.845</td>
<td>.853</td>
<td>.861</td>
<td>.868</td>
</tr>
<tr>
<td>8</td>
<td>.270</td>
<td>.284</td>
<td>.298</td>
<td>.313</td>
<td>.327</td>
<td>.342</td>
<td>.357</td>
<td>.372</td>
<td>.386</td>
<td>.401</td>
<td>.416</td>
<td>.431</td>
<td>.446</td>
<td>.461</td>
<td>.475</td>
</tr>
<tr>
<td>9</td>
<td>.163</td>
<td>.174</td>
<td>.185</td>
<td>.197</td>
<td>.208</td>
<td>.220</td>
<td>.233</td>
<td>.245</td>
<td>.258</td>
<td>.271</td>
<td>.284</td>
<td>.297</td>
<td>.311</td>
<td>.324</td>
<td>.338</td>
</tr>
<tr>
<td>10</td>
<td>.091</td>
<td>.098</td>
<td>.106</td>
<td>.114</td>
<td>.123</td>
<td>.131</td>
<td>.140</td>
<td>.150</td>
<td>.151</td>
<td>.170</td>
<td>.180</td>
<td>.190</td>
<td>.201</td>
<td>.212</td>
<td>.223</td>
</tr>
<tr>
<td>11</td>
<td>.047</td>
<td>.051</td>
<td>.056</td>
<td>.061</td>
<td>.067</td>
<td>.073</td>
<td>.079</td>
<td>.085</td>
<td>.092</td>
<td>.099</td>
<td>.106</td>
<td>.113</td>
<td>.121</td>
<td>.129</td>
<td>.138</td>
</tr>
<tr>
<td>12</td>
<td>.022</td>
<td>.025</td>
<td>.028</td>
<td>.031</td>
<td>.034</td>
<td>.037</td>
<td>.041</td>
<td>.045</td>
<td>.049</td>
<td>.053</td>
<td>.058</td>
<td>.063</td>
<td>.068</td>
<td>.074</td>
<td>.079</td>
</tr>
<tr>
<td>13</td>
<td>.010</td>
<td>.011</td>
<td>.013</td>
<td>.014</td>
<td>.016</td>
<td>.018</td>
<td>.020</td>
<td>.022</td>
<td>.025</td>
<td>.027</td>
<td>.030</td>
<td>.033</td>
<td>.036</td>
<td>.039</td>
<td>.043</td>
</tr>
<tr>
<td>14</td>
<td>.004</td>
<td>.005</td>
<td>.006</td>
<td>.007</td>
<td>.008</td>
<td>.009</td>
<td>.010</td>
<td>.012</td>
<td>.013</td>
<td>.014</td>
<td>.016</td>
<td>.018</td>
<td>.020</td>
<td>.022</td>
<td>.022</td>
</tr>
<tr>
<td>15</td>
<td>.002</td>
<td>.002</td>
<td>.002</td>
<td>.003</td>
<td>.003</td>
<td>.003</td>
<td>.004</td>
<td>.004</td>
<td>.005</td>
<td>.006</td>
<td>.007</td>
<td>.007</td>
<td>.008</td>
<td>.009</td>
<td>.010</td>
</tr>
<tr>
<td>16</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.002</td>
<td>.002</td>
<td>.003</td>
<td>.003</td>
<td>.004</td>
<td>.004</td>
<td>.005</td>
<td>.005</td>
</tr>
<tr>
<td>17</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.002</td>
<td>.002</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(n_o)</th>
<th>7.6</th>
<th>7.7</th>
<th>7.8</th>
<th>7.9</th>
<th>8.0</th>
<th>8.1</th>
<th>8.2</th>
<th>8.3</th>
<th>8.4</th>
<th>8.5</th>
<th>8.6</th>
<th>8.7</th>
<th>8.8</th>
<th>8.9</th>
<th>9.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.999</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>.996</td>
<td>.996</td>
<td>.996</td>
<td>.997</td>
<td>.997</td>
<td>.997</td>
<td>.997</td>
<td>.998</td>
<td>.998</td>
<td>.998</td>
<td>.998</td>
<td>.998</td>
<td>.998</td>
<td>.998</td>
<td>.999</td>
</tr>
<tr>
<td>3</td>
<td>.981</td>
<td>.983</td>
<td>.984</td>
<td>.985</td>
<td>.986</td>
<td>.987</td>
<td>.988</td>
<td>.989</td>
<td>.990</td>
<td>.991</td>
<td>.991</td>
<td>.992</td>
<td>.993</td>
<td>.993</td>
<td>.993</td>
</tr>
<tr>
<td>4</td>
<td>.945</td>
<td>.948</td>
<td>.952</td>
<td>.955</td>
<td>.958</td>
<td>.960</td>
<td>.963</td>
<td>.965</td>
<td>.968</td>
<td>.970</td>
<td>.972</td>
<td>.974</td>
<td>.976</td>
<td>.977</td>
<td>.979</td>
</tr>
<tr>
<td>5</td>
<td>.875</td>
<td>.882</td>
<td>.888</td>
<td>.895</td>
<td>.900</td>
<td>.906</td>
<td>.911</td>
<td>.916</td>
<td>.921</td>
<td>.926</td>
<td>.930</td>
<td>.934</td>
<td>.938</td>
<td>.942</td>
<td>.945</td>
</tr>
<tr>
<td>6</td>
<td>.769</td>
<td>.780</td>
<td>.790</td>
<td>.799</td>
<td>.809</td>
<td>.818</td>
<td>.826</td>
<td>.835</td>
<td>.843</td>
<td>.850</td>
<td>.858</td>
<td>.865</td>
<td>.872</td>
<td>.878</td>
<td>.884</td>
</tr>
<tr>
<td>7</td>
<td>.635</td>
<td>.649</td>
<td>.662</td>
<td>.674</td>
<td>.687</td>
<td>.699</td>
<td>.710</td>
<td>.722</td>
<td>.733</td>
<td>.744</td>
<td>.754</td>
<td>.765</td>
<td>.774</td>
<td>.784</td>
<td>.793</td>
</tr>
<tr>
<td>8</td>
<td>.490</td>
<td>.504</td>
<td>.519</td>
<td>.533</td>
<td>.547</td>
<td>.561</td>
<td>.575</td>
<td>.588</td>
<td>.601</td>
<td>.614</td>
<td>.627</td>
<td>.640</td>
<td>.652</td>
<td>.664</td>
<td>.676</td>
</tr>
<tr>
<td>$n_o$</td>
<td>7.6</td>
<td>7.7</td>
<td>7.8</td>
<td>7.9</td>
<td>8.0</td>
<td>8.1</td>
<td>8.2</td>
<td>8.3</td>
<td>8.4</td>
<td>8.5</td>
<td>8.6</td>
<td>8.7</td>
<td>8.8</td>
<td>8.9</td>
<td>9.0</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.235</td>
<td>.247</td>
<td>.259</td>
<td>.271</td>
<td>.283</td>
<td>.296</td>
<td>.309</td>
<td>.321</td>
<td>.334</td>
<td>.347</td>
<td>.360</td>
<td>.373</td>
<td>.386</td>
<td>.399</td>
<td>.413</td>
</tr>
<tr>
<td>11</td>
<td>.147</td>
<td>.156</td>
<td>.165</td>
<td>.174</td>
<td>.184</td>
<td>.194</td>
<td>.204</td>
<td>.215</td>
<td>.226</td>
<td>.237</td>
<td>.248</td>
<td>.259</td>
<td>.271</td>
<td>.282</td>
<td>.294</td>
</tr>
<tr>
<td>12</td>
<td>.085</td>
<td>.092</td>
<td>.098</td>
<td>.105</td>
<td>.112</td>
<td>.119</td>
<td>.127</td>
<td>.135</td>
<td>.143</td>
<td>.151</td>
<td>.160</td>
<td>.169</td>
<td>.178</td>
<td>.187</td>
<td>.197</td>
</tr>
<tr>
<td>13</td>
<td>.046</td>
<td>.050</td>
<td>.055</td>
<td>.059</td>
<td>.064</td>
<td>.069</td>
<td>.074</td>
<td>.079</td>
<td>.085</td>
<td>.091</td>
<td>.097</td>
<td>.104</td>
<td>.110</td>
<td>.117</td>
<td>.124</td>
</tr>
<tr>
<td>14</td>
<td>.024</td>
<td>.026</td>
<td>.029</td>
<td>.031</td>
<td>.034</td>
<td>.037</td>
<td>.041</td>
<td>.044</td>
<td>.048</td>
<td>.051</td>
<td>.056</td>
<td>.060</td>
<td>.064</td>
<td>.069</td>
<td>.074</td>
</tr>
<tr>
<td>15</td>
<td>.011</td>
<td>.013</td>
<td>.014</td>
<td>.016</td>
<td>.017</td>
<td>.019</td>
<td>.021</td>
<td>.023</td>
<td>.025</td>
<td>.027</td>
<td>.030</td>
<td>.033</td>
<td>.035</td>
<td>.038</td>
<td>.042</td>
</tr>
<tr>
<td>16</td>
<td>.005</td>
<td>.006</td>
<td>.007</td>
<td>.007</td>
<td>.008</td>
<td>.009</td>
<td>.010</td>
<td>.011</td>
<td>.013</td>
<td>.014</td>
<td>.015</td>
<td>.017</td>
<td>.018</td>
<td>.020</td>
<td>.022</td>
</tr>
<tr>
<td>17</td>
<td>.002</td>
<td>.003</td>
<td>.003</td>
<td>.003</td>
<td>.004</td>
<td>.004</td>
<td>.005</td>
<td>.005</td>
<td>.006</td>
<td>.007</td>
<td>.008</td>
<td>.009</td>
<td>.010</td>
<td>.011</td>
<td>.011</td>
</tr>
<tr>
<td>18</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.002</td>
<td>.002</td>
<td>.003</td>
<td>.003</td>
<td>.003</td>
<td>.004</td>
<td>.004</td>
<td>.005</td>
<td>.005</td>
<td>.005</td>
</tr>
<tr>
<td>19</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>.002</td>
<td>.002</td>
<td>.002</td>
<td>.002</td>
</tr>
<tr>
<td>20</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
</tr>
</tbody>
</table>

*Values of the cumulative Poisson distribution, depending on $\lambda = n \pi$ (sample size $\times$ theoretical proportion of occurrences in random trials) and $n_o$ (number of occurrences observed). Given $\lambda$ and $n_o$, the corresponding entry in the table body represents the probability that $n_o$ or more occurrences (or alternatively that $n_o/n$ proportion observed occurrences) would have been observed by chance alone. (Probability is always 0.000 for $n_o=0$.)
## Table VIII Signed-rank probabilities.\(^a\)

