

# PRACTICING STATISTICS

Guided  
Investigations  
for the  
Second  
Course



KUIPER • SKLAR

# Practicing Statistics

Guided Investigations for the Second Course

Shonda Kuiper

Grinnell College

Jeffrey Sklar

California Polytechnic State University  
San Luis Obispo

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam  
Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City  
Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor in Chief: Deirdre Lynch  
Acquisitions Editor: Marianne Stepanian  
Sponsoring Editor: Christina Lepre  
Developmental Editor: David Chelton  
Editorial Assistant: Sonia Ashraf  
Senior Managing Editor: Karen Wernholm  
Associate Managing Editor: Tamela Ambush  
Senior Production Project Manager: Sheila Spinney  
Digital Assets Manager: Marianne Groth  
Supplements Production Coordinator: Kerri Consalvo  
Media Producer: Audra Walsh  
Marketing Manager: Erin Lane  
Marketing Coordinator: Kathleen DeChavez  
Senior Author Support/Technology Specialist: Joe Vetere  
Rights and Permissions Advisor: Michael Joyce  
Image Manager: Rachel Youdelman  
Procurement Manager: Evelyn Beaton  
Procurement Specialist: Linda Cox  
Media Procurement Specialist: Ginny Michaud  
Associate Director of Design, USHE North and West: Andrea Nix  
Text Designer: Tamara Newman  
Cover Designer and Art Director: Barbara T. Atkinson  
Cover Image: Stepping stones in water © iStockphoto  
Production Coordination, Composition, and Illustrations: Laserwords Private Limited

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Pearson was aware of a trademark claim, the designations have been printed in initial caps or all caps.

**Library of Congress Cataloging-in-Publication Data**

Kuiper, Shonda, 1969–

Practicing statistics : guided investigations for the second course /

Shonda Kuiper, Jeffrey Sklar.—1st ed.

p. cm.

ISBN 0-321-58601-8

1. Statistics—Textbooks. I. Sklar, Jeffrey, 1972– II. Title.

QA276.12.K85 2013

519.5--dc23

2011028178

Copyright © 2013 Pearson Education, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. For information on obtaining permission for use of material in this work, please submit a written request to Pearson Education, Inc., Rights and Contracts Department, 501 Boylston Street, Suite 900, Boston, MA 02116, fax your request to 617-671-3447, or e-mail at <http://www.pearsoned.com/legal/permissions.htm>.

1 2 3 4 5 6 7 8 9 10—EB—14 13 12 11

**PEARSON**

[www.pearsonhighered.com](http://www.pearsonhighered.com)

ISBN 10: 0-321-58601-8

ISBN 13: 978-0-321-58601-8

To Mark, for always being there for me and having the valor  
to stand up for what is right, even when it is difficult. To Joshua  
and Caleb, my favorite boys in the entire world; you make me appreciate  
the adventure that exists within each new day. And to all my family and  
friends who have supported me throughout this project.

—Shonda Kuiper

To my wife Maria for her support and encouragement,  
and to my children Yazmin and Gabriel whose curiosity  
and fascination about their surroundings constantly remind  
me about the joys of teaching.

—Jeffrey Sklar

# About the Authors

**Shonda Kuiper** is an Associate Professor in the Mathematics and Statistics Department at Grinnell College. She teaches a variety of statistics courses, with an emphasis on the application of statistics in multiple disciplines. Shonda also serves as a statistical consultant for student and faculty research. She received her Bachelor of Arts and Sciences degree in mathematics from Wartburg College, and has an MS and a PhD in statistics from Iowa State University. She is also the recipient of two National Science Foundation grants to develop materials for undergraduate statistics courses. In addition to previously holding teaching positions, Shonda served as a senior engineer and later as a consulting statistician for Hallmark Cards, Inc.



**Jeffrey Sklar** is an Associate Professor in the Statistics Department at California Polytechnic State University, San Luis Obispo, and teaches various introductory statistics courses, as well as linear regression, multivariate statistics, and survival analysis courses. Before joining the faculty at Cal Poly in September 2005, he taught classes in the Department of Statistics and Applied Probability and the Gevirtz Graduate School of Education at the University of California, Santa Barbara. Jeffrey received his Bachelor of Arts and Sciences degree in mathematics and philosophy from the University of California, Davis, and his MA and PhD in statistics from University of California, Santa Barbara.



(Photos courtesy of the authors)

# Contents

Preface ix

## 1 Nonparametric Methods: Schistosomiasis 1

- 1.1 Investigation: Can a New Drug Reduce the Spread of Schistosomiasis? 2
  - 1.2 Statistical Inference Through a Randomization Test 4
  - 1.3 Performing a Randomization Test Using a Computer Simulation 5
  - 1.4 Two-Sided Tests 7
  - 1.5 What Can We Conclude from the Schistosomiasis Study? 8
  - 1.6 Permutation Tests versus Randomization Tests 9
  - 1.7 Permutation and Randomization Tests for Matched Pairs Designs 10
  - 1.8 The Bootstrap Distribution 12
  - 1.9 Using Bootstrap Methods to Create Confidence Intervals 14
  - 1.10 Relationship Between the Randomization Test and the Two-Sample t-Test 16
  - 1.11 Wilcoxon Rank Sum Tests for Two Independent Samples 17
  - 1.12 Kruskal-Wallis Test for Two or More Independent Samples 18
  - 1.13 Multiple Comparisons 20
- Research Project:** Gender Discrimination Among University Faculty 28

## 2 Making Connections: The Two-Sample t-Test, Regression, and ANOVA 30

- 2.1 Investigation: Do Distracting Colors Influence the Time to Complete a Game? 31
  - 2.2 The Two-Sample t-Test to Compare Population Means 32
  - 2.3 The Regression Model to Compare Population Means 36
  - 2.4 ANOVA to Compare Population Means 39
  - 2.5 Comparing Planned Variability to Random Variability 41
  - 2.6 Random Sampling and Random Allocation 41
  - 2.7 What Can We Conclude from the Game Study? 43
  - 2.8 Normal Probability Plots to Assess Normality 44
  - 2.9 Transformations 46
  - 2.10 Calculating Test Statistics 50
  - 2.11 Confidence Intervals 54
- Research Project:** Building a Better Paper Helicopter 63

## 3 Multiple Regression: How Much Is Your Car Worth? 67

- 3.1 Investigation: How Can We Build a Model to Estimate Used Car Prices? 68
- 3.2 Goals of Multiple Regression 69
- 3.3 Variable Selection Techniques to Describe or Predict a Response 70
- 3.4 Checking Model Assumptions 73
- 3.5 Interpreting Model Coefficients 80
- 3.6 Categorical Explanatory Variables 81
- 3.7 What Can We Conclude from the 2005 GM Car Study? 83

<b>3.8</b>	<i>F</i> -Tests for Multiple Regression	<b>83</b>
<b>3.9</b>	Developing a Model to Confirm a Theory	<b>87</b>
<b>3.10</b>	Interaction and Terms for Curvature	<b>88</b>
<b>3.11</b>	A Closer Look at Variable Selection Criteria	<b>92</b>
<b>Research Project:</b> Economic Growth in Third World Countries		<b>100</b>

## **4** The Design and Analysis of Factorial Experiments: Microwave Popcorn **102**

<b>4.1</b>	Investigation: Which Microwave Popcorn Is the Best?	<b>103</b>
<b>4.2</b>	Elements of a Well-Designed Experiment	<b>103</b>
<b>4.3</b>	Analyzing a Two-Way Factorial Design	<b>107</b>
<b>4.4</b>	Analyzing a Three-Way Factorial Design	<b>113</b>
<b>4.5</b>	What Can We Conclude from the Popcorn Study?	<b>115</b>
<b>4.6</b>	Paper Towels: Developing a Statistical Model for a Two-Way Factorial Design	<b>115</b>
<b>4.7</b>	Paper Towels: The Relationship Between Effects and ANOVA	<b>120</b>
<b>4.8</b>	Contrasts and Multiple Comparisons	<b>124</b>
<b>Research Project:</b> Testing for the Effect of Distracters		<b>134</b>

## **5** Block, Split-Plot, and Repeated Measures Designs: What Influences Memory? **137**

<b>5.1</b>	Investigation: What Influences Memory?	<b>138</b>
<b>5.2</b>	Elements of a Well-Designed Experiment	<b>139</b>
<b>5.3</b>	Statistical Analysis Based on the Experimental Design	<b>141</b>
<b>5.4</b>	Three Commonly Used Design Structures	<b>142</b>
<b>5.5</b>	Crossed and Nested Factors	<b>145</b>
<b>5.6</b>	Fixed and Random Factors	<b>147</b>
<b>5.7</b>	Model Assumptions	<b>149</b>
<b>5.8</b>	What Can We Conclude from the Memory Study?	<b>151</b>
<b>5.9</b>	Calculating Crossed and Nested Effects	<b>151</b>
<b>5.10</b>	Mathematical Calculations for ANOVA	<b>155</b>
<b>5.11</b>	Hasse Diagrams	<b>158</b>
<b>5.12</b>	Wash Your Hands: Analysis of Covariance (ANCOVA)	<b>161</b>
<b>Research Project:</b> What Impacts Memory?		<b>172</b>

## **6** Categorical Data Analysis: Is a Tumor Malignant or Benign? **176**

<b>6.1</b>	Investigation: Is Cell Shape Associated with Malignancy?	<b>177</b>
<b>6.2</b>	Summarizing Categorical Data	<b>178</b>
<b>6.3</b>	A Simulation Study: How Likely Is It That the Observed Sample Would Occur by Chance?	<b>179</b>
<b>6.4</b>	Fisher's Exact Test	<b>181</b>
<b>6.5</b>	Two-Sided Hypothesis Tests	<b>182</b>
<b>6.6</b>	The Chi-Square Test	<b>183</b>
<b>6.7</b>	What Can We Conclude from the Cancer Study?	<b>187</b>
<b>6.8</b>	Relative Risk and the Odds Ratio	<b>187</b>

- 6.9 Sampling Designs 189**
- 6.10 Comparing Tests of Homogeneity and Independence 192**
- 6.11 Chi-Square Goodness-of-Fit Tests 193**
- Research Project: Infant Handling in Female Baboons 204**

## **7 Logistic Regression: The Space Shuttle Challenger 211**

- 7.1 Investigation: Did Temperature Influence the Likelihood of an O-Ring Failure? 212**
- 7.2 Review of the Least Squares Regression Model 215**
- 7.3 The Logistic Regression Model 216**
- 7.4 The Logistic Regression Model Using Maximum Likelihood Estimates 218**
- 7.5 Interpreting the Logistic Regression Model 219**
- 7.6 Inference for the Logistic Regression Model 221**
- 7.7 What Can We Conclude from the Space Shuttle Study? 224**
- 7.8 Logistic Regression with Multiple Explanatory Variables 224**
- 7.9 The Drop-in-Deviance Test 226**
- 7.10 Measures of Association 228**
- 7.11 Review of Means and Variances of Binary and Binomial Data 229**
- 7.12 Calculating Logistic Regression Models for Binomial Counts 230**
- 7.13 Calculating Residuals for Logistic Models with Binomial Counts 231**
- 7.14 Assessing the Fit of a Logistic Regression Model with Binomial Counts 232**
- 7.15 Diagnostic Plots 236**
- 7.16 Maximum Likelihood Estimation in Logistic Regression 238**
- Research Project: Substance Abuse Among Youth 247**

## **8 Poisson Log-Linear Regression: Detecting Cancer Clusters 251**

- 8.1 Investigation: Are Cancer Rates Higher for People Living near a Toxic Waste Area? 252**
- 8.2 Comparing Count Data for Groups 252**
- 8.3 Building Models for Count Data 254**
- 8.4 The Binomial Model for Count Data 255**
- 8.5 The Poisson Model for Count Data 256**
- 8.6 Adding a Covariate to the Poisson Count Model 260**
- 8.7 Interpreting Poisson Regression Model Parameters 263**
- 8.8 Poisson Regression Models with More Than One Covariate 264**
- 8.9 Inference for Poisson Regression Models 266**
- 8.10 Assessing the Fit of the Poisson Regression Model 268**
- 8.11 What Can We Conclude from the Cancer Rate Study? 270**
- 8.12 Estimation Methods for Generalized Linear Models 270**
- 8.13 Do No-Smoking-at-Work Policies Keep Smoking at Home? 271**
- 8.14 Is the Number of Species on Archipelago Islands Related to Island Area, Elevation, and Neighboring Islands? 273**
- Research Project: Hitting a Grand Slam in Baseball 281**

## 9 Survival Analysis: Melting Chocolate Chips 284

- 9.1 Investigation: How Long Does It Take for Chocolate Chips to Melt? 285
  - 9.2 Overview of Survival Analysis Studies and Data 286
  - 9.3 The Survival Function 288
  - 9.4 Descriptive Statistics for Survival Data 294
  - 9.5 Confidence Intervals for Survival Probabilities 297
  - 9.6 Comparing Survival Functions 300
  - 9.7 What Can We Conclude About Melting Chocolate Chips? 305
  - 9.8 The Hazard Function 305
  - 9.9 The Cumulative Hazard Function 311
  - 9.10 Additional Types of Incomplete Data 317
- Research Project:** Shapesplosion: A Study of Reaction Time 329

## 10 Principal Component Analysis: Stock Market Values 332

- 10.1 Investigation: Can a Single Variable Explain Patterns in the Stock Market? 333
  - 10.2 A Visual Interpretation of PCA 333
  - 10.3 Calculating Principal Components for Two Variables 336
  - 10.4 Understanding Eigenvalues 342
  - 10.5 A Three-Dimensional Example 342
  - 10.6 What Can We Conclude from the Stock Market Investigation? 344
  - 10.7 The Impact of Standardizing Each Variable 344
  - 10.8 Determining the Number of Components to Retain 346
  - 10.9 Interpreting Principal Components 347
  - 10.10 Comparing Regression and Principal Components 350
  - 10.11 Incorporating Principal Components into Other Statistical Methods 351
  - 10.12 Calculating Eigenvectors and Eigenvalues Using Matrix Algebra 351
- Research Project:** The Global Warming Hockey Stick Controversy 359

## 11 Bayesian Data Analysis: What Colors Come in Your M&M's® Candy Bag? 369

- 11.1 Investigation: Do Prior Beliefs Improve Your Estimate of the Proportion of Brown or Orange M&M's? 370
  - 11.2 Combining Prior Information About  $\pi$  with Data 371
  - 11.3 Prior Distributions for  $\pi$  374
  - 11.4 Calculating the Posterior Distribution for  $\pi$  375
  - 11.5 The Posterior Mean 379
  - 11.6 What Can We Conclude About Colors of M&M's? 380
  - 11.7 Screening for the HIV Virus in the U.S. Blood Bank Supply: Applications of Bayes' Rule 380
  - 11.8 Ganzfeld Experiments: Continuous Prior Distributions for  $\pi$  384
  - 11.9 Return to M&M's: Bayesian Credible Intervals 394
- Research Project:** Do You Believe in ESP? 406

Appendix of Tables 409

Index 414

# Preface

## About This Book

The goal of this text is to help students discover the power, diversity, and broad applicability of statistics. The framework for the text is a collection of practical statistical methods used in the natural and social sciences. Within each chapter, guided activities assist students in working through the entire process of data analysis as they grapple with intriguing real-world problems.

The text includes sufficient material for a one- or two-semester second course in statistics and is accessible to students who have taken a one-semester, algebra-based introductory statistics course or a high school advanced placement statistics course. The text assumes a fundamental knowledge of the concepts of hypothesis testing and confidence intervals and familiarity with statistical inference by  $t$ - and  $z$ -procedures. However, these topics are briefly reviewed when referenced in the text.

In order to emphasize the process of data analysis from a statistician's perspective, we have collaborated with colleagues from various disciplines to develop authentic instructional materials that

1. broaden student understanding of the intellectual content and applicability of statistics as a discipline;
2. use interdisciplinary activities and projects to encourage communication between statisticians and those in other disciplines;
3. strengthen the ability of students to design, conduct, analyze, and present their own research;
4. incorporate the *Guidelines in Assessment and Instruction in Statistics Education*<sup>\*</sup>
  - emphasize statistical literacy and develop statistical thinking,
  - use real data,
  - stress conceptual understanding rather than mere knowledge of procedures,
  - foster active learning in the classroom,
  - use technology to develop conceptual understanding and analyze data, and
  - use assessments to improve and evaluate student learning.

## The Approach

Our intention is to introduce students to a wide range of statistical methods early in their college careers and, through active learning, let them experience how statisticians think and practice. Just as in an introductory statistics course, we don't necessarily expect students working through this material to become statisticians, but exposure to a broad set of topics will give them a starting point for analyzing data in any discipline. This text will give students the confidence they need to explore data and models (or at least recognize when to call a statistician and be equipped with the statistical vocabulary for that conversation).

Through using these materials, students from any discipline can develop an understanding of the basic conditions under which studies should be conducted and when particular statistical techniques should and should not be used. By researching the literature, planning and carrying out experiments, and presenting their results, students using this text will experience data analysis as it is actually practiced.

We have found that the use of guided activities and projects improves the overall quality of student work, yielding better results than a pure lecture setting. By taking an active learning approach, students are better able to develop research questions that are general enough to be interesting yet specific enough to be analyzed in the given time frame. They conduct more thorough analyses and better understand the importance of model checking. Finally, they tend to be much more enthusiastic about their work and are better able to communicate the key concepts to a diverse audience.<sup>†</sup>

## Organization

This highly adaptable text consists of 11 self-contained chapters, each of which focuses on a particular statistical method. Because of the modular nature of the text, chapters can be covered in a flexible sequence based on how an instructor wants to customize the course to the student audience and discipline.

<sup>\*</sup>American Statistical Association, "Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report," <http://www.amstat.org/education/gaise/> (accessed March 20, 2011).

<sup>†</sup>S. Kuiper and L. Collins, "Guided Labs That Introduce Statistical Techniques Used in Research From Multiple Disciplines," *The American Statistician*, 63 (2009): 343–347.

This text is designed for use in a one- or two-semester second course in statistics. However, there are several possible alternative uses for this text:

- covering an individual chapter in the final weeks of an AP introductory statistics course after the AP exam has been administered
- assigning a research project as a final project in an introductory statistics course
- using a selected subset of chapters as supplementary materials in existing courses on topics such as regression analysis, methods for categorical data, or the design of experiments

Additionally, this text can be used in a variety of classroom environments. The activities are designed so that they can be completed inside or outside of class, depending on the level of in-class active learning, group work, and discussion that instructors prefer in their courses. The materials work well in a lab setting, but they have also been successfully used in lecture courses with 50–100 students, where students worked through the activities outside of class while the extended activities and end-of-chapter exercises were used as the foundation for the course lectures. Note that at least some class time will be needed to discuss the activities as students progress through each chapter.

## Distinctive Features

Each chapter is presented in the context of a real-world research question, which gives students exposure to the application of statistical methods in a variety of disciplines. The chapter begins with an introduction that poses the research question and lays the groundwork for the activity-based investigation.

- A series of **Guided Activities** walk students through the entire process of data analysis, reinforcing statistical thinking and conceptual understanding.
- Optional **Extended Activities** build on the foundation presented in the Activities and provide more in-depth coverage in diverse contexts and theoretical backgrounds. These sections are particularly useful for more advanced courses that discuss the material in more detail.
- Throughout the chapter, three types of **Notes** provide additional guidance and support to students. *Notes* provide additional background information, and *Cautions* provide warnings or clarification of needed assumptions. *Mathematical Notes* provide conceptual insights and mathematical details for more advanced students, but can be ignored in lower level courses.
- **Exercises** at the end of each chapter offer students additional practice applying the methods they've learned in the chapter. Questions from multiple disciplines are included, which instructors can assign as homework.
- Each chapter concludes with a **Research Project**, which guides students as they apply the statistical concept taught in the chapter to a relatively complex research question. In most projects, students read and evaluate primary literature, then plan and carry out studies that emphasize the process of science and data analysis. The projects often ask student to transition from research questions to a statistical model, collect data or develop a simulation study, and give an oral or written report after peer review. Some projects include sample data sets in case students don't have time to collect their own.
- **Data sets and software instructions for Minitab® and R** that walk students through some of the activities are included on the companion CD-ROM.

## Student Support

Each new copy of this text includes a CD-ROM containing many useful resources for students and instructors. These resources can also be downloaded from <http://www.pearsonhighered.com/mathstatsresources>.

On the CD-ROM are the following:

- *Student Answer Bank*, which includes answers to selected activities from the text
- *R Manual*, with detailed instructions for performing the text's activities using R
- *Minitab Manual*, with detailed instructions for performing the text's activities using Minitab
- *A Review of Introductory Statistics*, a brief review of the topics covered in the introductory statistics course, for student reference
- Glossary
- Instructions on how to write a research paper and poster

- Instructions on how to access applets referenced in the text
- Data sets formatted as .csv and .txt files.

## Instructor Support

*Instructor's Answer Bank* contains answers to all of the text's activities, extended activities, and exercises. It is available for download from Pearson's Instructor Resource Center (<http://www.pearsonhighered.com/irc>).

## Instructor Notes

*Instructor Notes*, written by the text authors, guide instructors in teaching with this text's exciting new approach to a second course. The *Instructor Notes* are available for download on the Instructor Resource Center (<http://www.pearsonhighered.com/irc>).

## Acknowledgments

Partial support for this work was provided by the Course, Curriculum, and Laboratory Improvement program at the National Science Foundation under DUE #0510392. We would like to thank

- Tom Moore (Chapter 1), Julie Legler (Chapter 8), and Linda Collins (Chapter 10) for their original contributions. Linda's supplements, class testing, multiple reviews, and edits to all chapters are greatly appreciated.
- Sam Rebelsky, Henry Walker, Andrew Applebaum, Alex Cohn, Nathan Levin, Jeffrey Thompson, Arunabh Singh, Sarah Marcum, and Elizabeth Lorton for their development of the on-line games.
- Accuracy checkers Steven Garren (James Madison University) and James Surles (Texas Tech University) for their careful review of the entire text.
- Our reviewers, whose suggestions vastly changed and improved the overall quality of this text—particularly Chris Olsen, whose careful comments on every chapter were greatly appreciated. We'd like to extend our thanks to all who took the time to review our text:

Reza Abbasian, *Texas Lutheran University*

Ming-Wen An, *Vassar College*

Kathleen Arano, *Fort Hays State University*

Dipankar Bandyopadhyay, *Medical University of South Carolina*

Janet Winter-Becker, *Penn State University–Berks*

Beth Benzing, *Strath Haven High School*

Natalie Blades, *Brigham Young University*

Ann Cannon, *Cornell College*

Julie Clark, *Hollins University*

George Cobb, *Mount Holyoke College*

Philip Dixon, *Iowa State University*

Michael Dohm, *Chaminade University*

Michelle Everson, *University of Minnesota*

Steven Garren, *James Madison University*

James Godbold, *Mount Sinai School of Medicine*

Jo Hardin, *Pomona College*

Solomon Harrar, *University of Montana*

Nicholas Horton, *Smith College*

Lifang Hsu, *Le Moyne College*

Michael Hughes, *Miami University*

Patricia Humphrey, *Georgia Southern University*

Kyoungmi Kim, *University of California at Davis*

Lynn Kuo, *University of Connecticut*

Patrick E. McKnight, *George Mason University*

Hosik Min, *University of Hawaii*

Amy Nuzzolese, *Great Neck North High School*

Liam O'Brien, *Colby College*

Michael Posner, *Villanova University*

Song Qian, *Duke University*

Karl Ronning, *Davis Senior High School*

Celia Rowland, *W. G. Enloe Magnet High School*

Julia Soulakova, *University of Nebraska–Lincoln*

John Stevens, *Utah State University*

Linda Strauss, *Penn State University*

Rodney Sturdivant, *West Point Military Academy*

James Surles, *Texas Tech University*

Carla Thompson, *University of West Florida*

Lewis VanBrackle, *Keenesaw State University*

Steve C. Wang, *Stanford University*

JoAnna Crixell Whitener, *U.S. Military Academy*

Phil Yates, *Saint Michael's College*

Michael Zwilling, *University of Mount Union*

*This page intentionally left blank*

# Nonparametric Methods: Schistosomiasis

1

*Using statistics is no substitute for thinking about the problem.*

—Douglas Montgomery<sup>1</sup>

Randomization tests, permutation tests, and bootstrap methods are quickly gaining popularity as methods for conducting statistical inference. Why? These nonparametric methods require fewer assumptions and provide results that are often more accurate than those from traditional techniques using well-known distributions (such as the normal,  $t$ -, or  $F$ -distribution). These methods are typically based on computer simulations instead of assumptions about distributions and thus are particularly useful when the sample data are skewed or when the sample size is small. In addition, nonparametric methods can be extended to other parameters of interest, such as the median, whereas the well-known parametric methods described in introductory statistics courses are often restricted to inference for the population mean.

We begin this chapter by comparing two treatments for a potentially deadly disease called schistosomiasis (skis-tuh-soh-mahy'-uh-sis). We illustrate the basic concepts behind nonparametric methods by using randomization tests to

- Provide an intuitive description of statistical inference
- Conduct a randomization test by hand
- Use software to conduct a randomization test
- Compare one-sided and two-sided hypothesis tests
- Make connections between randomization tests and conventional terminology

After working through the schistosomiasis investigation, you will have the opportunity to analyze several other data sets using randomization tests, permutation tests, bootstrap methods, and rank-based nonparametric tests.

## 1.1 Investigation: Can a New Drug Reduce the Spread of Schistosomiasis?

Schistosomiasis is a disease occurring in humans caused by parasitic flatworms called schistosomes (skis'-tuh-sohms). Schistosomiasis affects about 200 million people worldwide and is a serious problem in sub-Saharan Africa, South America, China, and Southeast Asia. The disease can cause death, but more commonly results in chronic and debilitating symptoms, arising primarily from the body's immune reaction to parasite eggs lodged in the liver, spleen, and intestines.

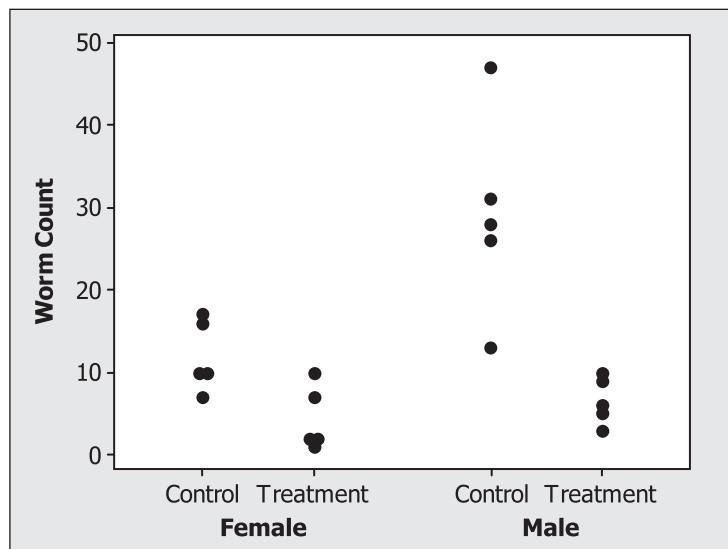
Currently there is one drug, praziquantel (pray'-zee-kwan-tel), in common use for treatment of schistosomiasis; it is inexpensive and effective. However many organizations are concerned about relying on a single drug to treat a serious disease that affects so many people worldwide. Drug resistance may have prompted an outbreak in the 1990s in Senegal, where cure rates were low. In 2007, several researchers published work on a promising drug called K11777, which, in theory, might treat schistosomiasis.<sup>2</sup>

In this chapter, we analyze data from a study in which the researchers wanted to find out whether K11777 helps to stop schistosome worms from growing. In one phase of the study, 10 female laboratory mice and 10 male laboratory mice were deliberately infected with the schistosome parasite. Starting seven days after being infected with schistosomiasis, each mouse was given an injection every day for 28 days. Within each sex, five mice were randomly assigned to a treatment of K11777 and the other five mice formed a control group injected with an equal volume of plain water. On day 49, the researchers euthanized the mice and measured both the number of eggs and the number of worms in their livers. Both numbers were expected to be lower in the treatment group if the drug was effective.

Table 1.1 gives the worm count for each mouse. An individual value plot of the data is shown in Figure 1.1. Notice that the treatment group has fewer worms than the control group for both females and males.

**Table 1.1** Worm count data for the schistosomiasis study. Treatment is a regimen of K11777 injections from day 7 to day 35. Control is the same regimen, but with a water solution only.

Female		Male	
Treatment	Control	Treatment	Control
1	16	3	31
2	10	5	26
2	10	9	28
10	7	10	13
7	17	6	47
Mean 4.40	12.00	6.60	29.00



**Figure 1.1** Individual value plot of the worm count data.

**NOTE**

There is a difference between individual value plots and dotplots. In **dotplots** (such as Figures 1.3 and 1.4 shown later in this chapter), each observation is represented by a dot along a number line ( $x$ -axis). When values are close or the same, the dots are stacked. Dotplots can be used in place of histograms when the sample size is small. **Individual value plots**, as shown in Figure 1.1, are used to simultaneously display each observation for multiple groups. They can be used instead of boxplots to identify outliers and distribution shape, especially when there are relatively few observations.

## Activity Describing the Data

1. Use Figure 1.1 to visually compare the number of worms for the treatment and control groups for both the male and the female mice. Does each of the four groups appear to have a similar center and a similar spread? Are there any outliers (extreme observations that don't seem to fit with the rest of the data)?
2. Calculate appropriate summary statistics (e.g., the median, mean, standard deviation, and range) for each of the four groups. For the female mice, calculate the difference between the treatment and control group means. Do the same for the male mice.

The descriptive analysis in Questions 1 and 2 points to a positive treatment effect: K11777 appears to have reduced the number of parasitic worms in this sample. But descriptive analysis is usually only the first step in ascertaining whether an effect is real; we often conduct a significance test or create a confidence interval to determine if chance alone could explain the effect.

Most introductory statistics courses focus on hypothesis tests that involve using a normal,  $t$ -, chi-square or  $F$ -distribution to calculate the  $p$ -value. These tests are often based on the central limit theorem. In the schistosomiasis study, there are only five observations in each group. This is a much smaller sample size than is recommended for the central limit theorem, especially given that Figure 1.1 indicates that the data may not be normally distributed. Since we cannot be confident that the sample averages are normally distributed, we will use a **distribution-free test**, also called a **nonparametric test**. Such tests do not require the distribution of our sample statistic to have any specific form and are often useful in studies with very small sample sizes.

**MATHEMATICAL NOTE**

For any population with mean  $\mu$  and finite standard deviation  $\sigma$ , the **central limit theorem** states that the sample mean  $\bar{x}$  from an independent and identically distributed sample tends to follow the normal distribution if the sample size is large enough. The mean of  $\bar{x}$  is the same as the population mean,  $\mu$ , while the standard deviation of  $\bar{x}$  is  $\sigma/\sqrt{n}$ , where  $n$  is the sample size.

We will use a form of nonparametric statistical inference known as a randomization hypothesis test to analyze the data from the schistosomiasis study. **Randomization hypothesis tests** are significance tests that simulate the random allocation of units to treatments many times in order to determine the likelihood of observing an outcome at least as extreme as the one found in the actual study.

**Key Concept**

**Parametric tests** (such as  $z$ -tests,  $t$ -tests or  $F$ -tests) assume that data come from a population that follows a probability distribution or use the central limit theorem to make inferences about a population. **Nonparametric tests** (such as randomization tests) do not require assumptions about the distribution of the population or large sample sizes in order to make inferences about a population.

We will introduce the basic concepts of randomization tests in a setting where units (mice in this example) are randomly allocated to a treatment or control group. Using a significance test, we will decide if an observed treatment effect (the observed difference between the mean responses in the treatment and control) is “real” or if “random chance alone” could plausibly explain the observed effect. The null hypothesis states that “random chance alone” is the reason for the observed effect. In this initial discussion, the alternative hypothesis will be one-sided because we want to show that the true treatment mean ( $\mu_{\text{treatment}}$ ) is less than the true control mean ( $\mu_{\text{control}}$ ). Later, we will expand the discussion to consider modifications needed to deal with two-sided alternatives.

## 1.2 Statistical Inference Through a Randomization Test

Whether they take the form of significance tests or confidence intervals, inferential procedures rest on the **fundamental question for inference**: “What would happen if we did this many times?” Let’s unpack this question in the context of the female mice in the schistosomiasis study. We observed a difference in means of  $7.6 = 12.00 - 4.40$  worms between control and treatment groups. While we expect that this large difference reflects the effectiveness of the drug, it is possible that chance alone could explain this difference. This “chance alone” position is usually called the null hypothesis and includes the following assumptions:

- The number of parasitic worms found in the liver naturally varies from mouse to mouse.
- Whether or not the drug is effective, there clearly is variability in the responses of mice to the infestation of schistosomes.
- Each group exhibits this variability, and even if the drug is not effective, some mice do better than others.
- The only explanation for the observed difference of 7.6 worms in the means is that the random allocation randomly placed mice with larger numbers of worms in the control group and mice with smaller numbers of worms in the treatment group.

In this study, the **null hypothesis** is that the treatment has no effect on the average worm count, and it is denoted as

$$H_0: \mu_{\text{control}} = \mu_{\text{treatment}}$$

Another way to write this null hypothesis is

$$H_0: \text{the treatment has no effect on average worm count}$$

The research hypothesis (the treatment causes a reduction in the average worm count) is called the **alternative hypothesis** and is denoted  $H_a$  (or  $H_1$ ). For example,

$$H_a: \mu_{\text{control}} > \mu_{\text{treatment}}$$

Another way to write this alternative hypothesis is

$$H_a: \text{the treatment reduces the average worm count}$$

Alternative hypotheses can be “one-sided, greater than” (as in this investigation), “one-sided, less-than” (the treatment causes an increase in worm count), or “two-sided” (the treatment mean is different, in one direction or the other, from the control mean). We chose to test a one-sided hypothesis because there is a clear research interest in one direction. In other words, we will take action (start using the drug) only if we can show that K11777 reduces the worm count.

### Key Concept

**The fundamental question for inference:** Every statistical inference procedure (parametric or non-parametric) is based on the question “How does what we observed in our data compare to what would happen if the null hypothesis were actually true and we repeated the process many times?” For a randomization test comparing responses for two groups, this question becomes “How does the observed difference between groups compare to what would happen if the treatments actually had no effect on the individual responses and we repeated the random allocation of individuals to groups many times?”

### Activity ➔ Conducting a Randomization Test by Hand

3. To get a feel for the concept of a  $p$ -value, write each of the female worm counts on an index card. Shuffle the 10 index cards, and then draw five cards at random (without replacement). Call these five cards the treatment group and the five remaining cards the control group. Under the null hypothesis (i.e. the treatment has no effect on worm counts), this allocation mimics precisely what actually happened in our experiment, since the only cause of group differences is the random allocation.

Calculate the mean of the five cards representing the treatment group and the mean of the five cards representing the control group. Then find the difference between the control and treatment group

means that you obtained in your allocation. To be consistent, take the control group mean minus the treatment group mean. Your work should look similar to the following simulation:

Control Group					Treatment Group				
1	7	7	10	17	2	10	16	2	10
<b>Mean = 8.4</b>					<b>Mean = 8</b>				

**Difference = 0.4**

- If you were to do another random allocation, would you get the same difference in means? Explain.
- Now, perform nine more random allocations, each time computing and writing down the difference in mean worm count between the control group and the treatment group. Make a dotplot of the 10 differences. What proportion of these differences are 7.6 or larger?
- If you performed the simulation many times, would you expect a large percentage of the simulations to result in a mean difference greater than 7.6? Explain.

The reasoning in the previous activity leads us to the randomization test and an interpretation of the fundamental question for inference. The fundamental question for this context is as follows: “If the null hypothesis were actually true and we randomly allocated our 10 mice to treatment and control groups many times, what proportion of the time would the observed difference in means be as big as or bigger than 7.6?” This long-run proportion is a probability that statisticians call the **p-value** of the randomization test. The p-values for most randomization tests are found through simulations. Despite the fact that simulations do not give exact p-values, they are usually preferred over the tedious and time-consuming process of listing all possible outcomes. Researchers usually pick a round number such as 10,000 repetitions of the simulation and approximate the p-value accordingly. Since this p-value is an approximation, it is often referred to as the **empirical p-value**.

#### Key Concept

Assuming that nothing except the random allocation process is creating group differences, the p-value of a randomization test is the probability of obtaining a group difference as large as or larger than the group difference actually observed in the experiment.

#### Key Concept

The calculation of an empirical p-value requires these steps:

- Repeat the random allocation process a number of times ( $N$  times).
- Record, each time, whether or not the group difference exceeds or is the same as the one observed in the actual experiment (let  $X$  be the number of times the group difference exceeds or is the same as the one observed).
- Compute  $X/N$  to get the p-value, the proportion of times the difference exceeds or is the same as the observed difference.

#### NOTE

Many researchers include the observed value as one of the possible outcomes. In this case,  $N = 9999$  iterations are typically used and the p-value is calculated as  $(X + 1)/(9999 + 1)$ . The results are very similar whether  $X/10,000$  or  $(X + 1)/(9999 + 1)$  is used. Including the observed value as one of the possible allocations is a more conservative approach and protects against getting a p-value of 0. Our observation from the actual experiment provides evidence that the true p-value is greater than zero.

## 1.3 Performing a Randomization Test Using a Computer Simulation

While physical simulations (such as the index cards activity) help us understand the process of computing an empirical p-value, using computer software is a much more efficient way of producing an empirical p-value based on a large number of iterations. If you are simulating 10 random allocations, it is just as easy to use

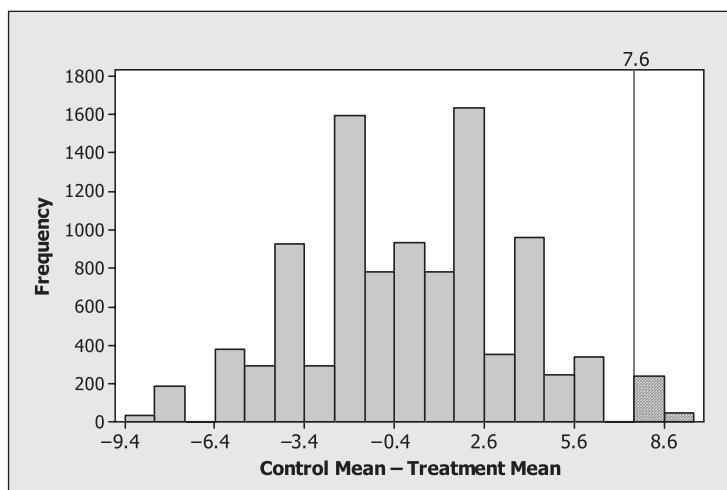
index cards as a computer. However, the advantage of a computer simulation is that 10,000 random allocations can be conducted in almost the same amount of time it takes to simulate 10 allocations. In the following steps, you will develop a program to calculate an empirical  $p$ -value.

## Activity Using Computer Simulations to Conduct a Hypothesis Test

7. Use the technology instructions provided on the CD to insert the schistosomiasis data into a statistical software package and randomly allocate each of the 10 female worm counts to either the treatment or the control group.
8. Take the control group average minus the K11777 treatment group average.
9. Use the instructions to write a program, function, or macro to repeat the process 10,000 times. Count the number of simulations where the difference between the group averages (control minus K11777) is greater than or equal to 7.6, divide that count by 10,000, and report the resulting empirical  $p$ -value.
10. Create a histogram of the 10,000 simulated differences between group means and comment on the shape of the histogram. This histogram, created from simulations of a randomization test, is called an **empirical randomization distribution**. This distribution describes the frequency of each observed difference (between the control and treatment means) when the null hypothesis is true.
11. Based on your results in Questions 9 and 10 and assuming the null hypothesis is true, about how frequently do you think you would obtain a mean difference as large as or larger than 7.6 by random allocation alone?
12. Does your answer to Question 11 lead you to believe the “chance alone” position (i.e., the null hypothesis that the mean worm count is the same for both the treatment and the control), or does it lead you to believe that K11777 has a positive inhibitory effect on the schistosome worm in female mice? Explain.

Figure 1.2 shows a histogram resulting from the previous activity. A computer simulation of Question 9 resulted in a  $p$ -value of  $281/10,000 = 0.0281$ . This result shows that random allocation alone would produce a mean group difference as large as or larger than 7.6 only about 3% of the time, suggesting that something other than chance is needed to explain the difference in group means. Since the only other distinction between the groups is the presence or absence of treatment, we can conclude that the treatment causes a reduction in worm counts.

We conducted four more simulations, each with 10,000 iterations, which resulted in  $p$ -values of 0.0272, 0.0282, 0.0268, and 0.0285. When the number of iterations is large, the empirical randomization distribution (such as the histogram created in Question 10) provides a precise estimate of the likelihood of all possible



**Figure 1.2** Histogram showing the results of a schistosomiasis simulation study. In this simulation, 281 out of 10,000 iterations resulted in a difference greater than or equal to 7.6.

values of the difference between the control and treatment means. Thus, when the number of iterations is large, well-designed simulation studies result in empirical  $p$ -values that are fairly accurate. The larger the number of iterations (i.e., randomizations) within a simulation study, the more precise the  $p$ -value is.

Because the sample sizes in the schistosomiasis study are small, it is possible to apply mathematical methods to obtain an **exact  $p$ -value** for this randomization test. An exact  $p$ -value can be calculated by writing down the set of all possibilities (assuming each possible outcome is equally likely under the null hypothesis) and then calculating the proportion of the set for which the difference is at least as large as the observed difference. In the schistosomiasis study, this requires listing every possible combination in which five of the 10 female mice can be allocated to the treatment (and the other five assigned to the control). There are 252 possible combinations. For each of these combinations, the difference between the treatment and control means is then calculated. The exact  $p$ -value is the proportion of times in which the difference in the means is at least as large as the observed difference of 7.6 worms. Of these 252 combinations, six have a mean difference of 7.6 and one has a mean difference greater than 7.6 (namely 8.8). Since all 252 of these random allocations are equally likely, the exact  $p$ -value in this example is  $7/252 = 0.0278$ . However, most real studies are too large to list all possible samples. Randomization tests are almost always adequate, providing approximate  $p$ -values that are close enough to the true  $p$ -value.

#### CAUTION

Conducting a two-sample  $t$ -test on the female mice provides a  $p$ -value of 0.011. This  $p$ -value of 0.011 is accurate only if the observed test statistic (i.e., the difference between means) follows appropriate assumptions about the distribution. Figure 1.2 demonstrates that the distributional assumptions are violated. While the randomization test provides an approximate  $p$ -value “close to 0.0278,” it provides a much better estimate of the exact  $p$ -value than does the two-sample  $t$ -test. Note that each of the five simulations listed gave a  $p$ -value closer to the exact  $p$ -value than the one given by the two-sample  $t$ -test. *Be careful not to trust a  $p$ -value provided by statistical software unless you are certain the appropriate assumptions are met.*

#### Key Concept

The larger the number of randomizations within a simulation study, the more precise the  $p$ -value is. When sample sizes are small or sample data clearly are not normal, a  $p$ -value derived from a randomization test with 10,000 randomizations is typically more accurate than a  $p$ -value calculated from a parametric test (such as the  $t$ -test).

Sometimes we have some threshold  $p$ -value at or below which we will reject the null hypothesis and conclude in favor of the alternative. This threshold value is called a **significance level** and is usually denoted by the Greek letter alpha ( $\alpha$ ). Common values are  $\alpha = 0.05$  and  $\alpha = 0.01$ , but the value will depend heavily on context and on the researcher’s assessment of the acceptable risk of stating an incorrect conclusion. When the study’s  $p$ -value is less than or equal to this significance level, we state that the results are **statistically significant at level  $\alpha$** . If you see the phrase “statistically significant” without a specification of  $\alpha$  the writer is most likely assuming  $\alpha = 0.05$ , for reasons of history and convention alone. However, it is best to show the  $p$ -value instead of simply stating a result is significant at a particular  $\alpha$ -level.

## 1.4 Two-Sided Tests

The direction of the alternative hypothesis is derived from the research hypothesis. In this K11777 study, we enter the study expecting a reduction in worm counts and hoping the data will bear out this expectation. It is our expectation, hope, or interest that drives the alternative hypothesis and the randomization calculation. Occasionally, we enter a study without a firm direction in mind for the alternative, in which case we use a two-sided alternative. Furthermore, even if we hope that the new treatment will be better than the old treatment or better than a control, we might be wrong—it may be that the new treatment is actually worse than the old treatment or even harmful (worse than the control). Some statisticians argue that a conservative objective approach is to always consider the two-sided alternative. For a **two-sided test**, the  $p$ -value must take into account extreme values of the test statistic in either direction (no matter which direction we actually observe in our sample data).

**Key Concept**

The direction of the alternative hypothesis does not depend on the sample data, but instead is determined by the research hypothesis before the data are collected.

We will now make our definition of the *p*-value more general to allow for a wider variety of significance testing situations. The ***p*-value** is the probability of observing a group difference as extreme as or more extreme than the group difference actually observed in the sample data, assuming that there is nothing creating group differences except the random allocation process.

This definition is consistent with the earlier definition for one-sided alternatives, as we can interpret *extreme* to mean either greater than or less than, depending on the direction of the alternative hypothesis. But in the two-sided case, *extreme* encompasses both directions. In the K11777 example, we observed a difference of 7.6 between control and treatment group means. Thus, the two-sided *p*-value calculation is a count of all instances among the 10,000 replications where the randomly allocated mean difference is either as small as or smaller than  $-7.6$  worms ( $\leq -7.6$ ) or as great as or greater than 7.6 worms ( $\geq 7.6$ ). This is often written as  $|\text{diff}| \geq 7.6$ .

### Activity ◉ A Two-Sided Hypothesis Test

13. Run the simulation study again to find the empirical *p*-value for a two-sided hypothesis test to determine if there is a difference between the treatment and control group means for female mice.
14. Is the number of simulations resulting in a difference greater than or equal to 7.6 identical to the number of simulations resulting in a difference less than or equal to  $-7.6$ ? Explain why these two values are likely to be close but not identical.
15. Explain why you expect the *p*-value for the two-sided alternative to be about double that for the one-sided alternative. Hint: You may want to look at Figure 1.2
16. Using the two-sided alternative hypothesis, the two-sample *t*-test provides a *p*-value of 0.022.\* This *p*-value would provide strong evidence for rejecting the assumption that there is no difference between the treatment and the control (null hypothesis). However, this *p*-value should not be used to draw conclusions about this study. Explain why.

For the above study, a simulation involving 100,000 iterations provided an empirical *p*-value of 0.0554. Again, because this particular data set is small, all 252 possible random allocations can be listed to find that the exact two-sided *p*-value is  $14/252 = 0.0556$ .

## 1.5 What Can We Conclude from the Schistosomiasis Study?

The key question in this study is whether K11777 will reduce the spread of a common and potentially deadly disease. The result that you calculated from the one-sided randomization hypothesis test should have been close to the exact *p*-value of 0.0278. This small *p*-value allows you to reject the null hypothesis and conclude that the worm counts are lower in the female treatment group than in the female control group. In every study, it is important to consider how random allocation and random sampling impact the conclusions.

*Random allocation:* The schistosomiasis study was an **experiment** because the units (female mice) were randomly allocated to treatment or control groups. To the best of our knowledge this experiment controlled for any outside influences and allows us to state that there is a cause and effect relationship between the treatment and response. Therefore, we can conclude that K11777 did *cause* a reduction in the average number of schistosome parasites in these female mice.

*Random sampling:* Mice for this type of study are typically ordered from a facility that breeds and raises lab mice. It is possible that the mice in this study were biologically related or were exposed to something that caused their response to be different from that of other mice. Similarly, there are risks in simply assuming that male mice have the same response as females, so the end-of-chapter exercises provide an opportunity

\*When we do not assume equal variances Minitab uses 7 degrees of freedom providing a *p*-value of 0.022 while R uses 7.929 degrees of freedom resulting in a *p*-value of 0.0194.

to conduct a separate test on the male mice. Since our sample of 10 female mice was not selected at random from the population of all mice, we should question whether the results from this study hold for all mice.

More importantly, the results have not shown that this new drug will have the same impact on humans as it does on mice. In addition, even though we found that K11777 does cause a reduction in worm counts, we did not specifically show that it will reduce the spread of the disease. Is the disease less deadly if only two worms are in the body instead of 10? Statistical consultants aren't typically expected to know the answers to these theoretical, biological, or medical types of questions, but they should ask questions to ensure that the study conclusions match the hypothesis that was tested. In most cases, drug tests require multiple levels of studies to ensure that the drug is safe and to show that the results are consistent across the entire population of interest. While this study is very promising, much more work is needed before we can conclude that K11777 can reduce the spread of schistosomiasis in humans.

## A Closer Look

## Nonparametric Methods

### 1.6 Permutation Tests versus Randomization Tests

The random allocation of experimental units (e.g., mice) to groups provides the basis for statistical inference in a randomized comparative experiment. In the schistosomiasis K11777 treatment study, we used a significance test to ascertain whether cause and effect was at work. In the context of the random allocation study design, we called our significance test a randomization test.

In **observational studies**, subjects are not randomly allocated to groups. In this context, we apply the same inferential procedures as in the previous experiment, but we commonly call the significance test a **permutation test** rather than a randomization test.\* More importantly, in observational studies, the results of the test cannot typically be used to claim cause and effect; a researcher should exhibit more caution in the interpretation of results.

#### NOTE

The permutation test does not require that the data (or the sampling distribution) follow a normal distribution. However, the null hypothesis in a permutation test assumes that samples are taken from two populations that are similar. So, for example, if the two population variances are very different, the *p*-value of a permutation test may not be reliable. However, the two-sample *t*-test (taught in most introductory courses) allows us to assume unequal variances.

#### Key Concept

Whereas in experiments units are randomly allocated to treatment groups, observational studies do not impose a treatment on a unit. Because the random allocation process protects against potential biases caused by extraneous variables, experiments are often used to show causation.

### Age Discrimination Study

Westvaco is a company that produces paper products. In 1991, Robert Martin was working in the engineering department of the company's envelope division when he was laid off in Round 2 of several rounds of layoffs by the company.<sup>3</sup> He sued the company, claiming to be the victim of age discrimination. The ages of the 10 workers involved in Round 2 were: 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64. The ages of the three people laid off were 55, 55, and 64.

Figure 1.3 shows a comparative dotplot for age by layoff category. This dotplot gives the impression that Robert Martin may have a case: It appears as if older workers were more likely to be laid off. But we know enough about variability to be cautious.

\*This text defines a randomization test as a permutation test that is based on random allocation. Some statisticians do not distinguish between permutation tests and randomization tests. They call simulation studies permutation tests, whether they are based on observational studies or experiments.

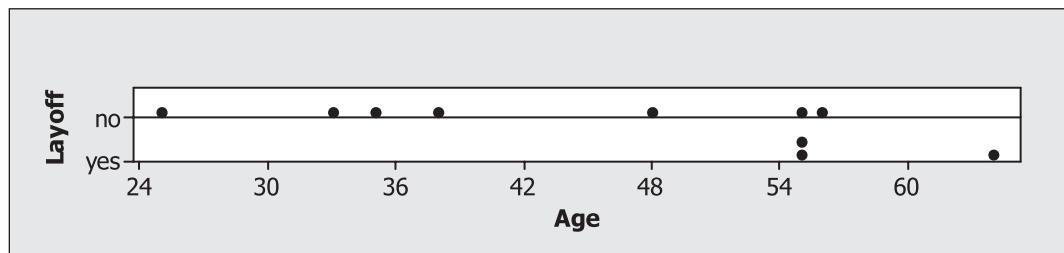


Figure 1.3 Dotplot of age in years of worker versus layoff (whether he or she was laid off).

## Extended Activity

### Is There Evidence of Age Discrimination?

Data set: Age

17. Conduct a permutation test to determine whether the observed difference between means is likely to occur just by chance. Use `Age` as the response variable and `Layoff` as the explanatory variable. Here we are interested in only a one-sided hypothesis test to determine if the mean age of people who were laid off is higher than the mean age of people who were not laid off.
18. Modify the program/macro you created in Question 17 to conduct a one-sided hypothesis test to determine if the median age of people who were laid off is higher than the median age of people who were not laid off. Report the  $p$ -value and compare your results to those in Question 17.

Since there was no random allocation (i.e., people were not randomly assigned to a layoff group), statistical significance does not give us the right to assert that greater age is *causing* a difference in being laid off. The null hypothesis in this context becomes “The observed difference could be explained *as if* by random allocation alone.” That is, we proceed as any practicing social scientist must when working with observational data. We “imagine” an experiment in which workers are randomly allocated to a layoff group and then determine if the observed average difference between the ages of laid-off workers and those not laid off is significantly larger than would be expected to occur by chance in a randomized comparative experiment.

While age could be the cause for the difference—hence proving an allegation of age discrimination—there are many other possibilities (i.e., extraneous variables), such as the educational levels of the workers, their competence to do the job, and ratings on past performance evaluations. Rejecting the “as if by random allocation” hypothesis in the nonrandomized context can be a *useful step* toward establishing causality; however, it cannot establish causality unless the extraneous variables have been properly accounted for.

In the actual court case, data from all three rounds of layoffs were statistically analyzed. The analysis showed some evidence that older people were more likely to be laid off; however, Robert Martin ended up settling out of court.

## 1.7 Permutation and Randomization Tests for Matched Pairs Designs

The ideas developed in this chapter can be extended to other study designs, such as a basic two-variable design called a **matched pairs design**. In a matched pairs design, each experimental unit provides both measurements in a study with two treatments (one of which could be a control). Conversely, in the completely randomized situation of the schistosomiasis K11777 treatment study, half the units were assigned to control and half to treatment; no mouse received both treatments.

### Music and Relaxation

Grinnell College students Anne Tillema and Anna Tekippe conducted an experiment to study the effect of music on a person’s level of relaxation. They hypothesized that fast songs would increase pulse rate more than slow songs. The file called `Music` contains the data from their experiment. They decided to use a person’s pulse rate as an operational definition of the person’s level of relaxation and to compare pulse rates for

two selections of music: a fast song and a slow song. For the fast song they chose “Beyond” by Nine Inch Nails, and for the slow song they chose Rachmaninoff’s “Vocalise.” They recruited 28 student subjects for the experiment.