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.250</td>
<td>.125</td>
<td>.063</td>
<td>.031</td>
<td>.016</td>
<td>.008</td>
<td>.004</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>2</td>
<td>.375</td>
<td>.188</td>
<td>.094</td>
<td>.047</td>
<td>.024</td>
<td>.012</td>
<td>.006</td>
<td>.003</td>
<td>.002</td>
</tr>
<tr>
<td>3</td>
<td>.625</td>
<td>.313</td>
<td>.156</td>
<td>.078</td>
<td>.039</td>
<td>.020</td>
<td>.010</td>
<td>.005</td>
<td>.003</td>
</tr>
<tr>
<td>4</td>
<td>.875</td>
<td>.438</td>
<td>.219</td>
<td>.109</td>
<td>.055</td>
<td>.027</td>
<td>.014</td>
<td>.007</td>
<td>.004</td>
</tr>
<tr>
<td>5</td>
<td>.625</td>
<td>.313</td>
<td>.156</td>
<td>.078</td>
<td>.039</td>
<td>.020</td>
<td>.010</td>
<td>.005</td>
<td>.005</td>
</tr>
<tr>
<td>6</td>
<td>.801</td>
<td>.438</td>
<td>.219</td>
<td>.109</td>
<td>.055</td>
<td>.027</td>
<td>.014</td>
<td>.007</td>
<td>.007</td>
</tr>
<tr>
<td>7</td>
<td>1.000</td>
<td>.563</td>
<td>.297</td>
<td>.149</td>
<td>.074</td>
<td>.037</td>
<td>.019</td>
<td>.009</td>
<td>.009</td>
</tr>
<tr>
<td>8</td>
<td>.688</td>
<td>.375</td>
<td>.195</td>
<td>.098</td>
<td>.049</td>
<td>.025</td>
<td>.012</td>
<td>.009</td>
<td>.012</td>
</tr>
<tr>
<td>9</td>
<td>.844</td>
<td>.469</td>
<td>.250</td>
<td>.129</td>
<td>.065</td>
<td>.032</td>
<td>.016</td>
<td>.016</td>
<td>.016</td>
</tr>
<tr>
<td>10</td>
<td>1.000</td>
<td>.578</td>
<td>.313</td>
<td>.164</td>
<td>.084</td>
<td>.042</td>
<td>.021</td>
<td>.021</td>
<td>.021</td>
</tr>
<tr>
<td>11</td>
<td>.688</td>
<td>.383</td>
<td>.203</td>
<td>.106</td>
<td>.054</td>
<td>.027</td>
<td>.017</td>
<td>.017</td>
<td>.017</td>
</tr>
<tr>
<td>12</td>
<td>.813</td>
<td>.461</td>
<td>.250</td>
<td>.131</td>
<td>.067</td>
<td>.034</td>
<td>.022</td>
<td>.022</td>
<td>.022</td>
</tr>
<tr>
<td>13</td>
<td>.938</td>
<td>.547</td>
<td>.301</td>
<td>.160</td>
<td>.083</td>
<td>.043</td>
<td>.029</td>
<td>.029</td>
<td>.029</td>
</tr>
<tr>
<td>14</td>
<td>.641</td>
<td>.359</td>
<td>.193</td>
<td>.102</td>
<td>.052</td>
<td>.029</td>
<td>.019</td>
<td>.019</td>
<td>.019</td>
</tr>
<tr>
<td>15</td>
<td>.742</td>
<td>.426</td>
<td>.233</td>
<td>.112</td>
<td>.064</td>
<td>.032</td>
<td>.021</td>
<td>.021</td>
<td>.021</td>
</tr>
<tr>
<td>16</td>
<td>.844</td>
<td>.496</td>
<td>.275</td>
<td>.148</td>
<td>.077</td>
<td>.040</td>
<td>.029</td>
<td>.029</td>
<td>.029</td>
</tr>
<tr>
<td>17</td>
<td>.945</td>
<td>.570</td>
<td>.322</td>
<td>.175</td>
<td>.092</td>
<td>.051</td>
<td>.031</td>
<td>.031</td>
<td>.031</td>
</tr>
<tr>
<td>18</td>
<td>.652</td>
<td>.375</td>
<td>.206</td>
<td>.110</td>
<td>.069</td>
<td>.041</td>
<td>.029</td>
<td>.029</td>
<td>.029</td>
</tr>
<tr>
<td>19</td>
<td>.734</td>
<td>.432</td>
<td>.240</td>
<td>.129</td>
<td>.076</td>
<td>.047</td>
<td>.031</td>
<td>.031</td>
<td>.031</td>
</tr>
<tr>
<td>20</td>
<td>.820</td>
<td>.492</td>
<td>.278</td>
<td>.151</td>
<td>.085</td>
<td>.054</td>
<td>.033</td>
<td>.033</td>
<td>.033</td>
</tr>
<tr>
<td>21</td>
<td>.910</td>
<td>.557</td>
<td>.320</td>
<td>.176</td>
<td>.099</td>
<td>.061</td>
<td>.040</td>
<td>.040</td>
<td>.040</td>
</tr>
<tr>
<td>22</td>
<td>1.000</td>
<td>.625</td>
<td>.365</td>
<td>.204</td>
<td>.104</td>
<td>.067</td>
<td>.044</td>
<td>.044</td>
<td>.044</td>
</tr>
<tr>
<td>23</td>
<td>.695</td>
<td>.413</td>
<td>.233</td>
<td>.123</td>
<td>.076</td>
<td>.050</td>
<td>.032</td>
<td>.032</td>
<td>.032</td>
</tr>
<tr>
<td>24</td>
<td>.770</td>
<td>.465</td>
<td>.266</td>
<td>.151</td>
<td>.085</td>
<td>.060</td>
<td>.040</td>
<td>.040</td>
<td>.040</td>
</tr>
<tr>
<td>25</td>
<td>.846</td>
<td>.520</td>
<td>.301</td>
<td>.176</td>
<td>.099</td>
<td>.069</td>
<td>.044</td>
<td>.044</td>
<td>.044</td>
</tr>
<tr>
<td>26</td>
<td>.922</td>
<td>.577</td>
<td>.339</td>
<td>.199</td>
<td>.111</td>
<td>.083</td>
<td>.056</td>
<td>.056</td>
<td>.056</td>
</tr>
<tr>
<td>27</td>
<td>1.000</td>
<td>.638</td>
<td>.380</td>
<td>.225</td>
<td>.128</td>
<td>.094</td>
<td>.062</td>
<td>.062</td>
<td>.062</td>
</tr>
<tr>
<td>28</td>
<td>.700</td>
<td>.424</td>
<td>.254</td>
<td>.151</td>
<td>.088</td>
<td>.061</td>
<td>.044</td>
<td>.044</td>
<td>.044</td>
</tr>
<tr>
<td>29</td>
<td>.765</td>
<td>.470</td>
<td>.282</td>
<td>.176</td>
<td>.111</td>
<td>.091</td>
<td>.069</td>
<td>.069</td>
<td>.069</td>
</tr>
<tr>
<td>30</td>
<td>.831</td>
<td>.519</td>
<td>.319</td>
<td>.201</td>
<td>.132</td>
<td>.103</td>
<td>.083</td>
<td>.083</td>
<td>.083</td>
</tr>
<tr>
<td>32</td>
<td>.966</td>
<td>.622</td>
<td>.390</td>
<td>.255</td>
<td>.180</td>
<td>.150</td>
<td>.133</td>
<td>.133</td>
<td>.133</td>
</tr>
<tr>
<td>33</td>
<td>.677</td>
<td>.473</td>
<td>.326</td>
<td>.223</td>
<td>.128</td>
<td>.092</td>
<td>.065</td>
<td>.065</td>
<td>.065</td>
</tr>
<tr>
<td>34</td>
<td>.733</td>
<td>.519</td>
<td>.361</td>
<td>.250</td>
<td>.158</td>
<td>.127</td>
<td>.107</td>
<td>.107</td>
<td>.107</td>
</tr>
<tr>
<td>35</td>
<td>.791</td>
<td>.567</td>
<td>.406</td>
<td>.277</td>
<td>.188</td>
<td>.157</td>
<td>.148</td>
<td>.148</td>
<td>.148</td>
</tr>
<tr>
<td>36</td>
<td>.850</td>
<td>.613</td>
<td>.452</td>
<td>.303</td>
<td>.219</td>
<td>.187</td>
<td>.179</td>
<td>.179</td>
<td>.179</td>
</tr>
<tr>
<td>37</td>
<td>.910</td>
<td>.659</td>
<td>.500</td>
<td>.331</td>
<td>.250</td>
<td>.219</td>
<td>.211</td>
<td>.211</td>
<td>.211</td>
</tr>
<tr>
<td>38</td>
<td>.970</td>
<td>.687</td>
<td>.548</td>
<td>.358</td>
<td>.281</td>
<td>.250</td>
<td>.243</td>
<td>.243</td>
<td>.243</td>
</tr>
</tbody>
</table>

\(^a\)Two-tailed probabilities for the distribution of \(T\), the signed-rank statistic. For a sample size \(n\) and value of \(T\), the entry gives the \(p\)-value. If a one-tailed test is appropriate, halve the entry. If \(n > 12\), go to Section 11.9.
### Table IX  Rank-sum $U$ probabilities.$^a$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>$n_1$:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.500</td>
<td>.200</td>
<td>.100</td>
<td>.400</td>
<td>.133</td>
<td>.056</td>
<td>.028</td>
<td>.333</td>
<td>.094</td>
<td>.036</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>.400</td>
<td>.200</td>
<td>.800</td>
<td>.267</td>
<td>.114</td>
<td>.058</td>
<td>.337</td>
<td>.190</td>
<td>.072</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.800</td>
<td>.400</td>
<td>.628</td>
<td>.534</td>
<td>.228</td>
<td>.114</td>
<td>.267</td>
<td>.190</td>
<td>.072</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.200</td>
<td>.700</td>
<td>.400</td>
<td>.800</td>
<td>.400</td>
<td>.200</td>
<td>.267</td>
<td>.142</td>
<td>.064</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.100</td>
<td>1.00</td>
<td>.858</td>
<td>.628</td>
<td>.342</td>
<td>.190</td>
<td>.096</td>
<td>.112</td>
<td>.056</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.286</td>
<td>.072</td>
<td>.024</td>
<td>.013</td>
<td>.004</td>
<td>.002</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.572</td>
<td>.142</td>
<td>.048</td>
<td>.020</td>
<td>.008</td>
<td>.004</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.856</td>
<td>.286</td>
<td>.096</td>
<td>.038</td>
<td>.018</td>
<td>.008</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.428</td>
<td>.166</td>
<td>.066</td>
<td>.030</td>
<td>.016</td>
<td>.008</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.642</td>
<td>.262</td>
<td>.114</td>
<td>.052</td>
<td>.026</td>
<td>.015</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.858</td>
<td>.380</td>
<td>.172</td>
<td>.082</td>
<td>.042</td>
<td>.016</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.548</td>
<td>.528</td>
<td>.126</td>
<td>.064</td>
<td>.016</td>
<td>.001</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.714</td>
<td>.352</td>
<td>.178</td>
<td>.094</td>
<td>.016</td>
<td>.001</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.904</td>
<td>.476</td>
<td>.246</td>
<td>.132</td>
<td>.066</td>
<td>.016</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.762</td>
<td>.428</td>
<td>.240</td>
<td>.132</td>
<td>.066</td>
<td>.016</td>
<td>.001</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.914</td>
<td>.536</td>
<td>.310</td>
<td>.190</td>
<td>.096</td>
<td>.032</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
<td>.002</td>
</tr>
</tbody>
</table>

$^a$ Probabilities for the rank-sum test. The table shows $U$ probabilities for different combinations of $n_1$ and $n_2$. Each cell represents the probability of obtaining a rank-sum $U$ value less than or equal to the one shown, given the specified $n_1$ and $n_2$. The probabilities are calculated using the rank-sum test, which is a non-parametric test used to compare two independent samples.
<table>
<thead>
<tr>
<th>$n_2$:</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$:</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>.700</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>.930</td>
<td>.588</td>
</tr>
<tr>
<td>17</td>
<td>.628</td>
<td>.382</td>
</tr>
<tr>
<td>18</td>
<td>.628</td>
<td>.382</td>
</tr>
<tr>
<td>19</td>
<td>.628</td>
<td>.382</td>
</tr>
<tr>
<td>20</td>
<td>.628</td>
<td>.382</td>
</tr>
<tr>
<td>21</td>
<td>.628</td>
<td>.382</td>
</tr>
<tr>
<td>22</td>
<td>.628</td>
<td>.382</td>
</tr>
<tr>
<td>23</td>
<td>.628</td>
<td>.382</td>
</tr>
<tr>
<td>24</td>
<td>.628</td>
<td>.382</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n_2$:</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.222</td>
<td>.044</td>
<td>.012</td>
<td>.004</td>
<td>.002</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>1</td>
<td>.444</td>
<td>.088</td>
<td>.024</td>
<td>.008</td>
<td>.004</td>
<td>.002</td>
<td>.001</td>
<td>.001</td>
</tr>
<tr>
<td>2</td>
<td>.666</td>
<td>.178</td>
<td>.048</td>
<td>.016</td>
<td>.006</td>
<td>.003</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>3</td>
<td>.888</td>
<td>.266</td>
<td>.082</td>
<td>.028</td>
<td>.010</td>
<td>.004</td>
<td>.003</td>
<td>.002</td>
</tr>
<tr>
<td>4</td>
<td>.400</td>
<td>.134</td>
<td>.048</td>
<td>.018</td>
<td>.008</td>
<td>.004</td>
<td>.002</td>
<td>.002</td>
</tr>
<tr>
<td>5</td>
<td>.534</td>
<td>.194</td>
<td>.072</td>
<td>.030</td>
<td>.012</td>
<td>.006</td>
<td>.003</td>
<td>.003</td>
</tr>
<tr>
<td>6</td>
<td>.712</td>
<td>.278</td>
<td>.110</td>
<td>.046</td>
<td>.020</td>
<td>.010</td>
<td>.004</td>
<td>.004</td>
</tr>
<tr>
<td>7</td>
<td>.888</td>
<td>.376</td>
<td>.154</td>
<td>.066</td>
<td>.030</td>
<td>.014</td>
<td>.006</td>
<td>.006</td>
</tr>
<tr>
<td>8</td>
<td>.496</td>
<td>.214</td>
<td>.094</td>
<td>.042</td>
<td>.020</td>
<td>.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>.630</td>
<td>.282</td>
<td>.138</td>
<td>.060</td>
<td>.028</td>
<td>.014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.774</td>
<td>.368</td>
<td>.170</td>
<td>.082</td>
<td>.040</td>
<td>.020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>.922</td>
<td>.460</td>
<td>.222</td>
<td>.108</td>
<td>.052</td>
<td>.028</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>.570</td>
<td>.284</td>
<td>.142</td>
<td>.072</td>
<td>.038</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>.682</td>
<td>.354</td>
<td>.182</td>
<td>.094</td>
<td>.050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>.808</td>
<td>.434</td>
<td>.228</td>
<td>.120</td>
<td>.064</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>.934</td>
<td>.524</td>
<td>.282</td>
<td>.152</td>
<td>.082</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n_2: )</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>( n_1: )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( U )</th>
<th>( 16 )</th>
<th>( 17 )</th>
<th>( 18 )</th>
<th>( 19 )</th>
<th>( 20 )</th>
<th>( 21 )</th>
<th>( 22 )</th>
<th>( 23 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( .622 )</td>
<td>( .724 )</td>
<td>( .832 )</td>
<td>( .944 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( .344 )</td>
<td>( .414 )</td>
<td>( .490 )</td>
<td>( .572 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( .190 )</td>
<td>( .232 )</td>
<td>( .280 )</td>
<td>( .336 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( .104 )</td>
<td>( .130 )</td>
<td>( .160 )</td>
<td>( .194 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For two samples of size \( n_1 \) and \( n_2 \) (\( n_2 > n_1 \)) and the value of \( U \), the entry gives the \( p \)-value. If a one-tailed test is appropriate, halve the entry. If \( n_2 > 8 \), go to Section 11.9. It can be seen from inspection that, if \( n_1 + n_2 < 7 \), a rejection of \( H_0 \) at \( \alpha = 0.05 \) is not possible.
# Symbol index

## Classes of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subscript</td>
<td>designates member of family indicated by symbol to which attached</td>
</tr>
<tr>
<td>Superscript</td>
<td>power of symbol to which attached</td>
</tr>
<tr>
<td>Superscript</td>
<td>(in parentheses) acts as subscript when subscript position is preempted</td>
</tr>
<tr>
<td>Greek letters</td>
<td>population parameters</td>
</tr>
<tr>
<td>Roman letters</td>
<td>corresponding to a Greek letter, the corresponding sample estimate</td>
</tr>
</tbody>
</table>

## Mathematical Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>!</td>
<td>(following an integer, e.g., ( k )) factorial; ( k! = k \times (k - 1) \times \cdots \times 2 \times 1 )</td>
</tr>
<tr>
<td>+</td>
<td>plus (add what appears before it to what appears after it)</td>
</tr>
<tr>
<td>−</td>
<td>minus (subtract what appears after it from what appears before it)</td>
</tr>
<tr>
<td>±</td>
<td>plus or minus, according to context</td>
</tr>
<tr>
<td>×</td>
<td>times (multiply what appears before it by what appears after it); by (indicating two dimensions of an entity, as a ( 2 \times 2 ) table)</td>
</tr>
<tr>
<td>÷</td>
<td>division (divide what appears before it by what appears after it)</td>
</tr>
<tr>
<td>=</td>
<td>equals (what appears before it is the same as what appears after it)</td>
</tr>
<tr>
<td>≠</td>
<td>unequal (what appears before it is different from what appears after it)</td>
</tr>
<tr>
<td>&lt;</td>
<td>less than (what appears before it is less than what appears after it)</td>
</tr>
<tr>
<td>≤</td>
<td>less than or equal to</td>
</tr>
<tr>
<td>&gt;</td>
<td>greater than (what appears before it is greater than what appears after it)</td>
</tr>
<tr>
<td>≥</td>
<td>greater than or equal to</td>
</tr>
<tr>
<td>≈</td>
<td>approximately equal to</td>
</tr>
<tr>
<td>≡</td>
<td>identical to (always equal to)</td>
</tr>
<tr>
<td>⇒</td>
<td>implies (e.g., A implies B is denoted as ( A \Rightarrow B ))</td>
</tr>
<tr>
<td>( \sqrt{} )</td>
<td>square root (taken of what comes under the bar following it)</td>
</tr>
<tr>
<td>( \circ )</td>
<td>in position of a superscript, sometimes indicates a ranked observation</td>
</tr>
<tr>
<td>%</td>
<td>percent (proportion of 100 units)</td>
</tr>
<tr>
<td>( 1 )</td>
<td>given (what comes before it is conditional upon what comes after it)</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>absolute (what appears within bars has minus sign, if any, dropped)</td>
</tr>
<tr>
<td>( e )</td>
<td>natural number (approximately 2.7183); base of exponential distribution</td>
</tr>
</tbody>
</table>
**APPENDIX 4 Symbol index**

\( f(\cdot) \) function of what appears within parentheses

\( \log(\cdot) \) logarithm of what appears within parentheses, often to base 10

\( \ln(\cdot) \) natural logarithm of what appears within parentheses to base \( e \)

\( \mathbf{M} \) (boldface upper case letter) symbol for matrix \( \mathbf{M} \); a rectangular array of quantities

\( \mathbf{M}_{(r,c)} \) matrix \( \mathbf{M} \) with dimensions \( r \times c \), that is, \( r \) rows and \( c \) columns

\( \binom{n}{k} \) combination of \( n \) things combined \( k \) at a time; \( \binom{n}{k} = n!(k!(n-k)!) \)

---

**STATISTICAL SYMBOLS AND TERMINOLOGY**

\( \sim \) is distributed as [e.g., \( x \sim N(0,1) \)]

\( \approx \) is approximately distributed as, where the approximation becomes increasingly better with a larger sample size [e.g., \( x \approx N(0,1) \)]

ANOVA analysis of variance

ANCOVA analysis of covariance

AR attributable risk

\( \alpha \) (alpha) the upper bound placed on the risk of a false positive outcome when planning a study; name of a particular probability distribution

\( \beta \) (beta) the risk of a false negative outcome under an assumed effect when planning a study; coefficient (multiplier) of a variable in a regression or other equation, often theoretical or modeled equation; name of a particular probability distribution

\( b \) often, a sample estimate of a \( \beta \) coefficient

\( \chi^2 \) (chi-square) a particular probability distribution; a value that obeys that distribution

\( c \) sometimes, column of a data table

CI confidence interval

\( \text{cov}(x,y) \) covariance of \( x \) with \( y \) (or whatever variables are indicated)

\( \delta \) a theoretical or putative difference

\( \Delta \) a clinically relevant or important difference

\( d \) a sample difference between two defined quantities; an estimate of \( \delta \)

\( df \) degrees of freedom

\( e \) expected value; also see under Mathematical Symbols

\( F \) a particular probability distribution; a value that obeys that distribution

FET Fisher’s exact test of contingency

\( H \) hypothesis designator, as \( H_0 \): null hypothesis, \( H_1 \) or \( H_a \): alternate hypothesis

\( i \) or \( j \) usually as subscript, indicator of member of set of variables

\( k \) as \( i \) or \( j \); as letter or superscript, indicator of position, number, or category

KS test Kolmogorov–Smirnov test

\( \lambda \) occurrence rate in a Poisson distribution

LR likelihood ratio

\( \mu \) (mu) mean of a population or a probability distribution

\( m \) mean of a sample
md  median
MS  mean square, as MSE: MS error, MSM: MS means, MST: MS total
mo  mode
MOE measure of effectiveness
n   number of members of a sample
NNB number needed to benefit
NNT number needed to treat
NPV negative predictive value
OR  odds ratio
π (pi) (1) a theoretical or population probability or proportion and
     (2) a product of terms following
P   among many meanings, (1) a probability; (2) a proportion; (3) as a “p-value”,
     given observed data from a study, the post hoc probability of observing data as or
     more indicative of an alternative hypothesis (e.g., a treatment effect) than when
     the null hypothesis (e.g., no treatment effect) is true; and (4) as a subscript,
     symbol to which it is attached is pooled over two or more samples
Parameter  a population-level characteristic of the distribution of response variable, for
            example, a population mean. A sample statistic (e.g., the sample mean) is used
            to estimate a parameter (e.g., the population mean)
PPV positive predictive value
ρ (rho) a theoretical or population correlation coefficient
r   (1) a sample correlation coefficient and (2) sometimes, number of rows
    of a data table
R^2  (1) coefficient of determination and (2) if a single variable, r^2
ROC receiver operating characteristic
RR  relative risk, risk ratio
Σ   (1) sum of values appearing after it; (2) sometimes, dispersion
    (variance—covariance) matrix in multivariate analysis; and (3) as otherwise
    defined
σ   population or theoretical standard deviation (square root of variance)
σ^2 population or theoretical variance (square of standard deviation)
s   sample standard deviation
s^2 sample variance
SED standard error of the difference
SEE standard error of the estimate
SEM standard error of the mean
SS  sum of squares, as SSE: SS error, SSM: SS means, SST: SS Total
t   (1) a particular probability distribution and (2) a value that obeys that distribution
T   (1) a sum of ranks in rank-order tests and (2) base of symbol for Hotelling’s
    distribution
TOST two one-sided tests (in equivalence testing)
U   the statistic of the Mann—Whitney form of the rank-sum test
x or y a variable; a value of that variable
x̄  alternative notation for \( m_x \), the mean of variable \( x \); also \( \bar{y} \), etc.
z   (1) often, the normal (or Gaussian) distribution and (2) a value that obeys that
distribution
Statistical Subject Index

Note: Page numbers followed by “t,” “f,” and “b” refer to figures, tables, and boxes, respectively.

A
A. See Riffenburgh’s A
Accuracy
  calculation, 220
  overview, 32
  10-year risk of ASCVD to ensure better accuracy, 2–3
  truth tables, 224
Additive rule, probability, 52
Adjustment variables in multiple regression analyses, 36–37
Administrative right censoring, 478–479
Agreement
  association in assessment, 338
  contingency as agreement, 344–345
  correlation as agreement, 345–346
  kappa for agreement among ratings
    example, 347–348
    method, 347b
    overview, 346–348
  multiple rankers
    example, 349–350, 352t
    Kendall’s W, 350
    Riffenburgh’s A, 350–352
    overview, 337
    reliability, 338
AI. See Artificial intelligence (AI)
α. See also Cronbach’s α; Type I error
  choosing α but not β in hypothesis testing, 146–147
  definition, 28
  ethical consequences, 25
Alternative hypothesis
  forms of hypothesis, 139
  means testing
    one- versus two-tailed test selection, 241
    specification, 241
  overview, 139
Analysis of covariance (ANCOVA)
  analysis of variance and regression, 395–396
  examples, 427–430
  purpose, 426–427
  table, 429b
Analysis of variance (ANOVA)
  Box-Cox transform, 651
  Brown-Forsythe test, 651
  equality of variances assumption, 651
  factorial designs, 434
  general linear models, 652
  Latin-square designs, 434
  means testing
    assumptions, 257
    means square, 257–259
    one-factor ANOVA
      example, 256–258
      method, 258b
      principles, 257
      table format, 258t
    sum of squares, 257–259
    trend test
      example, 266–267, 266t
      implementation, 266–267
      linear pattern testing, 265
      nonlinear pattern testing, 266
    pattern of change through means, 265
  missing data, 38
  nonadditivity testing, 652
  nonlinear models, 652
  one-factor ANOVA
    Bartlett’s test, 318b
    Levene’s test, 651
    multiple comparisons, 259–260
    post-hoc tests, 259
  sample size estimation, 558
  sparse-data designs, 434–435
  three-and-higher factor ANOVA
    examples, 430–431
    interpretation, 434
    method, 430b
    table, 431t
  two-factor ANOVA
    comparison with t test and one-factor ANOVA, 425
    examples, 402t, 420–423, 425–426
Analysis of variance (ANOVA) (Continued)
method, 422b

table interpretation, 422t, 423, 424t

Welch test, 651

ANOVA. See Analysis of covariance (ANCOVA)

ANOVA. See Analysis of variance (ANOVA)

AR. See Attributable risk (AR)

ARE. See Asymptotic relative efficiency (ARE)

Articles. See Journal articles

Artificial intelligence (AI), 666–667

Association
agreement assessment, 338
contingency tables
examples, 339, 341–342
overview, 338–339
r × c tables, 340–342
2 × 2 tables, 337
2 × c tables, 339–340
correlation as association
continuous or rank–order data, 342–344
element, 343–344
correlation coefficients
Kendall’s τ, 342–344
Pearson’s r, 342–346, 354
Spearman’s rs, 343–344, 376, 381–382
tetrachoric rs, 99–100, 184–185, 345
Cramér’s V, 340, 344–345
Cronbach’s α, 353–354, 629
F coefficient, 339
intraclasse correlation, 347, 354–355, 380
overview, 337
phi coefficient, 344–345
Association parameters, inference for, 445–447
odds ratio, confidence intervals for, 447, 447b
elements, 447
test of association, 445–447
elements, 446–447

Association testing, 45–49
a priori specification of a hypothesis, 45–46
setting up a test within a study, 45

Asymptotic relative efficiency (ARE), 557

Attributable risk (AR)
definition, 218t
overview, 222–223
relative risk relationship, 223–224
truth tables, 224

Autocorrelation, 519–523, 596–597

examples, 519, 521–522

Autoregressive covariance structure, 501

Average. See Mean; Median

Average deviation, 377

B

Bartlett’s test of homogeneity of variance, 318b
Bayes’ rule, probability, 53
Bayes’ Theorem, 591–592

Bayesian statistics
means testing
estimation methods, 614–615
elements, 613–616
shrinkage, 615
testing methods, 615–617
overview, 609–612
terminology, 612

Bayesian statistics, Markov chain Monte Carlo as, 658

Benefit, cost per unit of, 647

Benefits, statistics, 6

β. See also Type II error
choosing α but not β in hypothesis testing, 146–147
definition, 28

Bias
example, 13, 14f
experimental design in reduction, 8
overview, 8, 13–15
questionnaire, 629
regression variables, 652
representativeness increase by random samples, 14