Anne and Anna came up with the following experimental design. Their fundamental question involved two treatments: (1) listening to the fast song and (2) listening to the slow song. They could have randomly allocated 14 subjects to hear the fast song and 14 subjects to hear the slow song, but their more efficient approach was to have each subject provide both measurements. That is, each subject listened to both songs, giving rise to two data values for each subject, called a matched pairs. Randomization came into play when it was decided by a coin flip whether each subject would listen first to the fast song or the slow song.

### NOTE

There are several uses of randomness mentioned in this chapter. The emphasis of this chapter is on the use of **randomization tests** for statistical inference. Most introductory statistics courses discuss **random sampling** from a population, which allows the results of a specific study to be generalized to a larger population. In experiments, units are **randomly allocated to groups** which allows researchers to make statements about causation. In this example, Anne and Anna **randomize the order** to prescribe two conditions on a single subject.

Specifically, as determined by coin flips, half the subjects experienced the following procedure:

[one minute of rest; measure pulse (prepulse)] > [listen to fast song for 2 minutes; measure pulse for second minute (fast song pulse)] > [rest for one minute] > [listen to slow song for 2 minutes; measure pulse for second minute (slow song pulse)].

The other half experienced the procedure the same way except that they heard the slow song first and the fast song second.

Each subject gives us two measurements of interest for analysis: (1) fast song pulse minus prepulse and (2) slow song pulse minus prepulse. In the data file, these two measurements are called `Fastdiff` and `Slowdiff`, respectively.

Figure 1.4 shows a dotplot of the 28 `Fastdiff-minus-Slowdiff` values. Notice that positive numbers predominate and the mean difference is 1.857 beats per minute, both suggesting that the fast song does indeed heighten response (pulse rate) more than the slow song. We need to confirm this suspicion with a randomization test.

To perform a randomization test, we mimic the randomization procedure of the study design. Here, the randomization determined the order in which the subject heard the songs, so randomization is applied to the two measurements of interest for each subject. To compute a *p*-value, we determine how frequently we would obtain an observed difference as large as or larger than 1.857.

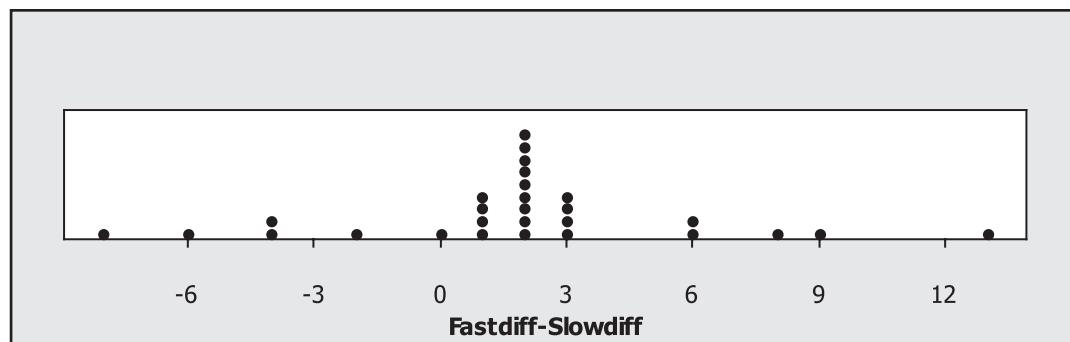


Figure 1.4 Dotplot of the difference in pulse rates for each of the 28 subjects.

## Extended Activity

### Testing the Effect of Music on Relaxation

Data set: Music

19. Before they looked at the data, Anne and Anna decided to use a one-sided test to see whether fast music increased pulse rate more than slow music. Why is it important to determine the direction of the test before looking at the data?
20. Create a simulation to test the `Music` data. Use the technology instructions provided to randomly multiply a 1 or a  $-1$  by each observed difference. This randomly assigns an order (`Fastdiff-Slowdiff` or `Slowdiff-Fastdiff`). Then, for each iteration, calculate the mean difference. The  $p$ -value is the proportion of times your simulation found a mean difference greater than or equal to 1.857.
  - a. Create a histogram of the mean differences. Mark the area on the histogram that represents your  $p$ -value.
  - b. Use the  $p$ -value to state your conclusions in the context of the problem. Address random allocation and random sampling (or lack of either) when stating your conclusions.

#### CAUTION

The type of randomization in Question 20 *does not* account for extraneous variables such as a great love for Nine Inch Nails on the part of some students or complete boredom with this band on the part of others (i.e., “musical taste” is a possible confounder that randomizing the order of listening cannot randomize away). There will always be a caveat in this type of study, since we are rather crudely letting one Nine Inch Nails song “represent” fast songs.

## 1.8 The Bootstrap Distribution

Bootstrapping is another simulation technique that is commonly used to develop confidence intervals and hypothesis tests. Bootstrap techniques are useful because they generalize to situations where traditional methods based on the normal distribution cannot be applied. For example, they can be used to create confidence intervals and hypothesis tests for any parameter of interest, such as a median, ratio, or standard deviation. Bootstrap methods differ from previously discussed techniques in that they sample **with replacement** (randomly draw an observation from the original sample and put the observation back before drawing the next observation).

Permutation tests, randomization tests, and bootstrapping are often called **resampling techniques** because, instead of collecting many different samples from a population, we take repeated samples (called **resamples**) from just one random sample.

## Extended Activity

### Creating a Sampling Distribution and a Bootstrap Distribution

Data set: ChiSq

21. The file `ChiSq` contains data from a highly skewed population (with mean 0.9744 and standard deviation 1.3153).
  - a. Take 1000 simple random samples of size 40 and calculate each mean ( $\bar{x}$ ). Plot the histogram of the 1000 sample means. The distribution of sample means is called the **sampling distribution**.
  - b. What does the central limit theorem tell us about the shape, center, and spread of the sampling distribution in this example?
  - c. Calculate the mean and standard deviation of the sampling distribution in Part A. Does the sampling distribution match what you would expect from the central limit theorem? Explain.
22. Take one simple random sample of size 40 from the `ChiSq` data.
  - a. Take 1000 resamples (1000 samples of 40 observations with replacement from the one simple random sample).
  - b. Calculate the mean of each resample ( $\bar{x}^*$ ) and plot the histogram of the 1000 resample means. This distribution of resample means is called the **bootstrap distribution**.
  - c. Compare the shape, center, and spread of the simulated histograms from Part B and Question 21 Part A. Are they similar?

23. Instead of using the sample mean, create a sampling distribution and bootstrap distribution of the standard deviation of the ChiSq data using a sample size of 40. Compare the shape, center, and spread of the simulated histograms and compare the mean and standard deviation of the distributions.

### Key Concept

The bootstrap method takes one simple random sample of size  $n$  from a population. Then many resamples (with replacement) are taken from the original simple random sample. Each resample is the same size as the original random sample. The statistic of interest is calculated from each resample and used to create a bootstrap distribution.

In many real-world situations, the process used in Question 21 is not practical because collecting more than one simple random sample is too expensive or time consuming. While the approach in Question 22 is computer intensive, it is simple and convenient since it uses only one simple random sample. The key idea behind bootstrap methods is the assumption that the original sample represents the population, so resamples from the one simple random sample can be used to represent samples from the population, as is done in Question 22. Thus, the **bootstrap distribution provides an approximation of the sampling distribution**.

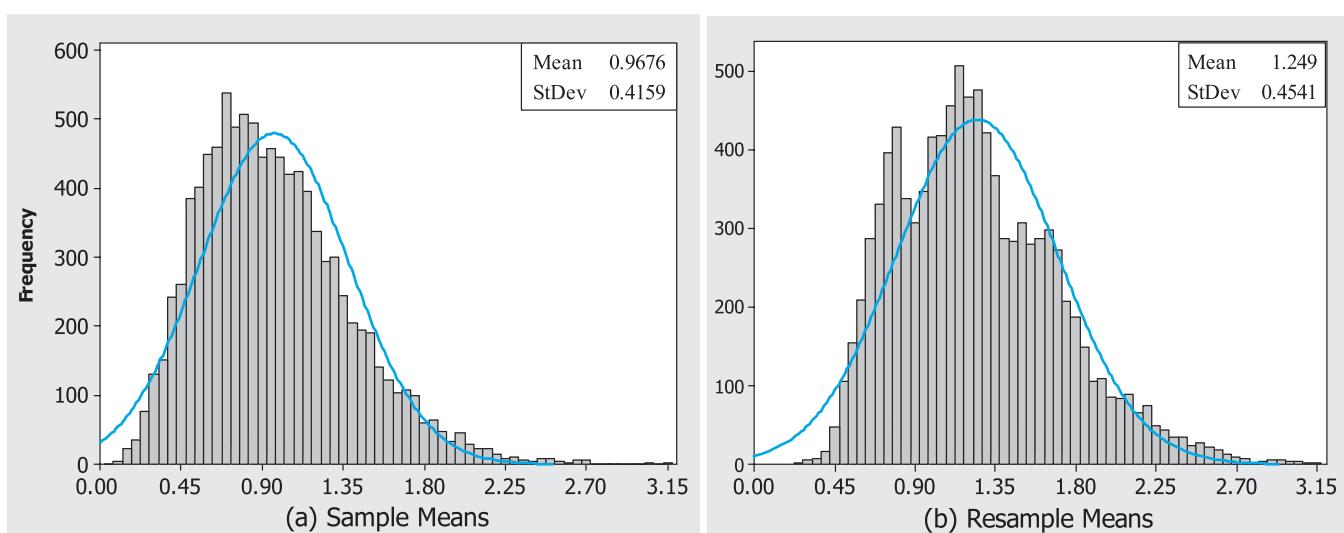
Most traditional methods of statistical inference involve collecting one sample and calculating the sample mean. Then, based on the central limit theorem, assumptions are made about the shape and spread of the sampling distribution. In Question 22 we used one sample to calculate the sample mean and then used the bootstrap distribution to estimate the shape and spread of the sampling distribution.

The central limit theorem tells us about the shape and spread of the sample mean. A key advantage of the bootstrap distribution is that it works for any parameter of interest. Thus, the bootstrap distribution can be used to estimate the shape and spread for any sampling distribution of interest.

### CAUTION

When sample sizes are small, one simple random sample may not represent the population very well. However, with larger sample sizes, the bootstrap distribution does represent the sampling distribution.

Figure 1.5 shows the sampling distribution and the bootstrap distribution when a sample size of 10 is used to estimate the mean of the ChiSq data. Notice that the spreads for both histograms are roughly equivalent. The central limit theorem tells us that the standard deviation of the sampling



**Figure 1.5** (a) The sampling distribution of the sample mean. The histogram is based on the mean of 10,000 samples of size 10 from the ChiSq data. (b) The bootstrap distribution of the sample mean. The histogram is based on the mean of 10,000 resamples (with replacement) of one simple random sample of the ChiSq data with mean 1.238 and standard deviation 1.490.

distribution (the distribution of  $\bar{x}$ ) should be  $\sigma/\sqrt{n} = 1.3153/\sqrt{10} = 0.4159$ . The standard deviation of the bootstrap distribution is 0.4541, which is a reasonable estimate of the standard deviation of the sampling distribution. In addition, both graphs have similar, right-skewed shapes. The strength of the bootstrap method is that it provides accurate estimates of the shape and spread of the sampling distribution. In general, histograms from the bootstrap distribution will have a similar shape and spread as histograms from the sampling distribution.

The bootstrap method does not improve our estimate of the population mean. The mean of the sampling distribution in Question 21 will typically be very close to the population mean. But the mean of the bootstrap distribution in Question 22 typically will not be as accurate, because it is based on only one simple random sample. Ideally, we would like to know how close the statistic from our original sample is to the population parameter. A statistic is biased if it is not centered at the value of the population parameter. We can use the bootstrap distribution to estimate the bias of a statistic. The difference between the original sample mean and the bootstrap mean is called the **bootstrap estimate of bias**.

### Key Concept

The estimate of the mean (or any parameter of interest) provided by the bootstrap distribution is not any better than the estimate provided by the observed statistic from the original simple random sample. However, the shape and spread of the bootstrap distribution will be similar to the shape and spread of the sampling distribution. The bootstrap technique can be used to estimate sampling distribution shapes and standard deviations that cannot be calculated theoretically.

## 1.9 Using Bootstrap Methods to Create Confidence Intervals

A **confidence interval** gives a range of plausible values for some parameter. This is a range of values surrounding an observed estimate of the parameter—an estimate based on the data. To this range of values we attach a level of confidence that the true parameter lies in the range. An alpha-level,  $\alpha$ , is often used to specify the level of confidence. For example, when  $\alpha = 0.05$ , we have a  $100(1 - \alpha)\% = 95\%$  confidence level. Thus, a  $100(1 - \alpha)\%$  confidence interval gives an estimate of where we think the parameter is and how precisely we have it pinned down.

### Bootstrap *t* Confidence Intervals

If the bootstrap distribution appears to be approximately normal, it is typically safe to assume that a *t*-distribution can be used to calculate a  $100(1 - \alpha)\%$  confidence interval for  $\mu$ , often called a **bootstrap *t* confidence interval**:

$$\bar{x} \pm t^*(S^*) \quad (1.1)$$

where  $S^*$  is the standard deviation of the bootstrap distribution and  $t^*$  is the critical value of the *t*-distribution with  $n - 1$  degrees of freedom.

The one simple random sample of size  $n = 10$  used to create the bootstrap distribution in Figure 1.5b has a mean of  $\bar{x} = 1.238$  and a standard deviation of  $s = 1.490$ . The bootstrap distribution in Figure 1.5b has a mean of  $\bar{x}^* = 1.249$  and a standard deviation of  $S^* = 0.4541$ . Notice that Formula (1.1) uses the mean from the original sample but uses the bootstrap distribution to estimate the spread. If we *incorrectly assume* that the sampling distribution in Figure 1.5 is normal, a 95% bootstrap *t* confidence interval for  $\mu$  is given by

$$\bar{x} \pm t^*(S^*) = 1.238 \pm 2.262(0.4541)$$

where  $t^* = 2.262$  is the critical value corresponding to the 97.5th percentile of the *t*-distribution with  $n - 1 = 9$  degrees of freedom. Thus, the 95% confidence interval for  $\mu$  is  $(0.211, 2.265)$ .

### ► MATHEMATICAL NOTE ▾

The bootstrap  $t$  confidence interval is similar to the traditional one-sample  $t$  confidence interval. The key difference is that the bootstrap distribution estimates the standard error of the statistic with  $S^*$  instead of  $s/\sqrt{n}$ . When the data are not skewed and have no clear outliers, parametric tests are very effective with relatively small sample sizes (10–30 observations may be enough to use the  $t$ -distribution). The following formula uses the  $t$ -distribution to calculate a  $100(1 - \alpha)\%$  confidence interval for the mean of a normal population:

$$\bar{x} \pm t^* \left( \frac{s}{\sqrt{n}} \right) \quad (1.2)$$

where  $s/\sqrt{n}$  is the standard error of  $\bar{x}$  and  $t^*$  is the critical value of the  $t$ -distribution with  $n - 1$  degrees of freedom. Using the original sample of size 10 with mean 1.238 and standard deviation 1.490, we find that a 95% confidence interval for  $\mu$  is (0.172, 2.304). However, this confidence interval is appropriate for sample means only when the sampling distribution is approximately normal. If the data are skewed, even sample sizes greater than 30 may not be large enough to make the sampling distribution appear normal.

With skewed data or small sample sizes (if the original data are not normally distributed), parametric methods (which are based on the central limit theorem) are not appropriate. In Figure 1.5 we see that the sampling distribution is skewed to the right. *Thus, with a sample size of 10, neither the traditional one-sample t confidence interval nor the bootstrap t confidence interval is reliable in this example.* However, with a sample size of 40, the histograms in Questions 21 and 22 should tend to look somewhat normally distributed.

## Bootstrap Percentile Confidence Intervals

**Bootstrap percentile confidence intervals** are found by calculating the appropriate percentiles of the bootstrap distribution. To find a  $100(1 - \alpha)\%$  confidence interval, take the  $\alpha/2 \times 100$  percentile of each tail of the bootstrap distribution.

For example, to find a 95% confidence interval for  $\mu$ , sort all the observations from the bootstrap distribution and find the values that represent the 2.5th and 97.5th percentiles of the bootstrap distribution. The 2.5th percentile of the bootstrap distribution in Figure 1.5b is 0.546, and the 97.5th percentile is 2.282. Thus, a 95% confidence interval for  $\mu$  is (0.546, 2.282).

Notice that the percentile confidence interval is not centered at the sample mean. Since the bootstrap distribution is right skewed, the right side of the confidence interval ( $2.282 - 1.238 = 1.044$ ) is wider than the left side of the confidence interval ( $1.238 - 0.546 = 0.692$ ). This lack of symmetry can influence the accuracy of the confidence interval.

### Key Concept

A bootstrap percentile confidence interval contains the middle  $100(1 - \alpha)\%$  of the bootstrap distribution. If the bootstrap distribution is symmetric and is centered on the observed statistic (i.e., not biased), percentile confidence intervals work well.<sup>4</sup>

## When to Use Bootstrap Confidence Intervals

Bootstrap methods are extremely useful when we cannot use theory, such as the central limit theorem, to approximate the sampling distribution. Thus, bootstrap methods can be used to create confidence intervals for essentially any parameter of interest, while the central limit theorem is limited to only a few parameters (such as the population mean).<sup>\*</sup> However, bootstrap methods are not always reliable.

Small sample sizes still produce problems for bootstrap methods. When the sample size is small, (1) the sample statistic may not accurately estimate the population parameter, (2) the distribution of sample means

\*Theoretical methods allow distributional tests for more than just the population mean. However, for purposes of this text it is sufficient to understand that distributional methods tend to be more complicated and are limited to testing only a few parameters that could be of interest.

is less likely to be symmetric, and (3) the shape and spread of the bootstrap distribution may not accurately represent those of the true sampling distribution.

In addition, bootstrap methods do not work equally well for all parameters. For example, the end-of-chapter exercises show that bootstrapping often provides unreliable bootstrap distributions for median values because the median of a resample is likely to have only a few possible values. Thus, confidence intervals for medians should be used only with large ( $n \geq 100$ ) sample sizes.

It is not easy to determine whether bootstrap methods provide appropriate confidence intervals. The bootstrap  $t$  and bootstrap percentile confidence intervals are often compared to each other. While the percentile confidence interval tends to be more accurate, neither of the two should be used if the intervals are not relatively close. If the bootstrap distribution is skewed or biased, other methods should be used to find confidence intervals. More advanced bootstrap methods (such as BCa and tilting confidence intervals) are available that are generally accurate when bias or skewness exists in the bootstrap distribution.<sup>5</sup>

## Extended Activity

### Estimating Salaries of Medical Faculty

Data set: MedSalaries

24. The file `MedSalaries` is a random sample of  $n = 100$  salaries of medical doctors who were teaching at United States universities in 2009.
- Create a bootstrap distribution of the mean by taking 1000 resamples (with replacement). Create a bootstrap  $t$  confidence interval and a bootstrap percentile distribution to estimate the mean salaries.
  - Create a bootstrap distribution of the standard deviation by taking 1000 resamples (with replacement). Create a bootstrap  $t$  confidence interval and a bootstrap percentile distribution to estimate the population standard deviation.
  - Use Formula (1.2) to create a 95% confidence interval for the mean. Compare this confidence intervals to those in Part A. Would you expect these intervals to be similar? Why or why not?
  - Explain why Formula (1.2) cannot be used to create a 95% confidence interval for the standard deviation.

## 1.10 Relationship Between the Randomization Test and the Two-Sample $t$ -Test

R.A. Fisher, perhaps the preeminent statistician of the 20th century, introduced the randomization test in the context of a two-group randomly allocated experiment in his famous 1935 book, *Design of Experiments*.<sup>6</sup> At that time he acknowledged that the randomization test was not practical because of the computational intensity of the calculation. Clearly, 1935 predates modern computing. Indeed, Efron and Tibshirani describe the permutation test as “a computer-intensive statistical technique that predates computers.”<sup>7</sup> Fisher went on to assert that the classical two-sample  $t$ -test (for independent samples) approximates the randomization test very well. Ernst cites references to several approximations to the randomization tests using classical and computationally tractable methods that have been published over time.<sup>8</sup>

If you have seen two-sample tests previously, it is likely to have been in the context of what Ernst calls the population model, which he distinguishes from the randomization model. In a **population model**, units are selected at random from one or more populations. Most observational studies are population models. One simple case of a population model involves comparing two separate population means. In this case, we can take two independent simple random samples and use the classic two-sample  $t$ -test to make the comparison.

In a **randomization model**, a fixed number of experimental units are randomly allocated to treatments. Most experiments are randomization models. In randomization models such as the schistosomiasis example, the two samples are formed from a collection of available experimental units that are randomly divided into two groups. Since there are a fixed number of units, the groups are not completely independent. For example, if one of the 10 male mice had a natural resistance to schistosomiasis and was randomly placed in the treatment group, we would expect the control group to have a slightly higher worm count. Since the two groups are not completely independent, the assumptions of the classic two-sample  $t$ -test are violated. Even if the sample sizes in the schistosomiasis study were much larger, the randomization test would be a more appropriate test than the two-sample  $t$ -test. However, empirical evidence has shown that the two-sample  $t$ -test is a very good

approximation to the randomization test when sample sizes are large enough. We are fortunate that, in this age of modern computing, we no longer have to routinely compromise by using the *t*-test to approximate the randomization test.

### Key Concept

Historically, the two-sample *t*-test was used to approximate the *p*-value in randomization models because randomization tests were too difficult to compute. However, now that computers can easily simulate random assignment to groups, randomization tests should be used to calculate *p*-values for randomization models, especially if sample sizes are fairly small.

## 1.11 Wilcoxon Rank Sum Tests for Two Independent Samples

The **Wilcoxon rank sum test**, also called the two-sample **Mann-Whitney test**, makes inferences about the difference between two populations based on data from two independent random samples. This test ranks observations from two samples by arranging them in order from smallest to largest.

Focusing on ranks instead of the actual observed values allows us to remove assumptions about the normal distribution. Rank-based tests have been used for many years. However, rank-based methods (discussed in this section and the next section) are much less accurate than methods based on simulations. In general, randomization tests, permutation tests, or bootstrap methods should be used whenever possible.

The following example examines whether pitchers and first basemen who play for National League baseball teams have the same salary distribution. The null and alternative hypotheses are written as

$H_0$ : the distribution of the salaries is the same for pitchers and first basemen

$H_a$ : the distribution of the salaries is different for pitchers and first basemen

Table 1.2 shows the salaries of five pitchers and five first basemen who were randomly selected from all National League baseball players. Table 1.3 ranks each of the players based on 2005 salaries.

Note that if two players had exactly the same salary, standard practice would be to average the ranks of the tied values.

**Table 1.2** Randomly selected pitchers and first baseman from 2005 National League baseball teams.

Team	Position	Name	Salary (\$)
Atlanta Braves	Pitcher	Sosa, Jorge	650,000
Atlanta Braves	Pitcher	Thomson, John	4,250,000
Milwaukee Brewers	Pitcher	Obermueller, Wes	342,000
Houston Astros	Pitcher	Backe, Brandon	350,000
New York Mets	Pitcher	Glavine, Tom	10,765,608
Cincinnati Reds	First Baseman	Casey, Sean	7,800,000
San Diego Padres	First Baseman	Nevin, Phil	9,625,000
Arizona Diamondbacks	First Baseman	Green, Shawn	7,833,333
Colorado Rockies	First Baseman	Helton, Todd	12,600,000
Philadelphia Phillies	First Baseman	Thome, Jim	13,166,667

**Table 1.3** Ranking the 10 randomly selected 2005 National League baseball players.

Position	P	P	P	P	FB	FB	FB	P	FB	FB
Salary (in \$1000s)	342	350	650	4250	7800	7833	9625	10,766	12,600	13,167
Rank	1	2	3	4	5	6	7	8	9	10

For the Wilcoxon rank sum test, we define the following terms:

- $n_1$  is the sample size for the first group (5 for the pitcher group in this example)
- $n_2$  is the sample size for the second group (5 for the first baseman group in this example)
- $N = n_1 + n_2$
- $W$ , the Wilcoxon rank sum statistic, is the sum of the ranks in the first group  
( $1 + 2 + 3 + 4 + 8 = 18$ )

If the two groups are from the same continuous distribution, then  $W$  has a mean

$$\mu_W = \frac{n_1(N+1)}{2} = \frac{5(11)}{2} = 27.5 \quad (1.3)$$

and standard deviation<sup>9</sup>

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}} = \sqrt{\frac{(5)(5)(11)}{12}} = 4.787 \quad (1.4)$$

If  $W$  is far from  $\mu_W$ , then the Wilcoxon rank sum test rejects the hypothesis that the two populations have identical distributions—that is, rejects  $H_0$  (no difference in distribution of salaries) in favor of  $H_a$  (salary distributions are different based on position). The  $p$ -value is the probability of observing a sample statistic,  $W$ , at least as extreme as the one in our sample. Since 18 is less than the hypothesized mean, 27.5, the  $p$ -value for the two-sided test in this example is found by calculating  $2 \times P(W \leq 18)$ .

#### ► MATHEMATICAL NOTE ▼

Computer software such as R, S-plus, or SAS tends to use the exact distribution of  $W$ , though Minitab uses a normal approximation for this test. If the data contain ties, the exact distribution for the Wilcoxon rank sum statistic changes and the standard deviation of  $W$  should be adjusted. Statistical software will typically detect the ties and use the normal distribution (using the adjusted standard deviation) instead of an exact distribution.<sup>10</sup>

## Extended Activity

### Wilcoxon Rank Sum Tests

Data set: NLBB Salaries

25. Using a software package, conduct the Wilcoxon rank sum test to determine if the distribution of salaries is different for pitchers than for first basemen.
26. Find  $2 \times P(W \leq 18)$  assuming  $W \sim N(27.5, 4.787)$ . How does your answer compare to that from Question 25?
27. Use a two-sided two-sample  $t$ -test (assume unequal variances) to analyze the data. Are your conclusions the same as in Question 25? Create an individual value plot of the data. Are any distributional assumptions violated? Which test is more appropriate to use for this data set?

At first it may seem somewhat surprising that first basemen tend to make more than pitchers. However, in 2005 there were 19 first basemen and 215 pitchers in the National League. Many pitchers did not play much and got paid a low salary, whereas all 19 first basemen were considered quite valuable to their teams.

## 1.12 Kruskal-Wallis Test for Two or More Independent Samples

The **Kruskal-Wallis test** is another popular nonparametric test that is often used to compare two or more independent samples. Like ANOVA, a more common parametric test that will be discussed in later chapters, the Kruskal-Wallis test requires independent random samples from each population. When the data clearly deviate from the normal distribution, the Kruskal-Wallis test will be more likely than a one-way ANOVA to identify true differences in the population. The null and alternative hypotheses for the Kruskal-Wallis test are

$H_0$ : the distribution of the response variable is the same for all groups

$H_a$ : some responses are systematically higher in some groups than in others

**Table 1.4** Randomly selected catchers from 2005 National League baseball teams.

Team	Position	Name	Salary (\$)
Washington Nationals	Catcher	Bennett, Gary	750,000
Atlanta Braves	Catcher	Perez, Eddie	625,000
Los Angeles Dodgers	Catcher	Phillips, Jason	339,000
Pittsburgh Pirates	Catcher	Ross, David	338,500
Pittsburgh Pirates	Catcher	Santiago, Benito	2,150,000

The Kruskal-Wallis test is also based on ranks. The ranks are summed for each group, and when these group sums are far apart, we have evidence that the groups are different. While the calculations for the Kruskal-Wallis test statistic are provided here, we suggest using statistical software to conduct this significance test. Continuing the baseball salaries example, Table 1.4 displays salaries of five randomly selected catchers from 2005 National League baseball teams.

For the Kruskal-Wallis test, we define the following terms:

- $n_1$  is the sample size for the first group (5 for the pitcher group)
- $n_2$  is the sample size for the second group (5 for the first baseman group)
- $n_3$  is the sample size for the third group (5 for the catcher group)
- $N = n_1 + n_2 + n_3$
- $R_i$  is the sum of the ranks for the  $i$ th group ( $R_1 = 35$ ,  $R_2 = 62$ , and  $R_3 = 23$ )

The Kruskal-Wallis test statistic is calculated as

$$H = \frac{12}{N(N+1)} \sum_i \left( \frac{R_i^2}{n_i} \right) - 3(N+1) = \frac{12}{(15)(16)} \left( \frac{35^2}{5} + \frac{62^2}{5} + \frac{23^2}{5} \right) - 3(16) = 7.98 \quad (1.5)$$

The exact distribution of  $H$  under the null hypothesis depends on each  $n_i$ , so it is complex and time consuming to calculate. Even most statistical software packages use the chi-square approximation with  $I - 1$  degrees of freedom to obtain  $p$ -values (where  $I$  is the number of groups).

**NOTE**

When the chi-square approximation is used, each group should have at least five observations.

**Extended Activity**
**Kruskal-Wallis Test**

Data set: NLBB Salaries

28. Using a software package, run the Kruskal-Wallis test (use all three groups with samples of size 5 per group) to determine if the distribution of salaries differs by position. Create an individual value plot of the data. Do the data look normally distributed in each group?

**MATHEMATICAL NOTE**

If the spread of each group appears to increase as the center (mean or median) increases, transforming the data—such as by taking the log of each response variable—will make the data appear much more normally distributed. Then parametric techniques can often be used on the transformed data. In the baseball salary example, the data are highly right skewed in at least two groups. While a log transformation on salaries is helpful, there is still not enough evidence that the transformed salaries are normally distributed. Thus, nonparametric methods are likely the most appropriate approach to testing whether there is a difference in the distribution of salaries based on position.

Nonparametric tests based on rank are usually less powerful (less likely to reject the null hypothesis) than the corresponding parametric tests. Thus, you are less likely to identify differences between groups when they really exist. If you are reasonably certain that the assumptions for the parametric procedure are satisfied, a parametric procedure should be used instead of a rank-based nonparametric procedure. Many introductory texts suggest that, in order to conduct a parametric test, you should have a sample size of 15 in each group and no skewed data or outliers.

## 1.13 Multiple Comparisons

In introductory texts, statistical inference is often described in terms of drawing one random sample, performing one significance test, and then stating appropriate conclusions—analysis done, case closed. However, there are many situations where inference is not that simple. Performing multiple statistical tests on the same data set can create several problems.

Using a significance level of  $\alpha = 0.05$  (i.e., rejecting  $H_0$  in favor of the alternative when the  $p$ -value is less than or equal to 0.05) helps to ensure that we won't make a wrong decision. In other words, one time out of 20 we expect to incorrectly reject the null hypothesis. But what if we want to do 20 or more tests on the same data set? Does this mean that we're sure to be wrong at least once? And if so, how can we tell which findings are incorrect? The following activities explore how researchers can protect themselves from drawing conclusions from statistical findings that could be the result of random chance.

### Extended Activity

#### Comparing Car Prices

Data set: `Car1`

The `Car1` data set contains prices of used 2005 General Motors cars. All cars in this sample are six-cylinder, four-door sedans in excellent condition with premium sound systems and leather interiors. The data set contains four makes of General Motors cars: Pontiac, Chevrolet, Cadillac, and Buick.

29. Open the `Car1` data set and conduct three two-sided hypothesis tests to determine if there is a difference in price. Compare the means: Pontiac versus Buick (test 1), Cadillac versus Pontiac (test 2), and Cadillac versus Buick (test 3). Provide the  $p$ -value for each of these three tests. Which tests have a  $p$ -value less than 0.05?
30. Assuming the null hypotheses are true, each of the three tests in Question 29 has a 5% chance of inappropriately rejecting the null hypothesis. However, the probability that *at least one* of the three tests will inappropriately reject the null hypothesis is 14.26%. Assuming that the null hypothesis is true and that each test is independent, complete the following steps to convince yourself that this probability (14.26%) is correct.
  - a. Each test will either reject (R) or fail to reject (F). List the rest of the eight possible outcomes in the table below.

Case	Test 1	Test 2	Test 3	Probability
1	F	F	F	
2	F	F	R	
3	F	R	F	
4				
5				
6				
7				
8				

- b. The probability that each hypothesis test rejects is  $P(R) = 0.05$ , and the probability that each hypothesis test fails to reject is  $P(F) = 0.95$ . You may recall that when events are independent, the probabilities can be multiplied. For example, the probability that all three tests fail to reject is  $P(F)P(F)P(F) = 0.95 \times 0.95 \times 0.95 = 0.8574$ . Similarly, the probability that the first two tests fail to reject and the third test does reject is  $0.95 \times 0.95 \times 0.05 = 0.0451$ . Complete the table by calculating the probabilities for all eight cases. Verify that the eight probabilities sum to one.

Notice that case 1 is the only case where no test is rejected. The probability that *at least one* test rejects is the sum of the probabilities for cases 2 through 8; more simply, the probability that at least one test rejects can be calculated as  $1 - 0.8574 = 0.1426$ .

31. Repeat Question 30 using  $\alpha = 0.10$ . Assuming that the null hypothesis is true and that each test is independent, what is the probability that *at least one* of the three tests will inappropriately reject the null hypothesis?
32. To compare all four groups of cars, six hypothesis tests will be needed. List all six null hypotheses. Assuming that the null hypothesis is true and that each test is independent, what is the probability that that *at least one* of the six tests will inappropriately reject the null hypothesis? Use  $\alpha = 0.05$ . With six tests, there are 64 cases. You do not need to list all 64 cases.

## Extended Activity

### The Least-Significant Differences Method and the Bonferroni Method

Data set: Car1

When the significance level is controlled for each individual test, as was done in Question 29, the process is often called the **least-significant differences method (LSD)**. Notice that using  $\alpha = 0.05$  for all tests has some undesirable properties, especially when a large number of tests being conducted. If 100 independent tests were conducted to compare multiple groups (and there really were no differences), the probability of incorrectly rejecting *at least one* test would be  $1 - 0.95^{100} = 0.994$ . Thus, using  $\alpha = 0.05$  as a critical value for 100 comparisons will almost always lead us to *incorrectly* conclude that some results are significantly different.

One technique that is commonly used to address the problem with multiple comparisons is called the **Bonferroni method**. This technique protects against the probability of false rejection by using a cutoff value of  $\alpha/K$ , where  $K$  is the number of comparisons. In Question 29, there are three comparisons (i.e., three hypothesis tests). Thus, a cutoff value of  $0.05/3 = 0.01667$  should be used. In other words, when there are three comparisons as in Question 29, the Bonferroni method rejects the null hypothesis when the *p*-value is less than or equal to 0.01667. Using the least-significant differences method ( $\alpha = 0.05$ ), as was done in Question 29, we would conclude that the prices of Buicks and Chevrolets are significantly different, but using the Bonferroni method we would fail to reject in all three tests.

33. Repeat Question 30 using the Bonferroni cutoff value of  $0.05/3 = 0.016667$  instead of  $\alpha = 0.05$ . Find the probability that *at least one* of the tests rejects.
34. Using all four groups of cars and  $\alpha = 0.05$  (cutoff of  $0.05/6$ ), do any of the six tests reject the null hypothesis with the Bonferroni method?
35. If there were seven groups, 21 hypothesis tests would be needed to compare all possible pairs. Using  $\alpha = 0.05$  and the Bonferroni's method (reject  $H_0$  if the *p*-value is less than  $0.05/21 = 0.00238$ ) what is the probability that at least one of the tests would reject?

#### MATHEMATICAL NOTE

Other terms that are commonly discussed with multiple comparisons are **familywise type I error** and **comparisonwise type I error**. Bonferroni's method is an example of a technique that maintains the familywise type I error. With the familywise type I error 0.05, assuming that there really is no difference between any of the  $K$  pairs, there is only a 5% chance that *any* test will reject  $H_0$ . The least-significant differences method is used to maintain a comparisonwise type I error rate: Assuming that a particular null hypothesis test is true, there is a 5% chance we will (incorrectly) reject that particular hypothesis. Montgomery's *Design and Analysis of Experiments* text provides more information on multiple comparisons.<sup>11</sup>

## Choosing a Critical Value

The  $\alpha$ -level represents the probability of a type I error. A **type I error** can be considered a false alarm: Our hypothesis test has led us to conclude that we have found a significant difference when one does not exist. However, it is important to recognize that it is also possible to make a **type II error**, which means our hypothesis test failed to detect a significant difference when one exists. In essence, a type II error can be thought of as an alarm that failed to go off.

Notice that if the Bonferroni method is used with all six tests, the critical value for each individual test is  $0.05/6 = 0.00833$ . Thus, this method often fails to detect real differences between groups, leaving us open to a high rate of type II error while protecting us against type I errors.

Neither the least-significant differences nor the Bonferroni method is ideal. Caution should be used with both techniques, and neither technique should be used with numerous comparisons. The key is to recognize the benefits and limitations of each technique and to properly interpret what the results of each technique tell us. Some researchers suggest limiting the number of tests, using both techniques, and letting the reader decide which conclusions to draw. Both techniques are commonly used when there are fewer than 10 comparisons. However, a researcher should always decide which comparisons to test *before* looking at the data.

## Chapter Summary

This chapter described the basic concepts behind randomization tests, permutation tests, bootstrap methods, and rank-based nonparametric tests. **Parametric tests** (such as  $z$ -tests,  $t$ -tests or  $F$ -tests) assume that data follow a known probability distribution or use the central limit theorem to make inferences about a population. **Nonparametric tests** do not require assumptions about the distribution of the population or the central limit theorem in order to make inferences about a population.

The **null hypothesis**, denoted  $H_0$ , states that in a study nothing is creating group differences except the random allocation process. The research hypothesis is called the **alternative hypothesis** and is denoted  $H_a$  (or  $H_1$ ). The **p-value** is the likelihood of observing a statistic at least as extreme as the one observed from the sample data when the null hypothesis is true. A threshold value, called a **significance level**, is denoted by the Greek letter alpha ( $\alpha$ ). When a study's  $p$ -value is less than or equal to this significance level, we state that the results are **statistically significant at level  $\alpha$** . **Exact p-values** are often difficult to calculate, but **empirical p-values** can often be simulated through a randomization or permutation test. The empirical  $p$ -value will become more precise as the number of randomizations within a simulation study increases.

The steps in a **randomization test** are as follows:

- An experiment is conducted in which units are assigned to a treatment and an observed sample statistic is calculated (such as the difference between group means).
- Software is used to simulate the random allocation process a number of times ( $N$  iterations).
- For each iteration, the statistic of interest (difference between group means) is recorded, with  $X$  being the number of times the statistic in the iteration exceeds or is the same as the observed statistic in the actual experiment.
- $X/N$  is computed to find the  $p$ -value, the proportion of times the statistic exceeds or is the same as the observed difference.

A **permutation test** is a more general form of the randomization test. The steps in both tests are identical, except that permutation tests do not require random allocation. Randomization tests and permutation tests can provide very accurate results. These tests are preferred over parametric methods when the sample size is small or when there are outliers in a data set. Since real data sets tend not to come from exactly normal populations, it is important to recognize that even  $p$ -values from parametric tests are approximate (but typically accurate as long as the sample sizes are large enough, the data are not skewed, there are no outliers, and the data are reasonably normal). A graph such as a boxplot or individual value plot should always be created to determine if parametric methods are appropriate. Randomization tests are gaining popularity because they require fewer assumptions and are just as powerful as parametric tests.

Bootstrap methods take many (at least 1000) resamples *with replacement* of the original sample to create a bootstrap distribution. If the bootstrap distribution is symmetric and unbiased, bootstrap  $t$  or bootstrap percentile confidence intervals can be used to approximate  $100(1 - \alpha)\%$  confidence intervals.

The steps in creating **bootstrap confidence intervals** are as follows:

- One sample of size  $n$  is taken from a population and the statistic of interest is calculated.
- Software is used to take resamples (with replacement) of size  $n$  from the original sample a number of times ( $N$  iterations). For each iteration, the statistic of interest is calculated from the resample.
- The **bootstrap distribution**, which is the distribution of all  $N$  resample statistics, is used to estimate the shape and spread of the sampling distribution.

- A **bootstrap  $t$  confidence interval** is found by calculating  $\bar{x} \pm t^*(S^*)$ , where  $S^*$  is the standard deviation of the bootstrap distribution and  $t^*$  is the critical value of the  $t(n - 1)$  distribution with  $100(1 - \alpha)\%$  of the area between  $-t^*$  and  $t^*$ .
- A  $100(1 - \alpha)\%$  **bootstrap percentile confidence interval** is found by taking the  $\alpha/2 \times 100$  percentile of each tail of the bootstrap distribution.

Bootstrap confidence intervals based on small samples can be unreliable. The bootstrap  $t$  or percentile confidence interval may be used if

- the bootstrap distribution does not appear to be biased,
- the bootstrap distribution appears to be normal, and
- the bootstrap  $t$  and percentile confidence intervals are similar.

Simulation studies can easily be extended to testing other terms, such as the median or variance, whereas most parametric tests described in introductory statistics classes (such as the  $z$ -test and  $t$ -test) are restricted to testing for the mean. Simulation studies are an extremely useful tool that can fairly easily be used to calculate accurate  $p$ -values for research hypotheses when other tests are not appropriate.

Before computationally intensive techniques were easily available, rank-based nonparametric tests, such as the **Wilcoxon rank sum test** and the **Kruskal-Wallis test**, were commonly used. These tests do not require assumptions about distributions, but they tend to be less informative because ranks are used instead of the actual data. Both the Mann-Whitney test and the Kruskal-Wallis test assume that sample data are from independent random samples whose distributions have the same shape and scale. Each sample in the Kruskal-Wallis test should consist of at least five measurements. Rank-based nonparametric tests tend to be less powerful (less likely to identify differences between groups) than parametric tests (when assumptions do hold) and resampling methods. When the sample sizes are small and there are reasons to doubt the normality assumption, rank-based nonparametric tests are recommended over parametric tests. Randomization tests and permutation tests are typically preferred over parametric and rank-based tests. Their  $p$ -values are often more reliable, and they are more flexible in the choice of parameter tested.

One final note of caution: Even though it is possible to analyze the same data with a variety of parametric and nonparametric techniques, statisticians should never search around for a technique that provides the results they are looking for. Conducting multiple tests on the same data and choosing the test that provides the smallest  $p$ -value will cause the results to be unreliable. If possible, determine the type of analysis that will be conducted before the data are collected.

## Exercises

---

- E.1. Is it important in the schistosomiasis study for all 20 mice to come from the same population of mice? Why or why not?
- E.2. Assume the researchers in this study haphazardly pulled the female mice from a cage and assigned the first five to the treatment and the last five to the control. Would you trust the results of the study as much as if five mice were randomly assigned to each group?
- E.3. A recent study in the northwest United States found that children who watched more television were more likely to be obese than children who watched less television. Can causation be inferred from this study?
- E.4. What is the difference between a random sample and a randomized experiment?
- E.5. Explain the difference between a population model and a randomization model.
- E.6. Explain how the independence assumption of the two-sample  $t$ -test is violated in a randomization model.
- E.7. If the sample size is large, will the histogram of the sample data have a shape similar to that of the normal distribution? Explain.
- E.8. If the sample size is large, will the sample mean be normally distributed? Explain.
- E.9. Why should boxplots or other graphical techniques be used to visualize data before a parametric test is conducted?
- E.10. Suppose that in our study of schistosomiasis in female mice the  $p$ -value was 0.85. Would you be able to conclude that there was no difference between the treatment and control means?

### E.11. Using Other Test Statistics

Data set: Mice

One major advantage of randomization/permuation tests over classical methods is that they easily allow the use of test statistics other than the mean.

- a. Modify the program/macro you created in Question 9 to measure a difference in group medians instead of a difference in means for the female mice. Report the *p*-value and compare your results to those for Question 9.
- b. You might also wonder if there is a difference in the variability in the groups. Modify the macro you created in Question 9 to test whether the variances of the female groups are equal. Report the *p*-value and state your conclusions.

### E.12. Testing Male Mice

Data set: Mice

- a. Using the data for the male mice, run a simulation to decide whether K11777 inhibits schistosome viability (i.e., reduces worm count) in male mice. Describe the results, including a histogram of the simulation results, the *p*-value, and a summary statement indicating your conclusion about the research question of schistosome viability.
- b. Modify the program/macro you created in Part A to measure a difference in group medians instead of a difference in means for the male mice. Report the *p*-value and compare your results to those for Part A.
- c. You might also wonder if there is a difference in the variability in the groups. Modify the macro you created in Part A to test if the variances of each male group are equal. Report the *p*-value and state your conclusions.

### E.13. Bird Nest Study

Data set: Birdnest

This data set was collected in the spring of 1999 for a class project by Amy Moore, a Grinnell College student. Each record in the data set represents data for a species of North American passerine bird. Passerines are “perching birds” and include many families of familiar small birds (e.g., sparrows and warblers) as well as some larger species like crows and ravens, but do not include hawks, owls, water fowl, wading birds, and woodpeckers. Moore took all North American passerines for which complete evolutionary data were available, which comprised 99 of the 470 species of passerines in North America (part of her study used this evolutionary information). One hypothesis of interest was about the relationship of body size to type of nest. Body size was measured as average length of the species, nest type was categorized as either closed or open. Although nests come in a variety of types (see the Nesttype variable), in this data set “closed” refers to nests with only a small opening to the outside, such as the tree-cavity nest of many nuthatches or the pendant-style nest of an oriole. “Open” nests include the cup-shaped nest of the American robin.

- a. Moore suspected that closed nests tend to be built by larger birds, but here we will treat the alternative as two-sided, since her suspicion was based on scanty evidence. Use comparative dotplots or boxplots and summary statistics to describe the relationship between average body length and nest type (the Closed variable). (Note: Closed = 1 for closed nests; Closed = 0 for open nests.) Does it appear that Moore’s initial suspicion is borne out by the data?
- b. Run a permutation test using a two-sided alternative to determine if type of nest varies by body length and interpret your results. Be sure to state your conclusions in the context of the problem and address how random allocation and random sampling (or lack of either) impact your conclusions.

### E.14. Twins Brain Study

Data set: Twins

In a 1990 study by Suddath et al., reported in Ramsey and Schafer,<sup>12</sup> researchers used magnetic resonance imaging to measure the volume of various regions of the brain for a sample of 15 monozygotic twins, where one twin was affected with schizophrenia and the other was unaffected. The twins were from North America and comprised eight male pairs, and seven female pairs ranging in age from 25 to 44 at the time of the study. The sizes in volume ( $\text{cm}^3$ ) of the hippocampus are in the file called Twins.

- a. Should the data be analyzed as match pairs or be treated as if there were two independent samples?
- b. Use appropriate graphics and summary statistics to describe the difference in brain volume for affected and unaffected twins.
- c. Use the appropriate permutation test to ascertain if the difference in brain volume described in Part B is the result of schizophrenia or if it could be explained as a chance difference. Report your  $p$ -value and summarize your conclusion.

#### E.15. Comparing Parametric and Nonparametric Tests

Data set: Birdnest and Music

- a. Using a  $t$ -test, compute the two-sided  $p$ -value for the bird nest study in Exercise E.13. and compare the results to what you found with the randomization test.
- b. Using a  $t$ -test, compute the one-sided  $p$ -value for the music study in Question 20 and compare the results to what you found with the randomization test.

#### E.16. Means versus Medians in Rank-Based Tests

Data set: SameMean

Rank-based nonparametric tests do not answer the same question as the corresponding parametric procedure. Many people assume that these nonparametric tests are testing for group medians. This is not always true. Rank-based tests can be interpreted as testing for the median only if the shapes and scales of the populations are the same. The following exercise illustrates this point by providing an example where the medians and the means are identical but nonparametric tests will reject the null hypothesis.

Use the SameMean data to conduct the Kruskal-Wallis test. Calculate the mean and median for each group. What conclusions can you draw from the data?

#### E.17. Rank Based Bird Nest Tests

Data set: Birdnest

- a. Use the Wilcoxon rank sum test to conduct a significance test for the bird nest study discussed in Exercise E.13.
- b. Use the Kruskal-Wallis test to conduct a significance test for the bird nest study. Determine whether the distribution of bird size (response is Length) is the same for each nest type. Note that when the chi-square approximation is used, each group should have at least five observations. You may need to create an “other” group to combine all nest types with sample sizes less than five.

#### E.18. Bootstrap Confidence Intervals

Data set: ChiSq

Take a simple random sample of size 40 from the ChiSq data file.

- a. Create a bootstrap distribution of the mean (or use the distribution you created in Question 22). Calculate a 95% bootstrap  $t$  confidence interval for the mean.
- b. Create a bootstrap distribution of the mean (or use the distribution you created in Question 22). Calculate a 95% bootstrap percentile confidence interval for the mean. Are the bootstrap  $t$  and percentile confidence intervals for the mean reliable?
- c. Create a bootstrap distribution of the standard deviation (or use the distribution you created in Question 23). Calculate a 95% bootstrap  $t$  confidence interval for the standard deviation.
- d. Create a bootstrap distribution of the standard deviation (or use the distribution you created in Question 23). Calculate a 95% bootstrap percentile confidence interval for the standard deviation. Are the bootstrap  $t$  and percentile confidence intervals for the standard deviation reliable?

#### E.19. Medians and Trimmed Means in Bootstrap Confidence Intervals

Data set: ChiSq

- a. Take a simple random sample of size  $n = 40$  from the ChiSq data. Create a bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the bootstrap distribution and explain why bootstrap confidence intervals are unlikely to be reliable.

- b. Take a second simple random sample of size  $n = 40$  from the `ChiSq` data. Create a second bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the second bootstrap distribution. With a sample size of 40, why are bootstrap distributions of medians unlikely to be normal?
- c. Bootstrap distributions for medians are unlikely to be normally distributed, and means tend to be influenced by outliers. The trimmed mean is a common measure of center that tends to better represent the average value with bootstrap methods. **Trimmed means** are calculated by first trimming the upper and lower values of the sample. For example, the 25% trimmed mean is the mean of the middle 50% of the sample data.

Take a simple random sample of size  $n = 40$  from the `ChiSq` data. Create a bootstrap distribution of the 25% trimmed mean by taking 1000 resamples (with replacement). In other words, for each resample calculate the mean of the middle 20 observations (remove the smallest 10 and largest 10 values in each resample). Create a histogram of the 1000 trimmed means and describe the shape of this bootstrap distribution. Create a bootstrap  $t$  confidence interval and a bootstrap percentile confidence interval to estimate the 25% trimmed mean.

#### E.20. Medians and Trimmed Means in Bootstrap Confidence Intervals

Data set: `MedSalaries`

- a. The file `MedSalaries` is a random sample of salaries of medical doctors who were teaching at United States universities in 2009. Create a bootstrap distribution of the median by taking 1000 resamples (with replacement). Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap  $t$  confidence interval or a bootstrap percentile confidence interval for the median?
- b. Create a bootstrap distribution of the 25% trimmed mean by taking 1000 resamples (with replacement). In other words, calculate the mean of the middle 50 observations from each resample. Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap  $t$  confidence interval or a bootstrap percentile confidence interval for the 25% trimmed mean?
- c. Create a bootstrap distribution of the 5% trimmed mean by taking 1000 resamples (with replacement). In other words, calculate the mean of the middle 90 observations from each resample. Describe the shape of the bootstrap distribution. Is it appropriate to create a bootstrap  $t$  confidence interval or a bootstrap percentile confidence interval for the 5% trimmed mean?
- d. Calculate a bootstrap  $t$  confidence interval and bootstrap percentile confidence interval for each of the preceding parts of this exercise if the bootstrap distribution indicates that it is appropriate.

#### E.21. Multiple Comparisons

Data set: `NLBB Salaries`

- a. Conduct a permutation test to determine if there is a difference in mean salaries between pitchers and first basemen. Report the  $p$ -value and your conclusions based on an individual  $\alpha$ -level of 0.05.
- b. Conduct a permutation test to determine if there is a difference in mean salaries between pitchers and catchers. Report the  $p$ -value and your conclusions based on an individual  $\alpha$ -level of 0.05.
- c. Conduct a permutation test to determine if there is a difference in mean salaries between first basemen and catchers. Report the  $p$ -value and your conclusions based on an individual  $\alpha$ -level of 0.05.
- d. If each of the previous three tests uses an  $\alpha$ -level of 0.05, what is the true probability that at least one of the tests will inappropriately reject the null hypothesis?
- e. What is the individual critical value if you use the Bonferroni method with an overall (familywise)  $\alpha$ -level of 0.05? Do any of your previous conclusions in the preceding parts of this exercise change if you test for an overall (familywise) comparison? Explain.

## Endnotes

---

1. Douglas Montgomery, *Design and Analysis of Experiments* (New York: Wiley, 2005), p. 21.
2. Maha-Hamadien Abdulla, Kee-Chong Lim, Mohammed Sajid, James H. McKerrow, and Conor R. Caffrey, "Schistosomiasis Mansoni: Novel Chemotherapy Using a Cysteine Protease Inhibitor," *PLoS Medicine* 4.1 (Jan. 2007). Online: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=17214506>. Note: Professor Conor R. Caffrey provided the raw data upon request.
3. This data set is from the text by Ann E. Watkins, Richard L. Scheaffer, and George W. Cobb, *Statistics in Action* (Emeryville, CA: Key Curriculum Press, 2004).
4. Bradley Efron, The Jackknife, the Bootstrap, and Other Resampling Plans, *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38 (1982).
5. Bradley Efron, "Nonparametric Standard Errors and Confidence Intervals, *Canadian Journal of Statistics*, 36 (1981).
6. R. A. Fisher, *Design of Experiments* (Edinburgh, UK: Oliver and Boyd, 1935).
7. Bradley Efron and Robert Tibshirani, *An Introduction to the Bootstrap* (New York: CRC Press, 1993), p. 202.
8. Michael D. Ernst, "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, 19.4 (2004), 676–685.
9. E. L. Lehmann, *Nonparametrics Statistical Methods Based on Ranks* (Upper Saddle River, NJ: Prentice Hall, 1998).
10. C. Bellera, M. Julien, and J. Hanley, "Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What  $N$ ? Graphical Insights," *Journal of Statistics Education*, 18.2 (2010).
11. Douglas Montgomery, *Design and Analysis of Experiments* (New York: Wiley, 2005).
12. Fred L. Ramsey and Daniel W. Schafer, *The Statistical Sleuth* (Pacific Grove, CA: Duxbury, 2002).
13. S. Petrellese, "Adelphi Settlement in Sex Discrimination Case a 'Step in Right Direction,'" *Garden City News*, March 29, 2009. Retrieved on Jan. 31, 2010, from [http://www.gcnews.com/news/2009-03-27/Front\\_page/004.html](http://www.gcnews.com/news/2009-03-27/Front_page/004.html).
14. L. W. Perna, "Sex and Race Differences in Faculty Rank and Tenure," *Research in Higher Education*, 42 (2001), 541–567; L. W. Perna, "Sex Differences in Faculty Salaries: A Cohort Analysis," *Review of Higher Education*, 24 (2001), 283–307.

# Research Project: Gender Discrimination Among University Faculty

In 2009, Adelphi University agreed to pay \$305,889 to 37 claimants in order to settle a pay discrimination lawsuit filed by the U.S. Equal Employment Opportunity Commission.

According to the EEOC's lawsuit, a class of female full-time professors was paid less than male professors of the same or lesser rank teaching within the same school. This violation had been ongoing since at least April 2004, the EEOC said. The lawsuit was filed in 2007 on behalf of Judith H. Cohen, Ph.D., an Adelphi education professor and attorney who still teaches at the university.<sup>13</sup>

Numerous studies have been conducted on wage disparities between men and women. Laura Perna cites several studies showing that even when factors such as rank, age, credentials, and field of study are controlled, full-time female faculty members earn less than their male counterparts.<sup>14</sup>

For this project, you will be asked to analyze a selection of 2009 faculty salaries. You should assume that your team has been hired by university administrators to determine if there is any evidence of gender discrimination at their university. The data set, *Faculty*, provides the 2009 salaries, ranks, and years since obtaining their Ph.D. for all faculty members in the statistics and English departments.

This project should be completed in teams of two or three. You are not allowed to discuss the results of your analysis with anyone except your teammates and your professor. Submit no more than a three-page summary of your analysis, including all tables, graphs, and results of the nonparametric analysis. Assume that your clients, the university administrators, have no more than an introductory statistics background and that they do not understand randomization/permuation tests. Thus, careful and concise explanations are needed. The primary question for this study is whether there is evidence of gender discrimination in faculty salaries. You should include the following items in your report:

1. Sources of bias that could exist in the data
2. An explanation of the nonparametric technique used, as well as an explanation of why this technique was selected
3. Appropriate plots of the data. Do not include too many graphs. Carefully select which graphs and statistics best describe the patterns in the data. All figures and tables should be well labeled and referenced within the text.
4. A description of any patterns in the plots. Are there any outliers? Can you give an explanation for these outliers?
5. An explanation of any assumptions you made in your analysis
6. A clear conclusion, stated within the context of the study, that addresses random allocation and random sampling.

Note: It is not necessary to conduct a randomization test that addresses all the potential biases or all the variables within the data set. However, it would likely be helpful to consider at least one explanatory variable in addition to gender. You might also consider creating fewer levels (i.e., grouping data) within potential explanatory variables.

## Other Project Ideas

The salaries of all government employees, including university faculty members, are required by law to be available to the public. A simple Google search (including words such as *state, employee, salary, year*) will allow you to find the salaries of many government employees by state. Submit an Excel file and codebook for your new data set and answer the following questions for the institution of your choice.

1. Who are the five highest paid employees at that institution?
2. How do administrative positions compare to faculty positions with respect to salary?

3. How does the type of position influence salary?
4. Compare the faculty from two departments, one that tends to be dominated by females and another that tends to be dominated by males. Identify (and possibly explain) any outliers you find in the data.
5. Are female-dominated fields characterized by lower wages than male-dominated fields?
6. Did you find evidence of gender discrimination in the institution you chose? Use a nonparametric technique to justify your answer.
7. Explain your data collection process. Did you have difficulty obtaining the data?

# Making Connections: The Two-Sample *t*-Test, Regression, and ANOVA

*In theory, there's no difference between theory and practice.  
In practice, there is.*

— Yogi Berra<sup>1</sup>

Statistics courses often teach the two-sample *t*-test, linear regression, and analysis of variance (ANOVA) as very distinct approaches to analyzing different types of data. However, this chapter makes connections among these three techniques by focusing on the statistical models. Statistical software has made it easy to calculate statistics and *p*-values. But without understanding the underlying model assumptions, it is easy to draw incorrect conclusions from the sample data. As studies become more complex, models become fundamental to drawing appropriate conclusions. In this chapter, a simple student experiment involving games and several additional studies are used to do the following:

- Compare the underlying statistical models for the two-sample *t*-test, linear regression, and ANOVA
- Discuss the model assumptions for each of these three tests
- Create and interpret normal probability plots
- Transform data in order to better fit the model assumptions
- Discuss the mathematical details of each hypothesis test and corresponding confidence interval

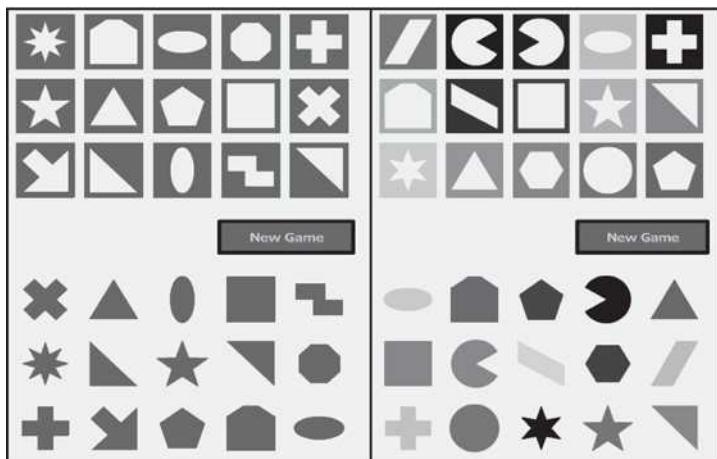
## 2.1 Investigation: Do Distracting Colors Influence the Time to Complete a Game?

In 1935, John Stroop published a paper presenting his research on the reaction time of undergraduate students identifying ink colors.<sup>2</sup> He found that students took a longer time identifying ink colors when the ink was used to spell a different color. For example, if the word “yellow” was printed in green ink, students took longer to identify the green ink because they automatically read the word “yellow.” Even though students were told only to identify the ink color, the automatized behavior of reading interfered with the task and slowed their reaction time. \* *Automatized behaviors* are behaviors that can be done automatically without carefully thinking through each step in the process. Stroop’s work, demonstrating that automatized behaviors can act as a distracter for other desired behaviors, is so well known that the effect is often called the *Stroop effect*.

Several students in an introductory statistics class wanted to develop a final project that would test the impact of distracters. They decided to conduct a study to determine if students at their college would perform differently when a distracting color was incorporated into a computerized game. This game challenges people to place an assortment of shaped pegs into the appropriate spaces as quickly as possible.

Before any data were collected, these students developed a clear set of procedures.

- 40 students would be randomly selected from the college.<sup>†</sup>
- 20 students would be assigned to the standard game and 20 would be assigned to a game with a color distracter. The student researchers would flip a coin to randomly assign subjects to a treatment. Once 20 subjects had been assigned to either group, the rest would automatically be assigned to play the other game.
- Subjects would see a picture of the game and have the rules clearly explained to them before they played the game. An example of both games is shown in Figure 2.1.
- Subjects would play the game in the same area with similar background noise to control for other possible distractions.
- The response variable would be the time in seconds from when the participant pressed the “start game” button to when he or she won the game.



**Figure 2.1** A black and white image of the electronic Shapeslosion game with and without color distracters. The instructions for the game were to click and drag each peg to the space with the matching shape.

\*Note that many psychologists would call this procedural knowledge instead of automatized behavior. Both are processes that can be done without conscious thought, but automatized behaviors are processes that cannot be slowed down, do not decline with age, and show no gender differences.

<sup>†</sup>Since it was not possible to force college students to be involved in this study, these researchers randomly selected students from an online college directory until they had 40 students who were willing to play the game.

**NOTE**

It is important to recognize that each subject in this study was assigned to exactly one treatment, either the standard game or the color distracter game. Some researchers may point out that a paired design (where each subject was assigned to both treatments) might have been more efficient. However, for the purposes of this chapter, this study will be treated as the students originally designed it: a study comparing two independent samples.

## Activity Understanding the Study Design

1. For this study, identify the units, the population for which conclusions can be drawn, the explanatory variable, and the response variable.
2. Is this study an experiment or an observational study? Explain.
3. The researchers hoped to determine if distracting colors influenced college students' response times when playing a computerized game. Write out in words and symbols appropriate null and alternative hypotheses. Let  $\mu_1$  represent the true mean response time of the color group and  $\mu_2$  the true mean response time of the standard group. Use a two-sided alternative hypothesis for this question.
4. Create an individual value plot or a boxplot of the `Games1` data from this study. Describe the graph. For example, does it look as if the groups have equal means or equal standard deviations? Are there any unusual observations in the data set? Calculate the mean and standard deviation of the color distracter responses,  $\bar{y}_1$  and  $s_1$ , as well as the mean and standard deviation of the standard game responses,  $\bar{y}_2$  and  $s_2$ .

Before they began the study, the student researchers compared it to other studies they had seen in the past. One student in the class was an economics major. He had seen hypothesis tests used in regression to analyze whether a certain variable belonged in a regression model. He suggested using regression to test if type of game (color distracter or standard) was important in modeling the time to play the game. Another student, a biology major, had previously completed an experiment in ecology in which her professor had told her to use an *F*-test and analysis of variance (ANOVA) to determine if a certain variable significantly influenced the results. A third student had seen the two-sample *t*-test explained in her statistics textbook and believed she should do what the textbook example did.

Each of these approaches seems fairly reasonable, so how can you decide which technique to use? Determining which technique to use depends on developing an appropriate statistical model. Instead of simply stating which student is right, this chapter will analyze the study with all three techniques. Throughout each analysis, the chapter will emphasize the statistical modeling process in order to determine which technique is appropriate.

## 2.2 The Two-Sample *t*-Test to Compare Population Means The Statistical Model

Generally, **statistical models** have the following form:

$$\text{observed value} = \text{mean response} + \text{random error}$$

The statistical model describes each observed value in a data set as the sum of a mean response for some subgroup of interest (often called a group mean) and a random error term. The mean response is fixed for each group, while the random error term is used to model the uncertainty of each individual outcome. The random error term for each individual outcome cannot be predicted, but in the long run there is a regular pattern that can be modeled with a distribution (such as the normal distribution).

The key question in this study is whether or not the two types of games have different average completion times. The two-sample *t*-test starts with the assumption that the two group means are equal. This is often written as the null hypothesis  $H_0: \mu_1 - \mu_2 = 0$  or, equivalently,  $H_0: \mu_1 = \mu_2$ .

The underlying model used in the two-sample *t*-test is designed to account for these two group means ( $\mu_1$  and  $\mu_2$ ) and random error. The statistical model for the first population, the color distracter group, is

$$\begin{array}{ccc} \text{observed} & \text{mean} & \text{error} \\ \text{value} & \text{response} & \text{term} \\ (\text{random}) & (\text{not random}) & (\text{random}) \\ \downarrow & \downarrow & \downarrow \\ y_{1,j} & = & \mu_1 + \varepsilon_{1,j} \quad \text{for } j = 1, 2, \dots, n_1 \end{array}$$

where  $j$  is used to represent each observation in the sample from the first population. For example,  $y_{1,9}$  represents the 9th observation in the first group (the color distracter group). In this data set, there were 20 observations taken from the first population; thus,  $n_1 = 20$ .

This model states that the color distracter game is expected to be centered at the constant value  $\mu_1$ . In addition, each observation is expected to have some variability (random error) that is typically modeled by a normal distribution with a mean equal to zero and a fixed variance  $\sigma^2$ .

Similarly, each observation from the second group, the standard game, can be modeled as the sum of  $\mu_2$  plus a random error term,  $\varepsilon_{2,j}$ :

$$y_{2,j} = \mu_2 + \varepsilon_{2,j} \quad \text{for } j = 1, 2, \dots, n_2$$

where  $n_2 = 20$ ,  $\mu_2$  is the mean of the standard group, and the  $\varepsilon_{2,j}$  are random variables (typically from a normal distribution) with a mean equal to zero and variance  $\sigma^2$ . Often, this statistical model is more succinctly written as

$$y_{i,j} = \mu_i + \varepsilon_{i,j} \quad \text{for } i = 1, 2 \quad \text{and } j = 1, 2, \dots, n_i \text{ where } \varepsilon_{i,j} \sim N(0, \sigma^2) \quad (2.1)$$

#### ► MATHEMATICAL NOTE ▼

You may recall from your introductory statistics course that adding a constant to each random variable in a population does not change the shape or spread of the population. Since each mean response ( $\mu_i$ ) is fixed (i.e., a constant value), Equation (2.1) can be used to show that  $y_{i,j} \sim N(\mu_i, \sigma^2)$ .

This model has one assumption that you may not have made when previously conducting a two-sample *t*-test. Equation (2.1) states that all  $\varepsilon_{i,j}$  come from a normally distributed population with a mean of zero and variance  $\sigma^2$ . This is called the **equal variance assumption**. Some introductory statistics courses discuss only a two-sample *t*-test that does not require the equal variance assumption. The equal variance assumption is made here because it makes sense for this experiment, the data support it ( $s_1$  is close to  $s_2$ ), and it allows a direct comparison to ANOVA and regression models.

In Equation (2.1), the mean response of the model is the population mean ( $\mu_1$  or  $\mu_2$ ). Just as a sample mean,  $\bar{y}_i$ , is used to estimate the population means,  $\mu_i$ , residuals are used to estimate the random error terms. **Residuals** are the difference between the observed response and the estimated mean response. For example, the random error term  $\varepsilon_{1,12} = y_{1,12} - \mu_1$  is estimated by  $\hat{\varepsilon}_{1,12} = y_{1,12} - \bar{y}_1$ .

#### ► NOTE ▼

A **statistic** is any mathematical function of the sample data. **Parameters** are actual population values that cannot be known unless the entire population is sampled. The mean response is based on population parameters. If a sample data set is used, we do not know the population parameters. Sample statistics (such as the sample mean,  $\bar{y}$ , and the sample standard deviation,  $s$ ) are used to estimate population parameters ( $\mu$  and  $\sigma$ ). Statisticians often use a hat on top of a parameter to represent an estimate of that parameter. For example, an estimate of the population standard deviation is written  $s = \hat{\sigma}$ , and an estimate for a mean is written  $\bar{y}_1 = \hat{\mu}_1$  or  $\bar{y}_2 = \hat{\mu}_2$ .

## Activity Statistical Models for the Two-Sample *t*-Test

5. Assume that we have two very small populations that can be written as  $y_{1,1} = 15, y_{1,2} = 17, y_{1,3} = 16, y_{2,1} = 11, y_{2,2} = 9, y_{2,3} = 10$ . Find  $\mu_1, \mu_2, \varepsilon_{1,1}, \varepsilon_{1,3}$ , and  $\varepsilon_{2,1}$ . Notice the double subscripts on the observed responses:  $y_{1,1}$  is read as “y one one.” The first subscript tells us that the observation was from the first group, and the second subscript tells us the observation number. For example,  $y_{1,j}$  is the  $j$ th observation from the first group.
6. Use the game study and the data in the file Games1 to identify  $n_1, n_2, y_{1,12}, y_{2,12}, \hat{\varepsilon}_{1,12}$ , and  $\hat{\varepsilon}_{2,12}$ , where  $y_{1,12}$  represents the 12th observation from group 1 (the color distracter group). Note that since this is a sample, not a population, we do not know  $\mu_1$  or  $\mu_2$ , but we can estimate them with  $\bar{y}_1 = \hat{\mu}_1$  and  $\bar{y}_2 = \hat{\mu}_2$ .

### Model Assumptions for the Two-Sample *t*-Test

Several implicit assumptions are built into the model for the two-sample *t*-test shown in Equation (2.1):

- Constant parameters: The population values in this model ( $\mu_1, \mu_2$ , and  $\sigma$ ) do not change throughout the study.
- **Additive terms:** The model described in Equation (2.1) shows that the observed responses are the sum of our parameters and error terms. For example, we are not considering models such as  $y_{i,j} = \mu_i \times \varepsilon_{i,j}$ .
- $\varepsilon_{i,j} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ . This assumption has many key components:
  - The error terms are independent and identically distributed (iid).
  - The error terms follow a normal probability distribution.
  - The error terms have a mean of zero. This implies that the average of several observed values will tend to be close to the true mean. In essence, there is no systematic bias in the error terms.
  - The population variance  $\sigma^2$  is the same for both groups (color distracter and standard games) being tested.

The first assumption tells us about the mean response. The parameter estimate ( $\bar{y}_i$ ) would not be meaningful if the true parameter value ( $\mu_i$ ) were not constant throughout the study. The second assumption simply states the types of models we are building. In later chapters with more complex models, we will discuss how to use residual plots to determine if the model is appropriate. In this chapter, we will focus on the assumptions about the error terms.

#### MATHEMATICAL NOTE

In later chapters, we will show that a curved pattern in a residual versus fit plot suggests that an additive model may not be appropriate. In this example, there are only two fitted values (i.e., expected values), so we cannot see any curved patterns. When the additive assumption is violated, residual plots may also indicate different standard deviations, a nonnormal distribution, or lack of independence. Transforming the data to a new scale can often make the additivity assumption (and several of the other assumptions) more appropriate.

The statistical model described in Equation (2.1) assumes that  $\varepsilon_{i,j}$  are modeled as **independent and identically distributed** (iid) random variables. The independent error term assumption states that there is no relationship between one observation and the next. For example, knowing that the 8th subject in a group played the game more quickly than average does not provide any information about whether the 7th or 9th person in the group will be above or below the average.

The identically distributed assumption states that each error is assumed to come from the same population distribution. Thus, each subject from a particular group is from the same population. If any error term based on a particular observation comes from a different population, the two-sample *t*-test will not be valid. For example, elementary school students may have different expected completion times for the Shapeslosion game than college students. It would be inappropriate to include younger students in a study where the population was assumed to be college students.

Model assumptions for the residuals should always be checked with plots of the data. The extended activities will describe normality tests in more detail, but in most situations a simple graph of the residuals will suffice. The two sample *t*-test actually requires only that the sample means (each  $\bar{y}_{i,j}$ ) be normally distributed. The central limit theorem allows us to assume this is true if group sample sizes are similar and large ( $n_1 \geq 15$  and  $n_2 \geq 15$ ) and there does not appear to be any extreme skewness or outliers in the residuals.

Since residuals are defined as the difference between each observed value and the corresponding group mean, they should always sum to zero. Thus, we cannot check residuals to determine whether each of the error terms is centered at zero. The assumption that the error terms are centered at zero is really stating that there are no other sources of variability that may be biasing our results. In essence, the only difference between the two population means is explained by the mean response.

To check the assumption that the two populations have the same variance, an informal test can be used. If the ratio of the sample standard deviations is less than 2, we can proceed with the analysis.\*

### Informal Test for Equal Variances

$$\text{If } \frac{\text{maximum}(s_1, s_2)}{\text{minimum}(s_1, s_2)} < 2 \quad \text{or, equivalently, if } \frac{\text{maximum}(s_1^2, s_2^2)}{\text{minimum}(s_1^2, s_2^2)} < 4$$

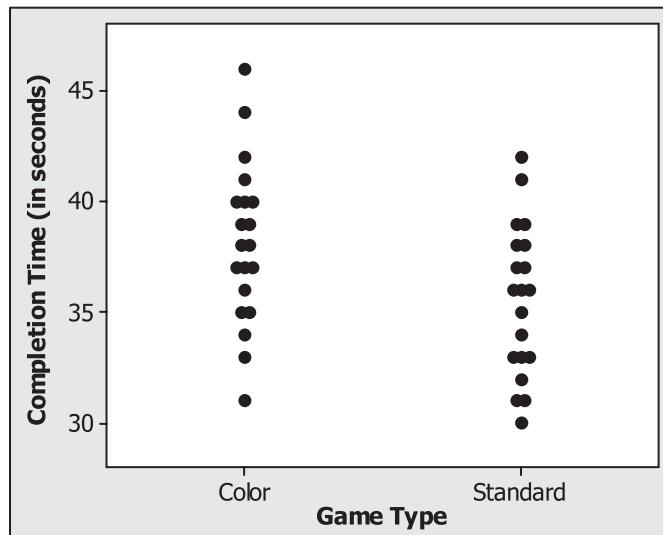
then we do not have enough evidence to conclude that the population variances are different.

Several key observations should be made about the individual value plot shown in Figure 2.2:

- The mean completion time is higher for the color distracter group than for the standard group.
- Neither group appears to have clear outliers, skewness, or large gaps.
- The spread (variance) of the two groups appears to be similar.

#### Key Concept

Every statistical hypothesis test has basic underlying conditions that need to be checked before any valid conclusions can be drawn.



**Figure 2.2** Individual value plot of the data from the color distracter and standard games.

\*Some texts suggest rejecting the equal variance assumption when the ratio is greater than 3 instead of 2. If the ratio is close to 2 (or 3), many statisticians would suggest conducting a more formal *F*-test for equal variances. This *F*-test is described in the review of introductory statistics on the CD.

## Activity ▶ Checking Assumptions for the *t*-Test

7. Calculate the residuals in the Games1 data. Plot a histogram of the residuals (or create a normal probability plot of the residuals). Do the residuals appear to be somewhat normally distributed?
8. Use the informal test to determine if the equal variance assumption is appropriate for this study.
9. The variable `StudentID` represents the order in which the games were played. Plot the residuals versus the order of the data to determine if any patterns exist that may indicate that the observations are not independent.
10. Use statistical software to conduct a two-sample *t*-test (assuming equal variances) and find the *p*-value corresponding to this statistic. In addition, use software to calculate a 95% confidence interval for the difference between the two means ( $\mu_1 - \mu_2$ ). Equation (2.7) and the extended activities provide details on conducting these calculations by hand. If  $H_0: \mu_1 = \mu_2$  is true, the ***p*-value** states how likely it is that random chance alone would create a difference between two sample means ( $\bar{y}_1 - \bar{y}_2$ ) at least as large as the one observed. Based on the *p*-value, what can you conclude about these two types of games?

## 2.3 The Regression Model to Compare Population Means The Linear Regression Model

The simple linear regression model discussed in introductory statistics courses typically has the following form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n \quad \text{where } \varepsilon_i \sim N(0, \sigma^2) \quad (2.2)$$

A **simple linear regression model** is a straight-line regression model with a single explanatory variable and a single response variable. For this linear regression model, the mean response ( $\beta_0 + \beta_1 x_i$ ) is a function of two parameters,  $\beta_0$  and  $\beta_1$ , and an explanatory variable,  $x$ . The random error terms,  $\varepsilon_i$ , are assumed to be independent and to follow a normal distribution with mean zero and variance  $\sigma^2$ .

In Equation (2.1), we used double subscripts:  $i = 1, 2$  was used to show that there were two distinct groups and  $j = 1, 2, \dots, n_i$  was used to identify each of the  $n_1 = n_2 = 20$  items within the two groups. In the regression model, there is only one set of subscripts:  $i = 1, 2, \dots, n$ , where  $n = 40 = n_1 + n_2$ . Instead of having two distinct means in the model ( $\mu_1$  and  $\mu_2$ ), as in the two-sample *t*-test, we have one regression model where the parameters,  $\beta_0$  and  $\beta_1$ , are fixed. The categorical explanatory variable,  $x$ , indicates game type.

A procedure commonly used to incorporate categorical explanatory variables, such as the game type, into a regression model is to define **indicator variables**, also called **dummy variables**, that will take on the role of the  $x$  variable in the model. Creating dummy variables is a process of mapping the column of categorical data into 0 and 1 data. For example, the indicator variable will have the value 1 for every observation from the color distracter game and 0 for every observation from the standard game. Most statistical software packages have a command for creating dummy variables automatically.

### NOTE

Typically an indicator variable is created for each category. Thus, there would be an indicator variable called `Color` equal to 1 for the color distracter game and 0 otherwise and another indicator variable called `Standard` equal to 1 for the standard game and 0 for all other categories. Notice that there is complete redundancy between the two indicator variables: Knowing the value of the `Color` variable automatically tells us the value of the `Standard` variable for each subject. Thus, only one of the indicator variables is needed in this model. Although this study has only two categories of games (color and standard), it is common for a categorical explanatory variable to have more than two categories. Chapter 3 provides the opportunity to use indicator variables when there are multiple categories.

**Key Concept**

Indicator variables can be created to incorporate categorical explanatory variables into a regression model.

## Activity Calculating a Regression Model and Hypothesis Test for the Slope

11. Use the software instructions and the `Games1` data to create indicator variables where  $x = 1$  represents the color distracter game and  $x = 0$  represents the standard game. Develop a regression model using `Time` as the response and the indicator variable as the explanatory variable.
12. Use statistical software to calculate the  $t$ -statistic and  $p$ -value for the hypothesis tests  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$ . In addition, construct a 95% confidence interval for  $\beta_1$ . Based on these statistics, can you conclude that the coefficient,  $\beta_1$ , is significantly different from zero? Details for calculating these statistics by hand are provided in the extended activities.
13. Repeat the two previous questions, but use an indicator variable where  $x = 1$  represents the standard game and  $x = 0$  represents the color distracter game. Compare the regression line, hypothesis test, and  $p$ -value to those from the previous questions. When there are only two categories (color distracter and standard), does the choice of indicator variable impact your conclusions? Why or why not?

In the previous questions, we assigned  $x$  to be the dummy variable that indicates the type of game. Notice that the mean response is still a constant (nonrandom) value for each of the two game categories. In other words, when  $x = 1$  the mean response is a fixed value, and when  $x = 0$  the mean response is a fixed value. In addition, the “slope” coefficient ( $\beta_1$ ) can be considered as an estimate of the average amount by which the response variable will change from the standard game ( $x = 0$ ) to the color distracter game ( $x = 1$ ).

Although the notation has changed, the regression model and the model used in the two-sample  $t$ -test are mathematically equivalent. When a subject is from the color distracter group, the mean response is  $\mu_1$  in the  $t$ -test and the mean response sets  $x = 1$  in the regression model. Thus,

$$\mu_1 = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 \quad (2.3)$$

When a subject is from the standard group, the mean response is  $\mu_2$  in the  $t$ -test and the mean response sets  $x = 0$  in regression. Thus,

$$\mu_2 = \beta_0 + \beta_1(0) = \beta_0 \quad (2.4)$$

Equations (2.3) and (2.4) can be combined to show the relationship between the two-sample  $t$ -test and regression hypotheses.

$$\mu_1 - \mu_2 = (\beta_0 + \beta_1) - \beta_0 = \beta_1 \quad (2.5)$$

Thus, stating that  $\mu_1 - \mu_2 = 0$  is equivalent to stating that  $\beta_1 = 0$ .

**Key Concept**

In testing the difference in two population means, testing the null hypothesis  $H_0: \beta_1 = 0$  for a regression model is equivalent to testing the two-sample  $t$ -test hypothesis  $H_0: \mu_1 - \mu_2 = 0$  when using the equal variance assumption.

## Model Assumptions for Regression

While no distributional assumptions are needed to create estimates of  $\beta_0$  and  $\beta_1$ , it is necessary to check the same model assumptions when conducting a hypothesis test for  $\beta_1$ . Just as in the two-sample  $t$ -test, the model assumes that the parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$  are constant. In addition, Equation (2.2) shows that our model

consists of the mean response *plus* the error term. The regression model also assumes that  $\varepsilon_{i,j} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ . This expression represents the following four assumptions:

- The error terms are independent and identically distributed (iid).
- The error terms follow a normal probability distribution.
- The error terms have a mean of zero.
- The error terms in the regression model are assumed to come from a single population with variance  $\sigma^2$  (i.e., the variance does not depend on  $x$ ).

In regression, assumptions about the error terms are also checked by residual plots. Here,  $y_i$  represents each observed response and  $\hat{y}_i = b_0 + b_1x_i$  represents the estimated mean response. So the residuals are simply the observed value minus the estimated value:  $\hat{\varepsilon}_i = y_i - \hat{y}_i$ .

Figure 2.3 shows a histogram of the residuals and a plot of the residuals by type of game. The histogram shows that the residuals approximately follow the shape of a normal distribution. The residual versus game type graph shows that there are no obvious outliers and that the spread of both groups is roughly equivalent. Since residuals are just the mean response subtracted from the observed value, the center of the residual plots has shifted to zero. However, the spread of the residual versus game plot is identical to the spread of the individual value plot in Figure 2.2.

### Key Concept

No assumptions are needed about the error terms to calculate estimates ( $b_1 = \hat{\beta}_1$  and  $b_0 = \hat{\beta}_0$ ) of the slope and intercept of the regression line. These estimates are simply well-known mathematical calculations. However, all the model assumptions should be satisfied in order to properly conduct a hypothesis test or create a confidence interval for  $\beta_1$ .

## Activity Checking Model Assumptions

14. Calculate the residuals from the regression line in Question 11. Plot a histogram of the residuals (or create a normal probability plot of the residuals). In addition, create a residual versus order plot and use the informal test to determine if the equal variance assumption is appropriate for this study. **Compare these plots to the residual plots created for the two-sample  $t$ -test. Why are these graphs so similar?**
15. Create a scatterplot with the regression line in Question 11. Use the graph to give an interpretation of the slope and  $y$ -intercept,  $b_1$  and  $b_0$ , in the context of the game study.

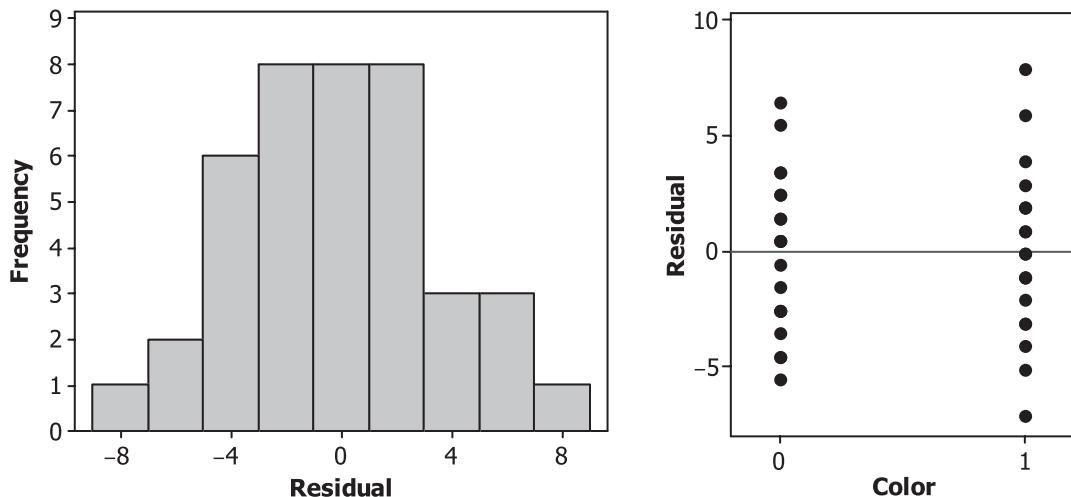


Figure 2.3 Histogram of residuals and plot of residuals versus color.

## 2.4 ANOVA to Compare Population Means

The term *ANOVA* is an acronym for *ANalysis Of VAriance*. ANOVA models often describe categorical explanatory variables in terms of factors and levels. The explanatory variable, also called a **factor**, in this study is the type of game; the two conditions, the two **levels** of the factor, are color distracter and standard.

### The ANOVA Model

The ANOVA model for the game study can be written as

$$y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \quad \text{for } i = 1, 2 \quad \text{and } j = 1, 2, \dots, n_i \quad \text{where } \varepsilon_{i,j} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad (2.6)$$

The mean response in the ANOVA model is  $\mu + \alpha_1$  for the color distracter group and  $\mu + \alpha_2$  for the standard group, where  $\mu$  is the mean of all the completion times in the study. This overall mean is often called the grand mean or the benchmark value;  $\alpha_1$  is the **effect**, or **main effect**, of the color distracter group. Effects are a measure of differences between group means. The effect  $\alpha_1$  represents the change in the response from the grand mean to the color distracter group mean.\*

To summarize, here is what the symbols in the model represent:

$y_{i,j}$ : observed completion time for subject  $j$  from group  $i$

$\mu$ : overall mean (the benchmark value)

$\alpha_i$ : effect of group  $i$  ( $i = 1, 2$ )

$\varepsilon_{i,j}$ : error for the  $j$ th subject ( $j = 1, 2, \dots, 20$ ) from the  $i$ th group ( $i = 1, 2$ )

Although the notation varies, the mean response for the ANOVA model is mathematically equivalent to the mean response in the  $t$ -test.

$\mu_1 = \mu + \alpha_1$ : population mean for the color distracter games

$\mu_2 = \mu + \alpha_2$ : population mean for the standard games

### Activity ◀ The ANOVA Model

16. Explain (or use equations to show) why the ANOVA hypothesis  $H_0: \alpha_1 = \alpha_2$  is equivalent to the two-sample  $t$ -test hypothesis  $H_0: \mu_1 = \mu_2$ .

#### ► MATHEMATICAL NOTE ▼

In the ANOVA model, there is the appearance that we are describing two means ( $\mu_1$  and  $\mu_2$ ) using three parameters ( $\mu$ ,  $\alpha_1$ , and  $\alpha_2$ ). Since it can be shown that  $\alpha_2 = -\alpha_1$ , there are actually just two parameters ( $\mu$  and  $\alpha_1$ ) that are estimated. Thus, the null hypothesis stating no effect size can also be written as  $H_0: \alpha_1 = \alpha_2 = 0$  or  $H_0: \mu_1 = \mu_2 = \mu$ .

17. Write the proper ANOVA model [provide the appropriate  $ij$  subscripts as in Equation (2.6)] for the observation representing the 3rd subject from the color distracter group. Also give the notation for the observation representing the 20th subject from the standard group.

18. Why doesn't  $\mu$  have any subscript in the ANOVA model?

After the data have been collected, the averages for all three meaningful groupings of the data can be calculated. The following mathematical notation is often used to represent the calculated sample averages:

$\bar{y}_{..}$ : **grand mean** (the overall average of the combined results)

$\bar{y}_1$ : average for the color distracter game sample results

$\bar{y}_2$ : average for the standard game sample results

\*In this text  $\mu$  is always considered the overall mean of the data. Also throughout this chapter, we are always assuming balanced data.

**NOTE**

Throughout this chapter,  $\bar{y}_{1..} = \bar{y}_1$  and  $\bar{y}_{2..} = \bar{y}_2$ . The dot notation is often used with more complex models to indicate that the average was taken over all values of that subscript. For example,  $\bar{y}_{2..}$  averages over all  $j = 1, 2, 3, \dots, n_2$ , observations from the standard game sample results.

The effect of the color distracter game,  $\alpha_1$ , can be estimated by  $\hat{\alpha}_1 = \bar{y}_{1..} - \bar{y}_{..}$ . Similarly,  $\hat{\alpha}_2 = \bar{y}_{2..} - \bar{y}_{..}$  estimates the standard game effect,  $\alpha_2$ . As in regression and the two-sample *t*-test, each residual  $\hat{\varepsilon}_{i,j}$  is the difference between an observed value and the corresponding mean response.

$$\begin{aligned}\hat{\varepsilon}_{i,j} &= \text{observed} - (\text{grand mean} + \text{effect of group}_i) \\ &= y_{i,j} - (\bar{y}_{..} + \hat{\alpha}_i) \\ &= y_{i,j} - [\bar{y}_{..} + (\bar{y}_{i..} - \bar{y}_{..})] \\ &= y_{i,j} - \bar{y}_{i..}\end{aligned}$$

**Key Concept**

Since the mean responses for the two-sample *t*-test, regression, and ANOVA are mathematically equivalent for this data set, the residual values are also identical for all three models.

## Activity Estimating the Model Values

19. Use the Games1 data to calculate  $\bar{y}_{..}$ ,  $\bar{y}_{1..}$ , and  $\bar{y}_{2..}$ .
20. Estimate the effect sizes for the color distracter game and the standard game.
21. The main effects are often visualized with a **main effects plot**. The main effects plot simply plots the average for each factor level and, in this example, shows that the color distracter group exhibited a higher average completion time than the standard group. Main effect plots are not very informative with just one explanatory variable. However, in more complex data sets with several explanatory variables, main effect plots can be quite useful in comparing effect sizes across all the explanatory variables. Use statistical software to create a main effects plot.
22. Calculate the residual for the 20th observation from the standard group,  $\hat{\varepsilon}_{2,20}$ .

## Model Assumptions for ANOVA

The model assumptions for ANOVA are equivalent to those for the two previous tests. In fact, the assumptions discussed in this section are called the six *Fisher assumptions*, after Ronald Fisher, who developed the ANOVA and the corresponding *F*-test.

- The parameters ( $\mu$ , each  $\alpha_i$ , and  $\sigma^2$ ) are constant throughout the study.
- Each term in the ANOVA model is added.
- The error terms are independent and identically distributed (iid).
- The error terms follow a normal probability distribution.
- The error terms have a mean of zero.
- Population variances within each factor level (each game type) are equal (i.e., the sample variances can be pooled).

The following questions provide an opportunity to use software to calculate an ***F*-statistic** (the test statistic for  $H_0: \alpha_1 = \alpha_2 = 0$  that is calculated using an ANOVA table) and corresponding *p*-value. In addition, you will use graphs to visualize the residuals to check the model assumptions. The extended activities will describe the ANOVA calculations in more detail.

## Activity ▶ Checking Assumptions

23. Use statistical software to calculate the  $F$ -statistic and find the  $p$ -value. Use the  $p$ -value to draw conclusions from this study.
24. How does the  $p$ -value in ANOVA compare to the  $p$ -value you found for the two-sample  $t$ -test and the regression model?
25. Take the square root of the  $F$ -statistic in the ANOVA table. Does this value look familiar? Explain.
26. Check the model assumptions by creating a histogram of the residuals, a plot of the residuals versus the type of game, and a plot of the residuals versus the order (the order in which data were collected). Are the residuals approximately normal? Are the residual variances similar for the two factor levels? Are there patterns in the residual plots that might indicate that the residuals are not iid?
27. Compare the three statistical models. Describe the connections among the  $t$ -test, ANOVA, and regression. Why are the  $p$ -values the same for all three models?

## 2.5 Comparing Planned Variability to Random Variability

The statistical model (observed value = mean response + random error) assumes that there are only two types of variability that can occur in a study. The difference between subgroup means (i.e., the difference between mean response values) represents the **planned variability** in the study. For example, in the game study we plan to find that the mean for the color distracter group is different from the mean for the standard group. The random error term is used to model the uncertainty of each individual outcome, called the **random variability**.

All three test statistics described in this chapter are based on a ratio. Each hypothesis test is based on comparing the planned variability to the random variability. The numerator in every test represents differences between group means. The denominator is a measure based on the variability of the residuals. If the subgroup means are far apart compared to the random variability, the null hypothesis is rejected and we conclude that the two population means are different.

Figure 2.4 shows boxplots for two fictitious data sets, **Results A** and **Results B**. Notice that the differences between group means are identical. In other words, the numerator of the test statistic (difference between group means) is the same for both data sets.

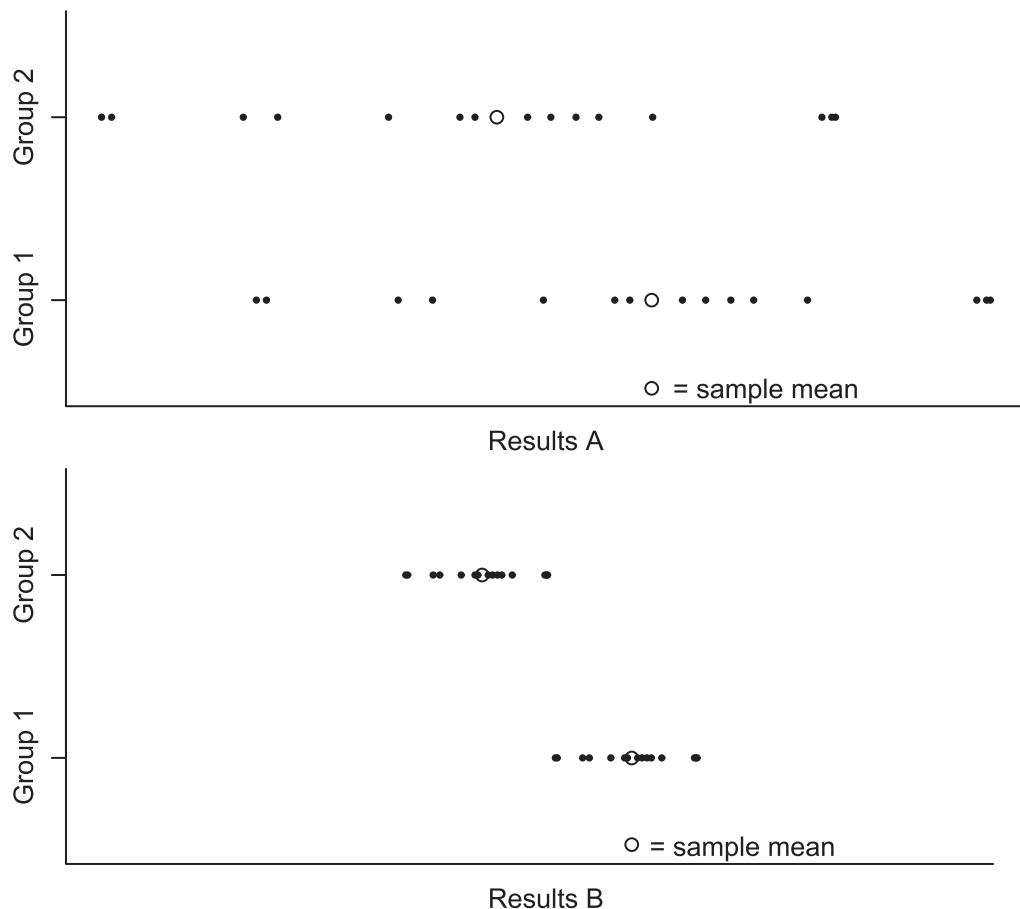
Even though the difference between group means (planned variability as described by the mean response) is the same, the variability within each group (random variability represented by the error term) is different. The residual variation (the denominator) is much larger for **Results A** than for **Results B**. Thus, the **Results B** data set will correspond to a larger test statistic and a smaller  $p$ -value, and we are more likely to reject the null hypothesis. Thus, **Results B** provides much stronger evidence that the difference between group means is not due simply to chance, but due to real differences in the two population means.

## 2.6 Random Sampling and Random Allocation

There is one more type of variability that is not included in the statistical model: **unplanned systematic variability**. This variability is caused by extraneous variables that can bias the results. **Extraneous variables** (such as time of day, prior computer game experience of the subject, number of pegs in the game, or amount of background noise) are not of interest in our study, but they may have an influence on the completion time of the game.

Essentially all studies have numerous extraneous variables that may be biasing the results. The problem is that we typically do not know all possible extraneous variables in a study or if they are biasing the results. Random sampling and random allocation are used to protect against the unwanted influence of extraneous variables:

- *Random sampling:* How was the sample collected? If the subjects in the sample were randomly selected from the population of interest, inferences can be drawn (generalized) to the entire population.



**Figure 2.4** Dotplots representing data from two studies. The difference between the group means is the same in both data sets, but the random variation is not the same. The variability in the residuals is much larger for Results A than for Results B.

- **Random allocation:** How were units assigned to treatments? If the units were randomly allocated to treatment groups, a statistically significant result in a well-designed study shows that the treatment *causes* changes in the response variable.

In the computer game study, students were “randomly” selected from the college. If the 40 students were truly a simple random sample of all students currently attending the college, the results of this study would hold for all students in the college. However, even if the researchers used a **sampling frame** (list of the population of all current students at their college) to randomly select 40 students, it would be unlikely that the first 40 subjects selected would agree to participate in the study. Thus, the population for the study would be all current college students who would agree to participate in the study. If the researchers’ version of “random sample” meant a collection of friends who agreed to participate in their study, the conclusions would hold only for the 40 students who volunteered.

The key point is that it is often very difficult to collect a true simple random sample from the population. If the sample is not reflective of the entire population (an appropriate random sample is not collected), the result may contain biases which may invalidate the results.

Random allocation is much easier to do appropriately in this study. Simply flipping a fair coin is enough to randomly assign subjects to a particular type of game. Therefore, since the sample data led us to reject the null hypothesis, we can be quite certain that the type of game *caused* a difference in the average completion time.

**Key Concept**

Random sampling and random allocation do not impact the type of statistical model or technique used, but they do impact the type of conclusions that can be drawn. When units are randomly sampled from a population, we can generalize the conclusions to that population. Well-designed experiments incorporate random allocation in a study and can be used to show causation.

Random sampling and random allocation can be used to convert unplanned systematic variability into random variability. For example, in the game study, the subjects' natural ability may bias the results if more talented subjects tend to play one game type over the other. However, if we randomly allocate subjects to a game type, we can expect each group to have an equivalent number of talented subjects. In addition, the variability in natural abilities now tends to look like the random variability that can be modeled with the error term.

In this chapter, we assume this was a well-designed study with no obvious biases. We focus on creating models and better understanding the random error term in order to determine if statistical techniques (two-sample *t*-test, regression, and ANOVA) are appropriate. Later chapters will discuss how to address extraneous variables and properly design studies.

**NOTE**

Later chapters will explain how studies can be designed to control for the influence of extraneous variables that are suspected of potentially biasing the results. Extraneous variables can be controlled by *limiting the study to conditions that are as consistent as possible*. For example, the researchers could decide to have the subjects play all games with the same number of pegs and play all games in a quiet room at the same time of day. Extraneous variables can also be controlled by *incorporating a new variable into the mean response*. Instead of simply testing for the type of game (color or standard), the researchers could include a second explanatory variable in the study. For example, the researchers could test each student's ability before the study, group students into experienced and inexperienced groups, and then, within each experience group, randomly assign the type of game each student should play.

## 2.7 What Can We Conclude from the Game Study?

Validation of model assumptions is essential before drawing conclusions from hypothesis tests. The residual plots created throughout this chapter appear to support the model assumptions. There are no clear trends or outliers in the residual plots. In general, the graphs do not give enough evidence to reject the assumption that the error terms are normally distributed with a mean of zero and a constant variance.

The *p*-value for all three hypothesis tests is 0.0279. When we assume the null hypothesis is true in an experiment, we are assuming that there is nothing creating group differences except the random allocation process. Under this assumption, a group difference at least as extreme as the one actually observed would occur only 2.79% of the time. This allows us to conclude that the type of game does cause a difference in completion times.

Under the conditions of this game study, we have shown that the statistical model and assumptions for the two-sample *t*-test (assuming equal variances), regression, and ANOVA models are mathematically equivalent. Thus, testing if there is a difference between means, if the regression slope is not zero, or if the factor effects are significant will lead to the same conclusion because they are exactly the same test.

The tests in this computer game study are identical because there were only two groups (two levels) of one explanatory variable and we assumed the variances of both groups were equivalent. Under these conditions, any of the three tests can be used to draw appropriate conclusions as long as the model assumptions are met. The extended activities and end-of-chapter exercises provide more details about the differences among the three tests.

## A Closer Look Statistical Models

### 2.8 Normal Probability Plots to Assess Normality

Figure 2.5 shows two histograms of the residuals calculated from the Games1 data. Both histograms use the same data and the same class widths (width of each bar); the only difference is that the bins start at different positions. Note that the histogram on the left looks somewhat skewed while the right graph is fairly symmetric.

These graphs are provided to illustrate that histograms are not always reliable for determining whether the residuals come from a normal distribution. Histograms are especially unreliable with small data sets, where the choice of class sizes can have a significant effect on the appearance of the graph.

An alternative to histograms is normal probability plots. A **normal probability plot** is a scatterplot of observed data versus the corresponding percentiles of the normal distribution. If the scatterplot forms a straight line, the percentiles of observed data match the percentiles of a normal distribution and we make the assumption that the observed data could have come from a population with a normal distribution.

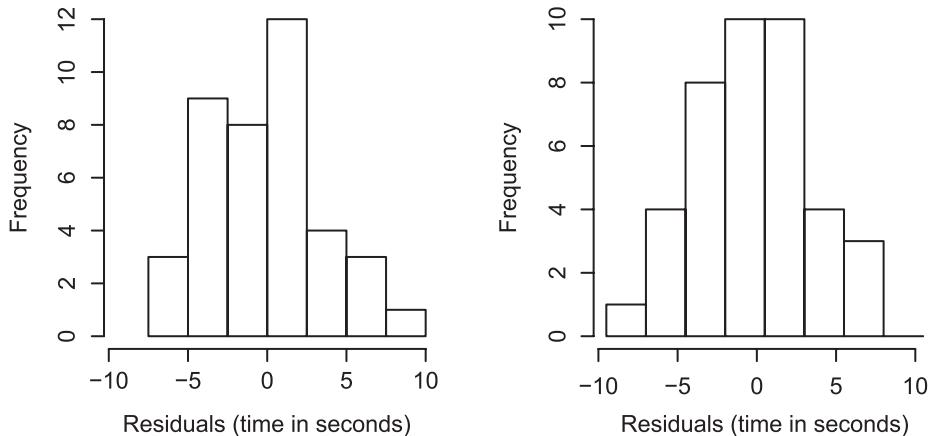


Figure 2.5 Two histograms of the computer game study residuals.

#### Extended Activity Creating Probability Plots

Data set: Normal

The following questions ask you to work through the process of calculating and interpreting probability plots.

28. **Calculating a Normal Probability Plot by Hand** Consider the following sample data set of  $n = 5$  observations: 14, 11, 17, 15, 13. Complete the following steps to create a normal probability plot.

- Sort the data from smallest to largest. Use a subscript in parentheses,  $(i)$ , to represent the ordered data. For example  $y_{(1)} = 11$  is the smallest observation and  $y_{(5)} = 17$  is the largest observed value.
- For each  $(i)$ , calculate the  $(i - 0.5)/n$  percentile of the standard normal distribution. For example, corresponding to  $(i) = (1)$ , the  $(1 - 0.5)/5 = 10$ th percentile of the standard normal distribution is  $-1.28$ , since  $P(Z \leq -1.28) = 0.10$  when  $Z \sim N(0, 1)$ . For  $(i) = (3)$ , the  $(3 - 0.5)/5 = 50$ th percentile (i.e., the median) of the standard normal distribution is 0. Repeat this process for the other ordered values,  $y_{(2)}$ ,  $y_{(4)}$ , and  $y_{(5)}$ .

- c. Make a normal probability plot by creating a scatterplot with the percentiles of the observed data along the  $x$ -axis and the percentiles of the standard normal distribution along the  $y$ -axis. If the data fall along a straight line, then the data are consistent with the hypothesis that they are a random sample from a population that is normally distributed.
- The data in this question are a little “heavier” toward the tails (the normal distribution has more observations in the center and fewer observations toward the tails than does this data set), so the probability plot has an S-shaped curve. With only five data points, the shape is not as clear as it would be for a data set with a larger sample size from a “heavy-tailed” population.
- d. If you standardized the data (subtracted the sample mean and then divided by the sample standard deviation), would you expect the shape of the normal probability plot to change?
- e. Does the shape of the normal probability plot change if you multiply each observation in the sample data set by 5?
- f. Does the shape of the normal probability plot change if you divide each observation in the sample data set by 3?
29. **Plotting Normal Data** For this problem, use the `Normal` data set. The first column of data actually is a random sample from a normal distribution.
- Use software to create a histogram and normal probability plot of the first column of the `Normal` data set.
  - Double the five largest observed values in the `Normal` data set. Create a histogram and normal probability plot of the “Largest 5 Doubled” data. Describe how the normal probability plot and the histogram change.
  - Now, double the five smallest observed values in the original `Normal` data set. Create a histogram and normal probability plot of the “Smallest 5 Doubled” data. Describe how the normal probability plot and the histogram change.
  - Draw (by hand) a picture of what a normal probability plot might look like for a data set with *fewer* observations in both tails than you would expect from a normal distribution.
  - Draw (by hand) a picture of what a normal probability plot might look like for a data set with *more* observations in both tails than you would expect from a normal distribution.

**NOTE**

As with any other hypothesis test, when we fail to reject  $H_0$  we do not prove  $H_0$  is true. Normal probability plots cannot be used to prove that the data came from a normal distribution ( $H_0$ ), but they can be used to show that the data are consistent with data from a normal population.

Assessing whether or not data could have come from a normal population by examining a normal probability plot requires some experience and may seem subjective. After all, even when data do come from a normal population, sampling variability (random variability) will sometimes lead to a normal probability plot where the data do not lie along a straight line.

**Extended Activity****Understanding Variability in Random Samples**

Data set: `Games1`

30. If you have no experience with probability plots, it can be helpful to look at several sample data sets that actually do come from a normal distribution. Use software to draw several random samples of the same size from an actual normal population and create normal probability plots. These plots can be compared to the normal probability plot from the actual data. If the real data plot looks similar to the plots where you know that the population is normal, then your data are consistent with the null hypothesis (i.e., the data came from a normal population). If the real data plot is “extreme” (quite different from the plots coming from a normal population), then the differences are not likely due to chance and you can reject the hypothesis that the data came from a normal population.
- Create a normal probability plot of the residuals of the `Games1` data from Question 26.

- b. Use software to draw a random sample of  $n = 40$  from an actual normal probability distribution with mean 0 and standard deviation 1. Create a normal probability plot of the sample data.
- c. Repeat the previous question eight more times, for a total of nine normal probability plots of “data” from an actual normal probability distribution. Does the plot in Part A resemble the nine plots with data sampled from a normal distribution? If you can’t distinguish the `Games1` residuals from the other plots, it would seem reasonable to assume that the `Games1` residuals are normally distributed.