sources, 14–15

survey, 623

Binomial distribution
large samples, 80–81, 132
overview, 80–81, 130

sample size effects, 131–132, 131f
table, 130–132, 131f

Binomial test, 196, 201–202, 617

Bivariate descriptive statistics, 43–45

Bland—Altman plot
examples, 643–647, 646f, 647f
overview, 642–647
Blinding, 573–574
Blinding, clinical trial, 10
Bonferroni method, multiple comparisons test, 259t, 262–263, 661
Bootstrap
assumptions, 640
examples, 641
mechanism, 639–640
overview, 639
Box-and-whisker chart, one variable descriptor, 106–110, 107f, 108t
Box–Cox transform, 651
Brown–Forsythe test, 651
Canonical correlation, 653–654
Case–control study
cohort study comparison, 9
design, 594
epidemiology, 589
overview, 4
Case–control study design, 591–593
measure of association in, 591–592
number and selection of controls, 592–593
Categorical data
association measures for, 215–224
categories versus ranks, 185
chi-square test of contingency
basis, 187
calculation, 187–188
percentage calculations, 189–190
sample size effects, 188–189, 189t
Yates’ correction, 188
Cochran’s Q test, 210–211
contingency table
degrees of freedom, 184b
larger tables, 193–195, 193b, 194t
row × column tables, 192–195
symbols, 183–184, 184t
2 × 2 table applications, 185–186
Fisher’s exact test of contingency
basis, 190b
examples, 191–192, 191t
overview, 186, 190–192
McNemar’s test, 205–209
organization
categorical methods, 183
inference, 152–153
overview, 33, 80
proportion testing
adjustment factor, 198–199
basis, 196
large samples, 196–197, 200
methods, 197b
overview, 195–201
rare events
basis, 201–202
examples, 201–204
large λ, 203–204
large samples, 201–205
method, 203b
no events observed, 202
π is near 1, 202
small λ, 204
three or more proportions, 197–198
two proportions, 197, 198b, 201
Royston’s ptrend test, 211–213
sensitivity, 220
sign test, 659
specificity, 220
terms and symbols, 183
Categorical variables, correlation coefficient, 99
Causation
causative processes versus data, 3
inference in epidemiology, 590–591
Censoring, 478–479
Central Limit Theorem, 83–86
Charts. See Descriptive statistics
Chernoff’s faces, 656
Chi-square
asymmetry effects, 127
contingency testing
basis, 187
calculation, 187–188
percentage calculations, 189–190
sample size effects, 188–189, 189t
Yates’ correction, 188
critical value, 79
distribution, 124–127, 125f, 126t
examples, 125–126
goodness-of-fit test of distribution normality
characteristics, 327
examples, 328–330, 328b, 331t
method, 328b
software, 323
graphical representation, 124, 125f
Chi-square (Continued)
overview, 78–79, 124–127
probabilities, 187, 227–228
table, 125–127, 126t
Chi-square and Fisher’s exact tests of contingency, 225
CI. See Confidence interval (CI)
Clinical relevance, 533, 536
Clinical trials
definition, 10
epidemiology, 589
fundamentals of, 568–573
comparison group, choosing, 572
evidence, statistical criteria for, 572–573
intervention, defining, 570
outcome, defining, 570–572
target population, defining, 569–570
interim analyses in, 577–582
fixed-sample methods, disadvantages of, 577–579
examples, 579–582
need for, 565–568
reducing bias in, 573–577
blinding, 573–574
randomization, 574–577
Cluster analysis, 656
Cluster sampling, 12
Cochran’s Q test, 210–211
Coefficient of correlation, 379–380
Coefficient of determination, regression assessment, 369, 371b
Cohen’s kappa. See Kappa
Cohort study
case-control study comparison, 9
epidemiology, 589
overview, 4
Comorbidity bias, 15
Comparison group, choosing, 572
Competing risk, 478–479
Conditional probability, 52–53
Confidence interval (CI)
applications, 118–119
asymmetry effects, 120–121
confusion with tolerance and prediction intervals, 162
correlation testing, 385–389
descriptive statistic use
correlation coefficient, 179–181
mean
estimated standard deviation, 169–171, 170t
population mean
known standard deviation, 166–169
population median, 175–176
proportions, 172–173
Rule of Three, 174
standard deviation, 176–179
variance, 176–179
Wald-based confidence intervals, performance of, 174
equivalence testing, 297–299
error rate, 162
hypothesis testing
providing, 147
relationship, 142
inference, 72–73
odds ratio, 232
origins, 118–119
overview, 161–162
for a parameter, 164–166
and asymmetry, 166
defined, 165
interval values and statistics, 165–166
regression assessment, 372b
sample size estimation
confidence intervals on means
β absent from confidence interval sample size estimation, 548–549
examples, 547–548
method, 547b
confidence intervals on proportions, 561–562
survival estimates, 482–484
Confidence intervals for rate ratio, 475, 475b
Confidence intervals on survival estimates, 482–484
Constant variance, 409
Contingency
as agreement, 344–345
chi-square test of contingency
basis, 187
calculation, 187–188
percentage calculations, 189–190
sample size effects, 188–189, 189t
Yates’ correction, 188
Fisher’s exact test of contingency
basis, 190b
examples, 191–192, 191r
overview, 186, 190–192

Contingency table

association
examples, 338–342
overview, 338–339
$r \times c$ tables, 340–342
$2 \times 2$ tables, 337
$2 \times c$ tables, 339–340

categorical data testing

Cochran’s Q test, 210–211
degrees of freedom, 184b
larger tables, 193–195, 193b, 194t
McNemar’s test, 205–209
row $\times$ column tables, 192–195
Royston’s $p_trend$ test, 211–213
symbols, 183–184, 184t
$2 \times 2$ table applications, 185–186
log odds ratio test, 226t, 231–232
sample size estimation and test of one
proportion, 549
symbols, 183–184, 184t
two variable descriptor, 99–100, 100t

Continuous autoregressive structure, 501–502

Continuous data
conversion to rank data, 154, 154t
means testing. See Means testing
overview, 32–33

Continuous random variables
frequency distributions for continuous variables
histogram, 58–60
internal selection, 61–62
relative frequencies, 61
sample size effects, 61
tallying data, 58, 60f
probability distribution, 57–58

Continuous variables, correlation coefficient,
96–97

Control group, 7

Controlled Oral Word Association Test
(COWAT)-FAS test, 467, 467t, 474t

Convergence with increasing sample size, 66–67

Cook’s distance, 453, 661
Correlated data, 491

Correlation
as agreement, 345–346
assessment
confidence interval, 385–389
examples, 385, 388
significance testing, 386b, 389
two correlation coefficient testing, 388–389
as association
continuous or rank-order data, 342–344
example, 343–344
assumptions, 381
autocorrelation, 519–523
examples, 519, 521–522
calculation, 382b
canonical correlation, 653–654
coefficient of correlation, 379–380
cross-correlation, 518
definition, 379–380
intra-class correlation coefficient, 380
nomenclature, 381–382
rank correlation, 381–382
regression relationship, 382–385
as reliability
intra-class correlation, 354–355
overview, 354
serial correlation, 518
small-but-significant correlation testing,
390
Spearman’s $r_s$, 383

Correlation coefficient. See also Intra-class
correlation (ICC)
calculation, 383–384
confidence interval on, 179–181
Kendall’s $t$, 342–344
Pearson’s $r$, 342–344, 354
sample size estimation tests, 555–557, 556t
Spearman’s $r_s$, 343–344, 376
tetrachoric $r$, 99–100, 183–184, 345
two correlation coefficient testing, 387b, 388–389
two variable descriptor
categorical variables, 99
continuous variables, 96–97
rank-order variables, 98–99, 98t

Correlation due to repeated measures, 493–495
example, 493

Cost-effectiveness, 640–641
element, 648

Covariance, two variable descriptor, 95–96

Covariance structures, 500–501
common choices of, 501–503
method for investigating, 502–503
COWAT-FAS test. See Controlled Oral Word Association Test (COWAT)-FAS test
Cox proportional hazards model, 594
Cox proportional hazards regression overview, 378–379, 378
survival curve analysis examples, 488, 490
log-rank test comparison, 486–487
model, 488–490
overview, 487–490
terminology, 488–489
Cramér’s V, 340, 344–345
Critical value
chi-square distribution, 78–79
hypothesis testing, 141, 146, 146f
normal distribution, 120–121
receiver operating characteristic curve, 234
t distribution, 123
Cronbach’s α, 353–354, 629
Cross-correlation, 518, 596–597
Crossover design, 10
Cross-sectional study, 590
Cumulative scale, 629
Curve fitting
goodness-of-fit test. See Chi-square
mathematical modeling. See Modeling
regression fits. See Regression
spline curve, 663
theoretical fits, 662–663
Curvilinear regression
analysis of variance and regression, 395–396
examples, 414–419
method, 414–419, 418f
more than three dimensions, 395
multiple dependent variables, case of, 395
overview, 392, 414
visualization of models in three dimensions, 394–395
visualization of models in two dimensions, 392–394
Data
imperfections and statistics, 632
method mismatch versus data, 631–632
missing. See Missing data
quantification, 31–32
question versus data mismatch, 631
ratings, 34
rounding, 34
string data, 35
survival
time data, 478
types
distinguishing, 33–34
overview, 32–35
Data exploration, graphing, 113
Data management
formatting, 37
missing data, 38
organization for inference
categorical data, 152–153
continuous data conversion to rank data, 154, 154f
continuous measurement data, 154–155
data type identification, 152–155
entry of data, 153, 153f
rank data, 153–154
planning, 37
quality of data, 38
raw data, 37
record keeping, 38
software
data management, 38–39
selection, 39
statistical analysis, 39
Data mining, 665
Data science, 666–667
Data-driven descriptive analysis, 40
Decision making. See Hypothesis testing
Decision theory. See Hypothesis testing
Degrees of freedom (df)
chi-square distribution, 78–79
contingency table, 183–184
F distribution, 128, 128f
t distribution, 77
Deming regression, 376–378
Dependence between categories in a 2 × 2 table
clinical relevance of, 231–232
Descriptive analyses, 40
Descriptive statistics See also specific terms
averages. See Mean; Median
bivariate descriptive statistics, 43–45
correlation coefficients. See Correlation coefficient
Design
experimental design concepts, 431–435
study. See Study design
Detecting patterns, 514–515
df. See Degrees of freedom (df)
Difference
hypothesis testing, difference versus equivalence
testing, 147–148, 295
inference, 73
symbols, 296–297
Discrete data, 32–33
Discriminant analysis, 654
Distribution
approximate distribution of a sample quantile,
86–87
averages, 65–66
binomial. See Binomial distribution
class characteristics, 65
chi-square. See Chi-square
critical values. See Critical value
degrees of freedom. See Degrees of freedom (df)
frequency distribution. See Frequency
distribution
Greek versus Roman letters, 65
joint distributions of two variables, 81–82
normal. See Normal distribution
Poisson. See Poisson distribution
probability distribution, 57–58, 741t, 742t, 744t, 745t, 746t, 747t, 750t, 754t, 755t
relative frequency distribution, 61–63, 118
sample distribution, 71
sample size methods. See Sample size
shape
common shapes, 69–71, 70f
standardizing distributions, 70–71
spread indicators, 68–69
standard normal. See Standard normal
statistical inference. See Inference
t. See t
tables of, 48t
tests
equality testing
Kolmogorov–Smirnov test, two-sample,
331–335, 334f, 335t
overview, 321–322
normality testing
chi-square goodness-of-fit test, 328–330, 328b
**Statistical Subject Index**

Distribution (*Continued*)
- Kolmogorov–Smirnov test, one-sample, 323–326, 327
- Lilliefors test of normality, 665
- overview, 321–322
- Shapiro–Wilk test, 322
- software, 323
- types of tests, 322
- types See specific types
- variable versus statistic, sampling distribution of, 71–72

Double censoring, 479

Dummy variables, 395–396

**E**

EBM. See Evidence–based medicine (EBM)
- Effect modifier, 36–37
- Effect size, 558, 590
- Empirical Bayes estimation, 511
- Endpoint, Cox proportional hazards regression, 488–489
- Epidemiology, 583
  - case–cohort study design, 594
  - case–control study design, 591–593
    - measure of association in, 591–592
    - number and selection of controls, 592–593
  - comparison with other branches of medicine, 583–584
  - definition, 583
  - historical perspective, 584
  - incidence, 586
  - mortality rate, 586–587
  - nested case–control study design, 593–594
  - odds ratio, 587–588
  - prevalence, 586
  - scientific knowledge stages, 583–584
  - study types
    - case–control study, 589
    - causation inference, 590–591
    - cohort study, 590–591
    - cross-sectional study, 590
    - experimental studies and intervention trials, 589
    - exposures and outcomes, 589
  - survival and causal factors, methods to analyze, 594–597
  - example, 595
  - graphing survival information, 596
- historical note, 598–599
- life tables list survival through time, 594–595
- serial correlation through time in epidemiology, 596–597
- autocorrelation, 597
- cross-correlation, 597
- transmission of disease
  - herd immunity, 585
  - host, 584–585
  - modes, 585
  - spectrum of disease, 584–585
- Equivalence
  - inference, 73
  - noninferiority comparison, 296
  - overview, 295
  - testing
    - assumptions, 297
    - confidence intervals, 297–299
    - difference symbols, 296–297
    - examples, 304–307
    - hypothesis testing differences, 147–148
    - method, 304b
    - null hypothesis, 296–297
    - one-sided versus two-sided tests, 297
    - sample size estimation, 558–559
    - TOST, 297
- Error probabilities, 258b, 302t, 533, 534f, 538f
- Estimation See specific estimates
- Ethics, medical studies
  - overview, 23–24
  - patient identifiers in data sheets, 24
  - patient protection requirements, 24
  - sample size consequences, 25–27
- Evidence
  - causality in epidemiology, 590–591
  - overview, 4
  - proof comparison, 4–5
- Evidence–based medicine (EBM), 5
- Evolving models. See Markov chain Monte Carlo (MCMC)
- Exchangeable covariance structure, 501
- Experimental design. See Study design

**F**

F distribution
- degrees of freedom, 128, 128f
- example, 129
- F test, 21, 400–402, 418
overview, 79–80, 127
sample size effects, 129–130
standard deviation, 127–128
table, 128–129, 129
Factor analysis, 655
False discovery rate (FDR), 636–637
False negative, 215–217, 218, 220, 223
False positive, 215–217, 218, 220, 223
False results, probability of multiple related tests that inflate, 21
FDR. See False discovery rate (FDR)
FET. See Fisher’s exact test (FET)
Fisher least significant differences, 259, 262
Fisher’s exact test (FET)
basis, 190
examples, 191–192, 191
NPTREND test comparison, 289–290
overview, 186, 190–192
row × column contingency tables, 192–193
Fixed effects, 511
Fixed-effects meta-analysis, 603–604
Fixed-sample methods, disadvantages of, 577–579
Fixed-sample testing procedure, 578–579
Forest plot, 607
Frailty models, survival analysis, 660–661
Framingham study, 439–440
Freeware, 39
Frequency distribution, continuous variables
histogram, 58–60
internal selection, 61–62
relative frequencies, 61
sample size effects, 61
tallying data, 58, 60
Friedman test
examples, 283, 284–287
method, 284
Functional form of predictors, assessing, 448–451
examples, 449–451
H
Hazard, Cox proportional hazards regression, 488–489
Health Insurance Portability and Accountability Act (HIPAA), 24
Heckman selection model, 652
Herd immunity, 585
Heterogeneity, tests for of estimated effects across studies, 606
Heteroscedasticity, accounting for, 412–414
empirical variance estimation, 413–414
transformations, 413
weighted least squares, 413
HIPAA. See Health Insurance Portability and Accountability Act (HIPAA)
Histogram, one variable descriptor, 101–102, 102f, 103f
Homoscedasticity, 143–144, 409
Hosmer–Lemeshow goodness-of-fit test, 453–457
examples, 455–457
procedure, 454b
influential observations, assessing, 451–453
examples, 451–453
Graphing. See Descriptive statistics specific graphs
Graphing survival information, 596
Group sequential testing, 577–582
fixed-sample methods, disadvantages of, 577–579
examples, 579–582
G
Gaussian distribution. See Normal distribution
General linear model
analysis of variance, 652
Gibbs sampler, 658
GLM. See General linear model
Goodness-of-fit, model diagnostics and, 448–457
Cook’s distance, 453
functional form of predictors, assessing, 448–451
examples, 449–451
Hosmer–Lemeshow goodness-of-fit test, 453–457
examples, 455–457
procedure, 454b
influential observations, assessing, 451–453
examples, 451–453
Hotelling T², 654
Hypothesis testing. See also Tests of hypothesis
alternate hypothesis, 139
applications, 142
assumptions, 142
clinical hypothesis, 137–144
confidence interval relationship, 142
critical value, 141
data organization for inference
categorical data, 152–153
continuous data conversion to rank data, 154, 154f
continuous measurement data, 154–155
data type identification, 152–155
entry of data, 153, 153f
### Hypothesis testing (Continued)