## 2.9 Transformations

### Transformations for ANOVA

It is common for data to not follow a normal distribution or for subgroups to have dramatically different variances. For example, in biological studies it is common for subgroups with larger means to also have larger variances. Consider measuring the weights of various animal species. We expect the weights of mice to have less variability than the weights of elephants, as measurement instruments often get less precise (more variable) as the measurements get larger.

In these types of situations, the data can often be transformed to fit model assumptions. **Transformations** are monotonic mathematical operations that change the scale of the explanatory variable, the response variable, or both. When groups within a data set have unequal variances or when data are skewed to the right, a square-root or natural-logarithm transformation on the response variable can often change the data to a scale where the equal variance and normality assumptions are more closely satisfied. Then the transformed data can be analyzed using traditional techniques such as ANOVA or regression.

#### → MATHEMATICAL NOTE →

**Monotonic functions** preserve the order of the original data. A monotonic increasing function maintains the direction of the data: For any two data points, when  $y_i > y_j$ , then  $f(y_i) > f(y_j)$ . A monotonic decreasing function reverses the direction of the data: For any two data points, when  $y_i > y_j$ , then  $f(y_i) < f(y_j)$ . If the transformation is not monotonic over the range of sample data (i.e., if the data set contains zeros or negative numbers), simply add a constant to each number to make all numbers positive or nonzero before transforming the data.

Although an infinite number of transformations could be tried, it is best to focus on commonly used transformations such as the ones listed below:

The **square-root transformation** ( $y^{1/2} = \sqrt{y}$ ) is commonly used when the response variable represents counts, such as a count of the number of observed species. Square-root transformations are also very useful when the variances are proportional to the means.

The **log transformation** is often used when the data represent size or weight measurements. In addition, it is useful when the standard deviations are proportional to the means. A common logarithm is based on the number 10 and written  $\log_{10}(x)$ . This log is defined as  $\log_{10}(10^x) = x$ . The natural logarithm,  $\ln(x)$ , is based on the number  $e = 2.71828$ , so  $\ln(e^x) = x$ . For statistical tests, it makes no difference whether you use log base 10 ( $\log_{10}$ ) or natural logs ( $\ln$ ), because they differ only by a constant factor. The log base 10 of a number equals 2.303 times the natural log of the number. Log transformations are often preferred over other transformations because the results tend to be easier to interpret.

The **reciprocal transformation** ( $y^{-1} = 1/y$ ) is often useful when the data represent waiting times, such as time until death or time until a battery fails. If most responses are relatively close to zero but a few responses are larger, this transformation will reduce the effect of large response values.

The **arcsine transformation** ( $\sin^{-1}(\sqrt{y})$ ) and **logit transformation** ( $\log[y/(1 - y)]$ ) are useful when measurements are proportions between 0 and 1. The arcsine transformation is often difficult to interpret and cannot be subjected to back transformation (described in the next section) to produce an informative interpretation. The logit function can be usefully interpreted and will be discussed in much more detail in Chapter 7.

## Extended Activity

### Transforming Emissions Data

Data set: Emission

31. The data set `Emission` provides hydrocarbon emission in parts per million (ppm) at idling speed for cars, based on the year each car was manufactured. These data were randomly sampled from a much larger study on pollution control in Albuquerque, New Mexico.
- Create individual value plots or side-by-side boxplots of `Emission` versus `Year`. Compare the mean and standard deviation of each group. Do the data within each group look consistent with data from a normal population?
  - Transform the response by taking the log of `Emission`. Create individual value plots or side-by-side boxplots of `log Emission` versus `Year`. Compare the plot of the transformed data to the plot in Part A. Which plot shows data that better fit the model assumptions?
  - Calculate an ANOVA table,  $F$ -test, and  $p$ -value to determine if the average `log (Emission)` varies based on `Year`. Note that the end-of-chapter exercises and Section 2.9 show that ANOVA can compare more than two groups. In this question,  $I = 5$  groups instead of 2 groups. However, the model and calculations are identical except that now  $i = 1, 2, 3, 4, 5$  instead of  $i = 1, 2$ . The null hypothesis is  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$  versus the alternative  $H_a$ : at least one group mean is different from another.
  - Create residual plots to evaluate whether the model assumptions for the  $F$ -test are violated.

Notice that although the log transformation was helpful, the data still have outliers. In addition, the equal variance and normality assumptions are still slightly violated. Some statisticians would consider the log-transformed data appropriate for the standard ANOVA. Others would try another transformation, such as taking the log of the transformed data again; this is called a **log log transformation**. Still others would suggest using a **nonparametric test**. (Nonparametric tests, such as the Kruskal-Wallis test, are described in Chapter 1). Nonparametric tests do not require error terms to follow the normal distribution. While any of these analyses would be appropriate, it would not be appropriate to conduct several analyses on the same data and then report only the conclusions corresponding to the test that gave the smallest  $p$ -value. For example, if we tried three or four hypothesis tests each with an  $\alpha$ -level = 0.10 and then simply picked the test with the smallest  $p$ -value, our chances of incorrectly rejecting the null hypothesis would actually be greater than 10%.

If there are very clear outliers, if data are skewed, or if the subgroup variances are clearly different, a transformation applied to the response variable may help the data fit the ANOVA model assumption. When the normality or equal variance assumption does not hold, the one-way ANOVA  $F$ -test still tends to be a fairly accurate method if there are equal sample sizes. The  $F$ -statistic is much less reliable when there are unbalanced sample sizes and one or more subgroups have a much larger variance than others.<sup>3</sup>

## Back Transformations

Transformations are not simply a way of playing around with the data until you get the answer you want. It is important to recognize that there is no reason to believe that the original scale used for the measurements is better than other scales. For example, in testing for differences in lengths, should it matter if the original data were collected in meters or in feet? One scale is not better than the other; we transform data simply so that it is easier for our audience to interpret. Some scales, such as pH levels,<sup>\*</sup> are always presented using a logarithmic scale.

For testing for differences between groups, the  $p$ -values of transformed data are reliable as long as model assumptions are satisfied. However, other statistical results, such as confidence intervals or the slope coefficient, are typically best understood in the original units. Thus, it is often desirable to back transform the results. **Back transformations** do the opposite of the mathematical function used in the original data transformation. For example, if the natural log transformation was used, a back transformation is conducted by taking the exponent of the number. Unfortunately, it can be very difficult to interpret some statistical results in either the transformed or the back-transformed scale.

---

<sup>\*</sup>pH is a measure of how acidic or how basic (alkaline) a solution is. It is measured as the negative logarithm (base 10) of the molar concentration of dissolved hydronium ions.

Consider conducting a *t*-test for the difference between the mean car emissions for the pre-63 and the 70–71 groups in Question 31. The standard deviation of the pre-63 group, 592, is more than double that of the 70–71 subgroup, 287.9. We will again assume equal variances in our *t*-test, but even if a different *t*-test were chosen, there would be clear evidence of nonnormality. Taking the natural log of Emission addresses the nonnormality problem and also makes the standard deviations very similar. The standard deviation is 0.57 for the transformed pre-63 group and is 0.678 for the transformed 70–71 group. The two-sided hypothesis test gives a *p*-value of 0.001, which provides strong evidence that there is a difference between group means. This test is valid since the model assumptions are met.

The 95% confidence interval for the transformed data is  $(-1.434, -0.411)$ . However, this transformed confidence interval is not easy to interpret in terms of actual car emissions. The back-transformed confidence interval is

$$(e^{-1.434}, e^{-0.411}) = (0.238, 0.663)$$

Note that the confidence limits are no longer symmetrical. In addition, this confidence interval no longer is interpreted as the difference between two means, but now represents the confidence interval for the ratio between the two means. The end-of-chapter exercises provide additional examples of interpreting results on the transformed scale (and back-transformed scale).

#### CAUTION

The back-transformed data do not have the same meaning as the original raw data. For two log-transformed means,  $\ln(\bar{y}_1) - \ln(\bar{y}_2) = \ln(\bar{y}_1/\bar{y}_2)$ . Thus, back transforming the data ( $e^{\ln(y_1/y_2)} = \bar{y}_1/\bar{y}_2$ ) results in the ratio of the two means. Results of back transformations based on the square-root, reciprocal, and arcsine transformations often have no practical interpretation.

#### Key Concept

It can be difficult to properly interpret slope coefficients or confidence intervals using either transformed or back-transformed data. Hypothesis tests for differences between groups do not need to be back transformed.

## Transformations for Regression

As with ANOVA, there are many situations in regression in which data are skewed, outliers exist, or the variability of the residuals tends to depend on the explanatory variable. Graphing the residuals is often the best way to identify the appropriate transformations. If the statistical model is correct, then no clear patterns (such as a strong curve or fan shape) should be seen in the plots. When there is no clear pattern in the residual plots, it is safe to assume that no statistical model based on the same explanatory variable will be a better fit for the data.

### Extended Activity

#### Transforming Brain and Body Weight Data

Data set: Weight

32. The Weight data set contains the brain weights (*y*) and body weights (*x*) of 32 species.
  - a. Create a scatterplot of *y* versus *x* with a regression line ( $\hat{y} = b_0 + b_1x$ ), a plot of residuals versus the explanatory variable, a plot of residuals versus predicted (or “fitted”) values ( $\hat{y}$ ), and either a normal probability plot or a histogram of the residuals.
  - b. Try various transformations of the explanatory and response variables to create a better linear regression model. Hint: Notice that since both the *x* and the *y* variable are right skewed and have outliers, both may need a transformation.

## Choosing the Right Transformation

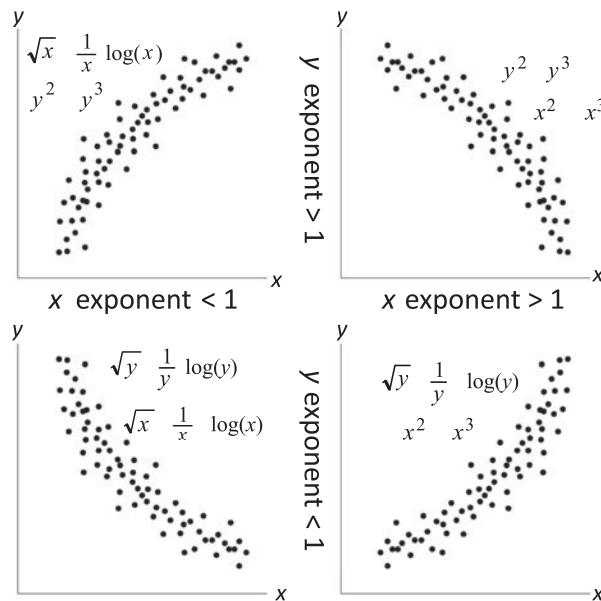
When a scatterplot of the data reveals a curved (nonlinear) shape, transformations are often used to straighten curved relationships so that a simple linear regression model will fit the data. In some cases, theoretical knowledge or previous studies can provide an indication of a suitable transformation. More formal methods, such as the Box-Cox method and the Box-Tidwell method,<sup>4</sup> can also be used to choose a transformation. However, the best indicators of an appropriate transformation are often found by viewing scatterplots and residual plots of the data.

Mosteller and Tukey introduced the ladder of powers and the bulging rule as a way to choose among the family of power transformations.<sup>5</sup> The following list of transformations is often referred to as the *ladder of powers* because power and logarithmic functions have a natural hierarchy:

$$\dots, y^{-2}, y^{-1}, y^{-1/2}, \log(y), y^{1/2}, y^1, y^2, \dots$$

Notice that  $\log(y)$  replaces the transformation  $y^0 = 1$ , since setting everything to a constant value is not useful. Exponents greater than one will cause the function to increase at a faster rate. Exponents less than one (and the log) will cause the function to bend downward. The curves become progressively steeper (sharper) as the exponent moves away from one.

The bulging rule provides a visual method for determining appropriate transformations. Figure 2.6 shows four different curves (bulges) and indicates which powers of  $y$  and  $x$  would likely straighten the line. For example, the upper left quadrant of Figure 2.6 shows a curve that tends to become more linear if  $y$  is transformed to a power greater than one (such as  $y^2$  or  $y^3$ ) and  $x$  is transformed to a power less than one (such as  $\sqrt{x}$  or  $x^{-1}$ ).



**Figure 2.6** Bulge rule showing appropriate transformations to linearize curved data.

Performing a transformation to control problems with unequal variances can increase the nonlinearity between the explanatory and response variables. Transforming the response variable influences both the variation and the linearity, but transforming the explanatory variable influences only the linearity. Thus, it is best to transform the response variable first to deal with nonconstant variance and then consider additional

transformations on the explanatory variable to make the model linear. The following steps are useful for choosing an appropriate transformation:

- Create a scatterplot of the original data.
- Use the ladder of powers or other methods to select a transformation for the explanatory variable, response variable, or both.
- Create a scatterplot of the transformed data. If the scatterplot is not linear, try a new transformation. If the scatterplot is linear, conduct the appropriate statistical analysis and create residual plots.
- If the residual plots are not random, try another transformation. If the residuals do appear random (the model assumptions about the error term are satisfied), then the statistical analysis is reliable.

Often there are no appropriate transformations that will satisfy all the model assumptions. Future chapters discuss more advanced techniques that can be used to allow for nonnormal residuals and for nonlinear relationships.

## Extended Activity Comparing Four ( $x, y$ ) Data Sets

Data set: RegrTrans

33. Do the following for each of the four data sets:

- a. Create a scatterplot of  $y$  versus  $x$  with a regression line ( $\hat{y} = b_0 + b_1x$ ), a plot of residuals versus the explanatory variable, a plot of residuals versus predicted (or “fitted”) values ( $\hat{y}$ ), and either a normal probability plot or a histogram of the residuals.
- b. By hand, sketch on the scatterplot a curve that would fit the data better than the regression line. Notice that the plot of residuals versus the explanatory variable emphasizes the patterns in the residuals much better than does the scatterplot of  $y$  versus  $x$ .
- c. Try various transformations of the explanatory and response variables to create a better linear regression model (as validated by graphical analysis of the residuals).

## 2.10 Calculating Test Statistics

This section is a rather terse description of the mathematical calculations behind the hypothesis tests and confidence intervals described in this chapter. Most introductory textbooks will dedicate an entire chapter to each of these techniques. The logic behind the calculations for regression and ANOVA will be described in more detail in later chapters of this text.

### The Two-Sample *t*-Test with the Equal Variance Assumption

The two-sample *t*-test can be used to test whether two population means are equal. The null hypothesis about the population means ( $H_0: \mu_1 = \mu_2$ ) is rejected if the difference between the sample means,  $\bar{y}_1$  and  $\bar{y}_2$ , is so large that it doesn’t appear reasonable to assume that the groups have the same mean.

The test statistic for the two-sample *t*-test is

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (2.7)$$

The above test statistic is a function of the following summary statistics from the sample data:

$$\begin{aligned} \bar{y}_1 &= \bar{y}_{1\cdot} = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1,j} & \bar{y}_2 &= \bar{y}_{2\cdot} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2,j} \\ s_1 &= \sqrt{\frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (y_{1,j} - \bar{y}_1)^2}, & s_2 &= \sqrt{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_{2,j} - \bar{y}_2)^2} \end{aligned}$$

The difference in population means ( $\mu_1 - \mu_2$ ) is not known, but comes from the statement of the null hypothesis:  $\mu_1 - \mu_2 = 0$  (or, equivalently,  $\mu_1 = \mu_2$ ). Thus, the test statistic is simply a ratio of the distance between the two sample means to a measure of variation.

The **pooled standard deviation** (denoted  $s_p$ ) uses a weighted average of the two sample variances in order to estimate the size of the variation in a typical random error term (i.e.,  $\sigma$ , the common standard deviation for the two populations).

Probability theory can be used to prove that if the model assumptions are true, the  $t$ -statistic in Equation (2.7) follows a  $t$ -distribution with  $(n_1 + n_2 - 2)$  degrees of freedom. If the  $t$ -statistic is large, the difference between the two means is large compared to the pooled standard deviation. We will reject the null hypothesis that the two means are equal ( $H_0: \mu_1 = \mu_2$ ) in favor of  $H_a: \mu_1 \neq \mu_2$  if the  $t$ -statistic is so large that it is unlikely to occur when  $\mu_1 = \mu_2$ . A large  $t$ -statistic corresponds to a small enough  $p$ -value, which is found with software or in a  $t$ -table.

## Extended Activity Calculating the Two-Sample $t$ -Test

Data set: Games1

34. Use Equation (2.7) to calculate the test statistic ( $t$ ) by hand (i.e., without statistical software) for the computer game study. Use software or a  $t$ -table with  $(n_1 + n_2 - 2)$  degrees of freedom to find the  $p$ -value.

## Regression

Introductory statistics textbooks describe how least squares techniques can be used to calculate the following statistics to estimate  $\beta_0$  and  $\beta_1$ :

$$b_1 = \hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}, \quad b_0 = \hat{\beta}_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} \quad (2.8)$$

where  $n = n_1 + n_2$ . In most introductory statistics texts, Equations (2.8) for the slope and intercept are simplified to

$$b_1 = r \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

where the sample correlation coefficient is

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{y_i - \bar{y}}{s_y} \right) \left( \frac{x_i - \bar{x}}{s_x} \right)$$

To test the null hypothesis  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$ , it can be shown that the  $t$ -statistic for the slope coefficient is

$$t = \frac{b_1 - \beta_1}{\hat{\sigma}} \quad \text{where} \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2}{n-2}} \quad (2.9)$$

Notice that  $\hat{\sigma}$  is an estimate of the standard deviation of the random errors. If the sample statistic  $b_1$  is far away from  $\beta_1 = 0$  relative to the size of the estimated standard deviation,  $\hat{\sigma}$ , then the  $t$ -statistic will be large and the corresponding  $p$ -value will be small.

Probability theory can be used to prove that if the regression model assumptions are true, the  $t$ -statistic in Equation (2.9) follows a  $t$ -distribution with  $n - 2 = n_1 + n_2 - 2$  degrees of freedom.

## Extended Activity Testing the Slope Coefficient

Data set: Games1

35. Without statistical software, use summary statistics and Equation (2.9) to calculate the test statistic under the null hypothesis that  $\beta_1 = 0$ . Use software or a *t*-table with  $n - 2 = n_1 + n_2 - 2$  degrees of freedom to find the *p*-value.
36. Compare the test statistic and *p*-values in Questions 34 and 35.

## Analysis of Variance (ANOVA)

Several calculations will be made to test the hypothesis  $H_0: \mu_1 = \mu_2$  in an ANOVA, but again the test statistic is a ratio of the spread between group sample means to the variability in the residuals.

If indeed  $\mu_1 = \mu_2$  (i.e.,  $H_0: \alpha_1 = \alpha_2 = 0$  is true), then we would expect the variation between the level means to be relatively small compared to the variability in the error terms. If the group means are relatively far apart, the *F*-statistic will be large and we will reject  $H_0: \mu_1 = \mu_2$  in favor of  $H_a: \mu_1 \neq \mu_2$ . While the logic is similar to that for the other tests described in this chapter, the test statistic for ANOVA, the *F*-statistic, requires many more calculations, as shown below.

**Sums of squares** (SS) are measures of spread, calculated in an ANOVA table like the one you saw in the software output for Questions 23 through 25. The three sums of squares calculated for the ANOVA table for the computer game study are described below.

**Group sum of squares** (SS<sub>Group</sub>) measures the difference between group means (also called level means). Group sum of squares represents the variability we want in the model. For the computer game, SS<sub>Group</sub> measures the spread between the two game type means, but ANOVA can be extended to more than just two groups.

Recall that the *i*th level mean is denoted as  $\bar{y}_i$ , the grand mean is denoted as  $\bar{y}_{..}$ , and the corresponding level effect is  $\hat{\alpha}_i = \bar{y}_i - \bar{y}_{..}$ .

$$\text{SS}_{\text{Group}} = \sum_{i=1}^I (\text{each level effect})^2 = \sum_{i=1}^I \left[ \sum_{j=1}^{n_i} (\bar{y}_{i..} - \bar{y}_{..})^2 \right]$$

(where  $I = \text{number of groups or levels}$ )

$$\begin{aligned} &= \sum_{i=1}^I [n_i \times (\bar{y}_{i..} - \bar{y}_{..})^2] = \sum_{i=1}^I n_i \hat{\alpha}_i^2 \\ &= \sum_{i=1}^2 [20 \times (\bar{y}_{1..} - \bar{y}_{..})^2] \text{ for the computer game study} \\ &= 20 \times (\bar{y}_{1..} - \bar{y}_{..})^2 + 20 \times (\bar{y}_{2..} - \bar{y}_{..})^2 \end{aligned}$$

**Error sum of squares** (SS<sub>Error</sub>) measures the spread of the observed residuals. Recall that each residual is defined as an observed value minus the estimated mean response:  $\hat{\epsilon}_{i,j} = y_{i,j} - \bar{y}_{i..}$ . In any ANOVA model with one explanatory variable, the mean response is the level average.

$$\begin{aligned} \text{SS}_{\text{Error}} &= \sum_{i=1}^I (\text{each residual effect})^2 = \sum_{i=1}^I \left[ \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{i..})^2 \right] \\ &= \sum_{i=1}^I [(n_i - 1) \times s_i^2] \quad \text{since } s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{i..})^2}{(n_i - 1)} \\ &= \sum_{i=1}^2 [(19) \times s_i^2] \quad \text{for the computer game study} \\ &= 19 \times s_1^2 + 19 \times s_2^2 \end{aligned}$$

**Total sum of squares** (SS<sub>Total</sub>) measures the overall spread of the responses in the full data set.

$$\begin{aligned} \text{SS}_{\text{Total}} &= \sum_{i=1}^I (\text{distance between each observation and the grand mean})^2 \\ &= \sum_{i=1}^I \left[ \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{..})^2 \right] \end{aligned}$$

$$\begin{aligned}
 &= (n - 1) \times s^2 \quad \text{since } s^2 = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{..})^2}{(n - 1)} \\
 &= (39) \times s^2 \quad \text{for the computer game study}
 \end{aligned}$$

Here,  $s^2$  is the overall sample variance and  $n = \sum_{i=1}^I n_i = \text{total sample size}$ .

It can be shown that  $\text{SS}_{\text{Total}} = \text{SS}_{\text{Group}} + \text{SS}_{\text{Error}}$ . While the specific formula for  $\text{SS}_{\text{Error}}$  is provided above, it is most easily calculated by subtracting  $\text{SS}_{\text{Group}}$  from  $\text{SS}_{\text{Total}}$ .

**Degrees of freedom** (df) for each sum of squares are calculated based on how many “free” pieces of information are summed. In this example, there are two levels of the game type factor. We can show that the weighted Type effects must sum to 0 ( $n_1\hat{\alpha}_1 + n_2\hat{\alpha}_2 = 0$ ). This implies that knowing the color distracter game effect automatically forces a known effect for the standard game. Thus,  $\text{SS}_{\text{Group}}$  has only  $I - 1 = 2 - 1 = 1$  df.  $\text{SS}_{\text{Total}}$ , like the usual one-sample variance, has  $n - 1 = 40 - 1 = 39$  df. It can also be shown that  $\text{df}_{\text{Total}} = \text{df}_{\text{Type}} + \text{df}_{\text{Error}}$ ; thus,  $\text{df}_{\text{Error}} = \text{df}_{\text{Total}} - \text{df}_{\text{Type}} = (n - 1) - (I - 1) = n - I = n - 2 = 38$ .

**Mean squares** (MS) are measures of “average” spread and are calculated by dividing a sum of squares (SS) by its associated degrees of freedom (df).

**Group mean squares** ( $\text{MS}_{\text{Group}}$ ) equal  $\text{SS}_{\text{Group}}/\text{df}_{\text{Group}}$ .  $\text{MS}_{\text{Group}}$  is a measure of variability between the levels of each factor and is often called **between-level variability**. It is actually just the variance of the level means.

**Mean square error** (MSE) equals  $\text{SS}_{\text{Error}}/\text{df}_{\text{Error}}$ . MSE is a pooled measure of the variability within each level, or the **within-level variability**. Remember that the variance is assumed to be the same for the responses within each level of the factor:  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . When  $I = 2$ , MSE is identical to the pooled variance ( $s_p^2$ ) used in the two-sample  $t$ -statistic in Equation (2.7).

If  $\text{MS}_{\text{Group}}$  is much larger than MSE, it is reasonable to conclude that there truly is a difference between level means and the difference we observed in our study was not simply due to chance variation (random error).

The **F-statistic** ( $\text{MS}_{\text{Group}}/\text{MSE}$ ) is a ratio of the between-level variability to the within-level variability. If indeed  $\mu_1 = \mu_2$  (i.e.,  $H_0: \alpha_1 = \alpha_2 = 0$  is true), then we would expect the variation between the level means in our sample data to be about the same as the typical variation within levels and the F-statistic would be close to one. Larger values of the F-statistic would imply that the level means were farther apart than chance error alone could explain. These calculations are often summarized in an **ANOVA table**, as shown in Table 2.1.

**Table 2.1** One-factor (“one-way”) ANOVA table (one factor with  $I$  levels).

Source	df	SS	MS	F-Statistic
<b>Group</b>	$I - 1$	$\sum_{i=1}^I n_i (\bar{y}_i - \bar{y}_{..})^2$	$\frac{\text{SS}_{\text{Group}}}{\text{df}_{\text{Group}}}$	$\frac{\text{MS}_{\text{Group}}}{\text{MSE}}$
<b>Error</b>	$n - I$	$\sum_{i=1}^I (n_i - 1)s_i^2$	$\frac{\text{SS}_{\text{Error}}}{\text{df}_{\text{Error}}}$	
<b>Total</b>	$n - 1$	$\sum_{i=1}^I \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{..})^2$		

Probability theory can be used to prove that if the model assumptions are true, the F-statistic follows an F-distribution with  $\text{df}_{\text{Group}}$  and  $\text{df}_{\text{Error}}$  degrees of freedom. The **p-value** gives the likelihood of observing an F-statistic at least this extreme (at least this large), assuming that the population means of the two game types are equal. Thus, when the p-value is small (e.g., less than 0.05 or 0.01), the effect size of the type of game is conventionally determined to be statistically significant.

## Extended Activity

### Calculating an ANOVA Table

Data set: Games1

37. Use  $n_1$ ,  $n_2$ ,  $\bar{y}_1$ ,  $\bar{y}_2$ ,  $s_1$ , and  $s_2$  to calculate  $SS_{Type}$  (i.e.,  $SS_{Group}$ ),  $SS_{Error}$ ,  $MS_{Type}$ , and  $MSE$  for the computer game study.

Since the group for the computer game study is game type (the *Type* variable in the data set), we will use the more descriptive labels  $MS_{Type}$ ,  $SS_{Type}$ , and  $df_{Type}$  instead of  $MS_{Group}$ ,  $SS_{Group}$ , and  $df_{Group}$ . This is common practice and is similar to how statistical software reports results based on variable names.

38. Calculate the overall variance of the completion times in the entire data set and use it to find the total sum of squares ( $SS_{Total}$ ) for the computer game study. Confirm that  $SS_{Total} = SS_{Type} + SS_{Error}$  for the computer game study.
39. Calculate the *F*-statistic and use software or an *F*-distribution table with  $df_{Type}$  and  $df_{Error}$  degrees of freedom to find the *p*-value.

## 2.11 Confidence Intervals

In previous sections, hypothesis tests were used with sample data to assess the evidence for a claim about a population. They were used to determine if there was evidence to support the claim that the two population means were different. An alternative approach is to use confidence intervals to create an interval estimate of a population parameter (such as the difference between two population means) and provide a level of confidence in the interval estimate.

Each confidence interval discussed in this chapter has the following form:

$$\text{estimate} \pm \text{critical value} \times \text{standard error of the estimate}$$

The **estimate** is a sample statistic used to estimate the population parameter. In the computer game study, a confidence interval for the population mean  $\mu_1$  would have an estimate of  $\bar{y}_{1.}$ . A confidence interval for  $\mu_1 - \mu_2$  is estimated by  $\bar{y}_{1.} - \bar{y}_{2.}$ .

A **confidence level** is the probability that the true parameter value will be captured by the confidence interval. In other words, a 95% confidence level ensures that the method used to calculate the confidence interval will successfully contain the true parameter value 95% of the time.

The **critical value** is a value from a distribution that is used to provide a confidence level for the interval. The critical values used for the two-sample *t*-test and regression are based on the *t*-distribution. The same model assumptions are used in both hypothesis tests and confidence intervals. Thus, the same distribution and degrees of freedom are used in the hypothesis test and confidence interval.

For the game study, a *t*-distribution with 38 degrees of freedom is used. The critical value  $t_{38}^*$  for a particular confidence level  $C$  is chosen so that  $C\%$  of the area under the *t*-distribution is between  $-t_{38}^*$  and  $t_{38}^*$ . For example, for a 95% confidence level ( $C = 95$ ),  $t_{38}^* = 2.02$ , since 95% of the area under a *t*-distribution with 38 df is between  $-2.02$  and  $2.02$ .

If the confidence level were chosen to be 99% ( $C = 99$ ), the confidence interval would be wider than before. A wider interval would have a higher probability of capturing the true mean. The critical value for a 99% confidence interval for the game study is  $t_{38}^* = 2.71$ .

The **standard error of the estimate** is a measure of the variability of the statistic. For example, a 95% confidence interval for  $\mu_1$  is

$$\bar{y}_{1.} \pm t_{19}^* \times \hat{\sigma}_{\bar{y}_{1.}}$$

$$38.1 \pm 2.09 \times \frac{3.65}{\sqrt{20}}$$

$$(36.39, 39.81)$$

A 95% confidence interval for  $\mu_1 - \mu_2$  is

$$\bar{y}_{1.} - \bar{y}_{2.} \pm t_{38}^* \times \hat{\sigma}_{\bar{y}_{1.} - \bar{y}_{2.}} \quad \text{Note that } s_p = \hat{\sigma}_{\bar{y}_{1.} - \bar{y}_{2.}}$$

$$38.1 - 35.55 \pm 2.02 \times \sqrt{\frac{(19)3.65^2 + (19)3.39^2}{20 + 20 - 2}}$$

$$(0.29, 4.81)$$

## Extended Activity

### Calculating a Confidence Interval for the Regression Coefficient

Data set: Games1

40. Note that the “standard error of the estimate” in the confidence interval is identical to the denominator in the test statistic for the corresponding hypothesis test. Use this information to *write out the formula* for a 95% confidence interval for  $\beta_1$ . Use the output from the corresponding  $t$ -test to find the estimate, the critical value, and the standard error of the regression coefficient. Use this information to calculate a 95% confidence interval for  $\beta_1$ .

#### MATHEMATICAL NOTE

The  $F$ -distribution used in ANOVA is not a symmetric distribution and all values are positive. Confidence intervals for the difference between two means are not calculated with an  $F$ -distribution. Note that the MSE in ANOVA =  $s_p$  from the two-sample  $t$ -test and the square root of the critical value  $\sqrt{F_{1,38}^*} = t_{38}^*$ .

#### Key Concept

Statistics based on sample data are estimates of population parameters. A confidence interval allows us to calculate an interval that has probability  $C$  (often  $C = 95\%$ ) of containing the true population parameter.

## Chapter Summary

When there is only one explanatory variable (factor) with two levels, a  **$t$ -test** is typically used to analyze the data. While this chapter has shown that **ANOVA** or **regression analysis** techniques provide equivalent results in this setting, they are typically used with different study designs.

The tests used in ANOVA and regression are developed under the assumption that the variances of each group are equal, while a two-sample  $t$ -test could be used without this assumption. Two-sample  $t$ -tests are limited to comparing two groups, while ANOVA and regression techniques are often used to analyze data sets with multiple explanatory variables, each having many levels. All three techniques are used when the response variable is quantitative.

While the  $t$ -test for  $\beta_1$  is appropriate, the scatterplot and regression line created in Question 15 show that a regression model does not accurately describe the data. For example, it would be meaningless to predict the expected completion time when  $x = 0.5$ . Simple linear regression models are typically used when the explanatory variables are quantitative.

Throughout this chapter, you were asked to evaluate **residual plots** to determine whether the model assumptions were met. When model assumptions are not met, the test statistics do not follow the corresponding  $t$ -distribution or  $F$ -distribution. Thus, the  $p$ -values may not be correct. No conclusions should be drawn from any statistical test without checking the appropriate assumptions.

The **statistical model** and corresponding hypothesis tests assume there are no **extraneous variables** that are biasing the study. Since it is typically impossible to identify all possible sources of bias during a study, **random sampling** and **random allocation** should be used.

In this chapter, we focused on testing the four assumptions about the **error terms** in each model. Table 2.2 shows the residual plots used in this chapter to check model assumptions.

**Table 2.2** Plots that can be used to evaluate model assumptions about the residuals.

Assumption	Plot
Normality	Histogram or normal probability plot
Zero mean	No plot (errors will always sum to zero under these models)
Equal variances	Plot of original data or residual vs. fits
Independence	Residuals vs. order
Identically distributed	No plot (ensure each subject was sampled from the same population within each group)

If the model assumptions are violated, **transformations** can often be used to rescale the data to better fit some model assumptions. Several sophisticated mathematical tests are also available to test these model assumptions. While these tests are useful, plots are often just as effective and better assist in understanding the data. Plots should always be used to visualize the data before any conclusions are drawn. Later chapters will provide much more detail on checking assumptions for more complex models.

## Exercises

- E.1. Assume you are conducting a *t*-test to determine if there is a difference between two means. You have the following summary statistics:  $\bar{x}_1 = 10$ ,  $\bar{x}_2 = 20$ , and  $s_1 = s_2 = 10$ . Without completing the hypothesis tests, explain why  $n_1 = n_2 = 100$  would result in a smaller *p*-value than  $n_1 = n_2 = 16$ .
- E.2. If the test  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  results in a small *p*-value, can we be confident that the regression model provides a good estimate of the response value for a given value of  $x_i$ ? Provide an explanation for your answer.
- E.3. What model assumptions (if any) need to be satisfied in order to calculate  $b_0$  and  $b_1$  in a simple linear regression model?
- E.4. Explain why the model  $y_i = \beta_0 + \beta_1 x_i$  is not appropriate, but  $\hat{y}_i = b_0 + b_1 x_i$  is appropriate.
- E.5. When there are only two levels (with equal sample sizes) being compared in an F-test, explain why  $\alpha_1 = -\alpha_2$ .
- E.6. Computer Games Again: Extending One-Way ANOVA to More Than Two Levels  
Data set: Games2

Assume that in the computer game study researchers were also interested in testing whether college students could play the game more quickly with their right or left hand. The data set Games2 shows a column Type2 with four types of games, based on distracter and which hand was used.

- Graph the data and compare the center and spread of the completion times for each of the four groups listed under Type2. Does any group have outliers or skewed data?
- Conduct an ANOVA with one explanatory variable that has the four levels listed under Type2. Notice that this data set has  $I = 4$  groups instead of 2 groups. However, the model and calculations are identical except that now  $i = 1, 2, 3, 4$  instead of  $i = 1, 2$  and now each group has a sample size of  $n_i = 10$ . The null hypothesis is  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$  versus the alternative  $H_a$ : at least one mean is different from another. Does the ANOVA show that a significant difference exists between group means?
- Create residual plots. Are the model assumptions satisfied?
- Are there any extraneous variables that might bias the results?
- Assuming that the data for this game were collected in the same way as the Game1 data, state your conclusions. Be sure to address random sampling and random allocation.

Note that we could consider modeling the completion times as a function of two explanatory variables: the hand used (right or left) and the game type (standard or with color distracter). This would require a two-way ANOVA analysis (discussed in later chapters).

### E.7. Paper Towel Experiment: Comparing Three Factor Levels

Data set: `PaperTowel`

As a final project in an introductory statistics class, several students decided to conduct a study to test the strength of paper towels. Television advertisements had claimed that a certain brand of paper towel was the strongest, and these students wanted to determine if there really was a difference. The students purchased rolls of towels at a local store and sampled 26 towels from 3 brands of paper towels: Bounty, Comfort, and Decorator.

Before any data were collected, these students determined that the following should be held as constant as possible throughout the study:

- 15 drops of water were applied to the center of each towel.
- Paper towels were selected that had the same size.
- The towels were held at all four corners by two people.
- Weights (10, 25, 50, or 100 grams) were slowly added to the center of each towel by a third person until it broke.
- The order in which the 26 paper towels were tested was randomized.

The file `PaperTowel` shows the breaking Strength of each towel. Breaking Strength is defined as the total weight that each towel successfully held. The next additional weight caused the towel to break.

- a. For this paper towel study, identify the explanatory variable, the observational units (experimental units or subjects), and the response variable. Write out (in words and symbols) appropriate null and alternative hypotheses.
- b. Graph the data and compare the center and spread of the breaking strength of each of the three brands. Does any group have outliers or skewed data?
- c. Conduct an ANOVA. The null hypothesis is  $H_0: \mu_1 = \mu_2 = \mu_3$  versus the alternative  $H_a$ : at least one mean is different from another. Does the ANOVA show that a significant difference exists between brands?
- d. Show that the equal variance assumption is violated in this study. Instead of using Strength as the response variable, use the natural log of Strength to conduct an ANOVA. The null hypothesis is still stated as  $H_0: \mu_1 = \mu_2 = \mu_3$  versus the alternative  $H_a$ : at least one mean is different from another. Does the ANOVA show that a significant difference exists between brands?
- e. Compare residual plots and the group standard deviations from the Strength and  $\ln(\text{Strength})$   $F$ -tests. Which test should be used to state your conclusions? Explain.
- f. Assume the students purchased one roll of Bounty paper towels and randomly selected 26 sheets from that one roll. The same holds true for other brands. What is the population for which the results of this study can hold?
- g. Assume the students randomly purchased 26 rolls of Bounty paper towels from various stores and randomly selected one sheet from each roll. The same holds true for other brands. What is the population for which the results of this study can hold?

### E.8. Dr. Benjamin Spock

Data set: `Jury`

Dr. Benjamin Spock was a well-known pediatrician who faced trial in 1968 for his activities as a Vietnam War protester. Specifically, he was charged with conspiring to violate the Selective Service Act by encouraging young men to resist the draft. As part of his defense, his counsel claimed that women were underrepresented on the jury. Women tend to be more sympathetic toward war protesters than men do. The defense counsel claimed that the judge had a history of choosing juries on which women were systematically underrepresented. At that time, jury members in Boston were chosen from a venire (a group of 30 to 200 individuals preselected from the population by the judge's clerk). By law, people were supposed to be selected for a venire at random. For Dr. Spock's trial, the judge's venire had 100 people and only 9 women, none of whom were selected to be on the actual jury.

Dr. Spock's defense counsel collected data on the percentages of women in venires from this judge's recent trials together with those of other judges in the Boston area.<sup>6</sup>

- a. Graph the data and compare the center and spread of the percentages of women in the venires of each group (Judge). Does any group have outliers or skewed data?
- b. Conduct an ANOVA with one explanatory variable and seven levels (judge). Notice that this data set now has groups with different sample sizes (i.e.,  $n_1 \neq n_2 \neq n_3$ , etc.). The null hypothesis is  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$  versus the alternative  $H_a$ : at least one mean is different from another. Does the ANOVA show that a significant difference exists between group means?
- c. Dr. Spock's defense counsel collected the data. What questions would you ask the defense about the data collection process before you were convinced that the data could be used in court?
- d. Assuming the data were properly collected, prepare a short statement advising the defense attorney whether the results of this study should be presented in court. If you suggest that the defense should use these results, provide an explanation of how these results are helpful. If you suggest that the defense should not use these results, provide an explanation of how these results are detrimental to the defense. It will be helpful to include a graph with a clear interpretation.

#### E.9. Tread Durability: Comparing Five Tire Brands

Data set: Tires

The data file Tires relates to five brands of tires chosen at random from local stores. Six tires of each brand were selected and placed in random order on a machine that tested tread durability in thousands of miles.

- a. Graph the data and compare the center and spread of the durability measurements for each group (Brand of tire). Does any group have outliers or skewed data?
- b. Conduct an ANOVA. Does the ANOVA show a significant difference exists between group means?
- c. Create residual plots and compare the group standard deviations. Are the model assumptions satisfied?
- d. Brand is used simply to identify different brands within the study. Explain why a simple linear regression model should not be used for the data, even though Brand could be treated as a quantitative variable.

#### E.10. Normal Probability Plots

Data set: Games1

- a. Create a normal probability plot and histogram of the residuals from the Games1 data (Question 26) and comment on the normality assumption for the random error term.
- b. Create a normal probability plot or a histogram of the observed responses (completion times  $y_i$ , where  $i = 1, 2, \dots, n$ ) from the computer game study.
- c. Explain why residuals should be used instead of the observed responses to test the normality assumption.

#### E.11. Comparing Normal Probability Plots

- a. Use software to draw five random samples, each of size  $n = 25$ , from an actual normal probability distribution with mean 0 and standard deviation 1. Create a normal probability plot and histogram for each “sample.”
- b. Use software to draw five random samples, each of size  $n = 50$ , from an actual normal probability distribution with mean 0 and standard deviation 1. Create a normal probability plot and histogram for each “sample.”
- c. Describe how changing sample size impacts our ability to determine if a sample is truly from a normal distribution.
- d. Use software to draw six random samples, each of size  $n = 50$ , from an actual normal probability distribution. Use means of -20 and 15 and standard deviations of 0.1, 3, and 300. Create a normal probability plot and histogram for each “sample.” Explain why the mean and standard deviation of a population do not impact the normal probability plots.

### E.12. Transforming ANOVA Data

Data set: `Hodgkins`

The data set `Hodgkins` contains plasma bradykininogen levels (in micrograms of bradykininogen per milliliter of plasma) in three types of subjects (normal, patients with active Hodgkin's disease, and patients with inactive Hodgkin's disease). The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation.

- Create individual value plots or side-by-side boxplots of the bradykininogen levels for each group. Compare the mean and standard deviation of each group. Do the data within each group look consistent with data from a normal population?
- Calculate the ANOVA model estimates in order to create a normal probability plot of the residuals. Do the error terms look consistent with data from a normal population?
- Transform the data by taking the log of the response. Create individual value plots or side-by-side boxplots of the transformed responses. Compare the mean and standard deviation of each group. Do the transformed data within each group look consistent with data from a normal population?
- Calculate the ANOVA model estimates using the transformed data in order to create a normal probability plot of the residuals. Do the error terms look consistent with data from a normal population? You may have to try out more than one transformation before you are satisfied with your answer to this part and Part C.
- Using the transformed data, calculate an ANOVA table,  $F$ -test, and  $p$ -value to determine if the three patient types differed in their mean bradykininogen levels.

### E.13. Transformations in Regression

Data set: `Weight`

In addition to brain weights and body weights, the `Weight` data set contains information on lifespan, gestation, and hours of sleep per day.

- Create a scatterplot of lifespan versus body weight with a regression line ( $\hat{y}_i = b_0 + b_1x_i$ ), a plot of residuals versus the explanatory variable, a plot of residuals versus predicted (or "fitted") values ( $\hat{y}$ ), and either a normal probability plot or a histogram of the residuals.
- Try various transformations of the explanatory and response variables to create a better linear regression model.
- Repeat Parts A and B using gestation as the response variable.
- Repeat Parts A and B using sleep per day as the response variable.

### E.14. Computer Games Again: Multiple Comparisons\*

Data set: `Games2`

- Conduct a  $t$ -test to determine if there is a difference in mean completion time between ColorLeft and StandardLeft. Report the  $p$ -value and your conclusions based on an individual  $\alpha$ -level = 0.05.
- Conduct a  $t$ -test to determine if there is a difference in mean completion time between ColorLeft and ColorRight. Report the  $p$ -value and your conclusions based on an individual  $\alpha$ -level = 0.05.
- Conduct a  $t$ -test to determine if there is a difference in mean completion time between ColorRight and StandardLeft. Report the  $p$ -value and your conclusions based on an individual  $\alpha$ -level = 0.05.
- List three other  $t$ -tests that could test for differences in mean completion time for different levels of Type2. In other words, what combinations of ColorRight, StandardRight, ColorLeft, and StandardLeft have not yet been tested?
- Assume all six hypothesis tests were compared. If each of these tests are independent and each of the tests used an  $\alpha$ -level = 0.05, what is the true probability that at least one of the tests would inappropriately reject the null hypothesis?
- What is the individual critical value if you use the Bonferroni method with an overall (familywise)  $\alpha$ -level = 0.10. Do any of your previous conclusions in Parts A through C change if you test for an overall (familywise) comparison? Explain.

---

\*Multiple comparisons are discussed in the extended activities in Chapter 1.

### E.15. Interpreting Back Transformations

Data set: Skinfold<sup>7</sup>

Celiac disease results in an inability to absorb carbohydrates and fats. Crohn's disease is another chronic intestinal disease in which the body's immune system attacks the intestines. Both Crohn's disease and celiac disease often result in malnutrition or impaired growth in children. A skinfold thickness measurement is a simple technique assessing body fat percentages by pinching the skin near the biceps and then using a calipers to measure the skin thickness.

- Transform the original data into  $\sqrt{\text{Thickness}}$ ,  $\ln(\text{Thickness})$ ,  $\log_{10}(\text{Thickness})$ , and the reciprocal ( $1/\text{Thickness}$ ).

For each transformation, conduct two-sample *t*-tests for differences between the mean of the Crohn's disease and celiac disease groups. Create a table with *p*-values and confidence intervals for the difference between the two means.

- Back transform the confidence intervals in Part A. For example, square the upper and lower bounds of the confidence interval created from the  $\sqrt{\text{Thickness}}$  data. Conduct the appropriate back transformation on the other three confidence intervals as well. Create a table of the four back-transformed confidence intervals. What do these back-transformed confidence intervals tell you?

Confidence limits for the difference between means often cannot be transformed back to the original scale. When reciprocal transformations have very small bounds, the back transformation provides unreasonably large bounds. For example, a skinfold transformation of  $1/0.022 = 45.5$  mm is not realistic. In addition, if the lower bound of a confidence interval is negative when the square-root transformation is used, back transforming the results (by squaring the bounds) will result in a confidence interval that does not contain zero. Thus, there are not reasonable practical interpretations of the back-transformed scales.

The log (this includes the natural log and  $\log_{10}$ ) is often preferable over other transformations because the back transformation has a practical interpretation. The  $\log_{10}$  back-transformed confidence interval (0.89, 2.03) provides results that can be interpreted, but not in the original units (millimeters). Notice that the confidence interval does not contain zero, but the results are not significant. Recall from your introductory statistics class that if a two-sided confidence interval contains zero, we fail to reject the null hypothesis for the corresponding two-sided hypothesis test. This is a 95% confidence interval for the ratio of the means. Thus, a value of one represents no difference between group means.

### E.16. Helicopters Again: Building Polynomial Regression Models

Data set: WingLength2

If you completed the helicopter research project, this series of questions will help you further investigate the use of regression to determine the optimal wing length of paper helicopters to maximize flight time. It is likely that your data from the project did not appear to lie along a straight line, but rather had a curved pattern. In this exercise, we will use a larger data set collected on six different wing lengths.

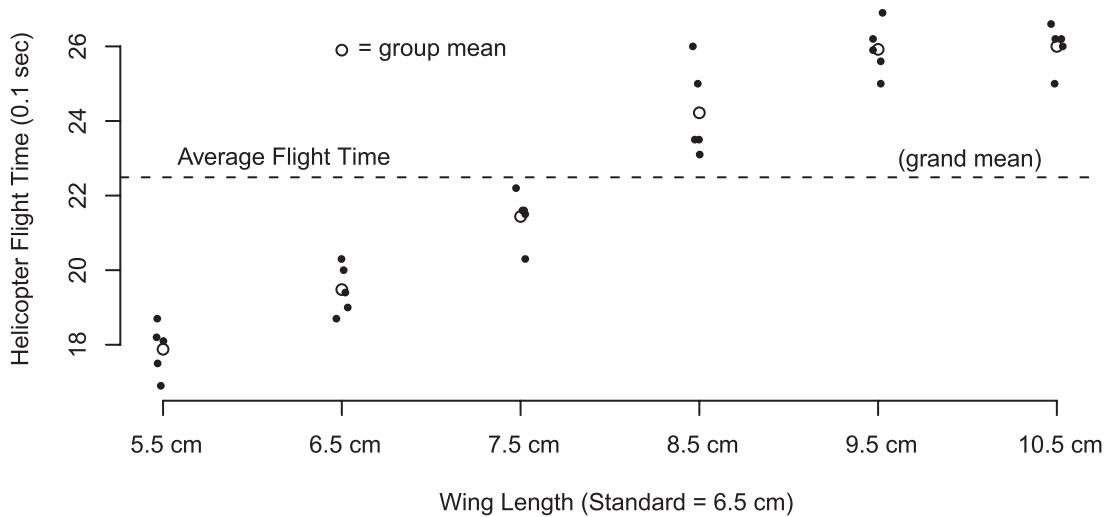
Figure 2.7 shows an individual value plot of the data, with five observations for each of the six wing length groups. The curved pattern is quite pronounced, and this makes sense. At some point, the wings are so long that the helicopter does not spin stably or is simply too heavy and falls faster. It appears that there is some optimal wing length around 9 or 10 cm.

- Using the WingLength2 data set, try several transformations of either the response or the explanatory variable to see if you can alleviate the problem of nonlinearity.
- It is likely that you cannot successfully find a transformation to solve the nonlinearity problem. A closer look at the residuals (Figure 2.8) helps to explain why. The residuals from the regression follow a somewhat sinusoidal pattern: down, then up, then down again. This is a pattern that is seen in a typical third-degree polynomial ( $y = ax^3 + bx^2 + cx + d$ ). To create the appropriate regression model, we will introduce a new method called *polynomial regression* (instead of simple linear regression).

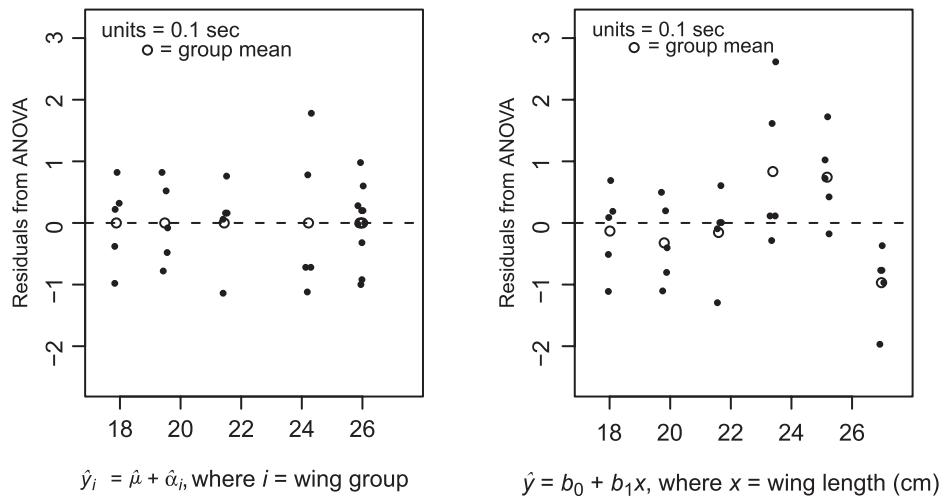
In polynomial regression, we can fit the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n \text{ where } \varepsilon_i \sim N(0, \sigma^2) \quad (2.10)$$

Using statistical computing software, fit the model in Equation (2.10) and report estimates for  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ .



**Figure 2.7** Flight times for paper helicopters when dropped from a height of 8 feet, each with a small paperclip attached at the bottom of the base of the helicopter.



**Figure 2.8** Residuals from an ANOVA and linear regression analysis of the flight times for paper helicopters from six groups: wing lengths 5.5 cm, 6.5 cm (standard), 7.5 cm, 8.5 cm, 9.5 cm, and 10.5 cm.

- Look at a plot of the residuals from the polynomial regression versus the predicted values ( $\hat{y}_i = b_0 + b_1x_i + b_2x_i^2 + b_3x_i^3$ ) and a normal probability plot of the residuals and comment on the validity of the regression model assumptions for these data.
- Given your answer to Part B and the fact that estimated flight times are considered to be a smooth (polynomial) function of wing length according to the relationship  $\hat{y}_i = b_0 + b_1x_i + b_2x_i^2 + b_3x_i^3$ , estimate the optimal wing length that will lead to maximum flight time. Note: Use calculus or visual inspection to identify a maximum that occurs within a reasonable range of wing lengths.
- Polynomial regression allows us to create a function relating the explanatory and response variables, and this can be useful for predicting responses for levels of the explanatory variable that were not actually measured as a part of the experiment. Of course, we should take care not to predict responses for levels of the explanatory variable that are quite far from those used in the experiment. The regression model may not extend past the domain we were able to analyze.

## Endnotes

---

1. Yogi Berra was an American League Baseball player and manager. This quote has also been attributed to computer scientist Jan L. A. van de Snepscheut.
2. J. R. Stroop, "Studies of Interference in Serial Verbal Reactions," *Journal of Experimental Psychology*, 18 (1935): 643–662.
3. The following articles provide more details on transformations: N. R. Draper and W. G. Hunter, "Transformations: Some Examples Revisited," *Technometrics*, 11 (1969): 23–40; G. E. P. Box and D. R. Cox, "An Analysis of Transformations (with Discussion)," *Journal of the Royal Statistical Society B*, 26 (1964): 211–252; M. S. Bartlett, "The Use of Transformations," *Biometrics*, 3 (1947): 39–52; J. L. Dolby, "A Quick Method for Choosing a Transformation," *Technometrics*, 5 (1963): 317–326.
4. Details of these methods are described in D.C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 3rd ed. (New York: Wiley, 2001).
5. F. Mosteller and J. W. Tukey, *Data Analysis and Regression* (Reading, MA: Addison-Wesley, 1977).
6. Data collected from H. Zeisel, "Dr. Spock and the Case of the Vanishing Women Jurors," *University of Chicago Law Review*, 37.1 (Autumn, 1969): 1–18.
7. Data from J. M. Bland and D. G. Altman, "Statistics Notes: The Use of Transformation When Comparing Two Means," *BMJ*, 312 (1996): 1153.
8. G. E. P. Box, "Teaching Engineers Experimental Design with a Paper Helicopter," *Quality Engineering*, 4 (1992): 453–459. Adapted with permission.