- **rank data**: 153–154
- **decision making**: 138
- **decision theory**: 137–138
- **description and prediction**: 156

#### Error

- **assumptions**
  - independence of errors, 143
  - normality of errors, 143
  - standard deviation equality, 143–144
- **probabilities**
  - choosing $\alpha$ but not $\beta$, 146–147
  - confidence interval, providing, 147
  - critical value, 146, 146f
  - difference versus equivalence testing, 147–148
  - null hypothesis acceptance versus rejection, 147
  - $p$-value, 144–145
  - relation among truth, decision and errors, 145–146
  - Type I error, 144, 145f
  - Type II error, 144, 145f
- **forms of hypothesis**, 139
- **multiplicity bias**, avoiding, 152
- **null hypothesis**, 138–139
- **numerical example**, 140–141, 141f, 141t
- **observed data to test pathway**, 138
- **operations research**, 138

#### Policies

- **approaches**, 149
- **computer calculation**, 149
- **historical perspective**, 148–149

**Sequential analysis.** See **Sequential analysis**

- **set up of test**, 157b
- **sidedness**, 26–27

### I

**ICC.** See **Intraclass correlation (ICC)**

- **Imputation, missing data**, 433, 660
- **Incidence rate**, 586
- **Independence, tests of**, 185–186, 190–192
- **Independent variable**, 7–8
- **Indicator symbol**, 31

**Inference.** See also **Hypothesis testing**

- **Bayesian statistics**, 612–613
- **causation inference in epidemiology**, 590–591
- **data organization for inference**
- **categorical data**, 152–153
- **continuous data conversion to rank data**, 154, 154t
- **continuous measurement data**, 154–155
- **data type identification**, 152–155
- **entry of data**, 153, 153t
- **rank data**, 153–154
- **difference**, 73
- **equivalence**, 73
- **robustness**, 74–75
- **statistic desirable properties**, 72
  - **steps**, 74
  - **via confidence interval**, 72–73

**Influential observations, assessing**, 451–453

**Informative descriptive tables and figures,**

- **principles of**, 114–115
- **graphing practices**, 114–115
- **table formatting practices**, 114

**Interacting variables**, 36–37

**Interim analysis**

- **sample size estimation**, 536
- **Interquartile range (IQR)**, 28
  - **one variable descriptor**, 94–95
- **Interval**, 161
- **Interval censoring**, 480–481

**Intervention, defining**, 570

**Intraclass correlation (ICC)**, 354–355, 380

**IQR.** See **Interquartile range (IQR)**

### J

- **Jackknife**, 640
- **Jargon, scientific**, 4

**Joint difference and equivalence testing**, 307–309

- **example**, 307–309
  - **history**, 308–309
  - **terminology**, 307

**Joint frequency distribution**, 82

**Journal articles**

- **meta-analysis.** See **Meta-analyses**
  - **reading**
    - **efficiency improvement**, 19
    - **goals**, 19
  - **shortcomings**
    - **confusing statistical versus clinical significance**, 20
    - **failure to distinguish a priori analyses from exploratory data-driven analyses**, 21–22
failures in statistical definitions, 21
generalizing from poorly behaved data, 20
inappropriate statistical tests, choosing, 22
multiple related tests that inflate the probability of false results, 21
violating assumptions underlying statistical methods, 20
writing
content selection, 23
format, 22
level of composition, 22–23

K
Kaplan–Meier curve, 481, 485f, 486f, 596
Kaplan–Meier estimator of survival curve, 481
Kappa, agreement among ratings
example, 347–348
method, 347
overview, 346–348
Kendall’s $t$, 342–344
Kendall’s $W$, agreement among multiple rankers, 350
Kolmogorov–Smirnov (KS) test, 408–409
one-sample test of normality
examples, 324–326, 327t
method, 323b
software, 322
two-sample test of distribution equality
examples, 331–335, 334f, 335t
large samples, 332–335
method, 331b
overview, 331–335
Kruskal–Wallis test, 281b
examples, 280–284, 280t, 283t
method, 280–284
KS test. See Kolmogorov–Smirnov (KS) test

L
L test. See Page’s $L$ test
Large sample–ranked outcomes, 276–279
large sample, 277
large-sample approximation, 276
single or paired data, 276–277
two large sample-independent ranked outcomes, 278–279
Law of Large Numbers, 545
Least squares, regression fit, 358f
Left censoring, 479
Levene’s test, 651
Leverage, 451
Life table
calculations for, for men with diabetes mellitus, 481
data and calculations required for, 480–481
estimates of survival, 482–484
Life tables list survival through time, 594–595
Likelihood ratio (LR)
caveats, 222
definition, 218t
overview, 222
truth tables, 224
Likert scale, 628
Lilliefors test of normality, 665
Line chart, one variable descriptor, 103–105, 104f
Linear regression, 489
assumptions, 361–363
Linear regression model, underlying assumptions of, 407–408
Location testing, 239–241
Log odds ratio test, 227–232
clinical relevance of dependence between categories in a $2 \times 2$ table, 231–232
contingency tables, 226t, 231–232
examples, 228–229
method, 228b
overview, 227–232
relative risk and odds ratio, choosing between, 230–231
Log OR (LOR) test, 225, 227–228
Log RR (LRR) test, 225
Logical Memory test, 504, 507, 510, 512
Logistic regression, 378–379, 378t
Logistic regression for binary outcomes, 437
association parameters, inference for, 445–447
odds ratio, confidence intervals for, 447, 447b
examples, 447
test of association, 445–447
examples, 446–447, 446t
goodness-of-fit, model diagnostics and, 448–457
Cook’s distance, 453, 454f
functional form of predictors, assessing, 448–451
examples, 449–451, 449t, 450f, 451f
Hosmer–Lemeshow goodness-of-fit test, 453–457
Logistic regression for binary outcomes (Continued)
examples, 455—457, 456t, 457t
procedure, 454b
influential observations, assessing,
451—453
examples, 451—453, 452f, 452t, 453t
interpretation, 443b
multiple logistic regression, 441—445
examples, 444—445
partial residual, 448b
simple logistic regression model, 438—440
examples, 438t, 439—440

Logit, 661

Log-rank test
survival curve analysis
comparison with other tests, 486—487
examples, 484—487, 485t
method, 485b

Longitudinal data, analysis of, 491, 493—514
admonition, 495—500
examples, 497
autocorrelation, 519—523
examples, 519, 521—522
correlation due to repeated measures, 493—495
example, 493
covariance structures, 500—501
common choices of, 501—503
cross-correlation, 518
detecting patterns, 514—515
general linear model, 503—507
examples, 503—504, 506—509, 512—514
interpretation, 495
moving averages
calculation of, 517—518
method for, 517
types of, 517
moving F method, 526—527
examples, 529—531
detecting patterns, 523
time-series methods, 515

Longitudinal data versus time-series data, 492
LR. See Likelihood ratio (LR)

Lung cancer, screening for
number needed to treat (NNT), 559

M
Machine learning, 666—667
Mahalanobis’ distance, 655
MANCOVA. See Multivariate analysis of
covariance (MANCOVA)
Mann—Whitney U. See Rank-sum test
MANOVA. See Multivariate analysis of variance
(MANOVA)
Mantel—Haenszel test, survival curve analysis, 486
Marginal effects, 511
Markov chain Monte Carlo (MCMC), 524—525,
657—659
as Bayesian statistics, 658
evolving models
examples of, 658—659
history, 658
ophthalmology, examples from, 659
processing, 658
Markov chains, 656—657
examples, 657
modeling, 657
shipboard flu epidemic, 657
Masking, clinical trial, 10
Mathematical model, 1—2
Matrix, 657
Maximum likelihood, 397
MCMC. See Markov chain Monte Carlo
(MCMC)
McNemar’s test, 205—209
Mean
approximate distribution of the mean, 83—86
confidence interval on
estimated standard deviation, 169—171, 170t
known standard deviation, 166—169
definition, 28, 30, 66
dependence of the variance on, 174—175
measure of variability in the sample mean, 84
one variable descriptor, 90
population standard error of, 85—86
sample standard error of the mean, 86
Mean-and standard-error chart, one variable
descriptor, 105—106
Means testing. See also Normal (z) test
alternative hypothesis
one- versus two-tailed test selection, 241
specification, 241
analysis of variance
assumptions, 257
“Memoryless” property of the Poisson distribution, 459–460
Meta-analyses, 601
   concept of, 601
   criteria for an acceptable meta-analysis, 602
   fixed-effects estimate, 603–604
   heterogeneity, tests for
      of estimated effects across studies, 606
      publication bias in, 602–603
      random-effects estimate, 603–605
      reporting the results of, 606–607
      steps to conduct, 601–602
Metropolis–Hastings algorithms, 658
Missing data
   bias, 15
   imputation, 433, 660
   management, 38
Mode
   definition, 66
   one variable descriptor, 91, 92f
Model diagnosis and goodness of fit, 407–412
   constant variance, 409
   influential observations, 409–411
   linear regression model, underlying assumptions
      of, 407–408
   normality, 408–409
Modeling
   Markov chain. See Markov chains; Markov
      chain Monte Carlo (MCMC)
   Monte Carlo simulation. See Markov chain Monte
      Carlo (MCMC)
Mortality rate, epidemiology, 586–587
Moving average
   calculation of, 517–518
   method for, 517
   types of, 517
Moving F method, 526–527
   examples, 529–531
   graphing, 528
Moving F statistic, 525
Multinomial regression, 365, 378t
Multiple comparisons test
   Bonferroni’s method, 262–263
   comparison of tests, 262, 262f
   examples, 260–265, 261t
   software, 260b
   types of tests, 259–260, 259t
null hypothesis specification, 241
    estimation of rate, 618
    example, 617–618
    t test
       single or paired means testing, 241–245
       two-sample means test, 245, 246t, 247–248,
          250, 253–256
    z test
       single or paired means testing, 241–245
       two-sample means test, 245, 246t, 247–252
Means versus rates, 460–461
   Poisson regression for count outcomes
      example, 461
Median
   confidence intervals, 175–176
   definition, 28, 66
   one variable descriptor, 91
Medical articles. See Journal articles
Multiple logistic regression, 441–445  
   examples, 444–445  
Multiple Poisson regression  
   model specification and interpretation, 469–472  
Multiple range test, 259–260  
Multiple regression  
   analysis of variance and regression, 395–396  
   caveat, 398  
   examples, 396–407  
   method, 396b  
   more than three dimensions, 395  
   multiple dependent variables, case of, 395  
   nominal variable management, 399–407  
   overview, 391–392  
   stepwise multiple regression, 398–399  
   visualization of models in three dimensions, 394–395  
   visualization of models in two dimensions, 392–394  
Multiple regression analyses, adjustment variables in, 36–37  
Multiple testing bias, 45–47  
Multiplicative rule, probability, 52–53  
Multiplicity bias. See also Multiple testing bias avoiding, 152  
Multivariable analysis, 35–36  
Multivariate analysis, 35–36  
   canonical correlation, 653–654  
   cluster analysis, 656  
   data display, 656  
   discriminant analysis, 654  
   factor analysis, 655  
   Hotelling T^2, 654  
   Mahalanobis’ distance, 655  
   overview, 35–36  
   principal components analysis, 655  
Multivariate analysis of covariance (MANCOVA), 655  
Multivariate analysis of variance (MANOVA), 655

Nominal data. See Categorical data
Noninferiority  
   equivalence comparison, 296  
   testing  
      critical values, 302t  
      examples, 300–304  
      method, 301b  
      steps, 302b  
Noninferiority/equivalence margin, choosing, 299–300  
Noninformative censoring, 478–479  
Nonparametric methods. See Rank-order data  
Nonparametric tests, 267–269  
Nonparametric tests, potential drawbacks to using, 291–293  
   choosing the test to use, 293  
   rank-sum test, nontransitivity of, 292–293  
   example, 292–293  
Non-superiority testing. See Noninferiority  
Nontransitive test, 292  
Nonuniform weighted sampling, 13  
Normal distribution  
   approximate distribution of a sample quantile, 86–87  
   critical value, 120–121  
   sample mean  
      continuous data and normal distribution, 83–84, 84f  
   sample size estimation  
      test on one mean, 541–542, 549  
      test on two means, 542–544  
   shorthand, 77  
   standard normal  
      overview, 75–77, 117–118, 118f  
      probability of certain ranges, 118–119  
      table, 118, 119t  
      use, 119  
   testing. See Distribution  
Normal (z) test  
   single or paired means testing, 241–245  
   two-sample means test, 245, 246t, 247–252  
Normality, tests of  
   chi-square goodness-of-fit test, 328–330, 328b  
   Kolmogorov–Smirnov test, one-sample, 324–326, 327t  
   Lilliefors test of normality, 665  
   overview, 321–322  
   Shapiro–Wilk test, 322
software, 323
types of tests, 322
Notation. See Symbols
NPTREND test
contingency test comparison, 289–290
examples, 286t, 287–289
method, 287b
NPV. See Negative predictive value (NPV)
Null hypothesis
acceptance versus rejection, 147
equivalence testing, 296–297
means testing specification, 241
overview, 138–139
Number needed to benefit (NNB)
assessing the benefit of a new treatment, 561–562
method for, 561b
Number needed to treat (NNT), 559–560, 560b
for additional detections, 560b
cost of, 560
detection compared to other methods of
efficacy, 561
method, 559b
Number needed to treat or benefit, 559–562
cost of number needed to treat, 560
detection compared to other methods of
efficacy, such as mortality, 561
number needed to benefit: assessing the benefit of a new treatment, 561–562
number needed to treat: screening for disease, 559–560

O
O’Brien–Fleming design, 580–582
Odds, probability comparison, 217
Odds ratio, confidence intervals for, 447, 447b
examples, 447
Odds ratio and relative risk, choosing between, 230–231
Odds ratio (OR). See also Log odds ratio test
certainty interval on, 232
definition, 218r
overview, 221
relative risk relationship, 221–222
truth tables, 224
One-factor ANOVA. See Analysis of variance (ANOVA)
One-sided hypothesis, 139

Operations research, 138
OR. See Odds ratio (OR)
Oral leukoplakia, treating
number needed to treat (NNT), 559–560
Ordered regression, 378–379, 378t
Ordinal data. See Rank–order data
Ordinary regression. See Regression
Outcome, defining, 570–572
Outlier, adjustments
Cook’s distance, 661
sample trimming, 662
Winsorization, 662

P
Page’s L test
certainty with other tests, 291
examples, 290–291, 290t
method, 290b
Parameter, 71
definition, 28, 63
Partial residual for logistic regression, 448b
Pattern detection. See Time-series
Pearson correlation coefficient, 342–344
Percentage, categorical data, 189–190
Percentile, definition, 28
Phi coefficient, 339, 344–345
Pie chart, one variable descriptor, 103, 104f
Placebo, in study design, 7
Planning. See Study design
Poisson distribution
appearance of, 134
graphical representation, 134f
large samples, 134–135
“memoryless” property of, 459–460
overview, 80–81, 132–133
table, 133, 133t
Poisson regression, 378–379, 378t
Poisson regression for count outcomes, 459
association parameters, inference for, 473–475
certainty intervals for rate ratio, 475
test of association, 473–474
comparing Poisson rates from two independent
samples, 463–465
eexample, 465
inference for the rate of a Poisson random
variable, 461–463
eexample, 463
basis, 201–202
two proportions, 197, 198b, 201
confidence interval on, 172–173
three or more proportions, 197–198
two proportions, 197, 198b, 201

\( \pi \) is near 1, 202

no events observed, 202

\( \lambda \), 203–204

large \( \lambda \), 203–204

large samples, 201–205

method, 203

randomization, 574–577

haphazard versus random assignment, 15

Randomized controlled trial (RCT), 10

Range, one variable descriptor, 91

Rank correlation, 381–382

Rank-order data
categories versus ranks, 185

continuous data conversion to rank data, 154, 154b

indications for use, 265, 267

methods, 80

organization for inference, 153–154

overview, 33

sample size estimation tests, 557

testing

\( \text{Friedman test} \)

examples, 283t, 284–287

method, 284b

\( \text{Kruskall–Wallis test} \)

examples, 280–284, 280t, 283t

method, 280–284

\( \text{log-rank test. See Log-rank test} \)

\( \text{NPTREND test} \)

contingency test comparison, 289–290

examples, 286t, 287–289

method, 287b

\( \text{Page's L test} \)

comparison with other tests, 291

examples, 290–291, 290t

method, 290b

\( \text{rank-sum test} \)

examples, 272–276

large samples, 279t

Mann–Whitney \( U \) calculation, 274

one-tailed test, 275

overview, 272–274

steps, 273b

synonyms, 275

\( \text{Runs Test}, 659 \)
Rank-order data (Continued)
signed-rank test
   examples, 269–272, 271t
   large-sample approximation, 276–279
   method, 270b
   normal approximation, 276b
   one tail versus two, 270–272
ties, 269
Rank-order variables, correlation coefficient,
   98–99, 98t
Rank-sum test
   examples, 272–276
   large samples, 279t
   for large samples, 279b
   Mann–Whitney U calculation, 274
   one-tailed test, 275
   overview, 272–274
   steps, 273b
   synonyms, 275
Rate
   incidence, 586
   one variable descriptor, 90
   prevalence, 586
   proportion. See Proportion
table test
   format, 47, 48t
Rate ratio, confidence intervals for, 475, 475b
Ratings, data, 34
RCT. See Randomized controlled trial (RCT)
Receiver operating characteristic (ROC) curve
   comparing two curves
      areas under the curve, 236b
      choosing better of two indicators, 235–237, 236f
      example, 237, 237f
critical value, 234
   overview, 232b
Record keeping, data management, 38
Registry, 8–9
Regression
   assessment tools
      coefficient of determination, 369, 371b
      confidence intervals, 372b
      examples, 370–376
      standard error of the mean, 369–370
   assumptions, 361–363, 363f
   bias in regression variables, 652
   confirmatory versus exploratory uses, 360–361
curvilinear regression
   analysis of variance and regression, 395–396
eqamples, 414–416
method, 414–419, 418f
more than three dimensions, 395
multiple dependent variables, case of, 395
overview, 392
visualization of models in three dimensions,
   394–395
visualization of models in two dimensions,
   392–394
definition, 359–360
Deming regression, 376–378
linear, 489
logit methods, 661
model, 378, 378t
multiple linear regression
   analysis of variance and regression, 395–396
caveat, 398
eamples, 396–407
method, 396b
more than three dimensions, 395
multiple dependent variables, case of, 395
nominal variable management, 399–407
overview, 391–392
stepwise multiple regression, 398–399
visualization of models in three dimensions,
   394–395
visualization of models in two dimensions,
   392–394
probit methods, 661
residuals and analysis, 652, 653f
simple regression
   concept of a line fitted to a set of points on a
tograph, 363–368
eamples, 363–365, 366f, 367f
method, 364b
survival times, adjusted comparison of
   Cox proportional hazards regression
      examples, 488, 490
      model, 488–490
      overview, 487–490
terminology, 488–490
tests of
      one slope, 363–368
      $R_2$, 369
types
      examples, 378–379
overview, 378—379, 378t
software, 379
Regression line, 357
Relative frequency
  frequency distributions for continuous variables, 61
  graphical relationship of relative frequency and
  probability, 64, 64f
  graphing, 56—57, 57f
  medicine, 56
  probability estimation, 55
Relative risk and odds ratio, choosing between, 230—231
Relative risk (RR)
  attributable risk relationship, 223—224
  definition, 218t
  odds ratio relationship, 221—222
  overview, 221
  truth tables, 224
Reliability
  as agreement, 338
  correlation as
    intraclass correlation, 354—355
    overview, 354
    Cronbach’s α, 353—354, 629
    example, 353
    internal consistency as, 353—354
Repeated measures ANOVA, 494—495, 499, 500t. See also Analysis of variance
(ANOVA)
Repeated measures data, 491
  longitudinal data, analysis of, 493—514
    admonition, 495—500
    examples, 497
    autocorrelation, 519—523
    examples, 519, 521—522
    calculation of moving average, 517—518
    common choices of covariance structures, 501—503
    correlation due to repeated measures, 493—495
    example, 493
    covariance structures, 500—501
    cross-correlation, 518
    detecting patterns, 514—515
    general linear model, 503—507
    examples, 503—504, 506—509, 512—514
    interpretation, 495
method for moving averages, 517
  moving F method, 526—527
    examples, 529—531
    graphing, 528
  need for smoothing processes, 515—517
    examples, 515
  serial correlation, 518
  testing patterns, 523
  time-series methods, 515
  types of moving average, 517
  longitudinal data versus time-series data, 492
Reporting bias, 15
Representativeness
  overview, 8
  random samples for increasing, 14
Residuals, analysis, 652, 653f
Retest effects, 503—504
Retrospective study designs, 591—593
  case—control study, measure of association in, 591—592
  number and selection of controls, 592—593
Ridit analysis, 653
Riffenburgh’s A, agreement among multiple
  rankers, 350—352
Right censoring, 478—479
Risk. See Attributable risk (AR); Relative risk (RR)
  Risk ratio, inference for, 225—227
  Robust test, 240
  Robustness, 74—75
  ROC curve. See Receiver operating characteristic
  (ROC) curve
Rounding, 34
Royston’s ptrend test, 211—213
RR. See Relative risk (RR)
Rule of Three, 174
Runs Test, 659
S
Safety factor, sample size estimation, 537
Sample
  bias, 13—15
  population comparison, 5, 8
  randomization. See Randomization
Sample average, 11
Sample quantile, approximate distribution of, 86—87
Sample size
Sample size (Continued)
- analysis of variance and sample size estimation, 558
- binomial distribution effects, 131–132, 131f
- bootstrapping. See Bootstrapping
- chi-square test of contingency effects, 188–189, 189r
- clinical relevance, 533, 536
- confidence intervals on means
  - β absent from confidence interval sample size estimation, 548–549
  - examples, 547–548
  - method, 547b
- convergence with increasing sample size, 66–67, 535
- correlation coefficient test, 555–557, 556f
- deviation from group mean effects, 11, 11r
- equivalence tests and sample size estimation, 558–559
- ethical consequences, 25–27
- F distribution effects, 129–130, 558
- importance, 533–536
- increase effects, 535
- interim analysis in relation to estimation, 536
- means tests. See Means testing
- minimum required size estimation, 533–535
- non-normal distribution sample size methods, 545–546
- normal distribution sample size methods
  - test on one mean, 541–542, 549
  - test on two means, 542–544
- patient care and clinical relevance, 533, 536
- power analysis
  - inputs, 539
  - logic behind method, 537–539, 538f
  - overview, 535
  - test sidedness and parameter selection, 539–540
  - value, 536–537
- probability estimation effects, 56
- proportions in estimation
  - confidence intervals on proportions, 561–562
  - test of one proportion, 549–551
  - test of two proportions, 551–553
- ranked data tests, 557
- safety factor in estimation, 537
- sequential analysis in relation to estimation, 536
- test with no objective prior data, 546–547
- very small sample value, 535
- Sample standard error of the mean, 86
- Sample trimming, 662
- Sampling bias. See Bias
- Sampling distribution of variable versus statistic, 71–72
- Sampling schemes
  - cluster sampling, 12
  - nonuniform weighted sampling, 13
  - purpose, 11
  - simple random sampling, 11
  - stratified sampling, 12
  - systematic sampling, 12
- Scatterplot, 110, 111f, 359f
- Scheffé test, 259r, 262t
- Schuirmann’s TOST. See Two one-sided test (TOST)
- Scientific goal, defining, 40–41
- Scientific knowledge
  - causative processes versus data, 3
  - clinical trial phases, 3–4
  - epidemiology stages, 583–584
  - stages, 1–3
- Scientific method, 3–4
- SD. See Standard deviation (SD)
- SEM. See Standard error of the mean (SEM)
- Sensitivity, 218r, 220, 224
- Sequential analysis, 663–665
  - advantages of, 664
  - disadvantages of, 664–665
  - example, 664
- Serial correlation, 518
  - through time in epidemiology, 596–597
    - autocorrelation, 597
    - cross-correlation, 597
- Shapiro–Wilk test, 322, 408–409
- Shrinkage, Bayesian statistics, 615
- Sidák test, 259r, 262t
- Sidedness of a test, 26–27
- Sign test, 659
- Signed-rank test
  - examples, 269–272, 271r
  - large-sample approximation, 276–279
    - method, 270b
    - normal approximation, 276b
    - one tail versus two, 270–272
- Significance
  - definition, 632
indicators, 632–633
multiple tests
approaches, 636
example, 638–639
false discovery rate, 636–637
rationale, 635–636
tests returning different significance levels, 633
purpose of estimating significance level, 632
Similarity testing. See Equivalence
Simple logistic regression model, 438–440
examples, 439–440
Simple Poisson regression model, 465–469
example, 467–469
Simulation, 632
Single Poisson regression parameter, tests for, 474b
Skewness, 108, 334
Smoothing processes, need for, 515–517
examples, 515
Software
analysis of variance, 260b
chi-square, 322
data management, 38–39
descriptive statistics help, 42
Kolmogorov–Smirnov test, 324
multiple comparisons test, 260b
regression, 379
selection factors, 39
Shapiro–Wilk test, 322
statistical analysis, 39
Spearman coefficient, 381
Specificity, 218t, 220, 224
Spline curve, fitting, 663
Standard deviation (SD)
confidence interval on, 176–179
definition, 28
one variable descriptor, 92–93
Standard error of the mean (SEM), 93
definition, 28, 84
one variable descriptor
two independent samples, 93–94
population standard error of the mean, 85–86
regression assessment, 369–370
sample standard error of the mean, 86
Standard normal
overview, 75–77, 117–118, 118f
probability of certain ranges, 118–119
table, 118, 119r
use, 119
Standardizing, distributions, 70–71
State-of-health bias, 15
Stationary models, 659
Statistic, 71
desirable properties, 72
observed data to test pathway, 138
Statistical and clinical significance, distinguishing between, 150–151
Statistical inference. See Inference
Statistical model, 1–2
Statistical results, reporting, 41–42
Statistics, 666
Straight-line model
example, 361f
line fitting, 358–359
Stratified blocked design, 577
Stratified randomization, 575–577
Stratified sampling, 12
String data, 35
Study design. See also Analysis of variance (ANOVA)
balanced design, 432–433
components, 7–8
control, 7
factorial designs, 434
factors influencing outcomes, 432
imputation, 433
placebo, 7
planning
professional guides, 17
steps, 16b
tips
Devil’s advocate perspective, 18
dummy data analysis, 18
working backward through logical process, 17
purpose, 431–435
representativeness and bias, 8
sparse-data designs, 434–435
study types See specific studies
variables, 7
Sum of squares. See Analysis of variance (ANOVA)
Survey
analysis, 622–623
bias, 623
data collection, 622
definition, 621
Survey (Continued)
  design, 621
  response rate, 623–624
  terminology, 624
Survival
  confidence intervals on survival estimates, 482–484
  Cox proportional hazards regression
    examples, 488, 490
    model, 488–490
    overview, 487–490
    terminology, 488–490
  data
    time data, 478
    definition, 477
  failure to survive, 477
  frailty models, 660–661
  Kaplan–Meier curve, 481, 485f, 486f
  life table generation, 480–484
  log-rank test
    comparison with other tests, 486–487
    examples, 484–487, 485t
    method, 485b
    time-dependent survival, 477–478
    time-series, 478
Survival and causal factors, methods to analyze, 594–597
  graphing survival information, 596
  historical note, 598–599
  life tables list survival through time, 594–595
  serial correlation through time in epidemiology, 596–597
  autocorrelation, 597
  cross-correlation, 597
Survival curve, Kaplan–Meier estimator of, 481
Symbols
  categorical data, 183
  categories, 1–3
  chi-square test of contingency, 187
  difference symbols, 296–297
  familiarity, 30
  formulas, 30
  Greek versus Roman letters, 65
  indicator symbols, 31
  purpose, 29
  ranks, 31
Systematic sampling, 12