# Research Project: Building a Better Paper Helicopter

## Collecting the Data

Figure 2.9 shows a “standard” paper helicopter design.<sup>8</sup> The wings are folded in opposite directions. Often a small weight, such as a small paperclip, is attached to the base of the helicopter to ensure that each helicopter spins properly. If designed well, the helicopter will spin when dropped and stay in the air for a relatively long time.

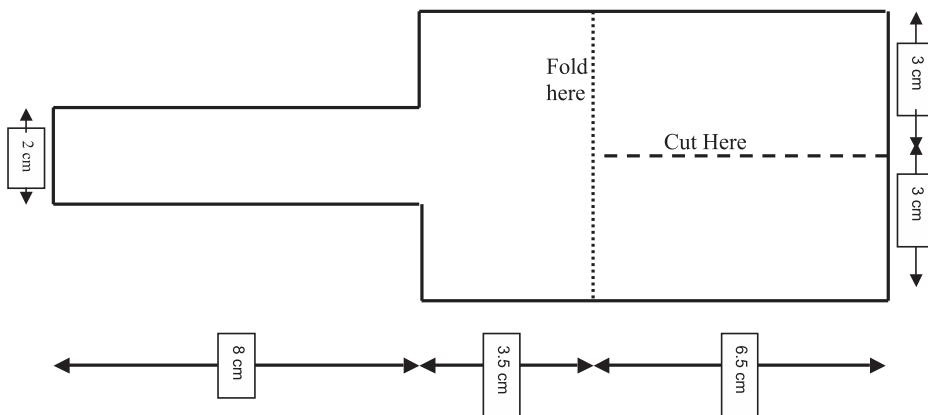


Figure 2.9 Standard paper helicopter design.

Many factors might influence the flight time (e.g., wing length, body width, body length, paper type, number of weights attached, type of weights attached, wing shape, number of folds, humidity, wind speed/drafts, how and where the helicopter is dropped). The objective of this project is to determine the optimal wing length needed to maximize the flight time of a paper helicopter. Throughout this project, we will also be comparing regression and ANOVA to better understand key differences between the methods.

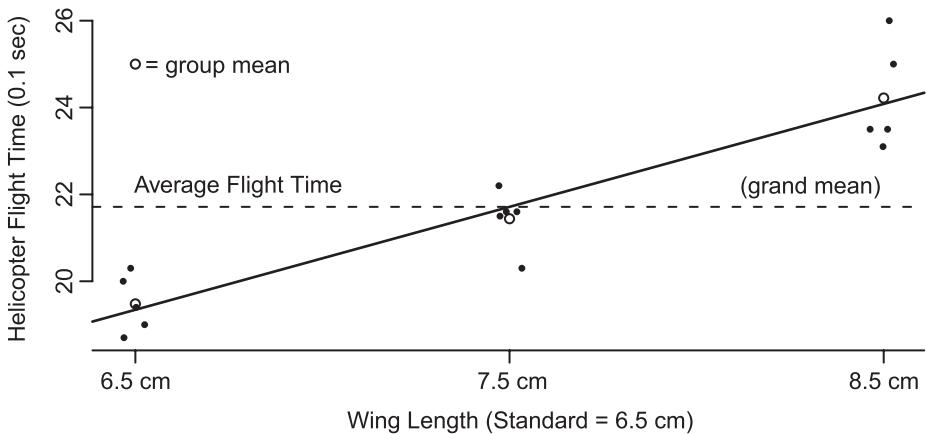
1. Cut out, fold, and drop the standard paper helicopter and several other shapes. Suggest fixed values for the following variables: body width, body length, paper type, number of weights attached, type of weights attached, wing shape, number of folds, humidity, wind speed/drafts, how and where the helicopter is dropped. These values should be held constant throughout your study to help avoid any potential biases.
2. After determining how the extraneous variables in the previous question will be held constant, suggest four potential wing lengths that seem “best” (i.e., have the longest flight times).

To give you an example of what your data might look like, we collected data on the following three wing lengths: 6.5 cm (standard), 7.5 cm, and 8.5 cm. Figure 2.10 shows an individual value plot of the data, with five observations per wing length group.

## Comparing ANOVA and Regression Models

Figure 2.10 indicates that the three groups have different means. We will use ANOVA and regression to test whether these differences are statistically significant. The ANOVA model for this project is similar to Equation (2.6), except that  $i = 1, 2, 3, 4$  and  $n_i = 5$  for each  $i$ :

$$y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \quad \text{for } i = 1, 2, 3, 4 \quad \text{and } j = 1, 2, 3, 4, 5 \quad \text{where } \varepsilon_{i,j} \sim N(0, \sigma^2) \quad (2.11)$$



**Figure 2.10** Flight times for paper helicopters when dropped from a height of 8 feet, each with a small paperclip attached at the bottom of the base of the helicopter.

Figure 2.10 also indicates that the three group means appear to lie along a line. A test that the slope is zero ( $H_0: \beta_1 = 0$ ) will also test for a difference in the three means. A linear regression model, as in Equation (2.2), may also be appropriate:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n \quad \text{where } \varepsilon_i \sim N(0, \sigma^2) \quad (2.12)$$

The data that we collected for Figure 2.10 have  $n = 15$ ,  $x_i = 6.5$  for five observations,  $x_i = 7.5$  for five more observations, and  $x_i = 8.5$  for another five observations. When you collect data with four groups,  $n = 20$  helicopter flights are recorded.

3. Using the four suggested wing lengths in Question 2, construct 20 helicopters—five helicopters for each wing length. Collect your own data on flight times. Be sure to randomize the order of your experiments before you begin. Record the wing length, flight time, and order in which the helicopters were dropped.
4. Explain the process that you used to do the randomization and why this is an important step in the experiment.
5. Create an individual value plot of your data. Does there appear to be an optimal wing length for attaining long helicopter flight times? If there are outliers, skewed data, or unequal variances, you may consider transforming the data. For example, take the log of the flight time.
6. Use statistical computing software to calculate an ANOVA table and test the hypothesis  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$  versus  $H_a$ : at least one of the group means is different from another. Report the  $MS_{\text{WingLength}}$ ,  $MSE$ ,  $F$ -statistic, and  $p$ -value. State your conclusion about the difference between group means.
7. If you found a significant difference in the  $F$ -test, use multiple comparison tests discussed in Chapter 1 to determine if one group is significantly better than others.
8. Use statistical computing software to calculate the least squares regression line. Conduct a hypothesis test to determine whether or not the slope coefficient is statistically significant. If you transformed the data in Question 6, use the same transformation for this question. At this time, do not try transformations to make the data fit a straight line.
9. Is the  $p$ -value for the ANOVA test that  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$  the same as or different from the  $p$ -value for the  $t$ -test that  $H_0: \beta_1 = 0$  (that the regression slope is zero)? Given your prior work in the computer game study, your answers may surprise you.

10. Notice that the statistical output also creates an ANOVA table for the regression model. Are the MSE values in Questions 6 and 8 the same or different? Are the degrees of freedom for error ( $df_{Error}$ ) the same for the ANOVA and linear regression models?

The ANOVA table for regression shown in Table 2.3 will be discussed in more detail in Chapter 3. At this time, notice that the tables shown in Table 2.1 and Table 2.3 are similar except for the degrees of freedom and sum of squares.

**Table 2.3** ANOVA table for a simple linear regression model, where  $y_i$  is an observed response,  $\hat{y}_i$  is the predicted response and  $\bar{y}$  is the grand mean.

Source	df	SS	MS	F-Statistic
<b>Group</b>	1	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\frac{SS_{Group}}{df_{Group}}$	$\frac{MS_{Group}}{MSE}$
<b>Error</b>	$n - 2$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\frac{SS_{Error}}{df_{Error}} = MSE$	
<b>Total</b>	$n - 1$	$\sum_{i=1}^n (y_i - \bar{y})^2$		

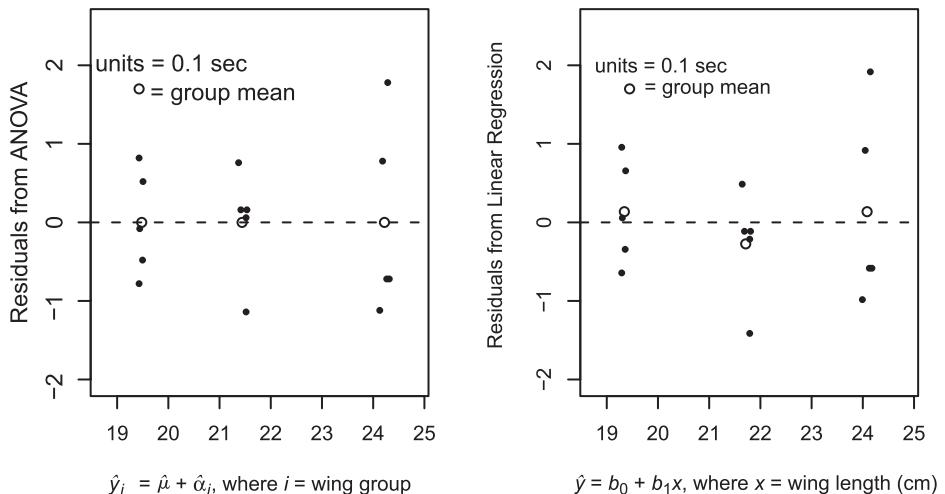
In both tests, the  $df_{Error}$  is the total sample size ( $n$ ) minus the number of free parameters estimated for the deterministic component of the model. For the ANOVA model, the number of parameters ( $\alpha_i$ , where  $i = 1, 2, \dots, I$ ) equals the number of groups ( $I$ ) being compared. In our example, we compared  $I = 3$  wing lengths, so  $df_{Error} = n - 3$ . Similarly, for your experiment with four wing lengths, the ANOVA  $df_{Error} = n - 4$ . For linear regression (with one explanatory variable), there are always just two parameters ( $\beta_0$  = intercept and  $\beta_1$  = slope), no matter how many levels or values the explanatory variable  $x$  takes on. Thus,  $df_{Error} = n - 2$ .

The second difference between Table 2.1 and Table 2.3 is the calculation of the sum of squares. The ANOVA test for linear regression forces the expected values to form a straight line. When there are just two groups, of course the linear regression can find a line that fits through the two means perfectly. This will generally not be true for data when there are more than two means, even if the regression model is correct.

The regression model assumes a linear relationship between the mean of the response and the value (i.e., level) of the explanatory variable. The ANOVA model calculates the mean response for each level of the explanatory variable, but does not require any particular pattern. The residual plots shown in Figure 2.11 are helpful in understanding this difference between the ANOVA and regression models (when comparing more than just two groups). Within each level, the residuals in the general ANOVA are centered around zero. However, the residuals for the regression model show that the regression model overestimates the time for the 7.5-cm group and slightly underestimates the time for the other two levels.

## Evaluating Your Own Model

11. Using computer software, draw the least squares regression line onto a scatterplot of flight time versus wing length from your own data.
12. Use the simple linear regression model to predict which wing length would cause the longest flight time. Is the answer similar to your answer to Question 6?
13. Do you think that the linear regression model is a good model for the relationship between flight time and wing length? Why or why not? Should you use the regression line to determine the optimal wing length? Why or why not?
14. Submit a 1- to 2-page summary of your work, including ANOVA table, graphs, and charts. This executive summary must also include a suggestion for the optimal helicopter wing length based on your experiment. You must prepare your optimal helicopter before class and have it ready to test during class.



**Figure 2.11** Residuals from an ANOVA and linear regression analysis of the flight times for paper helicopters from three groups: wing lengths 6.5 cm (standard), 7.5 cm, and 8.5 cm.

Despite the connections examined closely throughout the computer game study prior to this project, the ANOVA and linear regression models are indeed different, and one or the other might be more appropriate in different situations.

## Other Project Ideas

Several of the extended activities and end-of-chapter exercises can be used to motivate your own project ideas. In addition, there are many places where data are publicly available.

- Several other variables (such as base width, paper type, or base weight) could be tested to optimize flight time.
- Publically available data could be used to test a research question. Multiple websites are listed in the Chapter 3 project.
- The computer game projects provided in Chapters 4 and 5 could be used as long as only one explanatory and one response variable are tested.

# Multiple Regression: How Much Is Your Car Worth?

*Essentially, all models are wrong; some are useful.*

—George E. P. Box<sup>1</sup>

**M**ultiple regression is arguably the single most important method in all of statistics. Regression models are widely used in many disciplines. In addition, a good understanding of regression is all but essential for understanding many other, more sophisticated statistical methods.

This chapter consists of a set of activities that will enable you to build a multivariate regression model. The model will be used to describe the relationship between the retail price of 2005 used GM cars and various car characteristics, such as mileage, make, model, presence or absence of cruise control, and engine size. The set of activities in this chapter allows you to work through the entire process of model building and assessment, including

- Applying variable selection techniques
- Using residual plots to check for violations of model assumptions, such as heteroskedasticity, outliers, autocorrelation, and nonnormality distributed errors
- Transforming data to better fit model assumptions
- Understanding the impact of correlated explanatory variables
- Incorporating categorical explanatory variables into a regression model
- Applying  $F$ -tests in multiple regression

## 3.1 Investigation: How Can We Build a Model to Estimate Used Car Prices?

Have you ever browsed through a car dealership and observed the sticker prices on the vehicles? If you have ever seriously considered purchasing a vehicle, you can probably relate to the difficulty of determining whether that vehicle is a good deal or not. Most dealerships are willing to negotiate on the sale price, so how can you know how much to negotiate? For novices (like this author), it is very helpful to refer to an outside pricing source, such as the Kelley Blue Book, before agreeing on a purchase price.

For over 80 years, Kelley Blue Book has been a resource for accurate vehicle pricing. The company's Website, <http://www.kbb.com>, provides a free online resource where anyone can input several car characteristics (such as age, mileage, make, model, and condition) and quickly receive a good estimate of the retail price.

In this chapter, you will use a relatively small subset of the Kelley Blue Book database to describe the association of several explanatory variables (car characteristics) with the retail value of a car. Before developing a complex multiple regression model with several variables, let's start with a quick review of the simple linear regression model by asking a question: Are cars with lower mileage worth more? It seems reasonable to expect to see a relationship between mileage (number of miles the car has been driven) and retail value. The data set `Cars` contains the make, model, equipment, mileage, and Kelley Blue Book suggested retail price of several used 2005 GM cars.

### Activity A Simple Linear Regression Model

1. Produce a scatterplot from the `Cars` data set to display the relationship between mileage (`Mileage`) and suggested retail price (`Price`). Does the scatterplot show a strong relationship between `Mileage` and `Price`?
2. Calculate the least squares regression line,  $\text{Price} = b_0 + b_1(\text{Mileage})$ . Report the regression model, the  $R^2$  value, the correlation coefficient, the  $t$ -statistics, and  $p$ -values for the estimated **model coefficients** (the intercept and slope). Based on these statistics, can you conclude that `Mileage` is a strong indicator of `Price`? Explain your reasoning in a few sentences.
3. The first car in this data set is a Buick Century with 8221 miles. Calculate the residual value for this car (the observed retail price minus the expected price calculated from the regression line).

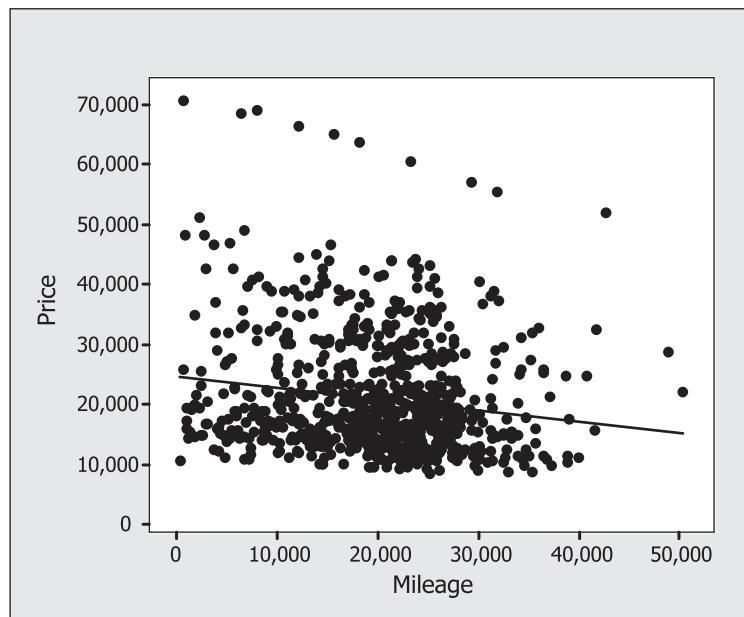
#### MATHEMATICAL NOTE

For any regression equation of the form  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ , the hypothesis test for the slope of the regression equation ( $\beta_1$ ) is similar to other  $t$ -tests discussed in introductory textbooks. (Mathematical details for this hypothesis test are described in Chapter 2.) To test the null hypothesis that the slope coefficient is zero ( $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$ ), calculate the following test statistic:

$$t = \frac{b_1 - \beta_1}{\hat{\sigma}_{b_1}} \quad (3.1)$$

where  $b_1$  is the estimated slope calculated from the data and  $\hat{\sigma}_{b_1}$  is an estimate of the standard deviation of  $b_1$ . Probability theory can be used to prove that if the regression model assumptions are true, the  $t$ -statistic in Equation (3.1) follows a  $t$ -distribution with  $n - 2$  degrees of freedom. If the sample statistic,  $b_1$ , is far away from  $\beta_1 = 0$  relative to the estimated standard deviation, the  $t$ -statistic will be large and the corresponding  $p$ -value will be small.

The  $t$ -statistic for the slope coefficient indicates that `Mileage` is an important variable. However, the  **$R^2$  value** (the percentage of variation explained by the regression line) indicates that the regression line is not very useful in predicting retail price. (A review of the  $R^2$  value is given in the extended activities.) As is always the case with statistics, we need to visualize the data rather than focus solely on a  $p$ -value. Figure 3.1 shows that the expected price decreases as mileage increases, but the observed points do not appear to be close to the regression line. Thus, it seems reasonable that including additional explanatory variables in the regression model might help to better explain the variation in retail price.



**Figure 3.1** Scatterplot and least squares regression model:  $\text{Price} = 24,765 - 0.1725(\text{Mileage})$ . The regression line shows that for each additional mile a car is driven, the expected price of the car decreases by about 17 cents. However, many points are not close to the regression line, indicating that the expected price is not an accurate estimate of the actual observed price.

In this chapter, you will build a linear combination of explanatory variables that explains the response variable, retail price. As you work through the chapter, you will find that there is not one technique, or “recipe,” that will give the best model. In fact, you will come to see that there isn’t just one “best” model for these data.

Unlike in most mathematics classes, where every student is expected to submit the one right answer to an assignment, here it is expected that the final regression models submitted by various students will be at least slightly different. While a single “best” model may not exist for these data, there are certainly many bad models that should be avoided. This chapter focuses on understanding the process of developing a statistical model. It doesn’t matter if you are developing a regression model in economics, psychology, sociology, or engineering—there are common key questions and processes that should be evaluated before a final model is submitted.

## 3.2 Goals of Multiple Regression

It is important to note that multiple regression analysis can be used to serve different goals. The goals will influence the type of analysis that is conducted. The most common goals of multiple regression are to describe, predict, or confirm.

- **Describe:** A model may be developed to describe the relationship between multiple explanatory variables and the response variable.
- **Predict:** A regression model may be used to generalize to observations outside the sample. Just as in simple linear regression, explanatory variables should be within the range of the sample data to predict future responses.
- **Confirm:** Theories are often developed about which variables or combination of variables should be included in a model. For example, is mileage useful in predicting retail price? Inferential techniques can be used to test if the association between the explanatory variables and the response could just be due to chance.

Theory may also predict the type of relationship that exists, such as “cars with lower mileage are worth more.” More specific theories can also be tested, such as “retail price decreases linearly with mileage.”

When the goal of developing a **multiple regression model** is **description or prediction**, the primary issue is often determining which variables to include in the model (and which to leave out). All potential explanatory variables can be included in a regression model, but that often results in a cumbersome model that is difficult to understand. On the other hand, a model that includes only one or two of the explanatory variables, such as the model in Figure 3.1, may be much less accurate than a more complex model. This tension between finding a simple model and finding the model that best explains the response is what makes it difficult to find a “best” model. The process of finding the most reasonable mix, which provides a relatively simple linear combination of explanatory variables, often resembles an exploratory artistic process much more than a formulaic recipe.

Including redundant or unnecessary variables not only creates an unwieldy model but also can lead to test statistics (and conclusions from corresponding hypothesis tests) that are less reliable. If explanatory variables are highly correlated, then their effects in the model will be estimated with more imprecision. This imprecision leads to larger standard errors and can lead to insignificant test results for individual variables that can be important in the model. Failing to include a relevant variable can result in biased estimates of the regression coefficients and invalid  $t$ -statistics, especially when the excluded variable is highly significant or when the excluded variable is correlated with other variables also in the model.<sup>2</sup>

### 3.3 Variable Selection Techniques to Describe or Predict a Response

If your objective is to describe a relationship or predict new response variables, variable selection techniques are useful for determining which explanatory variables should be in the model. For this investigation, we will consider the response to be the suggested retail price from Kelley Blue Book (the `Price` variable in the data). We may initially believe the following are relevant potential explanatory variables:

- Make (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn)
- Model (specific car for each previously listed Make)
- Trim (specific type of Model)
- Type (Sedan, Coupe, Hatchback, Convertible, or Wagon)
- Cyl (number of cylinders: 4, 6, or 8)
- Liter (a measure of engine size)
- Doors (number of doors: 2, 4)
- Cruise (1 = cruise control, 0 = no cruise control)
- Sound (1 = upgraded speakers, 0 = standard speakers)
- Leather (1 = leather seats, 0 = not leather seats)
- Mileage (number of miles the car has been driven)

#### Stepwise Regression

When a large number of variables are available, **stepwise regression** is an iterative technique that has historically been used to identify key variables to include in a regression model. For example, **forward stepwise regression** begins by fitting several single-predictor regression models for the response; one regression model is developed for each individual explanatory variable. The single explanatory variable (call it  $X_1$ ) that best explains the response (has the highest  $R^2$  value) is selected to be in the model.\*

In the next step, all possible regression models using  $X_1$  and exactly one other explanatory variable are calculated. From among all these two-variable models, the regression model that best explains the response is used to identify  $X_2$ . After the first and second explanatory variables,  $X_1$  and  $X_2$ , have been selected, the

---

\* An  $F$ -test is conducted on each of the models. The size of the  $F$ -statistic (and corresponding  $p$ -value) is used to evaluate the fit of each model. When models with the same number of predictors are compared, the model with the largest  $F$ -statistic will also have the largest  $R^2$  value.

process is repeated to find  $X_3$ . This continues until including additional variables in the model no longer greatly improves the model's ability to describe the response.\*

**Backward stepwise regression** is similar to forward stepwise regression except that it starts with all potential explanatory variables in the model. One by one, this technique removes variables that make the smallest contribution to the model fit until a “best” model is found.

While sequential techniques are easy to implement, there are usually better approaches to finding a regression model. Sequential techniques have a tendency to include too many variables and at the same time sometimes eliminate important variables.<sup>3</sup> With improvements in technology, most statisticians prefer to use more “global” techniques (such as best subset methods), which compare all possible subsets of the explanatory variables.

#### NOTE

Later sections will show that when explanatory variables are highly correlated, sequential procedures often leave out variables that explain (are highly correlated with) the response. In addition, sequential procedures involve numerous iterations and each iteration involves hypothesis tests about the significance of coefficients. Remember that with any multiple comparison problem, an  $\alpha$ -level of 0.05 means there is a 5% chance that each irrelevant variable will be found significant and may inappropriately be determined important for the model.

## Selecting the “Best Subset” of Predictors

A researcher must balance increasing  $R^2$  against keeping the model simple. When models with the same number of parameters are compared, the model with the highest  $R^2$  value should typically be selected. A larger  $R^2$  value indicates that more of the variation in the response variable is explained by the model. However,  $R^2$  never decreases when another explanatory variable is added. Thus, other techniques are suggested for comparing models with different numbers of explanatory variables.

Statistics such as the adjusted  $R^2$ , Mallows'  $C_p$ , and Akaike's and Bayes' information criteria are used to determine a “best” model. Each of these statistics includes a penalty for including too many terms. In other words, when two models have equal  $R^2$  values, each of these statistics will select the model with fewer terms. **Best subsets techniques** use several statistics to simultaneously compare several regression models with the same number of predictors.

#### MATHEMATICAL NOTE

$R^2$ , called the coefficient of determination, is the percentage of variation in the response variable that is explained by the regression line.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.2)$$

When the sum of the squared residuals  $(y_i - \hat{y}_i)^2$  are small compared to the total spread of the responses  $(\sum_{i=1}^n (y_i - \bar{y})^2)$ ,  $R^2$  is close to one.  $R^2 = 1$  indicates that the regression model perfectly fits the data.

Calculations for adjusted  $R^2$  and Mallows'  $C_p$  are

$$R_{\text{adj}}^2 = 1 - \left( \frac{n-1}{n-p} \right) \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.3)$$

\*An  $\alpha$ -level is often used to determine if any of the explanatory variables not currently in the model should be added to the model. If the  $p$ -value of all additional explanatory variables is greater than the  $\alpha$ -level, no more variables will be entered into the model. Larger  $\alpha$ -levels (such as  $\alpha = 0.2$ ) will include more terms while smaller  $\alpha$ -levels (such as  $\alpha = 0.05$ ) will include fewer terms.

$$C_p = (n - p) \left( \frac{\hat{\sigma}^2}{\hat{\sigma}_{\text{Full}}^2} \right) + (2p - n) \quad (3.4)$$

where  $n$  is the sample size,  $p$  is the number of coefficients in the model (including the constant term),  $\hat{\sigma}^2$  estimates the variance of the residuals in the model, and  $\hat{\sigma}_{\text{Full}}^2$  estimates the variance of the residuals in the full model (i.e., the model with all possible terms).

$R_{\text{adj}}^2$  is similar to  $R^2$ , except that  $R_{\text{adj}}^2$  includes a penalty when unnecessary terms are included in the model. Specifically, when  $p$  is larger,  $(n - 1)/(n - p)$  is larger, and thus  $R_{\text{adj}}^2$  is smaller.  $C_p$  also adjusts for the number of terms in the model. Notice that when the current model explains the data as well as the full model,  $\hat{\sigma}^2/\hat{\sigma}_{\text{Full}}^2 = 1$ . Then  $C_p = (n - p)(1) + 2p - n = p$ . Thus, the objective is to find the smallest  $C_p$  value that is close to the number of coefficients in the model. While any of these are appropriate for some models, adjusted  $R^2$  still tends to select models with too many terms. Mathematical details for these calculations are provided in the extended activities.

## Activity ▶ Comparing Variable Selection Techniques

4. Use the Cars data to conduct a stepwise regression analysis.
  - a. Calculate seven regression models, each with one of the following explanatory variables: Cyl, Liter, Doors, Cruise, Sound, Leather, and Mileage. Identify the explanatory variable that corresponds to the model with the largest  $R^2$  value. Call this variable  $X_1$ .
  - b. Calculate six regression models. Each model should have two explanatory variables,  $X_1$  and one of the other six explanatory variables. Find the two-variable model that has the highest  $R^2$  value. How much did  $R^2$  improve when this second variable was included?
  - c. Instead of continuing this process to identify more variables, use the software instructions provided to conduct a stepwise regression analysis. List each of the explanatory variables in the model suggested by the stepwise regression procedure.
5. Use the software instructions provided to develop a model using best subsets techniques. Notice that stepwise regression simply states which model to use, while best subsets provides much more information and requires the user to choose how many variables to include in the model. In general, statisticians select models that have a relatively low  $C_p$ , a large  $R^2$ , and a relatively small number of explanatory variables. (It is rare for these statistics to all suggest the same model. Thus, the researcher must choose a model based on his or her goals. The extended activities provide additional details about each of these statistics.) Based on the output from best subsets, which explanatory variables should be included in a regression model?
6. Compare the regression models in Questions 4 and 5.
  - a. Are different explanatory variables considered important?
  - b. Did the stepwise regression in Question 4 provide any indication that Liter could be useful in predicting Price? Did the best subsets output in Question 5 provide any indication that Liter might be useful in predicting Price? Explain why best subsets techniques can be more informative than sequential techniques.

Neither sequential nor best subsets techniques guarantee a best model. Arbitrarily using slightly different criteria will produce different models. Best subset methods allow us to compare models with a specific number of predictors, but models with more predictors do not always include the same terms as smaller models. Thus, it is often difficult to interpret the importance of any coefficients in the model.

Variable selection techniques are useful in providing a high  $R^2$  value while limiting the number of variables. When our goal is to develop a model to describe or predict a response, we are concerned not about the significance of each explanatory variable, but about how well the overall model fits.

If our goal involves confirming a theory, iterative techniques are not recommended. Confirming a theory is similar to hypothesis testing. Iterative variable selection techniques test each variable or combination of variables several times, and thus the  $p$ -values are not reliable. The stated significance level for a  $t$ -statistic is valid only if the data are used for a single test. If multiple tests are conducted to find the best equation, the actual significance level for each test for an individual component is invalid.

**Key Concept**

If variables are selected by iterative techniques, hypothesis tests should not be used to determine the significance of these same terms.

## 3.4 Checking Model Assumptions

The simple linear regression model discussed in introductory statistics courses typically has the following form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n \quad \text{where } \varepsilon_i \sim N(0, \sigma^2) \quad (3.5)$$

For this linear regression model, the mean response ( $\beta_0 + \beta_1 x_i$ ) is a linear function of the explanatory variable,  $x$ . The multiple linear regression model has a very similar form. The key difference is that now more terms are included in the model.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_{p-2} x_{p-2,i} + \beta_{p-1} x_{p-1,i} + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n \quad \text{where } \varepsilon_i \sim N(0, \sigma^2) \quad (3.6)$$

In this chapter,  $p$  represents the number of parameters in the regression model ( $\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1}$ ) and  $n$  is the total number of observations in the data. In this chapter, we make the following assumptions about the regression model:

- The model parameters  $\beta_0, \beta_1, \dots, \beta_{p-1}$  and  $\sigma$  are constant.
- Each term in the model is additive.
- The error terms in the regression model are independent and have been sampled from a single population (identically distributed). This is often abbreviated as iid.
- The error terms follow a normal probability distribution centered at zero with a fixed variance,  $\sigma^2$ .  
This assumption is denoted as  $\varepsilon_i \sim N(0, \sigma^2)$  for  $i = 1, \dots, n$ .

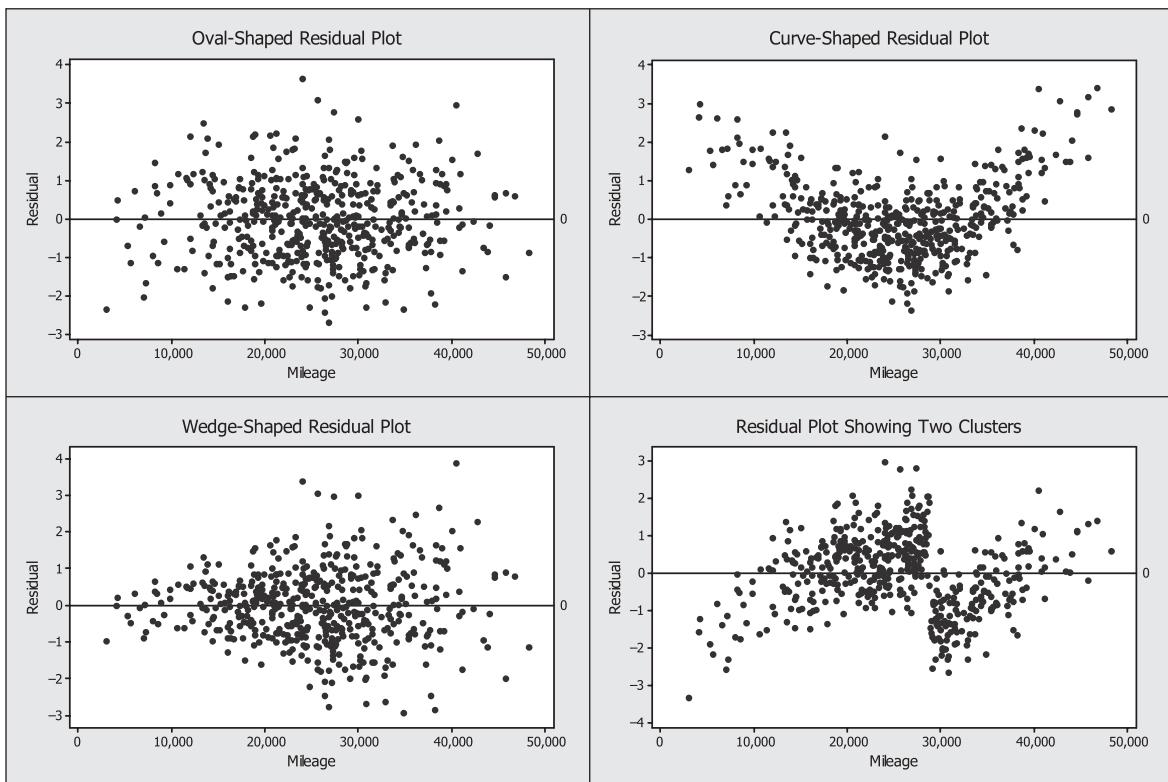
Regression assumptions about the error terms are generally checked by looking at the residuals from the data:  $y_i - \hat{y}_i$ . Here,  $y_i$  are the observed responses and  $\hat{y}_i$  are the estimated responses calculated by the regression model. Instead of formal hypothesis tests, plots will be used to visually assess whether the assumptions hold. The theory and methods are simplest when any scatterplot of residuals resembles a single, horizontal, oval balloon, but real data may not cooperate by conforming to the ideal pattern. An ornery plot may show a wedge, a curve, or multiple clusters. Figure 3.2 shows examples of each of these types of residual plots.

Suppose you held a cup full of coins and dropped them all at once. We hope to find residual plots that resemble the random pattern that would likely result from dropped coins (like the oval-shaped plot). The other three plots show patterns that would be very unlikely to occur by random chance. Any plot patterns that are not nice oval shapes suggest that the error terms are violating at least one model assumption, and thus it is likely that we have unreliable estimates of our model coefficients. The following section illustrates strategies for dealing with one of these unwanted shapes: a wedge-shaped pattern.

Note that in single-variable regression models, residual plots show the same information as the initial fitted line plot. However, the residual plots often emphasize violations of model assumptions better than the fitted line plot. In addition, multivariate regression lines are very difficult to visualize. Thus, residual plots are essential when multiple explanatory variables are used.

### Activity Heteroskedasticity

**Heteroskedasticity** is a term used to describe the situation where the variance of the error term is not constant for all levels of the explanatory variables. For example, in the regression equation  $\text{Price} = 24,765 - 0.173 (\text{Mileage})$ , the spread of the suggested retail price values around the regression line should be about the same whether mileage is 0 or mileage is 50,000. If heteroskedasticity exists in the model, the most common remedy is to transform either the explanatory variable, the response variable, or both in the hope that the transformed relationship will exhibit **homoskedasticity** (equal variances around the regression line) in the error terms.



**Figure 3.2** Common shapes of residual plots. Ideally, residual plots should look like a randomly scattered set of dropped coins, as seen in the oval-shaped plot. If a pattern exists, it is usually best to try other regression models.

7. Using the regression equation calculated in Question 5, create plots of the residuals versus each explanatory variable in the model. Also create a plot of the residuals versus the predicted retail price (often called a residual versus fit plot).
  - a. Does the size of the residuals tend to change as mileage changes?
  - b. Does the size of the residuals tend to change as the predicted retail price changes? You should see patterns indicating heteroskedasticity (nonconstant variance).
  - c. Another pattern that may not be immediately obvious from these residual plots is the right skewness seen in the residual versus mileage plot. Often economic data, such as price, are right skewed. To see the pattern, look at just one vertical slice of this plot. With a pencil, draw a vertical line corresponding to mileage equal to 8000. Are the points in the residual plots balanced around the line  $Y = 0$ ?
  - d. Describe any patterns seen in the other residual plots.
8. Transform the suggested retail price to  $\log(\text{Price})$  and  $\sqrt{\text{Price}}$ . Transforming data using roots, logarithms, or reciprocals can often reduce heteroskedasticity and right skewness. (Transformations are discussed in Chapter 2.)
 

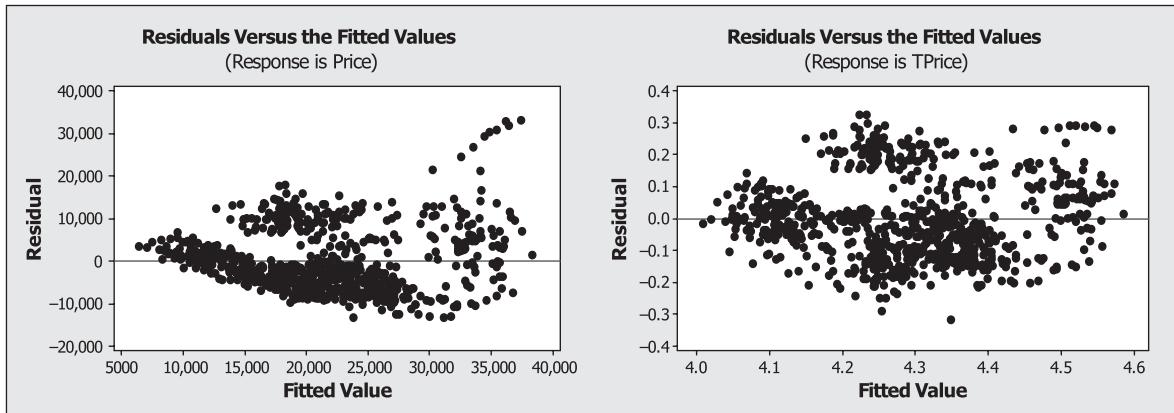
Create regression models and residual plots for these transformed response variables using the explanatory variables selected in Question 5.

  - a. Which transformation did the best job of reducing the heteroskedasticity and skewness in the residual plots? Give the  $R^2$  values of both new models.
  - b. Do the best residual plots correspond to the best  $R^2$  values? Explain.

While other transformations could be tried, throughout this investigation we will refer to the log-transformed response variable as `TPrice`.

Figure 3.3 shows residual plots that were created to answer Questions 7 and 8. Notice that when the response variable is `Price`, the residual versus fit plot has a clear wedge-shaped pattern. The residuals have

much more spread when the fitted value is large (i.e., expected retail price is close to \$40,000) than when the fitted value is near \$10,000. Using `TPrice` as a response did improve the residual versus fit plot. Although there is still a faint wedge shape, the variability of the residuals is much more consistent as the fitted value changes. Figure 3.3 reveals another pattern in the residuals. The following section will address why points in both plots appear in clusters.



**Figure 3.3** Residual versus fit plots using `Price` and `TPrice` (the  $\log_{10}$  transformation), as responses. The residual plot with `Price` as the response has a much stronger wedge-shaped pattern than the one with `TPrice`.

## Activity Examining Residual Plots Across Time/Order

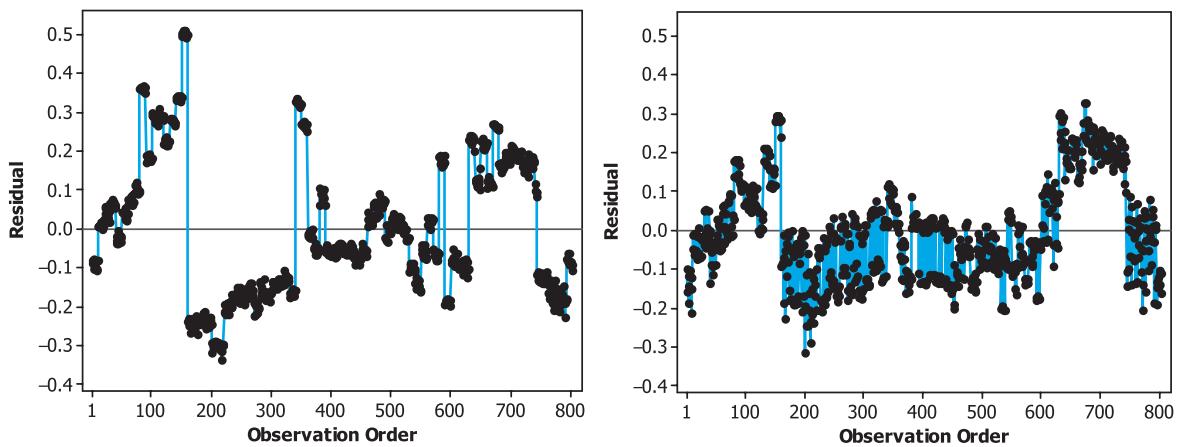
**Autocorrelation** exists when consecutive error terms are related. If autocorrelation exists, the assumption about the independence of the error terms is violated. To identify autocorrelation, we plot the residuals versus the order of the data entries. If the ordered plot shows a pattern, then we conclude that autocorrelation exists. When autocorrelation exists, the variable responsible for creating the pattern in the ordered residual plot should be included in the model.

### NOTE

Time (order in which the data were collected) is perhaps the most common source of autocorrelation, but other forms, such as spatial autocorrelations, can also be present. If time is indeed a variable that should be included in the model, a specific type of regression model, called a time series model, should be used.

9. Create a residual versus order plot from the `TPrice` versus `Mileage` regression line. Describe any pattern you see in the ordered residual plot. Apparently something in our data is affecting the residuals based on the order of the data. Clearly, time is not the influential factor in our data set (all of the data are from 2005). Can you suggest a variable in the `Cars` data set that may be causing this pattern?
10. Create a second residual versus order plot using `TPrice` as the response and using the explanatory variables selected in Question 5. Describe any patterns that you see in these plots.

While ordered plots make sense in model checking only when there is a meaningful order to the data, the residual versus order plots could demonstrate the need to include additional explanatory variables in the regression model. Figure 3.4 shows the two residual plots created in Questions 9 and 10. Both plots show that the data points are clearly clustered by order. However, there is less clustering when the six explanatory variables (`Mileage`, `Cyl`, `Doors`, `Cruise`, `Sound`, and `Leather`) are in the model. Also notice that the residuals tend



**Figure 3.4** Residual versus order plots using `TPrice` as the response. The first graph uses `Mileage` as the explanatory variable, and the second graph uses `Mileage`, `Cyl`, `Doors`, `Cruise`, `Sound`, and `Leather` as explanatory variables.

to be closer to zero in the second graph. Thus, the second graph (with six explanatory variables) tends to have estimates that are closer to the actual observed values.

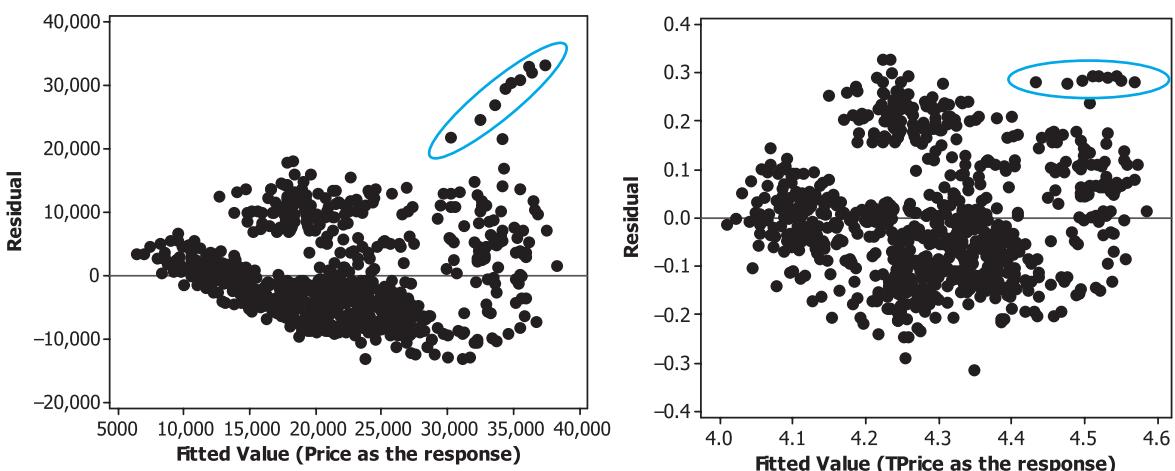
We do not have a time variable in this data set, so reordering the data would not change the meaning of the data. Reordering the data could eliminate the pattern; however, the clear pattern seen in the residual versus order plots should not be ignored because it indicates that we could create a model with a much higher  $R^2$  value if we could account for this pattern in our model. This type of autocorrelation is called **taxonomic autocorrelation**, meaning that the relationship seen in this residual plot is due to how the items in the data set are classified. Suggestions on how to address this issue are given in later sections.

## Activity Outliers and Influential Observations

11. Calculate a regression equation using the explanatory variables suggested in Question 5 and `Price` as the response. Identify any residuals (or cluster of residuals) that don't seem to fit the overall pattern in the residual versus fit and residual versus mileage plots. Any data values that don't seem to fit the general pattern of the data set are called **outliers**.
  - a. Identify the specific rows of data that represent these points. Are there any consistencies that you can find?
  - b. Is this cluster of outliers helpful in identifying the patterns that were found in the ordered residual plots? Why or why not?
12. Run the analysis with and without the largest cluster of potential outliers (the cluster of outliers corresponds to the Cadillac convertibles). Use `Price` as the response. Does the cluster of outliers influence the coefficients in the regression line?

If the coefficients change dramatically between the regression models, these points are considered **influential**. If any observations are influential, great care should be taken to verify their accuracy.

In addition to reducing heteroskedasticity, transformations can often reduce the effect of outliers. Figure 3.5 shows the residual versus fit plots using `Price` and `TPrice`, respectively. The cluster of outliers corresponding to the Cadillac convertibles is much more visible in the plot with the untransformed (`Price`) response variable. Even though there is still clustering in the transformed data, the residuals corresponding to the Cadillac convertibles are no longer unusually large.



**Figure 3.5** Residual versus fit plots using Price and TPrice as the response. The circled observations in the plot using Price are no longer clear outliers in the plot using TPrice.

In some situations, clearly understanding outliers can be more time consuming (and possibly more interesting) than working with the rest of the data. It can be quite difficult to determine if an outlier was accurately recorded or whether the outliers should be included in the analysis.

If the outliers were accurately recorded and transformations are not useful in eliminating them, it can be difficult to know what to do with them. The simplest approach is to run the analysis twice: once with the outliers included and once without. If the results are similar, then it doesn't matter if the outliers are included or not. If the results do change, it is much more difficult to know what to do. Outliers should never automatically be removed because they don't fit the overall pattern of the data. Most statisticians tend to err on the side of keeping the outliers in the sample data set unless there is clear evidence that they were mistakenly recorded. Whatever final model is selected, it is important to clearly state if you are aware that your results are sensitive to outliers.

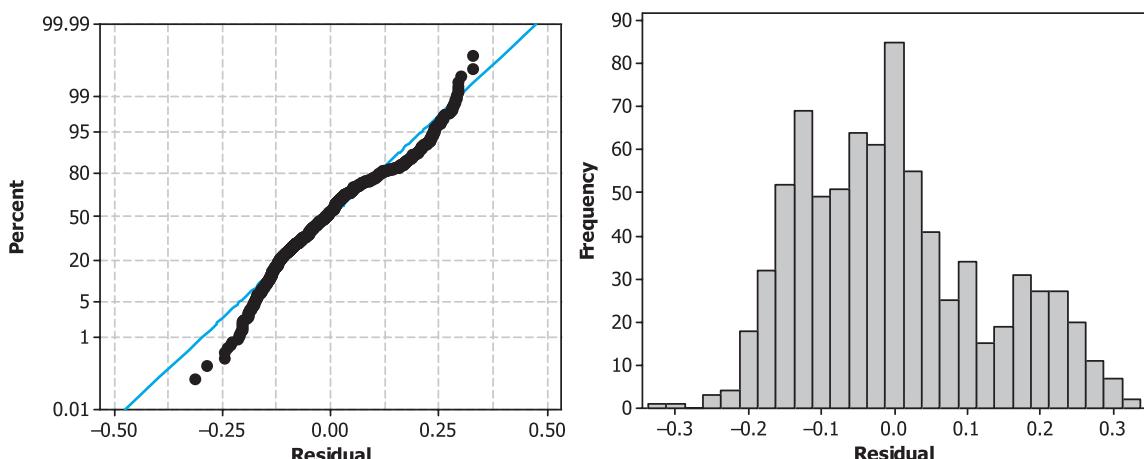
## Activity ▶ Normally Distributed Residuals

Even though the calculations of the regression model and  $R^2$  do not depend on the normality assumption, identifying patterns in residual plots can often lead to another model that better explains the response variable.

To determine if the residuals are normally distributed, two graphs are often created: a histogram of the residuals and a normal probability plot. Normal probability plots are created by sorting the data (the residuals in this case) from smallest to largest. Then the sorted residuals are plotted against a theoretical normal distribution. If the plot forms a straight line, the actual data and the theoretical data have the same shape (i.e., the same distribution). (Normal probability plots are discussed in more detail in Chapter 2.)

13. Create a regression line to predict TPrice from Mileage. Create a histogram and a normal probability plot of the residuals.
  - a. Do the residuals appear to follow the normal distribution?
  - b. Are the ten outliers visible on the normal probability plot and the histogram?

Figure 3.6 shows the normal probability plot using six explanatory variables to estimate TPrice. While the outliers are not visible, both plots still show evidence of lack of normality.



**Figure 3.6** Normal probability plot and histogram of residuals from the model using TPrice as the response and Mileage, Cyl, Doors, Cruise, Sound, and Leather as the explanatory variables.

At this time, it should be clear that simply plugging data into a software package and using an iterative variable selection technique will not reliably create a “best” model.

#### Key Concept

Before a final model is selected, the residuals should be plotted against fitted (estimated) values, observation order, the theoretical normal distribution, and each explanatory variable in the model. Table 3.1 shows how each residual plot is used to check model assumptions. If a pattern exists in any of the residual plots, the  $R^2$  value is likely to improve if different explanatory variables or transformations are included in the model.

**Table 3.1** Plots that can be used to evaluate model assumptions about the residuals.

Assumption	Plot
Normality	Histogram or normal probability plot
Zero mean	No plot (errors will always sum to zero under these models)
Equal variances	Plot of the residuals versus fits and each explanatory variable
Independence	Residuals versus order
Identically distributed	No plot (ensure each subject was sampled from the same population within each group)

#### Activity Correlation Between Explanatory Variables

**Multicollinearity** exists when two or more explanatory variables in a multiple regression model are highly correlated with each other. If two explanatory variables  $X_1$  and  $X_2$  are highly correlated, it can be very difficult to identify whether  $X_1$ ,  $X_2$ , or both variables are actually responsible for influencing the response variable,  $Y$ .

14. Create three regression models using Price as the response variable. In all three cases, provide the regression model,  $R^2$  value,  $t$ -statistic for the slope coefficients, and corresponding  $p$ -values.

- a. In the first model, use only `Mileage` and `Liter` as the explanatory variables. Is `Liter` an important explanatory variable in this model?
  - b. In the second model, use only `Mileage` and number of cylinders (`Cyl`) as the explanatory variables. Is `Cyl` an important explanatory variable in this model?
  - c. In the third model, use `Mileage`, `Liter`, and number of cylinders (`Cyl`) as the explanatory variables. How did the test statistics and *p*-values change when all three explanatory variables were included in the model?
15. Note that the  $R^2$  values are essentially the same in all three models in Question 14. The coefficients for `Mileage` also stay roughly the same for all three models—the inclusion of `Liter` or `Cyl` in the model does not appear to influence the `Mileage` coefficient. Depending on which model is used, we state that for each additional mile on the car, `Price` is reduced somewhere between \$0.152 to \$0.16. Describe how the coefficient for `Liter` depends on whether `Cyl` is in the model.
16. Plot `Cyl` versus `Liter` and calculate the correlation between these two variables. Is there a strong correlation between these two variables? Explain.

Recall that Question 4 suggested deleting `Liter` from the model. The goal in stepwise regression is to find the “best” model based on the  $R^2$  value. If two explanatory variables both impact the response in the same way, stepwise regression will rather arbitrarily pick one variable and ignore the other.

### Key Concept

Stepwise regression can often completely miss important explanatory variables when there is multicollinearity.

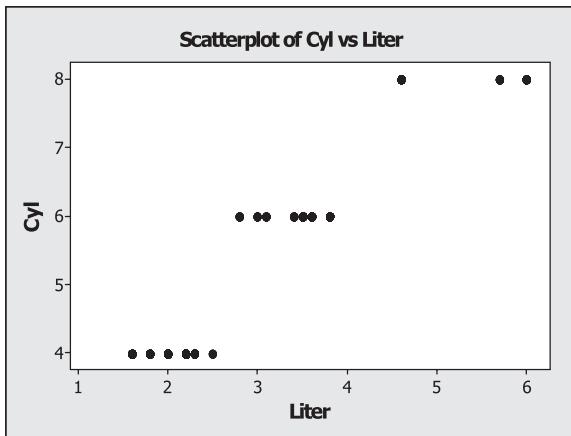
### ► MATHEMATICAL NOTE ▾

Most software packages can create a matrix plot of the correlations and corresponding scatterplots of all explanatory variables. This matrix plot is helpful in identifying any patterns of interdependence among the explanatory variables. An easy-to-apply guideline for determining if multicollinearity needs to be dealt with is to use the **variance inflation factor (VIF)**. VIF conducts a regression of each explanatory variable ( $X_i$ ) on the remaining explanatory variables, calculates the corresponding  $R^2$  value ( $R_i^2$ ), and then calculates the following function for each variable  $X_i$ :  $1/(1 - R_i^2)$ . If the  $R_i^2$  value is zero, VIF is one, and  $X_i$  is uncorrelated with all other explanatory variables. Montgomery, Peck, and Vining state, “Practical experience indicates that if any of the VIFs exceed 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.”<sup>4</sup>

Figure 3.7 shows that `Liter` and `Cyl` are highly correlated. Within the context of this problem, it doesn’t make physical sense to consider holding one variable constant while the other variable increases. In general, it may not be possible to “fix” a multicollinearity problem. If the goal is simply to describe or predict retail prices, multicollinearity is not a critical issue. Redundant variables should be eliminated from the model, but highly correlated variables that both contribute to the model are acceptable if you are not interpreting the coefficients. However, if your goal is to confirm whether an explanatory variable is associated with a response (test a theory), then it is essential to identify the presence of multicollinearity and to recognize that the coefficients are unreliable when it exists.