T
  t
    critical value, 123
    degrees of freedom, 77
    examples, 122–124
    graphical representation, 121, 122f
    nature of t, 121
    standard t, 77, 78f
    table, 122–124, 123t
    tetrachoric correlation coefficient, 99–100, 184, 345
  T². See Hotelling T²
  t test
    single or paired means testing, 241–245
    two-factor ANOVA comparison, 425
    two-sample means test, 245–256, 246t
  Table test. See Tests of hypothesis
  Tally, 58, 60f
  Target population, defining, 569–570
  Tattoo ink
    number needed to benefit (NNB) from switching, 561–562
  Tchebycheff’s Inequality, sample size effects, 66–68
  Testing patterns, 523
Tests of hypothesis. See also Hypothesis testing
  alternate hypothesis. See Alternate hypothesis
  ANCOVA. See Analysis of covariance (ANCOVA)
  ANOVA. See Analysis of variance (ANOVA)
    assumptions, 142
    Bartlett’s test, 318b
    binomial test, 196, 201–202, 617
    Bonferroni adjustment, 259t, 262–263, 661
    Box–Cox transform, 651
    Brown–Forsythe test, 651
    chi-square tests. See Chi-square
      choosing a test, 48t
    Cochran’s Q test, 210–211
    contingency test. See Contingency
    correlation coefficient test of significance. See Significance
  definition of test, 28
  difference test, 295–297, 304, 312, 321–322
  distribution tests. See Distribution
  equality of distributions. See Equivalence
  equivalence tests. See Equivalence
F test, 21, 400–402, 418
Fisher’s exact test. See Fisher’s exact test (FET)
Friedman test. See Friedman test
General linear model. See General linear model
Hotelling’s $T^2$ test, 654
Kolmogorov–Smirnov tests.
See Kolmogorov–Smirnov (KS) test
Kruskal–Wallis test. See Kruskall–Wallis test
Levene’s test, 651
Lilliefors test, 665
log odds ratio test. See Log odds ratio test
Mann–Whitney test. See Rank-sum test
McNemar’s test, 205–209
means. See Means testing
Moving $F$. See Moving $F$ method
noninferiority tests. See Noninferiority
normal. See Normal ($z$) test
normality of distribution. See Distribution
NPTREND test. See NPTREND test
null hypothesis. See Null hypothesis
Page’s $L$ test. See Page’s $L$ test
Poisson test, 617
posttest power, 633–634
power, 144, 634
proportion testing. See Proportion
PTREND test. See PTREND test
rank-sum test. See Rank-sum test
rates or averages
table format, 47, 48t
regression tests. See Regression
ROC curves. See Receiver operating characteristic (ROC) curve
Runs Test, 659
sample size. See Sample size
selection, 47, 48t
sequential analysis. See Sequential analysis
Shapiro–Wilk test. See Shapiro–Wilk test
sidedness in test design, 26–27
sign test, 659
signed-rank test. See Signed–rank test
steps in set up, 45
survival curve. See Survival
$t$ test. See $t$ test
tetrachoric correlation coefficient, 555–557, 556t
variability, 31–32
variances tests. See Variability testing

Wald–Wolfowitz runs test. See Runs Test
Welch’s test, 651
Wilcoxon-Mann-Whitney rank-sum test.
See Rank-sum test
Wilcoxon’s signed-rank test. See Signed-rank test
$z$ (normal) test. See Normal ($z$) test
Three-factor ANOVA. See Analysis of variance (ANOVA)

Time-series
pattern testing
Markov chain Monte Carlo. See Markov chain Monte Carlo (MCMC)
questions addressed by, 515
survival data, 478
Time-series data versus longitudinal data, 492
Time-series methods, 515
Tolerance interval, 161
certainty with prediction and confidence intervals, 162
example of, using $t$, 122–123
for patient measurements, 163–164
TOST. See Two one-sided test (TOST)
Transmission, disease
herd immunity, 585
modes, 585
spectrum of disease, 584–585
Trend test. See Analysis of variance (ANOVA)
Trivariate descriptive statistics, 43–45
True negative, 215–217
True positive, 215–217
Truly sequential testing, 579
Truth table
elements, 223, 225t
format, 215–217, 216t
statistics, 217–220
Tukey’s honestly significant differences, 259t, 262t
Two one-sided test (TOST), equivalence testing, 297
Two–factor ANOVA. See Analysis of variance (ANOVA)
Two-sided hypothesis, 139
Type I error
definition, 28
hypothesis testing, 144, 145t
Type II error
definition, 28
hypothesis testing, 144, 145t
Statistical Subject Index

U

U. See Rank-sum test
Univariate descriptive statistics, 43
Unstructured covariance matrix, 502

V

Variability testing
  comparison of two variances, 311, 433
  non-significant result interpretation, 312
  one sample variance testing
    examples, 312–315, 313t
    method, 312b
  purpose, 311
  three or more sample variance testing
    Bartlett’s test of homogeneity of variance, 318b
    examples, 318–321, 318b, 319f
  two sample variance testing
    examples, 315–317
    method, 315b
Variable, 7
Variable versus statistic, sampling distribution of, 71–72

Variance
  confidence interval on, 176–179
  dependence on mean, 174–175
  one variable descriptor, 92–93
  tests. See Variability testing

W

W. See Kendall’s W
Wald-based confidence intervals, performance of, 174
Wald–Wolfowitz test. See Runs Test
Welch test, 651
Wilcoxon rank-sum test. See Rank-sum test
Wilcoxon–Mann–Whitney test. See Rank-sum test
Winsorization, 662
Writing. See Journal articles

Y

Yates’ correction, 188

Z

z test. See Normal (z) test
Medical Subject Index

Note: Page numbers followed by “f,” “t,” and “b” refer to figures, tables, and boxes, respectively.

A
Accidents, frequency in boys versus girls, 209
Acquired immunodeficiency syndrome (AIDS) clinical trials and group sequential testing, 580
risks with HIV exposure, 134–135
time to contract syphilis, Cox regression, 490
AIDS. See Acquired immunodeficiency syndrome (AIDS)
Allergen testing, Friedman test, 283t, 285–286
Aluminum, Poisson table analysis of Alzheimer’s disease rate, 133, 133t
Alzheimer’s disease (AD) and amyloid-beta levels
descriptive analyses, 40–41
censoring, 479
Poisson table analysis of rate with aluminum exposure, 133, 133t
retest effects in general linear model, 503–504
Logical Memory test score, 508–509, 513–514
Analgesia. See Pain
Anesthesia
Nitronox in arm fracture in children, 181
patient satisfaction with score-based confidence interval, 175
patient satisfaction with oral surgery, confidence interval on proportions, 173–174, 553–554
surgical difficulty correlation with duration, 557
type effects on recovery time, 638–639
Ankle fracture. See Fracture
Ankle ligament repair
surgeon experience and Fisher’s exact test of improvement effects, 192, 192t
Anorexia, military versus public, 205
Antibiotics See specific antibiotics
Appendicitis
diagnosis accuracy and error, 215
truth table for prediction, 225t
Artificial knee, test on two means for range of motion, 542–543
ASCVD. See Atherosclerotic cardiovascular disease (ASCVD)
Aspirin. See Pain
Asthma, t test for training effects on acute care visits, 242–243
Atherosclerotic cardiovascular disease (ASCVD) to ensure better accuracy, 2–3
Azithromycin. See Theophylline, azithromycin effects on levels in emphysema patients
B
Back pain. See Pain
Basal cell carcinoma, confidence interval on proportion recurrence rate, 174
Bayes’ Theorem, 230
Benign prostatic hypertrophy. See Prostate volume; Prostate-specific antigen
Birth defect, large λ in drug studies, 203
Bleeding during abdominal surgery randomization strategy, 577
Blood pressure (BP)
change with position, 195, 195t
laryngoscopy intubation
correlation coefficient, 384
correlation testing, 389
regression analysis, 367–368
P-value, obtaining, 150–151
sampling distribution, 71
standard normal distribution, 121
Body temperature. See also Fever; Thermometer
lung congestion effects in infants

correlation coefficient, 383
correlation testing, 388–389
curvilinear regression, 417–419, 418f, 418t
multiple regression, 400–401
regression analysis, 365–366, 374–376
prediction by infant age, 365–366, 366f
Bone density, 95, 109, 255
standard deviation, 127
BP. See Blood pressure (BP)
Breast cancer, estrogen blocking treatment for trivariate descriptive statistics, 43–45
Medical Subject Index

C
CAEPs. See Combat arms ear plugs (CAEPs)
Caffeine, fecundity effects, 588
Cancer See also specific cancers
  bootstrapping, confidence interval on median size of tumor, 640–641
  rating severity
    Cochran’s Q for consistency, 210–211, 210t
    Cohen’s kappa, 345
    contingency as agreement, 345
    nonparametric tests, 290t
    rater experience impact, 291
  truth tables and survival rates, 223
Cardiac Arrhythmia Suppression Trial (CAST), 571–572
Cardiac experiment, repeated measures ANOVA table for, 497t
Cardiac index (CI), bioimpedance versus thermodilution, 304–305
Carinal resection. See Tracheal carina resection
  Royston’s P trend test, 211, 213
Cervical cancer, epidemiology study, 587, 587t
Chewing tobacco. See Smokeless tobacco
Cholesterol
  eating time effects, 211, 211t
  prediction without fasting, multiple regression, 405–407, 406t
CI. See Cardiac index (CI)
Coccioidiomycosis, dust exposure effects, 217, 217t, 221, 232
Cognitive testing study, interpretation of coefficient related to time in Logical Memory test score, 506, 512
Cognitively normal and demented patients comparison of the rate of word repetitions among
  simple Poisson regression model, 467–469
Cold. See Common cold
Combat arms ear plugs (CAEPs), Bland–Altman plots, 643–645, 646t
Common cold
  herbal remedy effectiveness, 546–547
  recovery time and physician impact, 158–159
Completely randomized trial design, 574–575
Conjugated equine estrogen and medroxyprogesterone acetate (E + P), 567–568
Coronary heart disease (CHD)
clinical trials, 566–567
  goodness-of-fit, assessing, 455–457, 456t
  simple logistic regression model, 438t, 439–440, 444–445, 446t
  testing the association between association parameters, inference for, 446–447
Coumadin. See International Normalized Ratio (INR)

D
Demerol, two sample variability testing, 315
Dental wash water, purity of, 521–522
Diabetes, survival
  Kaplan–Meier curves, 480–484, 483t, 484t
  life table, data and calculations required for, 480–481
  log rank test, 484–486, 486f
  men with diabetes mellitus
    Kaplan–Meier survival curve, 484
  women with diabetes mellitus
    probability, estimate of, 484
Diagnostic populations, comparing slopes over time across
  Logical Memory test score, 507–508, 512–513
Diastolic blood pressure. See Blood pressure (BP)
Digital rectal exam (DRE)
  pie chart of accuracy, 103, 104f
  prostate cancer diagnosis
    categorical data contingency testing overview, 185, 186t
    chi-square test of contingency, 186–188, 189t
    confidence interval on odds ratio to predict biopsy results, 232
  Fisher’s exact test, 190–191
  sample testing against theoretical proportions, 199–200
  significance of odds ratio in predicting a biopsy result, 228
  significance of relative risk in predicting a biopsy result, 227
DRE. See Digital rectal exam (DRE)
Dyspepsia, emergency treatment
  Bayesian statistics, 616–617
  t test for drug evaluation, 245, 253, 253t
  test on one mean, 542
  test on two means, 543–544
variability testing, one sample, 314–315
z test for drug evaluation, 244–245, 253, 253t