### Key Concept

If your goal is to create a model to describe or predict, then multicollinearity really is not a problem. Note that multicollinearity has very little impact on the  $R^2$  value. However, if your goal is to understand how a specific explanatory variable influences the response, as is often done when confirming a theory, then multicollinearity can cause coefficients (and their corresponding *p*-values when testing their significance) to be unreliable.



**Figure 3.7** Scatterplot showing a clear association between Liter and Cyl.

The following approaches are commonly used to address multicollinearity:

- *Get more information:* If it is possible, expanding the data collection may lead to samples where the variables are not so correlated. Consider whether a greater range of data could be collected or whether data could be measured differently so that the variables are not correlated. For example, the data here are only for GM cars. Perhaps the relationship between engine size in liters and number of cylinders is not so strong for data from a wider variety of manufacturers.
- *Re-evaluate the model:* When two explanatory variables are highly correlated, deleting one variable will not significantly impact the  $R^2$  value. However, if there are theoretical reasons to include both variables in the model, keep both terms. In our example, Liter and number of cylinders (Cyl) are measuring essentially the same quantity. Liter represents the volume displaced during one complete engine cycle; number of cylinders (Cyl) also is a measure of the volume that can be displaced.
- *Combine the variables:* Using other statistical techniques such as *principal components*, it is possible to combine the correlated variables “optimally” into a single variable that can be used in the model. There may be theoretical reasons to combine variables in a certain way. For example, the volume (size) and weight of a car are likely highly positively correlated. Perhaps a new variable defined as density = weight/volume could be used in a model predicting price rather than either of these individual variables.

In this investigation, we are simply attempting to develop a model that can be used to estimate price, so multicollinearity will not have much impact on our results. If we did re-evaluate the model in light of the fact that Liter and number of cylinders (Cyl) both measure displacement (engine size), we might note that Liter is a more specific variable, taking on several values, while Cyl has only three values in the data set. Thus, we might choose to keep Liter and remove Cyl in the model.

## 3.5 Interpreting Model Coefficients

While multiple regression is very useful in understanding the impacts of various explanatory variables on a response, there are important limitations. When predictors in the model are highly correlated, the size and meaning of the coefficients can be difficult to interpret. In Question 14, the following three models were developed:

$$\text{Price} = 9427 - 0.160(\text{Mileage}) + 4968(\text{Liter})$$

$$\text{Price} = 3146 - 0.152(\text{Mileage}) + 4028(\text{Cyl})$$

$$\text{Price} = 4708 - 0.154(\text{Mileage}) + 1545(\text{Liter}) + 2848(\text{Cyl})$$

The interpretation of model coefficients is more complex in multiple linear regression than in simple linear regression. It can be misleading to try to interpret a coefficient without considering other terms in the model. For example, when `Mileage` and `Liter` are the two predictors in a regression model, the `Liter` coefficient might seem to indicate that an increase of one in `Liter` will increase the expected price by \$4968. However, when `Cyl` is also included in the model, the `Liter` coefficient seems to indicate that an increase of one in `Liter` will increase the expected price by \$1545. The size of a regression coefficient and even the direction can change depending on which other terms are in the model.

In this investigation, we have shown that `Liter` and `Cyl` are highly correlated. Thus, it is unreasonable to believe that `Liter` would change by one unit but `Cyl` would stay constant. The multiple linear regression coefficients cannot be considered in isolation. Instead, the `Liter` coefficient shows how the expected price will change when `Liter` increases by one unit, after accounting for corresponding changes in all the other explanatory variables in the model.

#### Key Concept

In multiple linear regression, the coefficients tell us how much the expected response will change when the explanatory variable increases by one unit, *after accounting for corresponding changes in all other terms in the model*.

## 3.6 Categorical Explanatory Variables

As we saw in Question 9, there is a clear pattern in the residual versus order plot for the Kelley Blue Book car pricing data. It is likely that one of the categorical variables (`Make`, `Model`, `Trim`, or `Type`) could explain this pattern.

If any of these categorical variables are related to the response variable, then we want to add these variables to our regression model. A common procedure used to incorporate categorical explanatory variables into a regression model is to define **indicator variables**, also called **dummy variables**. Creating indicator variables is a process of mapping the one column (variable) of categorical data into several columns (indicator variables) of 0 and 1 data.

Let's take the variable `Make` as an example. The six possible values (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn) can be recoded using six indicator variables, one for each of the six makes of car. For example, the indicator variable for Buick will have the value 1 for every car that is a Buick and 0 for each car that is not a Buick. Most statistical software packages have a command for creating the indicator variables automatically.

### Activity Creating Indicator Variables

17. Create boxplots or individual value plots of the response variable `TPrice` versus the categorical variables `Make`, `Model`, `Trim`, and `Type`. Describe any patterns you see.
18. Create indicator variables for `Make`. Name the columns, in order, Buick, Cadillac, Chevrolet, Pontiac, SAAB, and Saturn. Look at the new data columns and describe how the indicator variables are defined. For example, list all possible outcomes for the Cadillac indicator variable and explain what each outcome represents.

Any of the indicator variables in Question 18 can be incorporated into a regression model.

However, if you want to include `Make` in its entirety in the model, do not include all six indicator variables. Five will suffice because there is complete redundancy in the sixth indicator variable. If the first five indicator variables are all 0 for a particular car, we automatically know that this car belongs to the sixth category. Below, we will leave the Buick indicator variable out of our model. The coefficient for an indicator variable is an estimate of the average amount by which the response variable will change. For example, the estimated coefficient for the `Saturn` variable is an estimate of the average difference in `TPrice` when the car is a `Saturn` rather than a `Buick` (after adjusting for corresponding changes in all other terms in the model).

### ► MATHEMATICAL NOTE ▶

For any categorical explanatory variable with  $g$  groups, only  $g - 1$  terms should be included in the regression model. Most software packages use matrix algebra to develop multiple regression models. If all  $g$  terms are in the model, explanatory variables will be 100% correlated (you can exactly predict the value of one variable if you know the other variables) and the needed matrix inversion cannot be done. If the researcher chooses to leave all  $g$  terms in the model, most software packages will arbitrarily remove one term so that the needed matrix calculations can be completed.

It may be tempting to simplify the model by including only a few of the most significant indicator variables. For example, instead of including five indicator variables for `Make`, you might consider only using `Cadillac` and `SAAB`. Most statisticians would recommend against this. By limiting the model to only indicator variables that are significant in the sample data set, we can overfit the model. Models are **overfit** when researchers overwork a data set in order to increase the  $R^2$  value. For example, a researcher could spend a significant amount of time picking a few indicator variables from `Make`, `Model`, `Trim`, and `Type` in order to find the best  $R^2$  value. While the model would likely estimate the mean response well, it unlikely to accurately predict new values of the response variable.

This is a fairly nuanced point. The purpose of variable selection techniques is to select the variables that best explain the response variable. However, overfitting may occur if we break up categorical variables into smaller units and then pick and choose among the best parts of those variables. (The Model Validation section in the extended activities discusses this topic in more detail.)

## Activity ◀ Building Regression Models with Indicator Variables

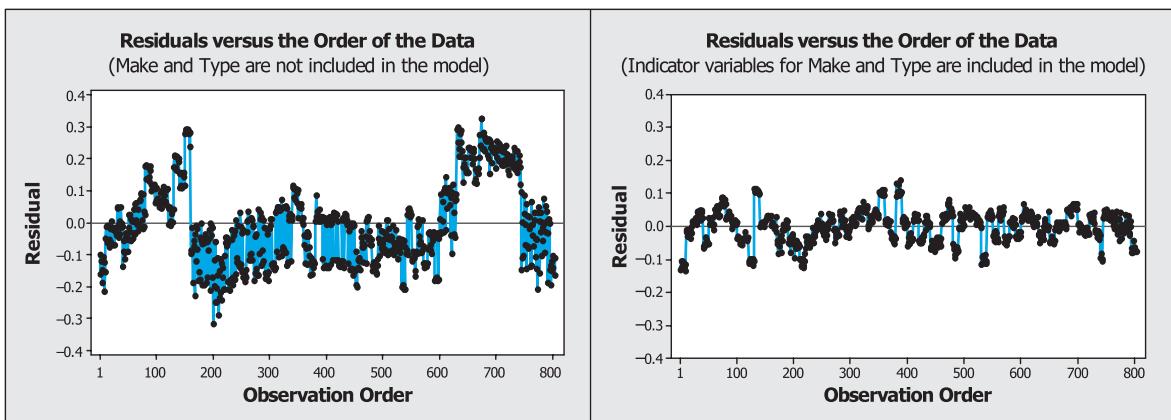
19. Build a new regression model using `TPrice` as the response and `Mileage`, `Liter`, `Saturn`, `Cadillac`, `Chevrolet`, `Pontiac`, and `SAAB` as the explanatory variables. Explain why you expect the  $R^2$  value to increase when you add terms for `Make`.
20. Create indicator variables for `Type`. Include the `Make` and `Type` indicator variables, plus the variables `Liter`, `Doors`, `Cruise`, `Sound`, `Leather`, and `Mileage`, in a model to predict `TPrice`. Remember to leave at least one category out of the model for the `Make` and `Type` indicator variables (e.g., leave `Buick` and `Hatchback` out of the model). Compare this regression model to the other models that you have fit to the data in this investigation. Does the normal probability plot suggest that the residuals could follow a normal distribution? Describe whether the residual versus fit, the residual versus order, and the residual versus each explanatory variable plots look more random than they did in earlier problems.

The additional categorical variables are important in improving the regression model. Figure 3.8 shows that when `Make` and `Type` are not in the model, the residuals are clustered. When `Make` and `Type` are included in the model, the residuals appear to be more randomly distributed. By incorporating `Make` and `Type`, we have been able to explain some of variability that was causing the clustering. In addition, the sizes of the residuals are much smaller. Smaller residuals indicate a better fitting model and a higher  $R^2$  value.

Even though `Make` and `Type` improved the residual plots, there is still clustering that might be improved by incorporating `Model` into the regression equation. However, if the goal is simply to develop a model that accurately estimates retail prices, the  $R^2$  value in Question 20 is already fairly high. Are there a few more terms that can be added to the model that would dramatically improve the  $R^2$  value?

To determine a final model, you should attempt to maximize the  $R^2$  value while simultaneously keeping the model relatively simple. For your final model, you should comment on the residual versus fit, residual versus order, and any other residual plots that previously showed patterns. If any pattern exists in the residual plots, it may be worth attempting a new regression model that will account for these patterns. If the regression model can be modified to address the patterns in the residuals, the  $R^2$  value will improve. However, note that the  $R^2$  value is already fairly high. It may not be worth making the model more complex for only a slight increase in the  $R^2$  value.

21. Create a regression model that is simple (i.e., does not have too many terms) and still accurately predicts retail price. Validate the model assumptions. Look at residual plots and check for heteroskedasticity, multicollinearity, autocorrelation, and outliers. Your final model should not have significant clusters, skewness, outliers, or heteroskedasticity appearing in the residual plots. Submit your suggested least squares regression formula along with a limited number of appropriate graphs that provide justification for your model. Describe why you believe this model is “best.”



**Figure 3.8** Residual versus order plots show that incorporating the indicator variables into the regression model improves the random behavior and reduces the sizes of the residuals.

## 3.7 What Can We Conclude from the 2005 GM Car Study?

The data are from an observational study, not an experiment. Therefore, even though the  $R^2$  value reveals a strong relationship between our explanatory variables and the response, a significant correlation (and thus a significant coefficient) does not imply a causal link between the explanatory variable and the response. There may be theoretical or practical reasons to believe that mileage (or any of the other explanatory variables) causes lower prices, but the final model can be used only to show that there is an association.

Best subsets and residual graphs were used to develop a model that is useful for describing or predicting the retail price based on a function of the explanatory variables. However, since iterative techniques were used, the  $p$ -values corresponding to the significance of each individual coefficient are not reliable.

For this data set, cars were randomly selected within each make, model, and type of 2005 GM car produced, and then suggested retail prices were determined from Kelley Blue Book. While this is not a simple random sample of all 2005 GM cars actually on the road, there is still reason to believe that your final model will provide an accurate description or prediction of retail price for used 2005 GM cars. Of course, as time goes by, the value of these cars will be reduced and updated models will need to be developed.

## A Closer Look    Multiple Regression

### 3.8 F-Tests for Multiple Regression

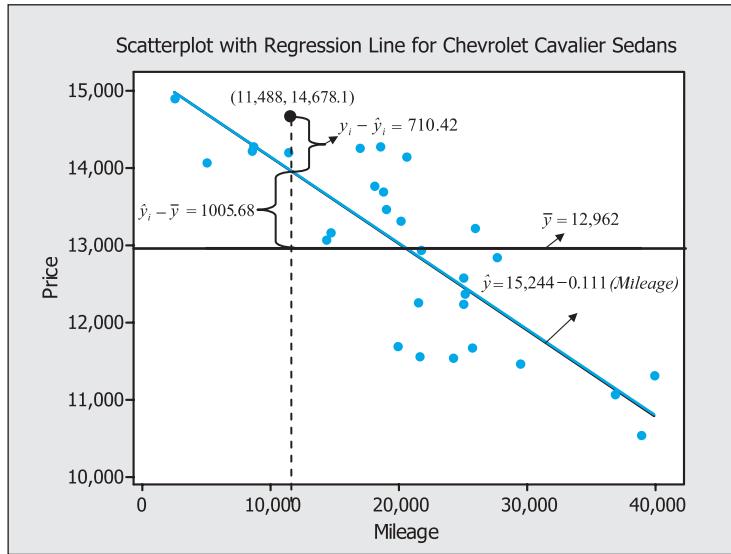
#### Decomposition of Sum of Squares

Many of the calculations involved in multiple regression are very closely related to those for the simple linear regression model. Simple linear regression models have the advantage of being easily visualized with scatterplots. Thus, we start with a simple linear regression model to derive several key equations used in multiple regression.

Figure 3.9 shows a scatterplot and regression line for a subset of the used 2005 Kelly Blue Book Cars data. The data set is restricted to just Chevrolet Cavalier Sedans. In this scatterplot, one specific observation is highlighted: the point where Mileage is 11,488 and the observed Price is  $y_i = 14,678.1$ .

In Figure 3.9, we see that for any individual observation the total deviation ( $y_i - \bar{y}$ ) is decomposed into two parts:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \quad (3.7)$$



**Figure 3.9** Scatterplot and regression line for Chevrolet Cavalier Sedans:  
Price = 15,244 – 0.111(Mileage).

Using our highlighted observation (11,488, 14,678.1), we see that

$$\begin{aligned} y_i - \bar{y} &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ 14,678.1 - 12,962 &= (14,678.1 - 13,967.68) + (13,967.68 - 12,962) \\ 1716.1 &= 710.42 + 1005.68 \end{aligned}$$

Squaring both sides of Equation (3.7) and then summing over all observations results in

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \end{aligned} \quad (3.8)$$

The key point of the previous calculations is to show that the total variability in the response,  $\sum_{i=1}^n (y_i - \bar{y})^2$ , can be decomposed into the following:

$$\begin{array}{ccc} \text{Total} & \text{Residual} & \text{Regression} \\ \text{sum of} & \text{sum of} & \text{sum of} \\ \text{squares} & \text{squares} & \text{squares} \\ (\text{SST}) & (\text{SSE}) & (\text{SSR}) \\ \downarrow & \downarrow & \downarrow \\ \sum_{i=1}^n (y_i - \bar{y})^2 & = & \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \end{array} \quad (3.9)$$

#### ► MATHEMATICAL NOTE ▼

To show that Equation (3.8) is true, we can write

$$2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 2 \sum_{i=1}^n \hat{y}_i(y_i - \hat{y}_i) - 2\bar{y} \sum_{i=1}^n (y_i - \hat{y}_i)$$

#### ► MATHEMATICAL NOTE ▼

Recall that the sum of residuals,  $\sum_{i=1}^n (y_i - \hat{y}_i)$ , equals zero. In addition, it can be shown that the sum of the residuals, weighted by the corresponding predicted value, always sums to zero:  $\sum_{i=1}^n \hat{y}_i(y_i - \hat{y}_i) = 0$ . (See Questions 25 through 29.)

## Extended Activity

### A Closer Look at Least Squares Regression Equations

Data set: Cavalier

Note that calculus is required for Activity Questions 25 through 29.

22. Create a regression model to predict Price from Mileage for the Cavalier data. Calculate the total sum of squares (SST), residual sum of squares (SSE), and regression sum of squares (SSR). Verify that  $SST = SSE + SSR$ .
23. Show that  $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$  for the model given in the previous question.
24. Using your final model in Question 21, calculate the total sum of squares (SST), residual sum of squares (SSE), and regression sum of squares (SSR). Verify that  $SST = SSE + SSR$ .
25. Set the partial derivative of the residual sum of squares with respect to  $b_0$  to zero, to show that  $b_0 n + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ . In other words, take the first derivative of  $\sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$  with respect to  $b_0$  and then set the first derivative equal to zero.
26. Set the partial derivative of the residual sum of squares with respect to  $b_1$  to zero, to show that  $b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$ .
27. The equations in Questions 25 and 26 are called the **normal equations** for simple linear regression. Use the normal equations to derive the least squares regression coefficients,  $b_0$  and  $b_1$ .
28. Use the fact that  $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$  and  $\hat{y}_i = b_0 + b_1 x_i$  to show that  $\sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) = b_1 (\sum_{i=1}^n x_i y_i - b_0 \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2)$ .
29. Use Questions 26 and 28 to show that  $\sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) = 0$ .

### The Analysis of Variance Table

The objective of regression is to create a model that best fits the observed points. Least squares regression models define a “best fit” as a model that minimizes the sum of squared residual values,  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ .

The coefficient of determination,  $R^2$ , is the percentage of variation in the response variable that is explained by the regression line:

$$R^2 = \frac{\text{variability of the fitted regression values}}{\text{total variability of the responses}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.10)$$

#### Key Concept

The coefficient of determination,  $R^2$ , is a measure of the usefulness of the explanatory variables in the model. If the explanatory variables are useful in predicting the response, the residual sum of squares,  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ , is small compared to the total spread of the responses,  $\sum_{i=1}^n (y_i - \bar{y})^2$ . In other words, the amount of variability explained by the regression model,  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ , is a large proportion of the total variability of the responses.

The sum of squares calculations are often summarized in an analysis of variance (ANOVA) table, as shown in Table 3.2.

**Table 3.2** ANOVA table for a least squares regression model, where  $n$  is the number of observations and  $p$  is the number of terms in the model (including the constant term).

Source	df	SS	MS	F-Statistic
Regression	$p - 1$	$\text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\frac{\text{SSR}}{\text{df}_{\text{Regr}}}$	$\frac{\text{MS}_{\text{Regr}}}{\text{MSE}}$
Error	$n - p$	$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\text{MSE} = \frac{\text{SSE}}{\text{df}_{\text{Error}}} = \hat{\sigma}^2$	
Total	$n - 1$	$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$		

## Testing the Significance of a Regression Model

Once a model has been developed, we are often interested in testing if there is a relationship between the response and the set of all explanatory terms in the model. To conduct an **overall test of model adequacy**, we test the following null and alternative hypotheses:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$$H_a: \text{at least one of the coefficients is not } 0$$

Notice that the  $\beta_0$  term in our regression model is not included in the null or the alternative hypothesis. Table 3.2 provides the details for the calculation of the  $F$ -statistic:

$$F = \frac{\text{MS}_{\text{Regr}}}{\text{MSE}} = \frac{\text{SSR}/(p - 1)}{\text{SSE}/(n - p)} \quad (3.11)$$

This statistic follows an  $F_{p-1, n-p}$  distribution, where  $n$  is the number of observations and  $p$  is the number of terms in the model (including  $\beta_0$ ). The same assumptions about the error terms,  $\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ , need to be checked before conducting the hypothesis test.

### NOTE

There are no model assumptions needed about the error terms to calculate estimates of the coefficients. However, all the model assumptions should be checked before conducting a hypothesis test.

## The Extra Sum of Squares F-Test

We are often interested in testing the contribution of a particular variable (or subset of variables) to the regression sum of squares. The **extra sum of squares F-test** can test the contribution of a specific set of variables by comparing the residuals of a full and a reduced model.

Suppose a model has been fit with  $k$  terms—we call this a **full model**. We may hypothesize that only  $p < k$  terms really contribute to the regression model—we call this smaller model the **reduced model**. In this situation, we want to test whether

$$H_0: \beta_p = \beta_{p+1} = \dots = \beta_{k-1} = 0$$

$$H_a: \text{at least one of the coefficients is not } 0$$

The previous ANOVA  $F$ -test can be modified to provide an  $F$ -test for this hypothesis. Notice that this hypothesis test makes no assumptions about the other terms,  $\beta_0, \beta_1, \dots, \beta_{p-1}$ , in the model. In addition, *every term in the reduced model must also be in the full model*.

$$F = \frac{(\text{SSR}_{\text{Full}} - \text{SSR}_{\text{Reduced}})/(k - p)}{\text{MSE}_{\text{Full}}} \quad (3.12)$$

This statistic follows an  $F$ -distribution with  $k - p$  and  $n - k$  degrees of freedom. The extra sum of squares  $F$ -test determines whether the difference between the sum of squared residuals in the full and reduced models is so large that it is unlikely to occur by chance.

## Extended Activity Testing Multiple Coefficients

Data set: Cavalier

Consider the Cavalier data set and the regression model  $y = \beta_0 + \beta_1(\text{Mileage}) + \beta_2(\text{Cruise}) + \varepsilon$ .

30. Submit the ANOVA table,  $F$ -statistic, and  $p$ -value to test the hypothesis  $H_0: \beta_1 = \beta_2 = 0$  versus  $H_a: \text{at least one of the coefficients is not } 0$ .
31. Conduct an extra sum of squares test to determine if `Trim` is useful. More specifically, use the reduced model in the previous question and the full model

$$y = \beta_0 + \beta_1(\text{Mileage}) + \beta_2(\text{Cruise}) + \beta_3(\text{LS Sport Sedan 4D}) + \beta_4(\text{Sedan 4D}) + \varepsilon$$

-  to test the hypothesis  $H_0: \beta_3 = \beta_4 = 0$  versus  $H_a: \text{at least one of the coefficients is not } 0$ .

## 3.9 Developing a Model to Confirm a Theory

If the goal is to confirm a theoretical relationship, statisticians tend to go through the following steps to identify an appropriate theoretical model.

- Verify that the response variable provides the information needed to address the question of interest. What are the range and variability of responses you expect to observe? Is the response measurement precise enough to address the question of interest?
- Investigate all explanatory variables that may be of importance or could potentially influence your results. Note that some terms in the model will be included even though the coefficients may not be significant. In most studies, there is often prior information or a theoretical explanation for the relationship between explanatory and response variables. Nonstatistical information is often essential in developing good statistical models.
- For each of the explanatory variables that you plan to include in the model, describe whether you would expect a positive or negative correlation between that variable and the response variable.
- Use any background information available to identify what other factors are assumed to be controlled within the model. Could measurements, materials, and the process of data collection create unwanted variability? Identify any explanatory variables that may influence the response; then determine if information on these variables can be collected and if the variables can be controlled throughout the study. For example, in the Kelley Blue Book data set, the condition of the car was assumed to be the same for all cars. The data were collected for GM cars with model year 2005. Since these cars were relatively new and the cars were considered to be in excellent condition, any model we create for these data would not be relevant for cars that had been in any type of accident.
- What conditions would be considered normal for this type of study? Are these conditions controllable? If a condition changed during the study, how might it impact the results?

After a theoretical model is developed, regression analysis is conducted one time to determine if the data support the theories.

### Key Concept

The same data should not be looked at both to develop a model and to test it.

## Extended Activity Testing a Theory on Cars

Data set: Cars

Assume that you have been asked to determine if there is an association between each of the explanatory variables and the response in the Cars data set.

32. Use any background information you may have (not the Cars data set) to predict how each explanatory variable (except Model and Trim) will influence TPrice. For example, will Liter or Mileage have a positive or negative association with TPrice? List each Make and identify which will impact TPrice most and in which direction.
33. Identify which factors are controlled in this data set. Can you suggest any factors outside the provided data set that should have been included? If coefficients are found to be significant (have small *p*-values), will these relationships hold for all areas in the United States? Will the relationships hold for 2004 or 2001 cars?
34. Run a regression analysis to test your hypothesized model. Which variables are important in your model? Did you accurately estimate the direction of each relationship? Note that even if a variable is not significant, it is typically kept in the model if there is a theoretical justification for it.

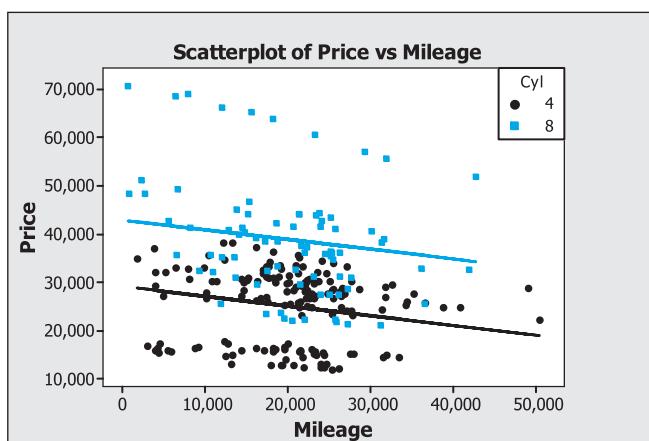
### 3.10 Interaction and Terms for Curvature

In addition to using the variables provided in a data set, it is often beneficial to create new variables that are functions of the existing explanatory variables. These new explanatory variables are often quadratic ( $X^2$ ), cubic ( $X^3$ ), or a product of two explanatory variables ( $X_1 \cdot X_2$ ), called **interaction terms**.

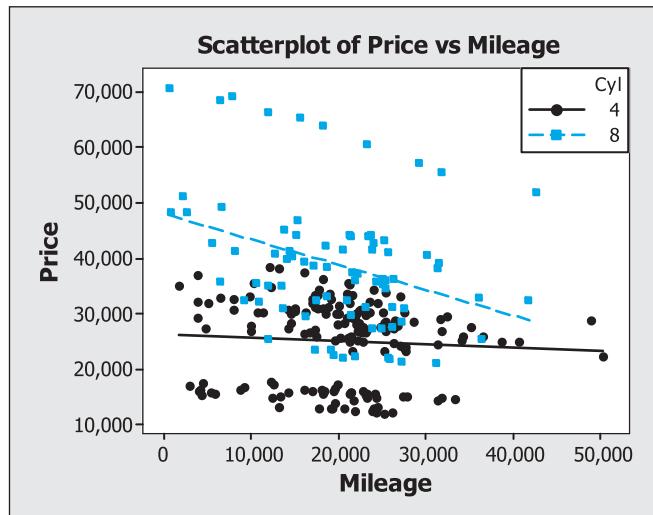
#### Interaction Terms

An **interaction** is present if the effect of one variable, such as Mileage, depends on a second variable, such as Cyl. If an interaction exists, the influence of Cyl changes for different Mileage values, and also the influence of Mileage will depend on Cyl.

The data set 4-8Cyl includes several four- and eight-cylinder cars from the original Cars data. Figure 3.10 shows a scatterplot and regression line to predict Price using both Mileage and Cyl. The regression model in Figure 3.10 has no interaction term. The parallel lines show that the estimated impact of changing cylinder



**Figure 3.10** Scatterplot and least squares regression line:  $\text{Price} = 15,349 - 0.20(\text{Mileage}) + 3443(\text{Cyl})$ . For each cylinder size, an increase of one mile is expected to reduce price by \$0.20.



**Figure 3.11** Scatterplot and least squares regression line:  $\text{Price} = 4533 + 0.340(\text{Mileage}) + 5431(\text{Cyl}) + 0.0995(\text{MileCyl})$ . If the interaction term ( $\text{MileCyl}$ ) is important, we expect to have regression lines that are not parallel.

size does not depend on mileage. Thus, for any given number of miles, when the number of cylinders changes from four to eight, we expect an increase in `Price` of  $4 \times \$3443 = \$13,772$ .

In the same way, the `Mileage` coefficient states that holding `Cyl` constant, we expect `Price` to decrease by \$0.20 for each additional mile on the car.

Figure 3.11 shows a scatterplot and regression line to predict `Price` using `Mileage`, `Cyl`, and a `Mileage*Cyl` interaction term (called `MileCyl`). The lack of parallel lines in the regression model  $\text{Price} = 4533 + 0.340(\text{Mileage}) + 5431(\text{Cyl}) - 0.0995(\text{MileCyl})$  indicates an interaction effect.

Caution should be used in interpreting coefficients when interaction terms are present. The coefficient for `Mileage` can no longer be globally interpreted as reducing `Price` by \$0.20 for each additional mile. Now, when there are four cylinders, `Price` is reduced by 0.058 [ $0.34(1) - 0.0995(1 \times 4) = -0.058$ ] with each additional mile. When there are eight cylinders, `Price` is reduced by 0.456 [ $0.34(1) - 0.0995(1 \times 8) = -0.456$ ] with each additional mile. Thus, an additional mile impacts `Price` differently depending on the second variable, `Cyl`.

## Extended Activity

### Understanding Interaction Terms

Data set: 4-8Cyl

35. Use the 4-8Cyl data set to calculate the two regression equations shown in Figures 3.10 and 3.11.
  - a. Does the  $R^2_{\text{adj}}$  value increase when the interaction term is added? Based on the change in  $R^2_{\text{adj}}$ , should the interaction term be included in the model?
  - b. For both models, calculate the estimated price of a four-cylinder car when `Mileage` = 10,000.
  - c. Assuming `Mileage` = 10,000, for both models explain how increasing from four to eight cylinders will impact the estimated price.
  - d. Conduct an extra sum of squares test to determine if the `MileCyl` interaction term is important to the model.
36. Use the 4-8Cyl data set to calculate the regression line  $\text{Price} = \beta_0 + \beta_1(\text{Mileage}) + \beta_3(\text{Cadillac}) + \beta_4(\text{SAAB})$ . You will need to create indicator variables for `Make` before calculating the regression line.
  - a. Create a scatterplot with `Mileage` as the explanatory variable and `Price` as the response. Overlay a second graph with `Mileage` as the explanatory variable and  $\hat{y}$  as the response. Notice that the

- predicted values (the  $\hat{y}$  values) form two separate lines. Do the parallel lines (no interaction model) look appropriate?
- Conduct one extra sum of squares test to determine if interaction terms (`MileCadillac` and `MileSAAB`) are important to the model (i.e., test the hypothesis  $H_0: \beta_5 = \beta_6 = 0$  versus  $H_a$ : at least one of the coefficients is not 0, where  $\beta_5$  and  $\beta_6$  are the coefficients for the two interaction terms). Create a scatterplot with the full regression model to explain the results of the hypothesis test.

## Quadratic and Cubic Terms

If a plot of residuals versus an explanatory variable shows curvature, the model may be improved by including a quadratic term. Is the relationship between mileage and retail price linear or quadratic for the Kelley Blue Book data? To test this, a quadratic term `Mileage*Mileage` can be created and included in a regression model.

### **| MATHEMATICAL NOTE |**

Even though models with quadratic ( $x^2$ ) or cubic ( $x^3$ ) terms are not linear functions of the original explanatory variables, the mean response is linear in the regression coefficients ( $\beta_0, \beta_1, \beta_2, \dots$ ). For example  $y = \beta_0 + z_1\beta_1 + z_2\beta_2 + \varepsilon$  would be considered a linear regression model when  $z_1 = x$  and  $z_2 = x^2$ .

## Extended Activity

### Understanding Quadratic Terms

Data set: MPG

The MPG data compare the miles per gallon of several cars against the speed the car was going as well as displacement. Displacement is a measure of the volume of all the cylinders within an engine. The larger the displacement, the more quickly fuel can move through an engine, giving the vehicle more power.

- Use the MPG data to create a regression model to predict MPG from Speed and Displacement:  

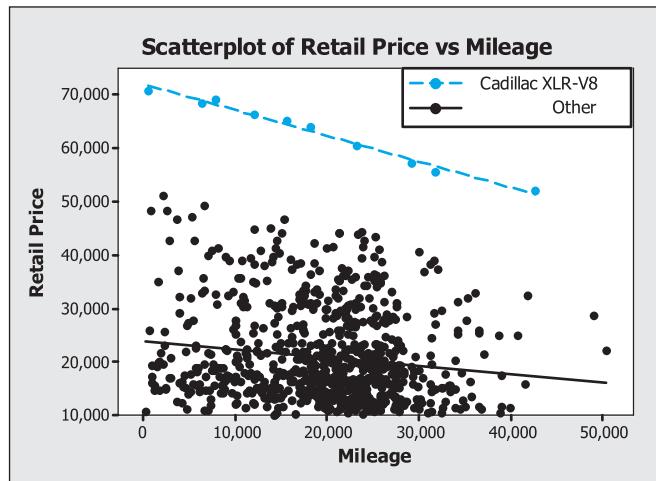
$$\text{MPG} = \beta_0 + \beta_1(\text{Speed}) + \beta_2(\text{Displacement}).$$
  - What are the regression equation and  $R^2$  value?
  - Look at residual versus Speed and residual versus Displacement plots. Describe any patterns you see.
  - What does the residual normal probability plot show?
- Create a regression model to predict MPG from Speed:  $\text{MPG} = \beta_0 + \beta_1(\text{Speed})$ .
  - What are the regression equation and  $R^2$  value?
  - Look at residual versus Speed and residual versus Displacement plots. Describe any patterns in the residual plots.
  - Describe any patterns in the residual normal probability plot.
  - Is Displacement an important explanatory variable? Use the residual plots and  $R^2$  to give an intuitive explanation.
- Create a model using displacement to predict MPG:  $\text{MPG} = \beta_0 + \beta_1(\text{Displacement})$ .
  - What are the regression equation and  $R^2$  value?
  - Look at residual versus Speed and residual versus Displacement plots. Describe any patterns in the residual plots.
- Create a  $(\text{Speed})^2$  term (called `Speed_sq`) and incorporate that term into your regression model to predict MPG:  $\text{MPG} = \beta_0 + \beta_1(\text{Speed}) + \beta_2(\text{Displacement}) + \beta_3(\text{Speed_sq})$ .
  - What are the regression equation and  $R^2$  value?
  - Look at residual versus Speed and residual versus Displacement plots. Describe any changes when  $(\text{Speed})^2$  is added to the model.
  - What does the residual normal probability plot show?

## Extended Activity

### Creating New Terms to Predict the Retail Price of Cars

Data set: Cars

The potential outliers identified in Question 11 can provide an interesting demonstration of an interaction. Figure 3.12 shows that the slope to predict Price from Mileage for the ten Cadillac XLR-V8s is much steeper than the slope found when using the other cars. This shows that depreciation for these high-end cars is almost 50 cents a mile, as opposed to 15 cents a mile on average for all car types combined.



**Figure 3.12** Scatterplot and regression lines: For the Cadillac XLR-V8, the regression line is  $\text{Price} = 71,997 - 0.4827(\text{Mileage})$ . This is a much steeper line than the regression line for all other cars:  $\text{Price} = 23,894 - 0.1549(\text{Mileage})$ .

41. Create a quadratic mileage term. Create two models to predict TPrice, one with only Mileage and another with both Mileage and  $(\text{Mileage})^2$  (called MileSq).
  - a. How much does the  $R^2$  value increase if a quadratic term is added to the model  $\text{TPrice} = \beta_0 + \beta_1(\text{Mileage})$ ?
  - b. Look at plots of residuals versus Mileage in both models. Did the addition of the MileSq term improve the residual plots?
42. Create an interaction term Mileage\*Cyl (called MileCyl). Use Mileage, Cyl, and MileCyl to predict TPrice. Does this interaction term appear to improve the model? Use residual plots and  $R^2$  to justify your answer.

While there is no “best” model, many final models developed by students in Question 21 tend to include the terms Cadillac, Convertible, and Liter. Since each of these terms is related to the Cadillac XLR-V8, it may be helpful to include an interaction term for Mileage\*Cadillac, Mileage\*Convertible, or Mileage\*Liter. Other Mileage, Make, or Type interactions may also be helpful additions to the model.

43. Develop additional quadratic and interaction terms. Determine if they improve the regression model in Question 42.
44. Submit a new regression model that best predicts TPrice. Does including quadratic or interaction terms improve your model from what was developed in Question 21?

Unless there is a clear reason to include them, researchers typically do not create interaction terms and test whether they should be included in a model. Most of the researcher’s effort should be spent on determining whether the original explanatory variables provided in the data set are related to the response. If an interaction

term ( $X_i * X_j$ ) is included in a final model, it is common practice to include each of the original terms ( $X_i$  and  $X_j$ ) in the model as well (even if the coefficients of the original terms are close to zero).

## 3.11 A Closer Look at Variable Selection Criteria

The growing number of large data sets as well as increasing computer power has dramatically improved the ability of researchers to find a **parsimonious model** (a model that carefully selects a relatively small number of the most useful explanatory variables). However, even with intensive computing power, the process of finding a “best” model is often more of an art than a science.

As shown earlier, stepwise procedures that use prespecified conditions to automatically add or delete variables can have some limitations:

- When explanatory variables are correlated, stepwise procedures often fail to include variables that are useful in describing the results.
- Stepwise procedures tend to overfit the data (fit unhelpful variables) by searching for any terms that explain the variability in the sample results. This chance variability in the sample results may not be reflected in the entire population from which the sample was collected.
- The automated stepwise process provides a “best” model that can be easily misinterpreted, since it doesn’t require a researcher to explore the data to get an intuitive feel for the data. For example, stepwise procedures don’t encourage researchers to look at residual plots that may reveal interesting patterns within the data.

### Adjusted $R^2$

While  $R^2$  is useful in determining how well a particular model fits the data, it is not very useful in variable selection. Adding new explanatory variables to a regression model will never increase the residual sum of squares; thus,  $R^2$  will increase (or stay the same) even when an explanatory variable does not contribute to the fit.

The **adjusted  $R^2$**  ( $R_{\text{adj}}^2$ ) increases only if the improvement in model fit outweighs the cost of adding an additional term in the model:

$$R_{\text{adj}}^2 = 1 - \left( \frac{n-1}{n-p} \right) \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \left( \frac{n-1}{n-p} \right) (1 - R^2) \quad (3.13)$$

where  $n$  is the sample size and  $p$  is the number of coefficients in the model (including the constant term).

#### ► MATHEMATICAL NOTE ▼

Intuitively, each additional term in a regression model means that one additional parameter value must be estimated. Each parameter estimate costs an additional degree of freedom. Thus,  $R_{\text{adj}}^2$  is an  $R^2$  value that is adjusted for degrees of freedom and can be written as

$$R_{\text{adj}}^2 = 1 - \frac{\text{MSE}}{\text{SST}/(n-1)}$$

$R_{\text{adj}}^2$  measures the spread of the residuals using MSE, while  $R^2$  measures the spread of the residuals using SSE.

### Mallows’ $C_p$

Another approach to variable selection is to use Mallows’  $C_p$  statistic

$$C_p = (n-p) \left( \frac{\hat{\sigma}^2}{\hat{\sigma}_{\text{Full}}^2} \right) + (2p-n) = (n-p) \left( \frac{\text{MSE}}{\text{MSE}_{\text{Full}}} \right) + (2p-n) \quad (3.14)$$

where  $n$  is the sample size,  $p$  is the number of coefficients in the model (including the constant term),  $\hat{\sigma}^2$  estimates the variance of the residuals in the model, and  $\hat{\sigma}_{\text{Full}}^2$  estimates the variance of the residuals in the full model (i.e., the model with all potential explanatory variables in the data set).

If the current model lacks an important explanatory variable,  $\hat{\sigma}^2$  is much larger than  $\hat{\sigma}_{\text{Full}}^2$  and  $C_p$  tends to be large. For any models where  $\hat{\sigma}^2$  is similar to  $\hat{\sigma}_{\text{Full}}^2$ ,  $C_p$  provides a penalty,  $2p - n$ , to favor models with a smaller number of terms. For a fixed number of terms, minimizing  $C_p$  is equivalent to minimizing SSE, which is also equivalent to maximizing  $R^2$ .

The  $C_p$  statistic assumes that  $\hat{\sigma}_{\text{Full}}^2$  is an unbiased estimate of the overall residual variability,  $\sigma^2$ . If the full model has several terms that are not useful in predicting the response (i.e., several coefficients are essentially zero), then  $\hat{\sigma}_{\text{Full}}^2$  will overestimate  $\sigma^2$  and  $C_p$  will be small.\*

When the current model explains the data as well as the full model,  $\hat{\sigma}^2/\hat{\sigma}_{\text{Full}}^2 = 1$ . Then  $C_p = (n - p)(1) + 2p - n = p$ . Thus, the objective is often to find the smallest  $C_p$  value that is close to  $p$ .

## Akaike's Information Criterion (AIC) and Bayes' Information Criterion (BIC)

Two additional model selection criteria are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).† Both of these criteria are popular because they are also applicable to regression models fit by maximum likelihood techniques (such as logistic regression), whereas  $R^2$  and  $C_p$  are appropriate only for least squares regression models.

Calculations for these two criteria are provided below. These statistics also include a measure of the variability of the residual plus a penalty term:

$$\text{AIC} = n[\log(\hat{\sigma}^2)] + 2p$$

$$\text{BIC} = n[\log(\hat{\sigma}^2)] + p[\log(n)]$$

where  $n$  is the sample size,  $p$  is the number of coefficients in the model, and  $\hat{\sigma}^2$  estimates the variance of the residuals in the model.

AIC and BIC are identical except for their penalties,  $2p$  and  $p[\log(n)]$ , respectively. Thus, AIC and BIC will tend to select slightly different models based on  $n$ . AIC and BIC both select models that correspond to the smallest value.

### Key Concept

No individual criterion ( $R_{\text{adj}}^2$ ,  $C_p$ , AIC, or BIC) is universally better than the other selection criteria. While these tools are helpful in selecting models, they do not produce a model that is necessarily "best."

Automated model selection criteria should not be used without taking time to explore the data. Stepwise procedures can be useful in initially screening a very large data set to create a smaller, more manageable data set. Once a manageable data set has been found, it is often best to spend time comparing many models from all potential terms.

## Model Validation

Often our goal is not just to describe the sample data, but to generalize to the entire population from which the sample was drawn. Even if a regression model is developed that fits the existing sample data very well and satisfies the model assumptions, there is no guarantee that the model will accurately predict new observations.

Variable selection techniques choose variables that account for the variation in the response. When there are many explanatory variables, it is likely that at least some of the terms selected don't explain patterns seen in the entire population; they are included simply because of chance variability seen in the sample.

\*  $\hat{\sigma}_{\text{Full}}^2 = \text{MSE}_{\text{Full}} = \text{SSE}_{\text{Full}}/(n - k)$ , where  $k$  is the number of terms in the full model. When  $k$  is unnecessarily large,  $\hat{\sigma}_{\text{Full}}^2$  is also large.

† BIC is also called the Schwarz criterion.

To validate that a regression model is useful for predicting observations that were not used to develop the model, do the following:

- Collect new data from the same population as the original data. Use the new data to determine the predictive ability of the original regression model.
- Split the original data. For example, randomly select 75% of the observations from the original data set, develop a model, and check the appropriate model assumptions. Test the predictive ability of the model on the remaining 25% of the data set. This is often called **cross-validation**.
- When the data set is not large enough to split, try **jackknife validation**. Hold out one observation at a time, develop a model using the  $n - 1$  remaining observations, and then estimate the value of the observation that was held out. Repeat this process for each of the  $n$  observations and calculate a predictive  $R^2$  value to evaluate the predictive ability of the model.
- Use theory and prior experience to compare your regression model with other models developed with similar data.

## Chapter Summary

In this chapter, we discussed techniques to **describe, predict, and test hypotheses** about the relationship between a quantitative response variable and multiple explanatory variables. The goals of a regression model will influence which techniques should be used and which conclusions can be drawn. The Cars activities in this chapter focused on developing a model that could describe the relationship between the explanatory variables and response variable as well as predict the value of the response based on a function of the explanatory variables.

Iterative techniques such as best subsets regression are often very useful in identifying which terms should be included in a model. The process of selecting explanatory variables to include in a model often involves **iterative techniques**, in which numerous models are created and compared at each step in the process. *Iterative techniques should not be used when the goal of multiple regression is to test hypotheses.* **Stepwise regression** is used to find the model with the highest  $R^2$  value; however, it does not provide much useful information about the model. For example, important variables (such as Liter in our investigation) are often not included in the stepwise regression models. **Best subsets regression** is a more useful iterative technique because it allows the researcher to better identify important explanatory variables, even if **multicollinearity** (highly correlated explanatory variables) exists. Table 3.3 lists the key measures used in variable selection.

No individual criterion ( $R^2_{\text{adj}}$ ,  $C_p$ , AIC, or BIC) is universally better than the other selection criteria. These tools are helpful in selecting models, but they do not produce a model that is necessarily “best.”

While iterative techniques are useful in reducing a large number of explanatory variables to a more manageable set, a researcher should ask the following questions to evaluate the resulting model:

- Were the techniques used to create the model appropriate based on the goals of the regression model?
- Do the coefficients make sense? Are the magnitudes of the coefficients reasonable? If the coefficients have the opposite sign than expected, **multicollinearity** may be present, the range of the explanatory variables may be too small, or important explanatory variables may be missing from the model.
- Do the residual plots identify any outliers or patterns that indicate unexplained structure that should be included in the model?

If the goal is to use hypothesis testing to determine how each of the explanatory variables impacts the response, iterative techniques are not appropriate. In addition, hypothesis tests about specific explanatory variables are not reliable when multicollinearity or lack of normality exists.

Model assumptions need to be met if the goal is to test hypotheses. While least squares regression models can be calculated without checking model assumptions, identifying patterns in residual plots that may indicate **heteroskedasticity, autocorrelation, outliers, or lack of normality** is important to creating a good model. If a pattern exists in any of the residual plots, it is likely that another model exists that better explains the response variable. Researchers need to be somewhat creative in deciding which graphs to create and how to adapt a model based on what they see.

**Table 3.3** Variable selection criteria.

Statistic	Selection Criteria
$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	Larger is better
$R_{\text{adj}}^2 = 1 - \left( \frac{n-1}{n-p} \right) \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	Larger is better
$C_p = (n-p) \left( \frac{\hat{\sigma}^2}{\hat{\sigma}_{\text{Full}}^2} \right) + (2p-n)$	Close to $p$ is better
$AIC = n[\log(\hat{\sigma}^2)] + 2p$	Smaller is better
$BIC = n[\log(\hat{\sigma}^2)] + p[\log(n)]$	Smaller is better

where

$n$  is the sample size

$p$  is the number of coefficients in the model (including  $\beta_0$ )

$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p}$  estimates the variance of the residuals in the current model tested  
 $\hat{\sigma}_{\text{Full}}^2$  estimates the variance of the residuals in the full model (the model with all explanatory variables)

Histograms or normal probability plots of residuals are used to determine if the residuals follow the normal distribution. Autocorrelation may exist when patterns appear in the ordered residual plots, indicating that each observation is not independent of the prior observation. Heteroskedasticity occurs when the variance of the residuals is not equal. If the variance increases as the expected value increases, a variance stabilizing transformation, such as the natural log or square root of the response, may reduce heteroskedasticity.

Tests of regression coefficients (as in Question 2) and **extra sum of squares tests** can be used for exploratory purposes or to test a theory. The individual  $t$ -test for coefficients can be unreliable if the explanatory variables are correlated. In addition,  $p$ -values for this  $t$ -test become less reliable as more tests are conducted.

The following hypothesis test can be analyzed with Table 3.2:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$H_a$ : at least one of the coefficients in the null hypothesis is not 0

The extra sum of squares test shown in Table 3.4 is used to compare a full model (with  $k$  terms) to a reduced model (with  $p$  terms), where  $k > p$ . Every term in the reduced model must also be in the full model.

$$H_0: \beta_p = \beta_{p+1} = \dots = \beta_{k-1} = 0$$

$H_a$ : at least one of the coefficients in the null hypothesis is not 0

Interaction terms and squared (or cubed) terms can be tested with the extra sum of squares test to determine if the additional terms improve the regression model. Testing models with all possible interaction and squared terms can become complex very quickly, so these terms are not typically tested unless there is some reason to include them.

**Table 3.4** The extra sum of squares  $F$ -statistic is the ratio of the mean square for the extra  $k - p$  terms to the  $MSE_{Full}$  with  $k - p$  and  $n - k$  degrees of freedom.

Source	df	SS	MS	$F$ -Statistic
<b>Reduced model</b>	$p - 1$	$SSR_{Reduced}$	$\frac{SSR_{Reduced}}{df_{Reduced}}$	$\frac{MS_{Reduced}}{MSE}$
<b>Extra <math>k - p</math> terms</b>	$k - p$	$SSR_{Full} - SSR_{Reduced}$	$\frac{SSR_{Full} - SSR_{Reduced}}{k - p}$	$\frac{MS_{Extra}}{MSE}$
<b>Error</b>	$n - k$	$SSE_{Full}$	$MSE_{Full} = \frac{SSE_{Full}}{n - k}$	
<b>Total</b>	$n - 1$	$SST = \sum_{i=1}^n (y_i - \bar{y})^2$		

Often the goal is not just to describe the sample data, but to generalize to the entire population from which the sample was drawn. Model validation techniques such as cross-validation and jackknife techniques are used to verify that the regression model applies to a larger population than just the sample data.

In multi-variable data sets, it is possible to obtain seemingly conflicting results. For example, we initially found a  $p$ -value indicating that mileage was a significant predictor of price, but the  $R^2$  value for the corresponding model was very small. Attempts to improve a regression model, such as by finding the appropriate data transformation or addressing multicollinearity, are often more of an art form than an automated process. Even highly technical statistical software packages cannot automatically develop a “best” model. Visual interpretation of residual plots typically provides a final model that is much better than a model developed by iterative (algorithmic) statistical software techniques.

Multiple regression involves a large set of tools for developing, assessing, and interpreting a regression model with one response variable and two or more explanatory variables. Numerous texts have been written about this topic, and this chapter does not attempt to be a comprehensive explanation of multiple regression. The goal of this chapter is to provide a foundation so that the reader will understand how a statistician approaches multiple regression as it is actually practiced in many disciplines.

## Exercises

- E.1. The initial regression model given in this chapter was  $Price = 24,765 - 0.1725(\text{Mileage})$ .
  - a. Give a practical explanation of the coefficient,  $b_1 = 0.1725$ , in the context of the `Cars` data.
  - b. The coefficient,  $b_1 = 0.1725$ , is a fairly small number. However, the  $p$ -value for the corresponding hypothesis test caused us to reject  $H_0: \beta_1 = 0$ . How could such a small value cause us to reject the null hypothesis?
  - c. The  $p$ -value caused us to conclude that mileage is important in predicting the retail price. However, the  $R^2$  value seemed to indicate that the regression model was not very useful in estimating the retail price. These results may originally appear somewhat contradictory. Explain how these statistics are measuring different aspects of the same regression model, and thus the results do not really contradict each other.
- E.2. Assume the basic forward stepwise regression is calculated with a data set that includes five quantitative explanatory variables. How many models were created and evaluated to determine the first term in the model?
- E.3. Assume the basic forward stepwise regression is calculated with a data set that includes five quantitative explanatory variables. If all five variables are included in the final model, how many models were created and evaluated to determine the final model?
- E.4. Assume your roommate has taken only one introductory statistics class, but is currently working on a multiple regression project in his economics class. He used best subsets regression on 50 explanatory

variables to find a good-fitting final model that included only four explanatory variables. He then used hypothesis tests to show that all four of these explanatory variables significantly impact the response. Write a short note to your roommate, explaining why the hypothesis tests are not valid.

- E.5. Explain why multicollinearity is a problem for researchers who are hoping to test whether specific explanatory variables impact the response (i.e., confirm a theory). Then explain why multicollinearity is not a problem for researchers who are trying to develop a model that accurately predicts their response.
- E.6. Explain why none of the previous activities using the `Cars` data can be used to show that `Mileage` or `Cylinder` size causes a change in `Price`.
- E.7. Select one residual plot that was created from any of the previous `Cars` activities. Explain how visual interpretation of the plot was more helpful than the  $R^2$  value in drawing appropriate conclusions about model validity.
- E.8. Explain why researchers often use statistics such as adjusted  $R^2$  or Mallows'  $C_p$  instead of the original  $R^2$  value.

#### **E.9. Brain and Body Weights: Transforming Variables**

Data set: `Weights`

In the process of studying the effect of sleep on mammals, T. Allison and D. Cicchetti collected the brain weights (in grams) and body weights (in kilograms) of 62 mammals.<sup>5</sup> In the data set, you can see the measurements for these two variables for each of the 62 animals studied.

- a. Create a scatterplot and regression line to predict `BrainWt` from `BodyWt` and appropriate residual plots. Even though the  $R^2$  value is reasonable, it is clear that there are extreme outliers on both the  $X$  and  $Y$  axes.
- b. Create a scatterplot and regression line to predict `log(BrainWt)` from `log(BodyWt)`. Create the appropriate residual plots and describe how the transformation impacted the validity of the linear regression model.
- c. The African elephant weighs 6654 kg. Use both models created in Parts A and B to estimate the African elephant's brain weight. Which model provides a more accurate estimate? Explain how this corresponds to the residual plots created in both questions.

#### **E.10. Arsenic and Toenails: Transforming Variables**

Data set: `Arsenic`

Karagas, Morris, Weiss, Spate, Baskett, and Greenberg collected toenail clippings of several individuals along with samples of each person's regular drinking water.<sup>6</sup> In this pilot study, the researchers attempted to determine if arsenic concentrations in people's toenails could be used to determine the level of arsenic in their water. Each of the 21 individuals in this study had a private well. High arsenic concentrations are related to cancer, and several studies have found a positive correlation between arsenic concentrations of drinking water and toenail samples.

- a. Create a scatterplot and regression line to predict `ArsWater` from `ArcToe`. There are extreme outliers along both the  $X$  and  $Y$  axes. However, 6 of the 21 water samples contained no arsenic, and taking the logarithm of `ArsWater` is not possible.
- b. Create a scatterplot and regression line to predict `log(Arswater + 1)` from `log(ArcToe)`. Create the appropriate residual plots and evaluate the validity of the model. While the transformations are helpful, model assumptions are still violated.
- c. Try using a few more variables or transformations to create a better model. This is an example where you are unlikely to find a satisfactory simple regression model.

#### **E.11. Socioeconomic Factors and HIV/AIDS**

Data set: `AIDS`

Is there any combination of socioeconomic factors that can be used to explain the prevalence of HIV/AIDS in the world's least developed countries? Least developed countries (or fourth world countries) are countries classified by the United Nations as meeting all three of the following criteria: a low income criterion, a human resource weakness criterion, and an economic vulnerability criterion.

The World Health Organization (WHO) attempts to measure the number of people living with HIV/AIDS. Regrettably, it is often difficult to get accurate counts.

The AIDS data set shows the prevalence of HIV/AIDS in adults aged 15 to 49 in least developed countries along with country-level data on several socioeconomic factors. Use this data set to create a multivariate regression model to predict the prevalence of HIV/AIDS from socioeconomic factors.

#### E.12. Partisan Politics

Data set: Politics

Do the population characteristics of each state (e.g., unemployment level, education level, age, income level, religious preferences, health insurance, and voter turnout) influence how people vote? Do these characteristics have an impact on the composition of state legislatures? In 2001, several college students collected data from 50 states on each of these factors from the U.S. Census Bureau.

- a. Calculate the percent Democrat in both the lower and the upper house for each state. Then calculate an average of these two percents as an overall Percent Democrat variable. (Data from Nebraska should not be included in the analysis since the state government is composed completely of nonpartisans.)
- b. Formulate hypotheses about how each of the explanatory variables (the population characteristics) will influence voting patterns.
- c. Form a regression model using the composition of the state governments in 2001 to test your theories.
- d. Plot the residuals and check the model assumptions. State your conclusions about each hypothesis.

#### E.13. Iowa Caucuses

Data set: Caucuses

The Caucuses data set contains the 2008 Democratic and Republican caucus results. Republicans count actual votes. Democrats don't record individual votes, but each precinct allocates its delegates based on the number of people in the precinct supporting each candidate during the caucus. Senator Clinton was expected to win in most polls, but ended up a surprising third after Senators Obama and Edwards. Many political analysts have found that females, less educated people, and older people voted for Senator Clinton, while younger people, more educated people, and African Americans preferred Senator Obama.

- a. Use the Caucuses data to test the hypothesis that education level is correlated to a higher percentage of votes for Senator Obama.
- b. Use the Caucuses data to create a multivariate model that "best" predicts the percentage of delegates Senator Obama would win.

#### E.14. Movies

Data set: 2008Movies

The 2008Movies file contains data on movies released in 2008.

- a. Calculate a regression model to predict box office from run time. Interpret the  $R^2$  value and test statistic for the slope in the context of this problem.
- b. Create indicator variables for genre and MPAA rating. Use best subsets regression to determine an appropriate regression model.
  - i. Validate the model assumptions.
  - ii. Look at residual plots and check for heteroskedasticity, multicollinearity, autocorrelation, and outliers. Transform the data if it is appropriate.
  - iii. Submit your suggested least squares regression formula along with a limited number of appropriate graphs that provide justification for your model. Describe why you believe this model is best.
- c. Conduct an extra sum of squares test to determine if one or more interaction terms (or quadratic terms) should be included in the model. You can choose any terms to test, but enough software output needs to be provided so that the instructor can verify your work.

## Endnotes

---

1. G.E.P. Box and N.R. Draper's *Empirical Model-Building and Response Surfaces* (New York: Wiley, 1987), p. 424.

2. More details are provided in more advanced textbooks such as M. H. Kutner, J. Neter, C. J. Nachtsheim, and W. Li, *Applied Linear Regression Models* (New York: McGraw-Hill, 2004).
3. See D. Montgomery, E. Peck, and G. Vining's text *Introduction to Linear Regression Analysis*, 3rd ed. (New York: Wiley, 2002) to get a better understanding of the details of variable selection techniques.
4. Ibid, p. 337.
5. T. Allison and D. Cicchetti, "Sleep in Mammals: Ecological and Constitutional Correlates," *Science*, 194 (Nov. 12, 1976): 732–734.
6. M. Karagas, J. Morris, J. Weiss, V. Spate, C. Baskett, and E. Greenberg, "Toenail Samples as an Indicator of Drinking Water Arsenic Exposure," *Cancer Epidemiology, Biomarkers and Prevention*, 5 (1996): 849–852.
7. W. Easterly, *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics* (Cambridge, MA: MIT Press, 2001), prologue.

# Research Project: Economic Growth in Third World Countries

Now that you have developed a multiple regression model using the `Cars` data, it is time to conduct your own research project. The following pages provide guided steps to conducting your own research project modeling economic growth in third world countries. Data for this project are found on the World Bank Website, <http://web.worldbank.org>, which has collected over 30 years of socioeconomic data on most countries.

## Reviewing the Literature

William Easterly, Senior Fellow at the Center for Global Development and the Institute for International Economics, is a prominent economist who has significant experience in economic development. In his book, Easterly shows that many attempted “solutions” to economic development in poor countries have violated basic principles of economics:

Fifty years ago, in the aftermath of World War II, we economists began our own audacious quest: to discover the means by which poor countries in the tropics could become rich like the rich countries in Europe and North America. Observing the sufferings of the poor and the comforts of the rich motivated us on our quest. If our ambitious quest was successful, it would be one of humankind’s greatest intellectual triumphs. . . . We thought we had found the elixir many different times. The precious objects we offered ranged from foreign aid to investment in machines, from fostering education to controlling population growth, from giving loans conditional on reforms to giving debt relief conditional on reforms. None has delivered on its promise. . . . This is a sad story, but it can be a hopeful one. We now have statistical evidence to back up theories of how the panaceas failed and how incentive-based policies can work.<sup>7</sup>

1. Read the “Cash for Condoms?” chapter from William Easterly’s book, *The Elusive Quest for Growth: Economists’ Adventures and Misadventures in the Tropics* (Cambridge, MA: MIT Press, 2001), pp. 87–99. If there are any words that you do not understand, look them up and provide a short definition for each. Identify each of the following and be ready to discuss this material in class:
  - a. Objective of the chapter, including key ideas or concepts
  - b. Response variable(s) suggested by Easterly
  - c. Potential explanatory variables suggested by Easterly
  - d. Variables that were assumed to be held constant during the study
  - e. Nuisance factors (i.e., factors that are not of key interest in the study, but may influence the results)

## Exploring the Data

The World Bank has for several years tracked socioeconomic data for most countries. The World Bank data set lists several variables for which data are available for most African countries for the year 2002. Be ready to submit your answers to the following as well as to discuss this material in class.

2. Which of the variables in the 2002 data set would be the most appropriate response variable representing economic growth?
  - a. Verify that the response variable provides the information needed to address the question of interest.
  - b. What are the range and variability of responses you expect to observe for this variable?
  - c. Is the response measurement precise enough to address your question of interest?
3. Identify which explanatory variables (or transformed variables) you would expect to influence economic growth. For each explanatory variable that you select, identify whether you would expect a positive or negative correlation between that variable and economic growth (i.e., your response variable).
4. Identify any missing factor(s) assumed to be controlled within the model.

5. Identify how measurements, material, and process may involve unwanted variability. For example, how would fertility, life expectancy, or CO<sub>2</sub> emissions be accurately recorded?
6. What conditions would be considered typical for this type of study? Are these conditions controllable? How might the results be impacted if any of these conditions changed during the study?
7. Use the data provided to create a preliminary model. You may want to talk to an economist and discuss questions that have arisen about model assumptions and variable selection. Compare models and  $R^2$  values among the working groups in the class.
  - a. Check for heteroskedasticity. Plot variables of interest against your response variable. Create a multiple regression model and look at residual plots. Identify any variables that exhibit patterns in their residual plots. Based on the plots you have created, make any appropriate transformation to the response and/or explanatory variables.
  - b. Check for multicollinearity. Are any of the explanatory variables highly correlated? If so, consider modifying your regression model. Explain the impacts multicollinearity may have on the interpretation of your results.

## Presenting Your Own Model

Based on the class discussion and your initial model, identify other variables that may be available that would be appropriate to include. For example, would you prefer to study one country for 30 years instead of multiple countries?

8. Use the website <http://devdata.worldbank.org/data-query> or your own resources to collect data to create a model to determine the relationship between population control and economic growth.
9. Once the data have been collected, meet with your professor to discuss your potential model assumptions, diagnostics, and analysis.
10. Analyze your data and use the discussions of prior work to write a 5- to 7-page research paper describing your analysis and discussing the results. (See “How to Write a Scientific Paper or Poster” on the accompanying CD.)
11. Bring three copies of your research paper to class. Submit one to the professor. The other two will be randomly assigned to other students in your class to review. Use the “How to Write a Scientific Paper or Poster” checklist to review each other’s papers and provide comments.

## Final Revisions

Make final revisions to the research paper. Then submit the first draft, other students’ comments and checklists, the data set you used (in electronic format) along with descriptions of the variables in the data set, and your final paper.

## Other Project Ideas

Several of the activities in the text can also be used to develop your own project ideas. In addition to the World Bank Website, there are many other places where data are publicly available.

- Information on a variety of sports can be found at <http://cbs.sportsline.com>, <http://sportsillustrated.cnn.com>, <http://www.nfl.com/stats/team>, <http://www.ncaa.org>, and <http://www.baseball-reference.com>.
- Information from many federal agencies can be found at <http://www.fedstats.gov>, the Bureau of Labor Statistics (<http://www.bls.gov/cpi>), the Behavioral Risk Factor Surveillance System (<http://www.cdc.gov/brfss>), the National Center for Health Statistics (<http://www.cdc.gov/nchs>), and the U.S. Census Bureau (<http://www.census.gov>).
- College and university information can be found at <http://www.collegeboard.com>, <http://www.act.org>, and <http://www.clas.ufl.edu/au>.
- The National Center for Educational Statistics Website is <http://nces.ed.gov>.
- Information about movies produced each year can be found at <http://www.the-numbers.com>, <http://www.boxofficemojo.com>, <http://www.imdb.com>, and <http://www.rottentomatoes.com>.

# The Design and Analysis of Factorial Experiments: Microwave Popcorn

*However beautiful the strategy, you should occasionally look at the results.*

—Winston Churchill<sup>11</sup>

Statistics ought to be viewed as a whole: understanding the process of formulating questions, properly designing a study, actively collecting meaningful data, and then deciding how to properly organize and draw conclusions from the data. Advancements in technology have made data collection and computationally intensive statistical techniques much more feasible. At one time, many statisticians had narrowly defined roles and were considered as primarily “number crunchers.” Today, statisticians characteristically work on interdisciplinary teams that emphasize scientific inference and understanding data in context.

Instead of emphasizing formulas, computation, and mathematical theory, this chapter uses a simple popcorn experiment to demonstrate the numerous challenges that can occur in designing experiments and collecting data.

The activities in this chapter will discuss the following:

- Key features of a well-designed experiment and proper data collection
- Proper determination of response variables, experimental factors, and levels
- Building on the one-way ANOVA discussed in Chapter 2 to describe multivariate factorial designs
- Evaluating multiple hypotheses based on main effects and interaction terms
- Calculating each of the between-group and within-group variances needed in ANOVA tables for balanced factorial designs
- Calculating effects and developing mathematical models
- Using multiple comparison tests with ANOVA tables

## 4.1 Investigation: Which Microwave Popcorn Is the Best?

Popcorn is a staple for many college students. While many students like popcorn because it is inexpensive and easy to prepare, it is also a whole grain food that's low in fat and calories. According to The Popcorn Institute, Americans consume an average of 54 quarts of popcorn a year per person.<sup>2</sup>

Two popcorn lovers, who also happened to be taking a statistics course, decided to test whether there is a difference in quality between microwave popcorn brands. Yvonne and Tue wanted to know if a cheaper brand of popcorn was just as good as more expensive popcorn. These students could have chosen to conduct a study that could be analyzed with a two-sample *t*-test if they had simply compared two brands of popcorn. However, if they did a two-sample *t*-test, they would need to hold many factors constant, such as the type of microwave, cooking time, and storage procedures. Since Yvonne and Tue believed that some of these factors could also impact the quality of the popcorn, they decided to include some of these additional factors in their study.

Modeling real-world phenomena often requires more than just one factor to explain changes in the response. **Factorial designs** are any statistical designs that are structured to use factors (i.e., explanatory variables) to organize meaningful groups of treatment conditions. A two-sample *t*-test can be considered a special case of a factorial design that has just one factor (popcorn brand in this case) and two levels (Brand A and Brand B). Factorial designs are very powerful statistical tools because they allow a researcher to simultaneously test the effects of multiple factor-level combinations on a response of interest.

### Key Concept

In factorial designs, each explanatory variable is called a **factor** and specific conditions within each factor are called **levels**. In any study, these factor-level combinations are called **conditions**; in experiments, they are often called **treatments**. Factorial designs are often used to test the effects of multiple factors simultaneously, where each factor has two or more levels.

## 4.2 Elements of a Well-Designed Experiment

Unfortunately, many people mistakenly believe that statistics is only a process of performing mathematical calculations on data in order to examine the validity of a hypothesis. Proper experimental design is just as important as, if not more important than, the choice of statistical calculations. In fact, designing experiments and collecting appropriate data are often the most difficult and time-consuming aspects of conducting experiments.

### Key Concept

A good design attempts to answer the question(s) of interest as clearly and efficiently as possible. Any statistical analysis is only as good as the quality of the data.

An **experiment** is defined as a study in which purposeful changes are made to controlled conditions in order to identify changes in a response. An experiment imposes a treatment on subjects or experimental units, while an **observational study** simply collects data under varying conditions without imposing any changes. Well-designed experiments are conducted under controlled conditions to make it easier to isolate the impact of each treatment combination. In observational studies, the conditions in the study are rarely the only characteristic that makes the two (or more) populations different. Thus, unknown factors that may bias the results are unfortunately built into an observational study.

### Key Concept

Both experiments and observational studies use sample data to draw conclusions about a larger population, process, or system. It is often much easier to show cause and effect relationships in a well-designed experiment because conditions are controlled.

**NOTE**

Some texts state that only experiments can be used to show cause and effect relationships. However, poorly designed experiments should not be used to show causation. In addition, observational studies (such as those testing a relationship between smoking and lung cancer) can be used to show causation if (1) there is a strong association between the explanatory and response variables, (2) higher doses are associated with stronger responses (e.g., more cigarettes increase the likelihood of getting cancer), (3) there are consistent results across many studies, and (4) there are credible explanations for the cause and effect relationship.

**Experimental design** is the process of planning an experiment that collects the right amount and type of data to answer a question of interest. Several decisions need to be made about how an experiment is to be constructed and executed to ensure that the experimental results are valid. Taking the time to properly design an experiment will improve the precision of answers to research questions. In addition, well-designed experiments often are much more efficient and obtain stronger conclusions than other studies.

The first step in designing an experiment is to clearly define a problem and state the objectives of the experiment. This is often much more difficult than it first appears. Before any data are collected, it is essential that everyone involved understand the objectives of the experiment, what measurements will be taken, what material is needed, and what procedures will be used. Good experimental design typically involves gaining background knowledge outside the field of statistics.

There are many possible ways to conduct an experiment to determine the effect of brand on the quality of popcorn. While microwave popcorn is something these students were quite familiar with, they needed to determine which brands to compare, the appropriate cooking time (which could vary by microwave), and how to define and measure “good” popcorn.

## Identifying a Response Variable

Many possible measurements could be taken on microwave popcorn. Yvonne and Tue could have created a taste rating or a texture rating, measured the volume of the kernels, counted the number of popped kernels, or calculated the percentage of “old maids,” the kernels that did not pop after the bag had been cooked.

Identifying the response variable corresponds to determining what measurements should be taken. Each experiment should ensure that the **response variable provides the information** needed and that the **response measurement is precise enough to address the question of interest**. Yvonne and Tue determined that their definition of “quality” popcorn would be popcorn that had the highest percentage of popped kernels per bag. Notice that if Yvonne and Tue had counted only the popped kernels, and not the unpopped kernels, they might have gotten a distorted response, since some brands may tend to have more kernels per bag.

Yvonne and Tue initially discussed randomly sampling 20 kernels from each popped bag and calculating the percentage of popped kernels. However, the size and shape differences between popped and un-popped kernels would have made it rather difficult to simply pull out a random sample. Thus, in order to ensure their counts were as accurate as possible, they decided to count every kernel in every bag of their experiment.

It is also useful to discuss the range and variability of responses expected to be observed. For example, if we conducted a study under conditions that typically gave only two outcomes, either 0% or 100%, the response would be categorical (such as yes/no or popped/not popped), and then an analysis based on categorical response variables should be used. Studies with categorical response variables can be analyzed with techniques such as the chi-square test or logistic regression, which are discussed in Chapters 6 and 7, respectively.

The percentage of popped kernels was considered a **quantitative response** variable in this experiment. Background research showed that some popcorn companies expected between 94% and 97% popped kernels, but based on their prior popcorn eating experience, Yvonne and Tue expected the percentage to be a little lower. In Yvonne and Tue’s study, they roughly estimated that responses should be between 60% and 99% popped kernels, with an average close to 90%.

### Key Concept

Care needs to be taken before a study is conducted to ensure that the response measurement is accurate and applicable to the research question.

## Identifying the Factors and Levels

The next step in designing an experiment is to investigate any factors that may be of importance or may potentially bias the results. Yvonne and Tue had two microwaves that they typically used to make popcorn, one in their dorm lounge and one in their room. The lounge microwave had a “popcorn” setting, which cooked for 1 minute 45 seconds, though the package instructions for each brand suggested varying cooking times. Most microwaves also have power settings. Should popcorn always be popped at the highest power setting?

With a little research, these students found that the quality of popcorn can also be affected by how it is stored. Popcorn stored in a moisture-rich environment, such as a refrigerator, tends to have a higher percentage of popped kernels. However, too much moisture may cause the popcorn to have a gummy texture. Finally, Yvonne and Tue wanted to compare a relatively expensive brand of popcorn (Pop Secret) to a relatively inexpensive brand (Fastco). Each of these brands also has a variety of flavors, such as butter, kettle corn, and caramel.

Notice that the discussion of factors and potential levels is based not on statistical calculations, but on nonstatistical knowledge. Nonstatistical knowledge is often essential in choosing factors, determining factor levels, and interpreting the results of a study.

Yvonne and Tue decided on three **factors of interest**, factors that would be included in the study to determine if different levels impact the results:

Factor 1: popcorn Brand at two levels, Fastco and Pop Secret

Factor 2: Microwave at two levels, Lounge and Room

Factor 3: cooking Time at two levels, 105 seconds and 135 seconds

It can sometimes be difficult to identify a reasonable range for each factor. Yvonne and Tue had noticed that some brands of popcorn tended to burn at around 150 seconds (2.5 minutes). Even though cooking popcorn longer than 135 seconds might increase the percentage of popped kernels, Yvonne and Tue decided to avoid cooking times likely to cause burning.

Yvonne and Tue then listed suspected **extraneous variables**, other factors that need to be controlled during the experiment to eliminate potential biases. Yvonne and Tue decided to hold some extraneous variables constant. In particular, they used only the highest power setting on each microwave, they stored all the popcorn on a shelf in their room, and they used only the butter flavor of each brand. There were other variables they could not control, such as age of the popcorn, which manufacturing plant prepared each bag of popcorn, and how different retail stores had stored the popcorn. To account for the extraneous variables they could not control (or had not even thought of), it would be best to randomly select bags of popcorn from the entire population. Instead, Yvonne and Tue did their best to randomly select several bags of Fastco and Pop Secret butter popcorn from a variety of stores in town. This was not a true random sample, and Yvonne and Tue had to be careful in making any statements about how the results of their study extended to a larger population. In addition, when possible, each bag of popcorn in the study was randomly allocated to a factor-level combination. While bags of popcorn can be randomly assigned to a cooking time and microwave, they cannot be randomly assigned to a popcorn brand.

### Key Concept

A good design controls for known extraneous variables (often by holding them constant throughout the study) and then uses random sampling and random allocation to control for any other unknown or uncontrollable extraneous variables.

## Choosing a Design

In addition to determining what conditions to test and what measurements to take, in order to create a good experimental design, a researcher must properly define units and determine how units are structured. An **experimental unit** is the smallest part of experimental material that is assigned (randomly, if possible) to a factor-level combination within a study. Since Yvonne and Tue counted every kernel of popcorn, some may incorrectly assume that each kernel is a unit. In an experiment, units are randomly assigned to treatments. In this study, each **kernel was not randomly assigned to a condition**, but each bag of popcorn was randomly assigned to be popped in a particular microwave for a particular length of time. **Thus, bags of popcorn are considered the units** for Yvonne and Tue’s study.

**Key Concept**

If (1) units are as similar as possible, (2) units are randomly assigned to treatment combinations, and (3) large enough sample sizes are used, then we can conclude that statistically significant differences in the response can be explained by the different treatment combinations.

This chapter will focus on **completely randomized factorial designs**. In completely randomized designs, each unit is assigned to exactly one factor-level combination. Only one measurement is collected for each unit. In the following section, we will use Yvonne and Tue's data to simultaneously test for the effects of two brands, two cooking times, and two microwaves on the percentage of popped kernels.

**NOTE**

While this chapter focuses only on completely randomized factorial designs, it is important to recognize the difference between completely randomized, **block** and **split-plot** or **(repeated measures)** designs. If each unit is assigned to only one factor-level combination, it is appropriate to use a completely randomized design. If each unit is assigned to several conditions or multiple measurements are taken on each unit, a more complex design structure, such as a block or **split-plot** (repeated measures) design, may be needed.

Block and split-plot (repeated measures) designs also work with different types of factors. The factors of interest in this chapter all have fixed effects. **Fixed effects** correspond to factors where each level is selected because it is of specific interest to the researcher. **Random effects** correspond to factors where the levels are randomly selected from a larger population. All factors in a completely randomized design are also **crossed**; this means that every level of each factor can be tested in combination with every level of every other factor. Alternatively, **nested factors** are factors for which some levels cannot occur simultaneously with other factor levels. Random effects and nested factors can be analyzed with block and split-plot (repeated measures) designs; they are described in Chapter 5.

**Key Concept**

If possible, a straightforward design and analysis are usually better than a complex design and analysis. If the design is too complicated and the data are not collected properly, even the most advanced statistical techniques may not be able to draw appropriate conclusions from an experiment.

## Determining Sample Sizes for Completely Randomized Designs

When more units are tested, it is more likely that the statistical analysis will identify true differences between conditions. However, every unit tested has a cost, and it is important to carefully determine how many units are practical to test. Is it worth testing additional units to gain a better understanding of the unit-to-unit variability?

Yvonne and Tue estimated that they could each count the kernels in one bag (popped and unpopped) in 10 to 15 minutes. They also thought that they could get a few close friends to count a few bags in exchange for free popcorn. Care should always be taken when measuring results. If the result is a subjective measurement (such as the taste of popcorn or the quality of artwork), very clear procedures should be written and, if possible, the same people should record each measurement. In this popcorn study, counting the percentage of popped kernels per bag is an **objective measurement**. As long as Yvonne and Tue have trustworthy friends, there should be no problem in having several people help count popcorn kernels.\* They estimated that they could conduct 32 tests in about four hours.

\* Ideally, they would have preferred to have at least two people count each bag to ensure against errors, but that would have doubled the time needed to conduct the experiment. Repeatability and reliability (often called gauge R&R) studies are a technique discussed in many quality control texts as a way to ensure that different people or processes provide similar results.

The choice of 32 tests, instead of a round number like 30, is also related to a well-designed experiment. Yvonne and Tue found that, based on their choice of factors and levels, there were a total of eight treatment combinations. Table 4.1 lists the eight possible treatment conditions that can be “assigned” to each bag of popcorn. Yvonne and Tue wanted to have a **balanced design**, a design where the same number of units is assigned to (or randomly selected from) each condition. Balanced designs are often easier to analyze and more likely to identify true differences in the effects of different conditions.<sup>3</sup> If Yvonne and Tue wanted to conduct a balanced design, they needed to conduct tests in a multiple of 8 (16, 24, etc.).

**Table 4.1** All possible treatment conditions for the three-factor popcorn study.

Treatment Combination	Popcorn Brand	Microwave Location	Cooking Time
1	Fastco	Lounge	105
2	Pop Secret	Lounge	105
3	Fastco	Room	105
4	Pop Secret	Room	105
5	Fastco	Lounge	135
6	Pop Secret	Lounge	135
7	Fastco	Room	135
8	Pop Secret	Room	135

#### ► MATHEMATICAL NOTE ▼

Table 4.1 lists the variables in standard order. Listing conditions in **standard order** is a simple technique that ensures that each factor combination is listed exactly once. The first variable alternates levels every row. For the second variable, levels alternate every other row. The third variable alternates every fourth row. This same process can work for multiple factors with multiple levels. If there were four factors, each with two levels, the fourth column would alternate every eight rows. For studies with more than two levels, simply ensure that the current variable lists each level exactly once for all prior combinations.

### Activity (▶) Determining the Number of Treatment Combinations

- 1. Assume we want to use three cooking times for popping the popcorn instead of two. List the possible treatment combinations that can be assigned. How many are there?
- 2. Without listing all possibilities, calculate how many treatment combinations would exist for a design that tested five brands with three microwaves at four cooking times.

## 4.3 Analyzing a Two-Way Factorial Design

Since several factors are included in this experiment, there are also several hypotheses to be tested. Yvonne and Tue actually had six research questions, which will be discussed in Section 4.4. But to keep the calculations simple, in this section we will assume that only two factors were tested (Brand and Time). This leads to three hypotheses corresponding to the Brand and Time factors.

1.  $H_{01}: \mu_{\text{Fastco}} = \mu_{\text{PopSecret}}$ , there is no difference in the mean response ( $\text{PopRate}$ ) between the two Brands.\*  
 $H_{a1}: \mu_{\text{Fastco}} \neq \mu_{\text{PopSecret}}$ , the two Brand means are different.
2.  $H_{02}: \mu_{105} = \mu_{135}$ , there is no difference in the mean PopRate for the two Times.  
 $H_{a2}: \mu_{105} \neq \mu_{135}$ , the two Time means are different.

\*This is often written as  $H_{01}$ : there is no Brand effect or  $H_{01}: \alpha_{\text{Fastco}} = \alpha_{\text{PopSecret}} = 0$ , where  $\alpha$  is called the **effect size**. For example,  $\alpha_{\text{Fastco}} = \mu_{\text{Fastco}} - \mu$ , where  $\mu$  is the overall **grand mean** of the responses.

3.  $H_{03}$ : Brand has no influence on how Time affects PopRate. This is equivalent to stating  $H_{03}$ : Time has no influence on how Brand affects PopRate or  $H_{03}$ : the effect of Time is the same for both Brands or  $H_{03}$ : there is no interaction between Time and Brand.
- $H_{a3}$ : Brand influences how Time affects PopRate or  $H_{a3}$ : there is an interaction between Time and Brand.

Factorial designs are efficient because much more information can be calculated, such as *p*-values for multiple hypothesis tests, without requiring more experimental units than for the typical two-sample *t*-test. Factorial designs are very beneficial in situations where experimental units are expensive or difficult to obtain. The next sections will discuss how to organize and draw conclusions for each of the above hypotheses in a factorial design with two factors, also called a **two-way factorial design**.

## Visualizing the Data

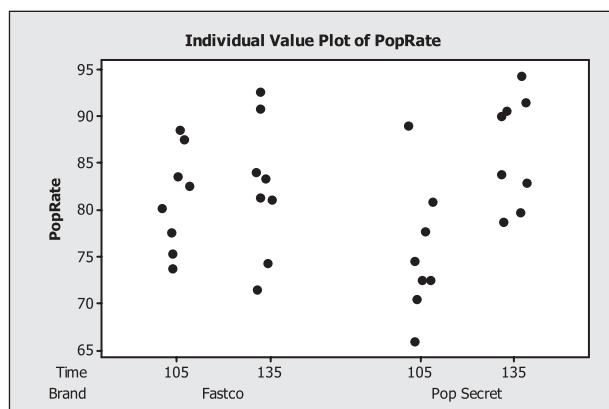
Before any formal analysis is done, we will carry out an informal analysis by looking at a graph. An individual value plot of the data is shown in Figure 4.1.

### Activity ▶ Visualizing and Summarizing Data

3. Use Figure 4.1 to compare the PopRate for each of the four factor-level combinations. Do the four groups appear to have similar means or similar standard deviations? Are there any outliers (extreme observations that don't seem to fit the rest of the data)? Describe any patterns you see in the data.
4. Calculate the average PopRate for each Brand and each Time. Calculate the overall average PopRate.
5. Use the data set labeled Popcorn to calculate appropriate summary statistics (the median, mean, standard deviation, range, etc.) for each of the four groups. For the Fastco brand, calculate the difference between the average PopRate for the two cooking times. Do the same for the Pop Secret brand.

## Notation for Multiple Explanatory Variables

Table 4.2 shows the Popcorn data set organized by the Brand and Time factors. Each of the four treatment combinations has eight observations. The data are also provided in the file Popcorn.



**Figure 4.1** Individual value plot of the PopRate (100 × count of popped kernels/total kernels) for each Brand and cooking Time factor-level combination. Points have jittering (small fluctuations) so that all values are visible.

**Table 4.2** Popcorn study data: the PopRate for each bag of popcorn with sample mean corresponding to each factor-level combination.

Microwave Popcorn Brand				
Fastco			Pop Secret	
Microwave Cooking Time		Microwave Cooking Time		
	105 seconds	135 seconds	105 seconds	135 seconds
	82.60	81.07	77.70	91.41
	87.50	90.80	72.54	83.78
	75.30	84.06	65.98	79.70
	80.20	71.50	72.56	90.05
	73.80	81.34	74.51	78.71
	77.60	92.60	80.83	90.56
	88.50	74.30	88.97	94.28
	83.56	83.36	70.44	82.90
Mean Response	81.13	82.38	75.44	86.42

Table 4.3 is useful in visualizing the differences and similarities among the meaningful groups within this data set: the overall average, the two Brand groups, the two Time groups, and the four factor-level combinations. Table 4.3 also includes mathematical notation representing each mean. For example,  $\bar{y}_{11..}$  represents the mean of the 8 responses from the first Brand, Fastco, and the first Time group, 105 seconds.  $\bar{y}_{2..}$  represents the mean of the 16 responses from the second Brand group, Pop Secret.  $\bar{y}_{...}$  represents the overall mean of all 32 responses.

**Table 4.3** Popcorn study sample means: the mean PopRate (per 100 kernels) for each of the nine meaningful groups of sample data.

		Factor-Level Group Means Cooking Time		
Brand		105 seconds	135 seconds	Brand Means
Fastco		$\bar{y}_{11..} = 81.13$	$\bar{y}_{12..} = 82.38$	$\bar{y}_{1..} = 81.8$
Pop Secret		$\bar{y}_{21..} = 75.44$	$\bar{y}_{22..} = 86.42$	$\bar{y}_{2..} = 80.9$
Cooking Time Means		$\bar{y}_{..1} = 78.3$	$\bar{y}_{..2} = 84.4$	$\bar{y}_{...} = 81.3$

#### MATHEMATICAL NOTE

The dot in the subscript indicates that the average was taken over all values of that subscript. The key is to recognize that groups are identified by their subscripts. Brand is the first subscript, and Time is the second. Each individual observation for each Brand and Time factor-level combination is represented by the third subscript. For example, the 4th observation in Table 4.2 for the Fastco brand (brand 1) and the 135-second time (time 2) group is  $y_{124} = 71.50$ . The average of the 8 observations in the Fastco brand (brand 1) and 135-second time group is represented by  $\bar{y}_{12..} = 82.38$ . In addition,  $\bar{y}_{1..}$  is the average response of all 16 of the Fastco brand (brand 1 observations), while  $\bar{y}_{..1}$  is the average response of all 16 of the 105-second times (time 1 observations).  $\bar{y}_{...}$  is the average PopRate, averaged over all observations for both Brand and cooking time. That is,  $\bar{y}_{...}$  is the overall average PopRate.

## Activity Understanding Notation

- 6. Which notation would be used to describe the sample average of the 135-second group?
- 7. Explain the difference between  $\bar{y}_{21}$  and  $\bar{y}_{12}$ .

### Comparing Variances\*

Figure 4.1 and Table 4.3 indicate that the difference between the Time means is much larger than the difference between the Brand means. In addition, the difference between Time means is much larger for the Pop Secret brand than for the Fastco brand. In this section, we will conduct an analysis, called analysis of variance, to find *p*-values for testing each of the three hypotheses stated earlier about the underlying mean responses.

**Analysis of variance (ANOVA)** is conducted by comparing the variability between groups to the variability within groups. For example, does the variability between Brand means (between groups) appear to be large compared to the variation of responses within the two Brand levels (within groups)? In ANOVA, these measures of variability are called **mean squares**. For example, the variability between brands is called mean square brand and is denoted  $MS_{\text{Brand}}$ . In actual practice, the following ANOVA calculations are done with computer software instead of by hand. The reason for working through these equations in detail is to better illustrate the logic behind using ANOVA to determine if between group variability is significant (i.e., to determine whether we can reject any of the null hypotheses).

**Between-Group Variability** To create a measure of the **variability between Brand means** ( $MS_{\text{Brand}}$ ) calculate the weighted variance of the Brand group means, using the size of each group as the weight. The weighted variance of the Brand group means is calculated with the following equation:

$$\begin{aligned} MS_{\text{Brand}} &= \frac{\sum_{i=1}^2 n_i \times (\bar{y}_{i..} - \bar{y}_{...})^2}{2 - 1} \\ &= \frac{16 \times (81.8 - 81.3)^2 + 16 \times (80.9 - 81.3)^2}{2 - 1} \end{aligned} \quad (4.1)$$

where  $n_i$  is the number of observations for brand  $i$ . In this study,  $n_i = 16$  observations are taken for each brand. Notice that Equation (4.1) looks similar to a typical variance calculation:

- There are two observed group means: 81.8 and 80.9.
- The spread is measured by summing the squared distance between each observed group mean and the overall mean and then dividing by the **number of group means minus one**.

As with any variance calculation, we are finding an average squared distance (mean squared distance, denoted  $MS_{\text{Brand}}$ ). The difference between this calculation and a typical variance calculation is the use of weights:

- Each observed mean is based on a group of size 16; this group sample size ( $n_i = 16$ ) is multiplied by each squared distance.

#### MATHEMATICAL NOTE

In our study, we have a balanced design (i.e., equal sample sizes in every group). However, the formulas throughout this section allow for studies with unequal sample sizes (called unbalanced designs).

The calculation of the variability between the Time means ( $MS_{\text{Time}}$ ) is very similar to Equation (4.1):

$$MS_{\text{Time}} = \frac{\sum_{j=1}^2 n_j \times (\bar{y}_{.j} - \bar{y}_{...})^2}{2 - 1} \quad (4.2)$$

---

\*If you have studied ANOVA tables before, you may find it surprising that we focus on mean squares (MS) and do not discuss sums of squares (SS) or degrees of freedom (df). The focus of this section is the concepts and logic behind ANOVA. ANOVA is the process of comparing between group and within group variability. These types of variability are represented by the mean squares. Chapter 2 and the extended activities discuss sums of squares and degrees of freedom.

The first two hypotheses at the beginning of this section correspond to questions about main factors incorporated into the experiment, Brand and Time. The third hypothesis focuses on whether the impact of one variable (Time) depends on a second variable (Brand). This is called an **interaction effect**.

Table 4.3 provides some evidence of interaction between Brand and Time. For the Fastco brand popcorn, the longer cooking time increases the percentage of popped kernels by  $\bar{y}_{12} - \bar{y}_{11} = 82.38 - 81.13 = 1.25$ , while the increase for the Pop Secret brand is many times larger:  $\bar{y}_{22} - \bar{y}_{21} = 86.42 - 75.44 = 10.98$ .

To test for an interaction effect (the third hypothesis), we first measure the variability between all four groups (each Brand and Time combination) and then subtract the squared values for the main factors.

$$MS_{\text{BrandTime}} = \frac{\sum_{i=1}^2 \sum_{j=1}^2 n_{ij}(\bar{y}_{ij.} - \bar{y}_{..})^2 - \sum_{i=1}^2 n_i(\bar{y}_{i..} - \bar{y}_{..})^2 - \sum_{j=1}^2 n_j(\bar{y}_{j..} - \bar{y}_{..})^2}{4 - 1 - 1 - 1} \quad (4.3)$$

The key aspect of Equation (4.3) is that it calculates the squared distance between the four factor-level group means and the overall mean after accounting for the main factor group means. Thus, this calculation is an estimate of how spread out the four group means are after accounting for any influence of the main factor means.

The denominator of the mean square for interaction is based on the denominators from  $MS_{\text{Brand}}$  in Equation (4.1) and  $MS_{\text{Time}}$  in Equation (4.2). In this example, there are four factor-level group means. Thus, the denominator is calculated as  $4 - 1 - (\text{denominator from } MS_{\text{Brand}}) - (\text{denominator from } MS_{\text{Time}}) = 4 - 1 - 1 - 1$ . Details for deriving mean squares are provided in the extended activities.

### Key Concept

The interaction term is not simply a measure of the spread between the four factor-level group means. It measures the remaining spread of the means after adjusting for differences between the main factor means.

**Within-Group Variability** The best estimate of the variability within each group (MSE) is simply a weighted average of the sample variances within each of the four factor-level groups:

$$\begin{aligned} MSE &= \frac{\sum_{i=1}^2 \sum_{j=1}^2 (n_{ij} - 1)s_{ij}^2}{(n_{11} - 1) + (n_{12} - 1) + (n_{21} - 1) + (n_{22} - 1)} \\ &= \frac{(8 - 1)s_{11}^2 + (8 - 1)s_{12}^2 + (8 - 1)s_{21}^2 + (8 - 1)s_{22}^2}{(8 - 1) + (8 - 1) + (8 - 1) + (8 - 1)} \end{aligned} \quad (4.4)$$

where  $s_{ij}^2$  is the sample variance for the group representing brand  $i$  and time  $j$ . The implicit assumption here is that the variances of the possible responses with each of the four group populations are all the same, so it makes sense to “pool” the sample variances into a single estimate of overall response variability. This **equal variance assumption** is key to the validity of the ANOVA statistical method. If the variability within each group is quite different, the MSE may not be an appropriate estimate. It is often useful to create individual value plots or side-by-side boxplots of the groups to check if the spreads of the sample groups are roughly similar.

### MATHEMATICAL NOTE

If groups of data from each factor-level combination have very different sample sizes and at least one group has a small sample size (e.g., less than 5 units per group), then ANOVA may not be appropriate. If the group(s) with the smallest sample size (s) has an unusually high variance, the MSE is likely to underestimate the true variance and ANOVA is likely to incorrectly reject the null hypothesis (conclude that there are differences when there really are no differences between group means). If the group(s) with the smallest sample size(s) has an unusually small variance, the MSE is likely to overestimate the true variance. The larger MSE may cause us to incorrectly fail to reject the null hypothesis (fail to detect true differences).

Equation (4.4) is often called the **mean square error (MSE)** of the responses, because “error” represents the unit-to-unit variability in the response that can’t be explained by any of the main factors or interactions. We are now ready to calculate a test statistic corresponding to each of the three hypotheses at the beginning of this section.

**The *F*-Statistic** The *F*-statistic is a ratio of the between-group variability (variation between factor-level averages) to the within-group variability (pooled estimate of variability within each factor-level combination):  $(\text{MS}_{\text{factor}})/\text{MSE}$ . Mathematical theory proves that if the assumptions of the ANOVA model hold, the *F*-statistic follows an *F*-distribution with degrees of freedom corresponding to the denominators of the MS for the factor being tested and the MSE. The ***p*-value** gives the likelihood of observing an *F*-statistic at least this large, assuming that the true population factor has equal level means. Thus, when the *p*-value is small, we conclude that there is a difference between the level means. Additional details are provided in the extended activities at the end of the chapter.

#### Key Concept

An *F*-statistic is simply the ratio of the between-group variability to the within-group variability.

### Activity ◀ Calculating *F*-Statistics

8. Use Equation (4.1) to estimate  $\text{MS}_{\text{Brand}}$ , the variability between Brand means.
9. Calculate the variability between Time means,  $\text{MS}_{\text{Time}}$ . Explain the key differences between Equation (4.1) and Equation (4.2).
10. Use Equation (4.3) to estimate  $\text{MS}_{\text{BrandTime}}$ .
11. Calculate the *F*-statistics corresponding to the three hypothesis tests:  $\frac{\text{MS}_{\text{Brand}}}{\text{MSE}}$ ,  $\frac{\text{MS}_{\text{Time}}}{\text{MSE}}$ , and  $\frac{\text{MS}_{\text{BrandTime}}}{\text{MSE}}$ .
12. What do you think are the largest and smallest possible values of any *F*-statistic?
13. Use the technology instructions provided on the CD to check your answers. Submit the software output. Note that a *p*-value for each *F*-statistic is provided. State your conclusions about each of the three-hypotheses based on these *p*-values.  
Don't be surprised if your hand calculations in Question 11 differ somewhat from the software output here. The data in Table 4.3 were rounded to one decimal place, so calculations in Questions 8 through 11 are not as accurate as statistical software.
14. Explain why a large *F*-statistic corresponds to a small *p*-value by referring to the definition of an *F*-statistic: a ratio of between-group variability to within-group variability.
15. **Checking Assumptions** As described in Chapter 2, assumptions need to be checked to ensure that the *p*-value for each ANOVA *F*-test is reliable:
  - The observations within each group (each factor-level combination) are independent and identically distributed.
  - Each group has equal variances.
  - The residual values follow a normal distribution with a mean of zero.
  - a. Examine the individual value plot in Figure 4.1 and comment on the assumptions for these hypothesis tests. Is there evidence of any skewness or outliers that may cause us to doubt the normal assumption for the PopRate within each factor-level combination of Brand and Time?
  - b. Does Figure 4.1 indicate that the spread of each group appears roughly similar, so the equal variance assumption seems reasonable? Another informal check of the equal variance assumption can be done by calculating the ratio of the maximum sample standard deviation to the minimum sample standard deviation. If this ratio is less than two, we can generally assume that there is not strong evidence against the equal variance assumption. Compare the standard deviations of the four treatment level combinations to determine if

$$\frac{\max(s_{ij})}{\min(s_{ij})} < 2$$

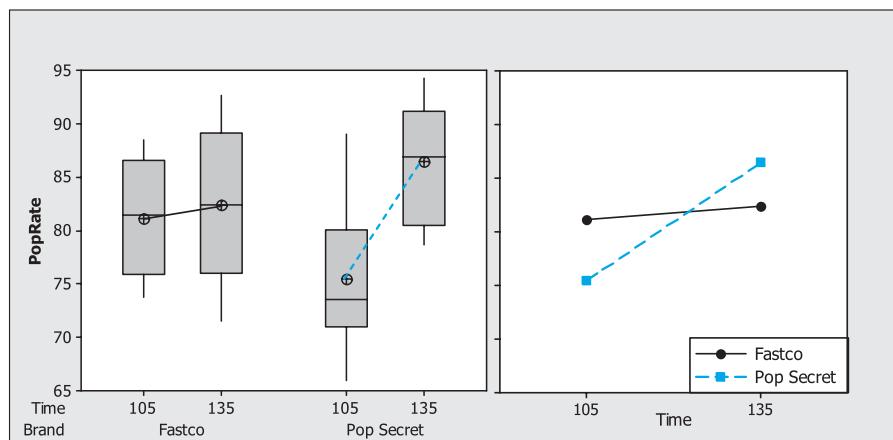
- c. Create a normal probability plot or histogram of the residuals from Question 13. Does it appear that the residuals follow a normal distribution?

**CAUTION**

Some statisticians will reject the equal variance assumption when the ratio of standard deviations is greater than 3 instead of 2. Others recommend that formal tests be used to test for equal variances. However, some tests, such as Bartlett's test, are very sensitive to nonnormality. Box criticized using Bartlett's test as a preliminary test for equal variances, saying "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port."<sup>4</sup> Levene's test of homogeneity of variance is less sensitive to departures from normality.<sup>5</sup>

## Interpreting Interaction Terms

In Question 13, the  $p$ -value corresponding to the third hypothesis test listed at the beginning of this section was 0.04. This demonstrates an **interaction**: the effect of one variable (`Time`) on the response depends on a second variable (`Brand`). Figure 4.2 provides a side-by-side boxplot and an interaction plot of the Popcorn data. An **interaction plot** is simply a plot of the four factor-level group means shown in Table 4.3. These plots show that for both brands, the average `PopRate` increases when the cooking time changes from 105 to 135 seconds. However, the change in means for the Fastco brand is very small compared to the change observed in the Pop Secret brand.



**Figure 4.2** Side-by-side boxplots and an interaction plot of the `PopRate` for each `Brand` and cooking `Time` factor-level combination.

The interaction plot is helpful in visualizing how the effect of one factor can depend on another factor, especially when there are multiple factors in the study. When the lines in an interaction plot are essentially parallel, the effect of the first variable is not influenced by a second variable. Nonparallel lines indicate an interaction between main factors (e.g., the effect of `Time` depends on `Brand`). However, the interaction plot does not show the within group variability, so only the  $p$ -value from the ANOVA can be used to determine if the interaction is significant. The  $p$ -value of 0.04 shows that the observed interaction effect is so large that it is unlikely to have occurred just by chance. We conclude that  $H_{a3}$  is true: `Brand` influences the effect of `Time` on `PopRate`.

## 4.4 Analyzing a Three-Way Factorial Design

One advantage of ANOVA is that the analysis can easily be extended to multiple factors with many levels. In this section, all three factors in the popcorn study (`Brand`, `Time`, and `Microwave`) will be

simultaneously examined for their influence on `PopRate` with a three-way ANOVA (also called a **three-factor ANOVA**).

Using only the 32 observations from the `Popcorn` data, a three-way ANOVA will allow us to simultaneously test the following six hypotheses:

1.  $H_{01}: \mu_{\text{Fastco}} = \mu_{\text{PopSecret}}$   
 $H_{a1}: \mu_{\text{Fastco}} \neq \mu_{\text{PopSecret}}$
2.  $H_{02}: \mu_{105} = \mu_{135}$   
 $H_{a2}: \mu_{105} \neq \mu_{135}$
3.  $H_{03}$ : Brand has no influence on how Time affects `PopRate`  
 $H_{a3}$ : there is an interaction between Brand and Time
4.  $H_{04}: \mu_{\text{Room}} = \mu_{\text{Lounge}}$   
 $H_{a4}: \mu_{\text{Room}} \neq \mu_{\text{Lounge}}$
5.  $H_{05}$ : Microwave has no influence on how Brand affects `PopRate`  
 $H_{a5}$ : there is an interaction between Microwave and Brand
6.  $H_{06}$ : Microwave has no influence on how Time affects `PopRate`  
 $H_{a6}$ : there is an interaction between Microwave and Time

NOTE

It is also reasonable to test for a three-way interaction.  $H_{a7}$ , the size of the effect of Time for each level of Brand, also depends on a third variable, Microwave. In practice, the three-way interaction effect may be difficult to interpret and some researchers choose not to include them in their analysis. The impacts of including additional tests are described in Chapter 5.

## Activity Conducting a Three-Way ANOVA

16. Create individual value plots of the eight possible factor-level groups listed in Table 4.1.
  - a. Do you see any patterns in the `PopRate` among these groups?
  - b. Does the spread of the responses within each group look roughly similar?
  - c. Are there any outliers or unusual observations for any group(s)?
17. Calculate the eight group standard deviations. If the largest standard deviation is no more than two times the smallest standard deviation, it is typically appropriate to assume equal variances for the population of responses within each group. Is it appropriate to assume equal variances in this `Popcorn` study?
18. Use a statistical software package to simultaneously test all six hypotheses given at the beginning of this section.
  - a. Submit the appropriate software output.
  - b. Create a normal probability plot (described in Chapter 2) or a histogram of the residuals. Are the residuals consistent with the assumption of a normal distribution?
  - c. Use the  $p$ -value corresponding to each hypothesis to state your conclusions.
  - d. Address how random sampling and random allocation influence your conclusions.
19. Determine whether  $\text{MS}_{\text{Brand}}$ ,  $\text{MS}_{\text{Time}}$ , and  $\text{MS}_{\text{BrandTime}}$  are the same as in Question 13. Explain why the  $F$ -statistics corresponding to these three hypotheses have changed.

In completely randomized designs, all  $F$ -statistics corresponding to the tests for each main factor and interaction use the same denominator. The mean square for each main factor ( $\text{MS}_{\text{Brand}}$ ,  $\text{MS}_{\text{Time}}$ , and  $\text{MS}_{\text{Microwave}}$ ) is a measurement of the variability between level means. Since this is a balanced design, when the level means for a factor are farther apart, the corresponding mean square and  $F$ -statistics are larger. Thus, the main effects plot shown in Figure 4.3 allows us to quickly see that the `Time` factor is the most significant (has the smallest  $p$ -value) and the `Brand` factor is the least significant.

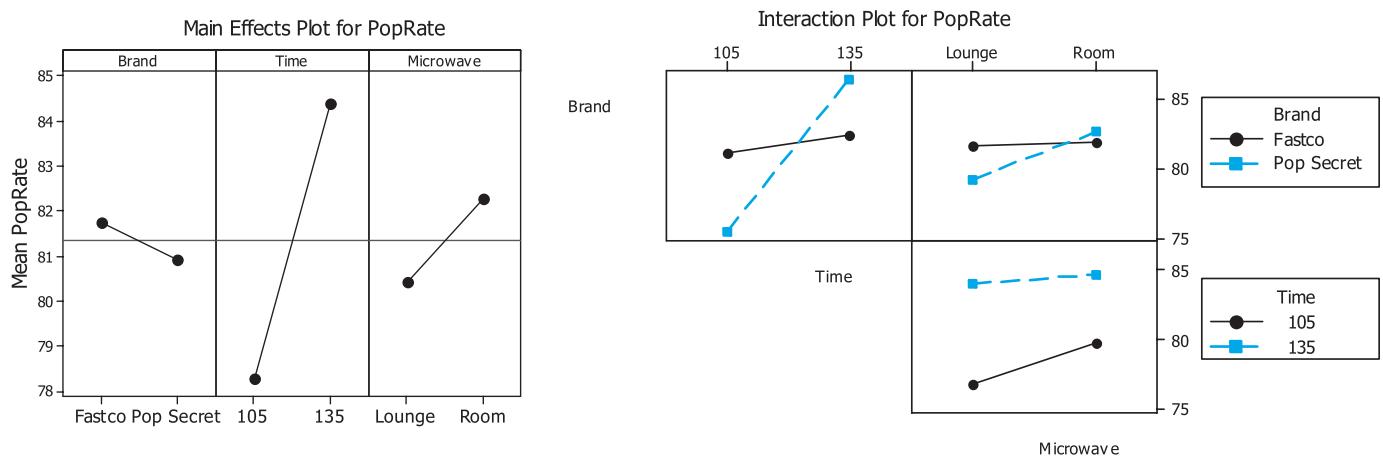


Figure 4.3 Main effect plots and interaction plots for a three-factor ANOVA.

Figure 4.3 also provides interaction plots corresponding to the three hypotheses tests about interactions. Using the same logic, Figure 4.3 shows that the hypothesis test corresponding to the Brand and Time interaction will have the smallest  $p$ -value. While both the effect of Time and the effect of Brand are somewhat influenced by Microwave, the effect of Time is most influenced by changing Brand.

## 4.5 What Can We Conclude from the Popcorn Study?

Yvonne and Tue's study illustrated how essential it is to carefully plan out a study before any data are collected. If the data are not collected properly, typically there is *no* statistical analysis that can draw accurate conclusions. When the study is well designed and the data are reliable, analysis is often straight forward with statistical software.

The units in this study (Bags) were randomly assigned to Time and Microwave factor-level combinations. The ANOVA results allowed us to conclude that Time causes a difference in PopRate. In addition, we found evidence that there is a Brand and Time interaction.

The bags of popcorn in this study were not a true random sample of all popcorn produced by these two brands. Thus, we need to be careful about making any conclusions that extend to a larger population. Because of the efforts the students made to properly collect random samples from various stores around their college town, the author of this chapter would feel fairly comfortable stating that the conclusions hold for these two brands of butter-flavored microwave popcorn in their town at the time of this study.

### A Closer Look Completely Randomized Factorial Studies

## 4.6 Paper Towels: Developing a Statistical Model for a Two-Way Factorial Design

As a final project in an introductory statistics class, several students decided to conduct a study to test the strength of paper towels. Several television advertisements had claimed that a certain brand of paper towel was the strongest, and these students wanted to determine if there really was a difference. The students sampled 26 towels from two brands of paper towels, Comfort and Decorator.

**NOTE**

Recall that random sampling is needed to extend the results to a larger population. If the students randomly sampled 26 towels from just one roll, the conclusions would hold only for that roll. Ideally they should have randomly purchased 26 rolls of each brand from multiple locations and then randomly selected one towel per roll.

Before any data were collected, these students determined that the following conditions should be held as constant as possible throughout the study:

- Paper towels were selected that had the same size.
- The towels were held at all four corners by two people.
- Weights (10, 25, 50, 100, or 250 grams) were slowly added to the center of each towel by a third person until it broke.

In this study, there are two factors. One has two levels, Comfort (Brand C) or Decorator (Brand D), and the other has three levels (0, 5, or 15 drops of water applied to the center of the paper towel). This leads to  $2 \times 3 = 6$  conditions, called **factor-level combinations** or **factorial combinations**:

Brand C and 0 drops of water  
 Brand C and 5 drops of water  
 Brand C and 15 drops of water  
 Brand D and 0 drops of water  
 Brand D and 5 drops of water  
 Brand D and 15 drops of water

Twenty-six sheets were tested at each of the six factor-level combinations. Thus, there are 156 experimental units used in this study. The response variable is the breaking strength of each paper towel in grams. Breaking strength is defined as the total weight that each towel successfully held. The next additional weight caused the towel to break.