E
Ear infection, Friedman test for gentamicin treatment, 285t, 286
Ear protection. See Combat arms ear plugs (CAEPs)
Ears
difference in distribution centers for the two ears rank-sum test, normal approximation to, 279
Edema. See Rhinoplasty; Snake bite
Electrostimulation, instrument performance, 282t, 283–284, 283r
Emphysema. See Theophylline
Epileptic patients, seizures among estimating the rate of
Poisson regression, 461
hypothesis test for the rate of
Poisson random variable, inference for the rate of, 463
testing the rate of
Poisson regression, 465
Erythropoietin, confidence interval on correlation with hematocrit, 181
Estrogen blocking treatment for breast cancer trivariate descriptive statistics, 43–45
Estrogen plus progesterone (E + P), 567–568

F
Fasciotomy. See Snake bite
Fever, test of two proportions for sex differences in reporting, 553
Flu
Markov chain modeling, 657
Fracture
ankle repair
age and plantar flexion correlation confidence interval on, 179–180
test on correlation coefficient, 555–556
hardware impact and rank-order data testing, 272
two sample Kolmogorov–Smirnov test of equality, 335
blood flow after orthopedic casting, two-factor ANOVA, 433
certainty interval on mean treatment time for arm, 168–169
Nitronox in arm fracture in children, 181

G
GAG levels. See Glycosaminoglycan (GAG) levels
Gentamicin, Friedman test for ear infection treatment, 285t, 286–287
GFR. See Glomerular filtration rate (GFR)
Glaucoma
binomial table of laser trabeculoplasty efficacy, 130–131, 131t
intraocular pressure, testing of drug effects, 545–546
Glomerular filtration rate (GFR), Bayesian statistics of baseline levels, 613–616
Glycosaminoglycan (GAG) levels
F distribution, 130
standard error of the mean calculation, 86
variance and standard deviation calculation, 69
Group A Streptococcus
certainty interval for median in outbreak, 176
length of hospital stay multiple regression, 401–405, 402t
receiver operating characteristic curve in identification with white blood cell count, 235, 235f

H
Handwashing between patients sequential analysis, 664
Heart rate
certainty interval on mean, 171
variability, 179
by disease and by time of day
time-of-day variability, 497
two-factor analysis of variance (ANOVA), 493, 493t
eating effects
one-way analysis of variance, 420–421
t test, 420
time of day effects, 420t, 421
two-factor ANOVA, 421–423, 424t
signed rank test of drug effects, 269–270, 271t
test on one mean following trauma, 542
by type of anesthetic in trauma, 529–531
Hematocrit, 165
certainty interval on correlation with erythropoietin, 181
distributional parameters, uncertainty in, 163
Hematocrit, pyloromyotomy, open versus laparoscopic surgery, 275–276
Medical Subject Index

Hematocrit (Continued)
   Kolmogorov–Smirnov test of distribution  
      normality, 326  
   tolerance interval, 163
Hepatitis C, blood bank screening, 205
HIV. See Human immunodeficiency virus (HIV)
   Human immunodeficiency virus (HIV). See also  
      Acquired immunodeficiency syndrome  
      (AIDS)
   acquired immunodeficiency syndrome risks with  
      exposure, 134–135
   analysis of variance of CD4 cell counts with  
      years of disease, 266–267
   confidence interval on median viral load,  
      175–176
   protease inhibitor reduction of pulmonary  
      complications, 189
   vaccine  
      normal distribution testing, 326, 327t, 330,  
      330r
   two sample Kolmogorov–Smirnov test of  
      equality, 334–335, 334f, 335r
   two sample means testing, 255t, 256
Idiopathic thrombocytopenic purpura (ITP),  
   immunoglobulin treatment effects, 82
Infant malaria, life table for, 595
INR. See International Normalized Ratio (INR)
   Institutional review board (IRB), protocol  
      evaluation, 245, 245r
   Intelligence quotient (IQ), mental hospital stay  
      duration  
      correlation coefficient, 384
   correlation testing, 389
   multiple regression, 399–400
   regression analysis, 367–368, 367f, 377–378
   International Normalized Ratio (INR)  
      clinic versus laboratory readings  
         Bland–Altman plots, 646–647, 647f
         Cohen’s kappa, 348
         contingency as association, 341–342, 341t
         correlation as agreement, 346
         out-of-range readings, 206–207
   confidence interval, 164, 169
   data distributions, 62
   degrees of freedom in readings, 81
   test of two sample proportions, 201
   Intraocular pressure (IOP). See also Glaucoma  
      testing of drug effects, 545–546
   Intubation  
      blood pressure effects. See Blood pressure (BP)  
      rate, effect of prior surgery, 306–307
   IOP. See Intraocular pressure (IOP)
   IQ. See Intelligence quotient (IQ)
   IRB. See Institutional review board (IRB)
   ITP. See Idiopathic thrombocytopenic purpura  
      (ITP)
K  
   Keratotomy. See Radial keratotomy
   Kidney, cooling before surgery  
      multi-factor ANOVA, 430–431
      two-factor ANOVA, 425–426, 425t
   Knee. See Artificial knee
L  
   Laser trabeculoplasty, binomial table efficacy in  
      glaucoma, 130–131, 131f
   Lead, confidence interval on proportion levels in  
      children, 172
   Leg performance, postoperative  
      correlation as association, 343–344, 343t
      correlation coefficient, 98–99, 98t
      data organization for inference, 155
      multiple regression, 406
      t distribution, 124
   Ligament repair  
      log RR (LRR) test, 227
   Lung cancer  
      alcohol consumption and  
         bivariate descriptive statistics, 43
      McNemar’s test of smoking association,  
         206–208
   Lung congestion  
      body temperature effects in infants  
         correlation coefficient, 383–384
         correlation testing, 388–389
         curvilinear regression, 400–401, 417–419,
         418f, 418t
         multiple regression, 400–401
         regression analysis, 366, 374–376
M  
   Malaria  
      infant malaria, life table for, 595
      rat survival after treatment  
         curvilinear regression, 414–416, 415f
      residuals, 652, 653f
Malarial rats, survival of  
curvilinear regression, 414–417

Medical technical support, level of, 525–526, 529
Mesh, absorbable, 200

Migraine
nausea treatment and Page’s L test, 290–291
repeated measures ANOVA, 499, 500

Mortality and extent of carinal resection
Royston’s Prend test, 211–213

N
Nasal congestion
rank-sum test, nontransitivity of, 292–293

Nausea
contingency testing of ondansetron efficacy, 189
statistical testing setup for drug impact, 159
migraine-associated nausea treatment and Page’s L test, 290–291
postoperative
Bayesian statistics of rate, 617
contingency table, 99, 100

Neonate, confidence interval on variability in time to full feeding, 179

Neuropsychological testing
CI for the rate ratio comparing mild cognitive impairment to cognitively normal patients in
Poisson regression, 475
predicting the rate of repetitions in
Poisson regression, 472–473
testing the association between cognitive status and the rate of repeats in
Poisson regression, 474–475

Nitric oxide (NO), exercise effects on breath levels
charting, 108–110
correlation coefficient, 98
data organization for inference, 155
hypothesis testing, 149–150
inference, 75
interquartile range, 94–95
moving F method, 526–527
non-inferiority testing, 303–304
regression analysis, 367–368

NO. See Nitric oxide (NO)
Norwalk virus, epidemiology study, 588

O
Odansetron. See Nausea

Ophthalmology, examples from
Markov chain Monte Carlo, 659

P
Pain
drug dose as an independent variable, 35
t test for drug comparison with unequal variances, 254–256, 254

Pneumococcal vaccine
cost-effectiveness of, 648

Prick testing, Friedman test, 283, 284–285

Prostate biopsy patients, bias in early sampling of large sample-ranked outcomes, 276–277

Prostate cancer. See also Digital rectal exam (DRE);
Prostate volume; Prostate-specific antigen
ANOVA of age relationship, 256, 261, 261
positive biopsy, test of one rate, 550, 558–559
prediction of nodal involvement in prostate cancer patients
model diagnostics and goodness-of-fit, 449–453, 449, 452

Prostate volume
age effects
correlation coefficient, 96–97
covariance, 95–97
line chart of problems with age, 103–105, 104
mean-and-standard-error chart, 105–106, 106
scatterplot, 110, 111
three-dimensional bar chart, 112, 112
chi-square distribution, 125–126
confidence intervals on mean, 169–171
digital rectal exam accuracy, 103, 104
F distribution, 129
frequency distribution, 58, 60, 70, 102, 103

O
Prostate volume (Continued)
hormonal therapy for benign prostatic hyperplasia, 541
hypothesis testing, 140–141
Kolmogorov–Smirnov test of distribution normality, 323–326
proportion positive biopsies, 173
standard deviation, 176–179
tolerance interval, 122–123
two sample Kolmogorov–Smirnov test of equality, 331–332
two sample means testing of age and volume risks of cancer
t test, 249–252
z test, 247–252, 249f
variance and standard deviation calculation, 69
Prostate-specific antigen (PSA), 515, 516, 519, 520f
age dependence, box-and-whisker chart, 108, 109f
bias in early sampling of biopsy patients, 276–279
early versus later testing
Bayesian statistics, 617
normal test of averages, 242–244
Kruskal–Wallis test, 276–279, 279f, 280f, 281–283
large sample–ranked outcomes, 277t, 278t
mean between cancerous and healthy patients, 144
median, 91
mode, 91, 92f
normal distribution, 83, 84f
normality testing, 327
for patients with versus without BPH
large sample–ranked outcomes, 278–279
rank–order data example, 267t
rank–sum test for patients with versus without benign prostatic hyperplasia, 278–279
signed rank test, 270b, 271
rank–sum test for positive versus negative biopsy, 273t, 274
rare event testing, 202–204
receiver operating characteristic curve of PSA and PSA density, 234–237, 236f
versus remainder
large sample–ranked outcomes, 277
sample size effect on mean, 11t
standard error of the mean, 93
three-dimensional bar chart of frequency, PSA, and PSA density, 112–113, 113f
two sample Kolmogorov–Smirnov test of equality, 332–334, 333t
variability testing
one sample, 312–313
three samples and cancer risks, 317–321, 319f
two samples, 315
variance calculation, 92–93
PSA. See Prostate-specific antigen
Psychotherapy, therapist selection, 194–195, 195t, 340
PTSD. See Nitric oxide (NO)
Pyloromyotomy
confidence interval on proportions, 555
confidence interval on time to full feeding, 171
hematocrit for open versus laparoscopic surgery, 275–276
R
Radial keratotomy
multi-factor ANOVA of bandage lens comfort, 429, 430t
surgeon experience and outcomes, 229–230
test of two proportions, 553
Radiation therapy
Cox regression of survival in laryngeal cancer, 490
sample testing against theoretical proportions in survival, 199–200
Respiration rate, age dependence, 359
Rhinoplasty
data on edema following, 263t
steroid reduction of edema, 263–264, 320–321
S
Sarcoma, injection site and incidence, 204
Schistosomiasis, test of one rate, 559–562
Scurvy, cause of epidemiology study, 599
SEPS. See Subfascial endoscopic perforating vein surgery (SEPS)
Shipborne flu
Markov chain modeling, 657
Silicone, plasma levels after implantation
confidence interval on the mean, 171
confidence interval on variability, 178–179
hypothesis testing, 148
non-inferiority testing, 303
Poisson distribution, 81
Smokeless tobacco, ethnicity effects on use, 192, 193t, 339–340
Smoking, McNemar’s test of lung cancer association, 206–208, 208t
Snake bite
   NPTrend test of location by child age, 286t, 287–288
   results analysis in children, 639
Snakebite edema, fasciotomy and antivenin to relieve
   fasciotomy and antivenin means, 497
Snoring, loudness dependence on body mass index, 289
Stomach cancer, McNemar’s test of genetic predisposition, 208–209, 209t
Streptococcus. See Group A Streptococcus
Streptococcus pneumoniae, vaccination for, 648
Stretch marks, postpartum, confidence interval on the mean, 170
Subfascial endoscopic perforating vein surgery (SEPS), ulcer, 132
Surgeon experience
   Fisher’s exact test of improvement effects, 191–192, 191t
   radial keratotomy outcomes, 229–230
Syphilis, Cox regression of time to contraction in AIDS, 490
Systolic blood pressure (SBP)
   P-value, obtaining, 150–151

T
Tattoo
   titanium ink and removal, 189
Temperature. See Body temperature; Fever; Thermometer
Testosterone levels among hypogonadism patients clinical trials, 582
Theophylline, azithromycin effects on levels in emphysema patients
   charting of data, 110
   confidence interval on the mean, 171
   correlation coefficient, 382–383, 383t
   correlation testing, 385–389
   equivalence testing, 307
   hypothesis testing, 148
   inference, 75
   regression analysis, 363–365
   evaluation of prediction, 370
   sex and age effects using ANCOVA, 427–430
   t test of sex differences, 255
Thermometer
   confidence interval on precision, 177–178
   rank-order data testing, 272
   reliability, 353
   tympanic thermometer performance
      rank-sum test, 272–276, 275t
      t test, 275–276
      variability testing, 315
Thrombosis, categorical data of recurrence, 183, 184t
Tobacco. See Smokeless tobacco; Smoking
Tracheal carina resection
   confidence interval on mean, 547–548
   Poisson distribution, 81
   standard normal distribution of patient age, 120–121
   t test of survival, 275–276
   tolerance interval of extent of resection, 164, 166–169
Tretinoin, confidence interval on proportions for sample size estimation, 554–555
Tumor Board, agreement among ratings, 352

U
Ulcer, subfascial endoscopic perforating vein surgery, 132

V
Vaccination. See Human immunodeficiency virus (HIV); Pneumococcal vaccine

W
Warfarin. See International Normalized Ratio (INR)
Weight, estimation in emergency department, 346
White blood cell count (WBC)
   bias example, 13, 14f
   confidence interval on standard deviation, 178–179
   infectious agent relationship modeling, 358f
   receiver operating characteristic curve in Group A Streptococcus identification with white blood cell count, 235, 235f
   statistical testing setup, 156–158