The three null hypotheses corresponding to this two-factor design are as follows:

1.  $H_{01}$ : there is no difference in mean strength between the two brands of paper towel  
 $H_{a1}$ : the two brand means are different
2.  $H_{02}$ : there is no difference in mean strength when 0, 5, or 15 drops of water are used  
 $H_{a2}$ : the mean strength of at least one water amount group is different from the others
3.  $H_{03}$ : the amount of water has no influence on how brand affects strength  
or  $H_{03}$ : the effect of the amount of water on strength is the same for both brands  
or  $H_{03}$ : there is no interaction between brand and water  
 $H_{a3}$ : there is an interaction between brand and water

Table 4.4 represents some of the data for the paper towel study. Each of the six cells has 26 observations. The complete data set is in the file `PaperTowels`. While not all observations are shown, Table 4.4 helps us understand the data structure. After the data have been collected, the averages for all meaningful groups of the data can be calculated as shown in Table 4.5.

## Extended Activity Algebraic Notation

Data set: `PaperTowels`

20. What values in Table 4.4 are represented by  $y_{213}$  and  $y_{122}$ ?
21. Give the proper algebraic notation for the observation representing the 3rd paper towel with Brand D and 15 drops of water.
22. What are the values of  $\bar{y}_{.3.}$  and  $\bar{y}_{21.}$ ?
23. Complete Table 4.5 by calculating the three missing averages.

**Table 4.4** Strength of paper towels (grams of weight added before the towel broke). The Brand factor has two levels:  $i = 1$  represents Comfort and  $i = 2$  represents Decorator towels. The Water factor has three levels:  $j = 1, 2$ , and  $3$  represent 0 drops, 5 drops, and 15 drops of water, respectively.

		Amount of Water		
		Breaking Strength of Paper Towels (grams)	0 drops ( $j = 1$ )	5 drops ( $j = 2$ )
Brand of Towel	Comfort, Brand C ( $i = 1$ )	3200	2000	375
	Comfort, Brand C ( $i = 1$ )	3400	1800	475
	Comfort, Brand C ( $i = 1$ )	2800	1700	500
	Comfort, Brand C ( $i = 1$ )	⋮	⋮	⋮
	Comfort, Brand C ( $i = 1$ )	3100	1800	325
	Decorator, Brand D ( $i = 2$ )	2400	875	400
	Decorator, Brand D ( $i = 2$ )	2400	600	450
	Decorator, Brand D ( $i = 2$ )	2000	825	325
	Decorator, Brand D ( $i = 2$ )	⋮	⋮	⋮
	Decorator, Brand D ( $i = 2$ )	1700	700	300

**Table 4.5** Average Strength of paper towels (grams).

		Amount of Water			Brand Average
Strength		0 drops	5 drops	15 drops	
Brand of Towel	Brand C	$\bar{y}_{11} = 3205.8$			$\bar{y}_{1..} = 1772.8$
	Brand D	$\bar{y}_{21} = 2219.2$	$\bar{y}_{22} = 704.8$	$\bar{y}_{23} = 446.2$	$\bar{y}_{2..} = 1123.4$
Water Average		$\bar{y}_{1..} = 2712.5$		$\bar{y}_{3..} = 423.6$	$\bar{y}_{...} = 1448.1$

## Calculating Effects

As the data structure becomes more complex, a statistical model becomes more useful for describing the population(s) from which the data may have come. Generally, statistical models consist of a **mean response** and a **random error term** (details are provided in Chapter 2). The mean response describes the expected (mean) breaking strength. Figure 4.4 is useful in visualizing the meaningful groups within this model that contribute to the mean response of the model: the grand mean, the two brand groups, the three water amount groups, and the six factor-level combination groups.

The random error term follows an overall pattern that can be modeled with a probability distribution (e.g., the normal distribution). The error term incorporates the reality that observations will vary within each factor-level combination. Even when the same weights are applied to the same brand of paper towel, using the same water amount, the observed breaking strength may not be the same.

The labels (e.g., the grand mean, group effects, and random errors) and symbols in Figure 4.4 are described below:

$y_{ijk}$ : the  $k$ th observed breaking strength ( $k = 1, 2, \dots, 26$ ) for brand  $i$  and water amount  $j$

$\mu$ : overall mean breaking strength of the entire population of paper towels across brands and water amounts (also called the grand mean)

$\alpha_i$ : brand effect ( $i = 1, 2$ ), where  $\alpha_1$  is the effect of Brand C and  $\alpha_2$  is the effect of Brand D

$\beta_j$ : amount of water effect ( $j = 1, 2, 3$ ), where  $\beta_1$  represents the effect of 0 drops of water

$(\alpha\beta)_{ij}$ : interaction effect, where  $(\alpha\beta)_{23}$  represents the Brand D/15 drops of water interaction effect

$\varepsilon_{ijk}$ : the random error—the difference between the  $k$ th observed value ( $k = 1, 2, \dots, 26$ ) and the population mean breaking strength for brand  $i$  and water amount  $j$

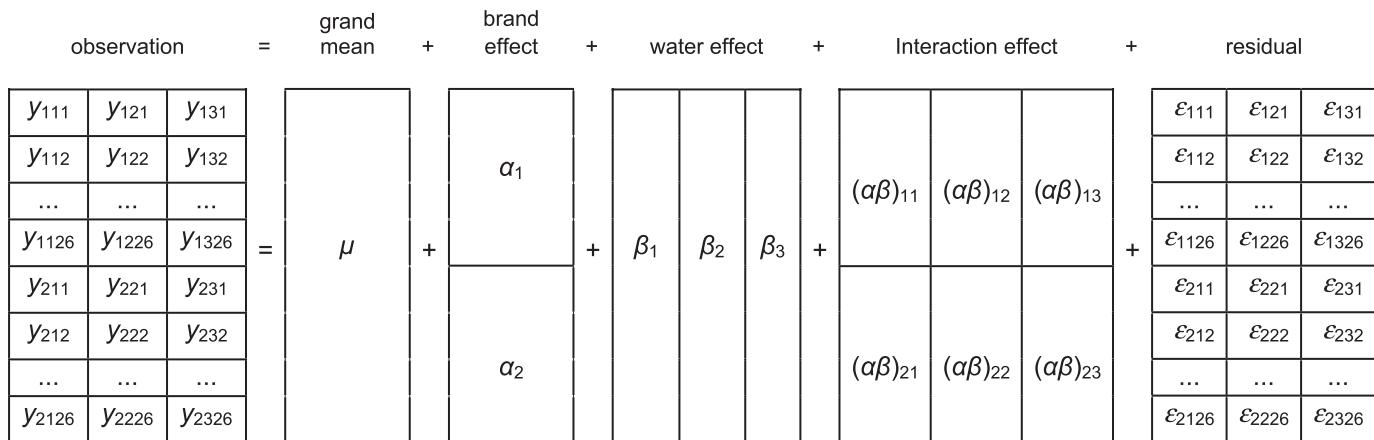


Figure 4.4 Two-way factorial diagram.

Notice that each of the  $2 \times 3 \times 26 = 156$  observed strength measurements represents one of the 156 equations in Figure 4.4. The structure of the data is now used to write down a statistical model for the data:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \text{ for } i = 1, 2, j = 1, 2, 3, \text{ and } k = 1, 2, \dots, 26 \quad (4.5)$$

#### Key Concept

Identifying the meaningful groups within each data set (the data structure) is the first step in developing a statistical model of the population(s) from which the data come.

Table 4.5 shows that the strength average changes from 1772.8 to 1123.4 with a change from Brand C to Brand D paper towels. This difference is smaller than the differences due to changes in the water amount. **Main effects** are calculated to measure the impact of changing the levels of each factor in the model. A main effect is the difference between the factor-level average and the grand mean. For example,

$$\hat{\alpha}_1 = \text{effect of Brand C} = \text{Brand C mean} - \text{grand mean}$$

$$= \bar{y}_{1..} - \bar{y}_{..} = 1772.8 - 1448.1 = 324.7$$

$$\hat{\beta}_3 = \text{effect of 15 drops of water} = \text{15 drops mean} - \text{grand mean}$$

$$= \bar{y}_{.3} - \bar{y}_{..} = 423.6 - 1448.1 = -1024.5 \quad (4.6)$$

#### NOTE

$\mu$ ,  $\alpha_i$ ,  $\beta_j$ , and  $(\alpha\beta)_{ij}$  in Equation (4.5) are population parameters. Statistics such as  $\hat{\alpha}_1$  and  $\hat{\beta}_3$  in Equation (4.6) are used to estimate the population effect sizes.

### Extended Activity

#### Estimating Main Effects

Data set: PaperTowels

24. Use the PaperTowels data to estimate the effect of Brand D and explain any symmetry that you find with the effect of Brand C calculated above.

25. A **main effects plot** is a graph that plots the average response for each level of each factor. To properly compare the effect sizes, the vertical axis should be the same for each factor. Use statistical software to create a main effects plot. Identify and label the following values on the plot:  $\bar{y}_{1..}$ ,  $\bar{y}_{3..}$ ,  $\bar{y}_{...}$ ,  $\hat{\alpha}_1$ , and  $\hat{\beta}_3$ .

 **NOTE**

This is still a balanced design, since  $n_{ij}$  is the same for each factor-level combination. However, since the sample sizes are not the same in every group level (78 towels for each Brand mean and 52 towels for each Water level), the factor with the largest difference between means does not necessarily correspond to the smallest  $p$ -value.

In addition to determining the main effects for each factor, it is often critical to identify how multiple factors interact in affecting the results. An interaction occurs when one factor affects the response variable differently depending on a second factor. To calculate the effect of the brand and water interaction, take the average for a particular factor combination minus the grand mean and the corresponding main effects.

$$\begin{aligned}
 & \text{Interaction effect of Brand C and 15 drops of water} \\
 &= \text{average of Brand C, 15 drops group} \\
 &\quad - (\text{effect of Brand C} + \text{effect of 15 drops} + \text{grand mean}) \\
 &= \bar{y}_{13..} - [\hat{\alpha}_1 + \hat{\beta}_3 + \bar{y}_{...}] \\
 &= \bar{y}_{13..} - [(\bar{y}_{1..} - \bar{y}_{...}) + \bar{y}_{3..} - \bar{y}_{...}] + \bar{y}_{...} \\
 &= 401.0 - [324.7 + (-1024.5) + 1448.1] \\
 &= -347.3
 \end{aligned} \tag{4.7}$$

The estimate of the Brand C and 15 drops of water interaction effect in Equation (4.7) tells us that the best estimate of any paper towel strength from this group should be reduced by an additional 347.3 after we take into account all other influencing factors (the grand mean and main effects).

**Extended Activity** 

### Calculating Interaction Effects

Data set: PaperTowels

26. Show that  $\bar{y}_{ij..} - \bar{y}_{i..} - \bar{y}_{j..} + \bar{y}_{...}$  is equivalent to the  $ij$ th interaction effect.
27. Calculate the other five interaction effects. Hand draw Figure 4.4 and fill out the effect sizes with observed values (i.e., replace  $\mu$ ,  $\alpha_i$ ,  $\beta_j$ , and  $(\alpha\beta)_{ij}$  with estimates from the data). Do not fill out the observations or the error terms ( $y_{ijk}$  or  $\varepsilon_{ijk}$ ).
28. Create an interaction plot. Does there appear to be evidence of an interaction effect?
29. Draw a diagram similar to Figure 4.4 for a two-way factorial design with four levels of the first factor, three levels of the second factor, and two observations per factor-level combination in Equation (4.5).
30. Residuals, or observed random error terms, are defined as the observed responses,  $y_{ijk}$ , minus the estimate for the mean response (the sum of the grand mean, the two main effects, and the interaction effect). Calculate the residual values for  $y_{213}$  and  $y_{122}$ .

The effect for Brand C might be positive (Brand C has a higher average breaking strength than Brand D), but then the effect for Brand D must be negative and exactly the same size as the effect for Brand C. The two effects sum to zero. This is called a **restriction** on the model terms. The entire set of restrictions for the model in Equation (4.5) is provided below.

$$\sum_{i=1}^2 \alpha_i = \alpha_1 + \alpha_2 = 0, \quad \sum_{j=1}^3 \beta_j = \beta_1 + \beta_2 + \beta_3 = 0$$

$$\begin{aligned}\sum_{i=1}^2 (\alpha\beta)_{ij} &= (\alpha\beta)_{1j} + (\alpha\beta)_{2j} = 0 \text{ for all } j \\ \sum_{j=1}^3 (\alpha\beta)_{ij} &= (\alpha\beta)_{i1} + (\alpha\beta)_{i2} + (\alpha\beta)_{i3} = 0 \text{ for all } i\end{aligned}\quad (4.8)$$

More specifically, the restrictions state that the interaction effects involving Brand C must sum to zero (i.e.,  $(\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{13} = 0$ ). In the same way, all interactions corresponding to the 15 drops of water groups must also sum to zero. All six restrictions corresponding to the interactions can be checked in Question 27. The residual values sum to zero within each group of interest. This will always be true whenever calculating effects. This is not surprising, since effects measure the deviation of a particular group mean from the overall mean.

## 4.7 Paper Towels: The Relationship Between Effects and ANOVA

Each of the three hypothesis tests in the paper towel study is tested on a separate line in an ANOVA table. The  $F$ -statistics for an ANOVA were already calculated earlier in this chapter by comparing between group and within group variability. This section will show the relationship between calculating effects and the ANOVA table. For each null hypothesis, the statement “the means are equal for all levels of a factor” is equivalent to the statement “factor effects are zero.”

The **sum of squares** (SS) for a main factor in the multi-factor ANOVA is identical to the one-factor SS described in Chapter 2. The sum of squares (the numerator of the mean square calculation) is the sum of all squared effects corresponding to that factor. For the `Brand` effect, this is written mathematically as

$$\begin{aligned}SS_{\text{Brand}} &= \sum (\text{Brand effect on each of the 156 observations})^2 \\ &= \sum_{i=1}^2 78(\bar{y}_{i..} - \bar{y}_{...})^2\end{aligned}$$

This equation can be generalized. Instead of 78 elements in each `Brand` group, we can indicate  $n_i$  elements. Instead of 2 levels for `Brand`, we can indicate  $I$  levels. The first factor, `Brand`, can be labeled factor  $A$ ; the second factor, `Water`, can be labeled factor  $B$ ; etc. Then

$$SS_{\text{Brand}} = SS_A = \sum_{i=1}^I n_i(\bar{y}_{i..} - \bar{y}_{...})^2 \quad \text{for } I = 2 = \text{number of Brand levels} \quad (4.9)$$

Similarly,  $SS_{\text{Water}} = SS_B$  is the sum of squares for the `water` effect on each observation.

$$\begin{aligned}SS_{\text{Water}} &= SS_B = \sum (\text{Water effect on each of the 156 observations})^2 \\ &= \sum_{j=1}^3 52(\bar{y}_{j..} - \bar{y}_{...})^2 \\ &= \sum_{j=1}^J n_j(\bar{y}_{j..} - \bar{y}_{...})^2 \quad \text{for } J = 3 = \text{number of water levels}\end{aligned}\quad (4.10)$$

The sum of squares for the interaction term,  $SS_{AB}$ , is

$$\begin{aligned}SS_{AB} &= \sum (\text{interaction effect on each of the 156 observations})^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J n_{ij}(\text{ijth level effect})^2 \\ &= \sum_{i=1}^2 \sum_{j=1}^3 26(\bar{y}_{ij..} - \bar{y}_{i..} - \bar{y}_{j..} + \bar{y}_{...})^2\end{aligned}\quad (4.11)$$

The **error sum of squares** ( $SS_{\text{Error}}$ ) measures the spread of the observed residuals. Each residual is defined as an observed value minus the estimated value:  $\hat{\varepsilon}_{ijk} = y_{ijk} - \bar{y}_{ij.}$

$$\begin{aligned}
 SS_{\text{Error}} &= \sum (\text{each residual effect})^2 \\
 &= \sum_{i=1}^I \sum_{j=1}^J \left[ \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij.})^2 \right] \\
 &= \sum_{i=1}^I \sum_{j=1}^J [(n_{ij} - 1) \times s_{ij}^2] \\
 &= 25 \times s_{11}^2 + 25 \times s_{12}^2 + 25 \times s_{13}^2 + 25 \times s_{21}^2 + 25 \times s_{22}^2 + 25 \times s_{23}^2 \quad (4.12)
 \end{aligned}$$

The **total sum of squares** ( $SS_{\text{Total}}$ ) measures the overall spread of the responses in the full data set.

$$\begin{aligned}
 SS_{\text{Total}} &= \sum (\text{distance between each observation and the grand mean})^2 \\
 &= \sum_{i=1}^I \sum_{j=1}^J \left[ \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{...})^2 \right] \\
 &= (N - 1) \times s^2 \quad (4.13)
 \end{aligned}$$

#### ► MATHEMATICAL NOTE ▼

The variance within each factor-level group is calculated as

$$s_{ij}^2 = \frac{\sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij.})^2}{n_{ij} - 1}$$

and the overall sample variance of the response variable is

$$s^2 = \frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{...})^2}{N - 1}$$

where  $N = 156$  is the total sample size.

## Degrees of Freedom

**Degrees of freedom** (df) are determined by how many “free” pieces of information are available when calculating effects. For example, Equation (4.8) shows that each of the main effects must sum to zero. Thus, knowing the effects of any two levels of Water forces a known effect for the last level. In our example, the effect of 0 drops of water increases the expected mean strength by 1264.4. Similarly, the effect of using 5 drops of water is  $-239.9$ . The the effects must sum to zero ( $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 1264.4 - 239.9 - 1024.5 = 0$ ).

#### Key Concept

For any main factor with  $J$  levels, one effect is fixed if we know the other  $J - 1$  effects. Thus, when there are  $J$  levels for a main factor of interest, there are  $J - 1$  degrees of freedom (free pieces of information).

## Extended Activity

### Calculating Degrees of Freedom for Interaction Terms

Data set: PaperTowels

31. Table 4.6 is a table with two rows and three columns, similar to the interaction effect term in the two-way factorial diagram in Figure 4.4. However, for this question we will assume that only two effects are known:  $(\widehat{\alpha\beta})_{11} = 2$  and  $(\widehat{\alpha\beta})_{12} = -5$ .

**Table 4.6** Interaction effects.

	0	5	15
Brand C	2	-5	
Brand D			

- a. Equation (4.8) states that all the  $AB$  effects within Brand C add up to zero [ $(\widehat{\alpha\beta})_{11} + (\widehat{\alpha\beta})_{12} + (\widehat{\alpha\beta})_{13} = 0$ ]. Use this rule to calculate  $(\widehat{\alpha\beta})_{13}$ .
  - b. Equation (4.8) also states that all the  $AB$  effects within 0 water amount add up to zero (the same for 5 and 15 drops of water). Use this rule to calculate  $(\widehat{\alpha\beta})_{21}$ ,  $(\widehat{\alpha\beta})_{22}$ , and  $(\widehat{\alpha\beta})_{23}$ .
  - c. Consider a different interaction table with two rows and three columns. Explain why it is not possible to have effects of  $(\widehat{\alpha\beta})_{11} = 4$ ,  $(\widehat{\alpha\beta})_{13} = -4$ , and  $(\widehat{\alpha\beta})_{22} = 6$  and still follow the restrictions in Equation (4.8).
  - d. What are the degrees of freedom corresponding to any interaction term (in a balanced completely randomized design) with two levels of factor  $A$  and three levels of factor  $B$ ? In other words, under the restrictions in Equation (4.8), what is the number of free pieces of information (the number of cells in Table 4.6 that are not fixed)?
32. Table 4.7 is another table of interaction effects, with five rows and three columns (five levels of factor  $A$  and three levels of factor  $B$ ). Again, we will assume that only some of the effects are known.

**Table 4.7** Interaction effects.

3	1	
-2	6	
1	4	
3	5	

- a. Use Equation (4.8) to calculate the effect corresponding to each of the remaining cells.
  - b. In Table 4.7, eight cells are filled. If only seven cells were filled, would it be possible to calculate the effects corresponding to all remaining cells?
  - c. What are the degrees of freedom corresponding to any interaction term (in a balanced completely randomized design) with five levels of factor  $A$  and three levels of factor  $B$ ? In other words, what is the minimum number of cells that must be filled in order to allow us to use Equation (4.8) to estimate all other effects?
33. Use the previous two questions to determine the degrees of freedom for an interaction term for a balanced completely randomized design with three levels of factor  $A$  and four levels of factor  $B$ .

For the  $AB$  interaction term, there are  $I \times J$  effects that are calculated. In the popcorn study,  $I \times J = 2 \times 3 = 6$ . Each effect represents one piece of information. In addition:

- All the  $AB$  effects within Brand C add up to zero [ $(\widehat{\alpha\beta})_{11} + (\widehat{\alpha\beta})_{12} + (\widehat{\alpha\beta})_{13} = 0$ ]. Within Brand C, if two effects are known, the third will be fixed (the same holds for Brand D). Thus, these restrictions eliminate  $I = 2$  free pieces of information (free cells in an interaction effects table).

- Similarly, all the  $AB$  effects within 0 water amount add up to zero [ $(\widehat{\alpha\beta})_{11} + (\widehat{\alpha\beta})_{21} = 0$ ]. The same restriction holds for all other levels of factor  $B$  (for 5 and 15 drops of water). Thus, an additional  $J = 3$  pieces of information are no longer free. However, one piece of information is already fixed from the requirement that the sum of all brand effects is zero. Thus, only  $J - 1 = 3 - 1 = 2$  free pieces of information are taken for water amounts.

The degrees of freedom for the interaction effect are

$$\begin{aligned}
 df_{AB} &= \text{number of interaction effects} - [df_A + df_B + 1] \\
 &= IJ - [(I - 1) + (J - 1) + 1] \\
 &= IJ - I - J + 1 \\
 &= (I - 1)(J - 1)
 \end{aligned} \tag{4.14}$$

#### ► MATHEMATICAL NOTE ▼

Calculating interaction degrees of freedom as  $(I - 1)(J - 1)$  in Equation (4.14) is quite easy. However, the reason Equation (4.14) also shows interaction  $df = IJ - [(I - 1) + (J - 1) + 1]$  is that this follows the calculation of the interaction effect shown in Equation (4.7):  $\bar{y}_{ij..} - [(\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{j..} - \bar{y}_{...}) + \bar{y}_{...}]$ . The key point is to recognize that knowing how the effects are calculated drives formulas for both sum of squares and degrees of freedom. This is also true for more complex designs beyond the scope of this chapter.

#### Key Concept

For each term in a model, degrees of freedom represent the number of cells in a factor diagram that must be filled before all other cells can be filled with no added information. Thus, degrees of freedom are the number of “free” effects before the restrictions allow us to predict all other effects.

Assuming each of the  $IJ$  groups has  $K$  observations, these calculations are summarized in Table 4.8.

**Table 4.8** Two-way ANOVA table.

Source	df	SS	MS	F-Statistic
$A$	$I - 1$	$\sum_{i=1}^I n_i (\bar{y}_{i..} - \bar{y}_{...})^2$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MSE}$
$B$	$J - 1$	$\sum_{j=1}^J n_j (\bar{y}_{j..} - \bar{y}_{...})^2$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MSE}$
$AB$	$(I - 1)(J - 1)$	$\sum_{i=1}^I \sum_{j=1}^J n_{ij} (\bar{y}_{ij..} - \bar{y}_{i..} - \bar{y}_{j..} + \bar{y}_{...})^2$	$\frac{SS_{AB}}{df_{AB}}$	$\frac{MS_{AB}}{MSE}$
Error	$IJ(K - 1)$	$\sum_{i=1}^I \sum_{j=1}^J [(n_{ij} - 1) \times s_{ij}^2]$	$MSE = \frac{SSE}{df_{Error}}$	
Total	$N - 1$	$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (\bar{y}_{ijk} - \bar{y}_{...})^2$		

## Extended Activity ► Analyzing the Paper Towel Data

Data set: PaperTowels

34. **Checking Assumptions** In the statistical model in Equation (4.5), the following assumptions need to be validated about the random error terms,  $\varepsilon_{ijk}$ , before any formal hypothesis test can be developed:
- The error terms are independent and identically distributed.
  - The error terms follow a normal probability distribution, denoted as  $\varepsilon \sim N(0, \sigma^2)$ .

Note that the second assumption includes an equal variance assumption about the random errors from the different factor-level groups:  $\sigma_{11}^2 = \sigma_{12}^2 = \sigma_{13}^2 = \sigma_{21}^2 = \sigma_{22}^2 = \sigma_{23}^2$ .

- The independence assumption implies that there is no relationship between one observation and the next. The identically distributed assumption means that each observation sampled within each brand/water combination is from a population with the same mean and variance. If all 26 paper towels were sampled from one roll to assess the Brand C/5 drops factor combination, would you be concerned about violating the independence and/or the identically distributed assumption? Why or why not?
  - Calculate the sample means and standard deviations of all six factor-level combinations. Clearly, some groups have much larger variation than others. In addition, the variation within each group increases as the average breaking strength increases. To address this issue, a transformation of the data can often be used that will “stabilize” the variances so that the equal variance assumption is reasonable on this new scale. (Chapter 2 describes transformations in more detail.)
  - Transform the response variable using natural log (Strength) and  $\sqrt{\text{Strength}}$ . Did both transformations improve the equal variance assumption?
35. **Visualizing the Data** Draw individual value plots or side-by-side boxplots of the square-root transformed responses in the six factor-level groups.
- Is there any extreme skewness or outliers that would cause us to question the normality assumption?
  - Without doing any statistical calculations, do you expect to reject the three null hypotheses for the paper towel study? Justify your answer by visually comparing the variation (i.e., the spread) in the strength between groups to the variation within groups.
36. **Analyzing the Data** Use computer software to conduct an analysis on the square-root transformed PaperTowels data to test for differences between brands, water levels, and interactions. Use the ANOVA as well as appropriate graphs to state your conclusions about the paper towel study.

## 4.8 Contrasts and Multiple Comparisons

Chapter 2 describes a study where researchers tested whether a color distracter influenced the completion time of an online computer game. In addition to a color distracter, they were also interested in whether subjects could play the game more quickly with their right or left hand. Chapter 2 was restricted to one-factor ANOVAs. Thus, in that chapter the data were sorted into four groups: StandardRight, ColorRight, StandardLeft, and ColorLeft. Instead of testing for evidence against a general hypothesis test ( $H_0: \mu_{SR} = \mu_{CR} = \mu_{SL} = \mu_{CL}$ ), the two-way ANOVA allows us to test three more specific hypotheses of interest.

### Extended Activity

#### Comparing One-Way and Two-Way ANOVA

Data set: Games2

- The data set Games2 shows a column Type2 with four types of games based on distracter and which hand was used. Conduct an ANOVA using Type2 (just one explanatory variable with four levels) to test for differences in completion time. What is the  $p$ -value corresponding to the null hypothesis  $H_0: \mu_{SR} = \mu_{CR} = \mu_{SL} = \mu_{CL}$  versus the alternative  $H_a$ : at least one mean is different from another?
- Conduct a two-way ANOVA using Type, Hand, and the Type\*Hand interaction to test for differences in completion time. List the three null and alternative hypotheses and provide a  $p$ -value for each test.
- Since the same response variable is used for both Question 37 and Question 38, it should not be surprising that the total sum of squares is identical for both questions. Compare the other sums of squares in the ANOVAs from Questions 37 and 38. How is  $SS_{\text{Type2}}$  related to  $SS_{\text{Type}}$ ,  $SS_{\text{Hand}}$ , and  $SS_{\text{TypeHand}}$ ?

Since Question 37 leads us to reject  $H_0: \mu_{SR} = \mu_{CR} = \mu_{SL} = \mu_{CL}$ , it seems reasonable to conduct a multiple comparisons test (conduct multiple tests to identify differences between each group mean and every other group mean).

**NOTE**

There are six possible comparisons when there are four group means. Chapter 1 discussed familywise type I error and comparisonwise type I error. The **least-significant differences method** is a technique using comparisonwise type I error: If the  $p$ -value is less than  $\alpha$ , reject  $H_0$  in favor of  $H_a$ . Assuming that a particular null hypothesis is true, the least-significant differences method has an  $\alpha\%$  chance of (incorrectly) rejecting that hypothesis. When multiple tests are conducted on the same data set, the least-significant differences method leads to type I errors: rejecting null hypotheses when they should not be rejected.

**Bonferroni's method** is an example of a technique that maintains familywise type I error: If the  $p$ -value is less than  $\alpha/K$  (where  $K$  is the number of pairs), reject  $H_0$  in favor of  $H_a$ . With the familywise type I error =  $\alpha$ , assuming that there really is no difference between any of the  $K$  pairs, there is only an  $\alpha\%$  chance that any test will reject  $H_0$ . This leads to type II errors: failing to reject null hypotheses when they should be rejected. Chapter 1 describes the need for multiple comparison procedures and describes both the least-significant difference and the Bonferroni method in more detail.

Question 38 can be thought of as testing for orthogonal contrasts. An **orthogonal contrast** is a linear combination of treatment means where the coefficients add up to zero. For example, before they collected any data, the researchers in the game study were interested in several specific comparisons:

- Comparing standard to color games:  $H_{01}: (1)\mu_{SR} + (-1)\mu_{CR} + (1)\mu_{SL} + (-1)\mu_{CL} = 0$ . This is mathematically equivalent to  $H_{01}: (\frac{1}{2})\mu_{SR} + (\frac{1}{2})\mu_{SL} = (\frac{1}{2})\mu_{CR} + (\frac{1}{2})\mu_{CL}$  or  $H_{01}: \mu_S = \mu_C$ .
- Comparing right to left hand:  $H_{02}: (1)\mu_{SR} + (1)\mu_{CR} + (-1)\mu_{SL} + (-1)\mu_{CL} = 0$ . This is mathematically equivalent to  $H_{02}: \mu_R = \mu_L$ .
- Testing for an interaction:  $H_{03}: (1)\mu_{SR} + (-1)\mu_{CR} + (-1)\mu_{SL} + (1)\mu_{CL} = 0$ .

Each of these linear combinations of population means can be estimated by a contrast:

$$\text{Contrast 1 (C1)} = (1)\bar{y}_{SR} + (-1)\bar{y}_{CR} + (1)\bar{y}_{SL} + (-1)\bar{y}_{CL}$$

$$\text{Contrast 2 (C2)} = (1)\bar{y}_{SR} + (1)\bar{y}_{CR} + (-1)\bar{y}_{SL} + (-1)\bar{y}_{CL}$$

$$\text{Contrast 3 (C3)} = (1)\bar{y}_{SR} + (-1)\bar{y}_{CR} + (-1)\bar{y}_{SL} + (1)\bar{y}_{CL}$$

The coefficients of each of the linear combinations sum to zero:

$$\text{Coefficients for contrast 1} = C_{11} + C_{21} + C_{31} + C_{41} = (1) + (-1) + (1) + (-1) = 0$$

$$\text{Coefficients for contrast 2} = C_{12} + C_{22} + C_{32} + C_{42} = (1) + (1) + (-1) + (-1) = 0$$

$$\text{Coefficients for contrast 3} = C_{13} + C_{23} + C_{33} + C_{43} = (1) + (-1) + (-1) + (1) = 0$$

**Key Concept**

For any null hypothesis test comparing  $G$  group means  $H_0: \mu_1 = \mu_2 = \dots = \mu_G$  versus the alternative  $H_a$ : at least one group mean is different from another, a contrast is an estimate of a linear combination of the group means: Contrast 1 (C1) =  $(C_{11})\bar{y}_1 + (C_{21})\bar{y}_2 + \dots + (C_{G1})\bar{y}_G$ , where the coefficients sum to zero:  $C_{11} + C_{21} + \dots + C_{G1} = 0$ .

**MATHEMATICAL NOTE**

Any set of  $G$  group means (four group means in our case;  $H_0: \mu_{SR} = \mu_{CR} = \mu_{SL} = \mu_{CL}$ ) being compared in a between group sum of squares can be used to create  $G - 1$  mutually orthogonal contrasts ( $H_{01}, H_{02}, H_{03}$ ). Any two contrasts are said to be orthogonal if the dot product (sum of the cross products) of their coefficient vectors is zero. Contrasts must be orthogonal to ensure that they are independent. This independence allows us to partition the variation in ANOVA, so that the sum of squares corresponding to all  $G - 1$  contrasts will sum to the between group sum of squares.

Often contrasts are incorporated into the ANOVA analysis. The  $F$ -statistic for a contrast is simply the mean square for a particular between groups measure (for example,  $MS_{C1}$  is the mean square for  $C1$ ) divided by the pooled within group variances ( $MSE$ ). We can write the mean square for a contrast as

$$MS_{C1} = \frac{(C1)^2}{\sum_{g=1}^G \frac{C_{g1}^2}{n_g}} \quad (4.15)$$

where  $C_{g1}$  is the contrast coefficient and  $n_g$  is the sample size for each group. In the game study,

$$\begin{aligned} C1 &= (1)\bar{y}_{SR} + (-1)\bar{y}_{CR} + (1)\bar{y}_{SL} + (-1)\bar{y}_{CL} \\ &= (1)34 + (-1)36 + (1)37.1 + (-1)40.2 \\ &= -5.1 \end{aligned}$$

Thus, the mean square for contrast 1 ( $MS_{C1}$ ) based on Question 37 is identical to the mean square for game type ( $MS_{Type}$ ) in Question 38:

$$\begin{aligned} MS_{C1} &= \frac{(-5.1)^2}{\frac{1^2}{10} + \frac{(-1)^2}{10} + \frac{1^2}{10} + \frac{(-1)^2}{10}} \\ &= \frac{(-5.1)^2}{0.4} \\ &= 65.025 = MS_{Type} \end{aligned}$$

## Extended Activity Calculating Contrasts

Data set: Games2

40. Use Equation (4.15) to calculate  $MS_{C2}$  and  $MS_{C3}$ . Show your work.
41. Compare  $MS_{C2}$  and  $MS_{C3}$ , the mean squares found in Question 38.

### Key Concept

Orthogonal contrasts allow multiple comparisons of linear combinations of group means. The key advantage of contrasts is that they do not have inflated type I or type II errors. However, contrasts should always be determined before any data are collected. Looking at the data in order to develop contrasts will bias your results.

### MATHEMATICAL NOTE

There are a few other common techniques for multiple comparisons. **Scheffé's method** produces simultaneous confidence intervals for any and all contrasts, including contrasts suggested by the data (this is often called post hoc data exploration). Instead of the traditional formula for confidence intervals, Scheffé suggested using a wider confidence interval (i.e., one less likely to reject the null hypothesis) to account for the multiple testing. This method often fails to reject null hypotheses even when there are differences between groups, but it can be useful when other pairwise comparison tests are not appropriate. Remember from your introductory statistics course: If zero is in the confidence interval, you fail to reject the corresponding hypothesis test; if zero is not in the confidence interval, you should reject the corresponding hypothesis test. **Tukey's honest significant difference (HSD)** uses a studentized range distribution instead of the  $F$ -distribution to create confidence intervals for differences between meaningful pairs. When there are a large number of pairwise comparisons, Tukey's method is typically preferred over Bonferroni's method.<sup>6</sup>

## Chapter Summary

This chapter emphasized the importance of a well-designed experiment. A statistician often needs to communicate with people in other fields in order to properly define the research question, choose appropriate factors and levels, and determine the number of samples needed for the study.

A *p*-value never tells the whole story; *p*-values can be meaningless if assumptions are not met or if there are extraneous variables in the data. Before any conclusions are drawn from a statistical analysis using ANOVA, it is important to use graphs or formal tests to validate the equal variance and normality assumptions.

When equal variance or normality assumptions are violated, the *F*-statistics do not follow an *F*-distribution and the *p*-values may not be accurate. Empirical studies have shown that ANOVA tends to be fairly “robust” to departures from the assumptions of equal variances and normality. If the model assumptions are not met, researchers should try transforming the data to better fit the model assumptions. If no transformation appears to help, researchers should clearly explain that the *p*-values may not be reliable.

The independence and identically distributed assumptions are also essential. A good experimental design has the following characteristics:

- It avoids systematic error (bias is minimized by controlling for extraneous variables and using randomization).
- It has broad validity (results hold for more than just the units tested in the study).
- It allows for direct comparison between treatment conditions.
- It is precise (the chance unit-to-unit variability is small).
- It allows estimation of unit-to-unit variability.
- It can show causation (most observational studies cannot).

ANOVA tables are used to test for differences in means among meaningful groups of data. The analysis is called an analysis of variance because each mean square value is an estimate of a meaningful group within the data. The extended activities demonstrated how to calculate an ANOVA table; they emphasized **main effects** and **interaction effects**. An interaction between two variables occurs when the effect of one variable depends on the second variable. While this chapter emphasized a two-factor ANOVA, the same process holds for all **completely randomized designs with fixed factors**.

This chapter introduced the basics of a very powerful statistical technique. The ability to simultaneously test for the effects of multiple variables on a response allows statisticians to better model real-world situations. A well-designed experiment can test multiple hypotheses with a relatively small sample size. If used properly, these techniques efficiently and reliably help us better understand the world we live in. The end-of-chapter exercises and future chapters provide the opportunity for you to experience for yourself how these techniques are used in biology, chemistry, engineering, psychology, and many other disciplines.

## Exercises

### E.1. Design Your Own Study 1

Assume that you are asked to help design a study for an owner of four organic grocery stores, each located in a different city. The owner is interested in knowing whether placing an advertisement in the main Sunday paper will promote business (increase total sales) in the stores. You are considering the following three options: no advertisement, offering a coupon for one free item (with purchase), offering a coupon for special prices on a several items. For simplicity, we will assume that each coupon is effective for seven days (Sunday through Saturday). Write a brief outline for an experimental design.

- a. Identify any extraneous variables that may potentially bias your results. How will these potential biases be addressed?
- b. List the two factors and corresponding levels in your experimental design.
- c. Specify the response variable you will use.
- d. Specify the units and sample size for this study.
- e. List each factor-level combination and the order in which things will be tested at each location. Describe how you determined the order of each factor-level combination.
- f. List the factors and the corresponding degrees of freedom that will occur in the ANOVA table.

## E.2. Design Your Own Study 2

Assume that you have a small garden and are interested in knowing whether particular species of tomato plants will yield more tomatoes. You are also interested in knowing if adding fertilizer will increase the growth of tomatoes. You will try three species of tomatoes with and without fertilizer and measure the total weight of the tomatoes produced. Write a brief outline for an experimental design.

- Identify any extraneous variables that may potentially bias your results. How will these potential biases be addressed?
  - Specify the units and sample size for this study.
  - List each factor-level combination and the order in which things will be tested. Describe how you determined the order of each factor-level combination.
  - List the factors and the corresponding degrees of freedom that will occur in the ANOVA table.
- E.3. Assume that you are working on a study with two factors, each with two levels. Sketch an interaction plot where there is an interaction effect but there are no main effects. Hint: Assume that both levels of the first factor have a mean of 50. Also assume that both levels of the second factor have a mean of 50. However, all four factor-level combination means will need to be something other than 50.

## E.4. Microwave Popcorn Again: Three-Way Factorial Design

Data set: Popcorn

- Use the data structure of the three-way factorial design to write a statistical model for the microwave popcorn study. Since there are three factors of interest in the study, the model will need three terms, such as  $\alpha_i$ ,  $\beta_j$ , and  $\gamma_k$ , to represent each main effect and interaction effect. Use  $y_{ijkl}$  to represent an individual observation.
- Table 4.9 gives a three-way ANOVA table. Fill in the formulas for  $SS_C$ ,  $SS_{BC}$ ,  $MS_C$ ,  $MS_{BC}$ , and the  $F$ -statistic.
- Calculate the effect sizes for Brand, Microwave, and Time. Create a main effects plot and explain how the graph helps us to visualize the effect size.
- Calculate the effects for the Brand\*Microwave interaction. Show your work in calculating only the Fastco\*Room interaction effect by hand. Clearly interpret the meaning of these effects as it relates to this experiment.

**Table 4.9** Three-way ANOVA table.

Source	df	SS	MS	F-Statistic
A	$I - 1$	$\sum_{i=1}^I n_i (\bar{y}_{i...} - \bar{y}_{....})^2$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MSE}$
B	$J - 1$	$\sum_{j=1}^J n_j (\bar{y}_{.j.} - \bar{y}_{....})^2$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MSE}$
C	$K - 1$			
AB	$(I - 1)(J - 1)$	$\sum_{i=1}^I \sum_{j=1}^J n_{ij} (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j.} + \bar{y}_{....})^2$	$\frac{SS_{AB}}{df_{AB}}$	$\frac{MS_{AB}}{MSE}$
AC	$(I - 1)(K - 1)$	$\sum_{i=1}^I \sum_{k=1}^K n_{ik} (\bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....})^2$	$\frac{SS_{AC}}{df_{AB}}$	$\frac{MS_{AC}}{MSE}$
BC	$(J - 1)(K - 1)$			
Error	subtraction	subtraction	$MSE = \frac{SSE}{df_{Error}}$	
Total	$N - 1$	$\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L (\bar{y}_{ijkl} - \bar{y}_{....})^2$		

### E.5. Movies: Unbalanced Data

Data set: `Movies`

The file `Movies` contains the ratings and genres of movies that came out in 2008. This is simply an observational study, as clearly movie producers do not try to create an equal number of G, PG, and R movies each year.

- a. Write a statistical model that predicts gross earnings based on `Genre` and `Rating`. Use  $\alpha_i$  to represent the effect of `Rating` and  $\beta_j$  to represent the effect of `Genre`. Notice that the sample size varies for every group.
- b. Use statistical software to test for the effects of `Rating` and `Genre`. Do not include an interaction term. Check the model assumptions, perform any necessary transformations, and state your conclusions.
- c. Try to use statistical software to test for the effects of `Rating`, `Genre`, and `Rating*Genre`. Note that the analysis cannot be conducted because many of the interaction factor combinations have no observations. Group several of the `Genre` variables so that there are at least two observations in each `Rating*Genre` group. Now use your new groupings to test for the effects of `Rating`, `Genre`, and `Rating*Genre`.

### E.6. Cholesterol

Data set: `Cholesterol`

`Cholesterol` is a waxy substance found in blood and cell membranes. All animals need some cholesterol in their system; however, too much cholesterol can cause heart attacks and strokes. A study was conducted to determine how exercise, diet, and three types of drugs impact cholesterol levels. Seventy-two patients at a nearby hospital who had been diagnosed with high cholesterol (a level greater than 240 milligrams per deciliter) consented to be in the study. Each of the 72 patients was randomly assigned to a specific treatment, and after six months the patients' cholesterol levels were measured again. These measurements are provided in the `Cholesterol` data set.

- a. List two potential nuisance variables for the study. Suggest methods you would use to account for these two nuisance variables.
- b. Create appropriate graphs/charts to check for equal variances and normality. Does it appear that it is appropriate to use ANOVA to analyze these data? Show your work and give concise but appropriate explanations.
- c. Make the assumption that ANOVA is appropriate for the data above (i.e., do not transform any data). Using software, conduct an ANOVA for the data that will analyze all main effects, as well as all two-way interactions. Clearly and concisely interpret the results. Create appropriate graphs/charts that describe the data. Interpret these charts/graphs. Be selective in which graphs/charts to supply—too many graphs/charts can detract from your presentation.

### E.7. Don't Touch That!

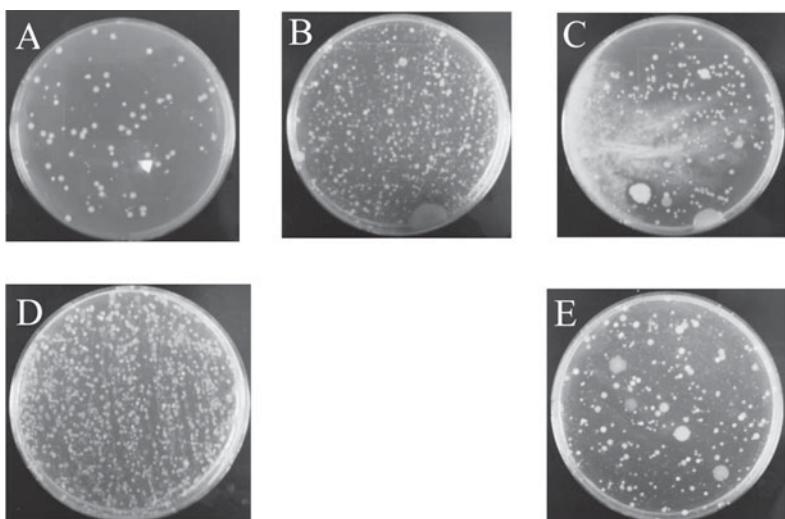
Data set: `Bacteria`

Antibacterial agents have become very popular in the marketplace, in products from gels to plastic children's toys. However, even now, you are surrounded by bacteria. Scientists have known for many years that bacteria are able to adhere to solid surfaces and form a resistant coat; however, it was not until the 1970s that the concept of biofilms became prevalent in the scientific community.<sup>7</sup> These films make completely removing bacteria from a surface nearly impossible.

Why should you care? Before the discovery of vaccines and antibiotics, humans succumbed to a myriad of acute infectious diseases. However, the majority of these have been eradicated or controlled, making bacterial infections from biofilms among the most threatening.<sup>8</sup> Today, battles against biofilm-forming bacteria are on the front lines of medical research. Interestingly, biofilms coat every surface that you see around you.

Two students, Isaac and Courtney, sampled surfaces around their campus to analyze the prevalence of bacteria. They compared different types of surfaces in both residential and academic buildings.

Data were collected by wiping surfaces with a wet Q-tip and swabbing the result on a standard nutrient agar plate. Locations were all tested on April 24, 2009 over the course of a two-hour period. The plates were incubated at 37°C for 48 hours before colony-forming units (CFUs) were counted as a measure of bacteria levels. Figure 4.5 shows several of the plates from this study. When CFUs exceeded 400 per plate, one fourth of the plate was counted and the total was calculated from that sample.



**Figure 4.5** Pictures of representative plates from Isaac and Courtney's samples around campus: (A) a bathroom door handle in Norris, (B) a desk in Noyce (the large cluster is a mold spot, not counted as a CFU), (C) a desk in Yellow House (the white film is a fungus, not counted as a CFU), (D) a desk in ARH, and (E) a bathroom door handle in the Cowles apartment. Photos courtesy of Derek R. Blanchette.

Six different buildings were swabbed, with two buildings representing each type of facility:

- Academic buildings: Noyce and ARH
- Public residential buildings: Norris and Dibble
- Private residences: Cowles apartment 7110 and Yellow House (1478 Park St.)

Within each building, two faucets, two door handles, and two desk surfaces were swabbed.

- a. Make the assumption that ANOVA is appropriate for the data (i.e., do not transform any data). Using software, conduct an ANOVA for the data that will analyze all main effects, as well as all two-way interactions. Treat the six buildings as six levels of the Building factor. Treat faucet, desk, and door as three levels of the Location factor.
- b. Create appropriate graphs/charts to check for equal variances and normality. Does it appear that it is appropriate to use ANOVA to analyze these data?
- c. Transform the response to the natural log of Count and redo the analysis. How do the results change? Create appropriate graphs to display the main effects and the interaction effects. Describe how the academic buildings (ARH and Noyce) compare to the other buildings.

#### E.8. Soda Fizz

Data set: Soda

Soda fizz is caused by carbon dioxide that is dissolved into liquid under pressure of up to 1200 pounds per square inch. When the consumer opens the package, pressure is released, the carbon dioxide gas is liberated from the liquid, and gas bubbles rise to the surface. This creates the desired tingling taste.<sup>9</sup> Since soda fizz is caused by the release of a gas, temperature may have an effect on amount of fizz, as gas volume generally increases with increased temperature. Furthermore, since the fizz is caused by gas bubbles being released from the liquid, the tilt of the cup being poured into may also impact the amount of fizz.

Two students, Julie and Daphne, conducted an experiment to test the effects of soda type (Pepsi vs. 7-Up), angle of the cup (cup flat on the table or slightly tipped), and soda temperature (refrigerated at 5°C vs. room temperature at 21°C) on the height of fizz produced when soda was poured out of a can into a cup. For each of the 24 trials, these students poured a can of either Pepsi or 7-Up into a clear cup and measured the peak fizz (in centimeters) produced, using a ruler on the outside of the cup.

- a. Create appropriate graphs/charts to check for equal variances and normality. Does it appear that it is appropriate to use ANOVA to analyze these data? Show your work and give concise but appropriate explanations.

- b. Conduct a transformation on the data. Instead of using `Fizz` as the response, use the natural log of `Fizz`. Using software, conduct an ANOVA for the `Soda` data. Analyze all main effects, as well as all two-way interactions. Create appropriate graphs that describe the data. Clearly and concisely interpret the results.

### E.9. Age and Memory

Data set: `MemoryA`

Michael Eysenck tested 100 subjects (50 people between the ages of 55 and 65, 50 younger people) to determine if there was a relationship between age and memory.<sup>10</sup> Each subject was shown 27 words and asked to recall as many of those words as possible. He also tested whether five different techniques impacted memory. Each subject was given one of five types of instructions:

- Counting: count the number of letters in each word
- Rhyming: think of a word that rhymes with each word
- Adjective: think of an adjective to describe each word
- Imagery: create an image of each word
- Intentional: remember as many words as possible

The subjects in the first four groups were not aware that they would later be asked to recall each word. The data set called `Age` provides the number of words that each subject properly wrote down after being asked to recall the list.

- a. Create appropriate graphs/charts to check for equal variances and normality. Does it appear that it is appropriate to use ANOVA to analyze these data? Show your work and give concise but appropriate explanations.
- b. Take the square root of the response and conduct an ANOVA on the transformed data to analyze all main effects and two-way interactions. Create appropriate graphs that describe the data. Clearly and concisely interpret the results.

### E.10. More Paper Towels

Data set: `Towels2`

- a. In fact, three brands of paper towels were actually tested with three amounts of water. The data for the complete study are provided in `Towels2`. Write a statistical model corresponding to this data set.
- b. Conduct an ANOVA on the `Towels2` data set. Transform the data if appropriate. Explain any changes from the degrees of freedom and sum of squares values found in Question 36. Create main effects and interaction plots and use the *p*-values to state your conclusions.

### E.11. Paper Cups: Fractional Factorial Designs

Data set: `Cups`

Have you ever wondered how paper cups are made? During the process, different temperatures, adhesives, pressure settings, paper stocks, types of machine, machine speeds, and many other variables can impact the quality of the cup that is made. Some cup-making machines can produce over 300 cups per minute. However, it can take a few hours to find the optimal running conditions. This can lead to a significant amount of wasted time and materials, as operators adjust the machine settings until good-quality cups are produced.

A manager of a manufacturing company, shift managers, cup machine operators, and a lone statistician decided to identify which factors were most influential in keeping their cups from leaking. Over 30 possible factors were identified, but after some thoughtful discussions the group settled on six variables that should be tested for their effects on leaking cups. One of the six factors of interest was which paper supplier to use. Since the company was considering changing suppliers, funds were available to do some product testing before a purchase was made. However, each trial (each run of production under specified factor conditions) would cost the company thousands of dollars in lost production time, material costs, and employee costs. The following is a list of questions representing the main factors in this study.

- Should the side-seam temperature be set at 70% or 90% when the cup is folded and the sides are sealed together?
- Should the side seam be sealed with an additional adhesive? This adds to production cost, but could be worthwhile if it improves the quality of the cup.
- Should the bottom temperature be set at 80% or 94% when the bottom is attached to each cup?
- Should the bottom pressure be set at 1000 or 1120 when the bottom is attached to each cup?
- Which supplier, Royal or Imperial, should be used? The suppliers have very similar prices.
- Should the paper stock come with an additional coating? This coating adds to production cost, but could be worthwhile if it improves the quality of the cup.

The company agreed to conduct 32 tests, but wanted to test all six factors and all corresponding two-way interactions. Fractional factorial designs are very useful for this type of exploratory data analysis. The details of fractional factorial designs are beyond the scope of this text. However, balanced data are a key concept behind these designs. For example, in the Cups data, every factor has two levels and each level has 16 observations. In addition, within the first factor level (the 16 observations where side-seam temperature is set to 70%), every other factor is still balanced (every other factor has 8 observations at each level).

- a. Use software to conduct an ANOVA to analyze the Cups data set. This ANOVA should include six terms corresponding to the six main factors and 15 interaction terms.
- b. Create a main effects plot comparing the effects of each main factor. Explain how the *p*-values correspond to what you see in the main effects plot.
- c. Create a graph of all interaction plots. Explain how the *p*-values correspond to what you see in the interaction plots.
- d. Create a probability plot or histogram of the residuals to determine if the residuals are consistent with data from a normal distribution. Do residuals appear to follow the normal distribution?
- e. State your conclusions for this study. Provide an answer to each of the six key questions listed above, using *p*-values and plots. In addition, provide an interpretation of any significant interaction effects. Your explanations should be understandable to both managers and machine operators. Assume you are the lone statistician involved in this study. Be sure to clearly identify the population for which these conclusions hold and also state whether causation has been shown.

## Endnotes

---

1. Sir Winston Churchill (1874–1965) was the British prime minister during World War II, won the Nobel Prize in Literature in 1953, and was made an honorary U.S. citizen in 1963.
2. <http://www.popcorn.org>
3. D. Montgomery, *Design and Analysis of Experiments*, 7th ed. (New York: Wiley, 2009) is one of several design of experiments texts that will describe that balanced designs are more powerful (more likely to reject the null hypothesis when there truly is a difference between groups) than unbalanced designs.
4. G. E. P. Box, “Non-normality and Tests on Variances,” *Biometrika*, 40 (1953): 318–335.
5. H. Levene, I. Olkin et al. (eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (Stanford: Stanford University Press, 1960), pp. 278–292. In addition, see M. B. Brown and A. B. Forsythe, “Robust Tests for Equality of Variances,” *Journal of the American Statistical Association*, 69 (1974): 364–367.
6. D. Montgomery, *Design and Analysis of Experiments*, 7th ed. (New York: Wiley, 2009) is one of several design of experiment texts that describe multiple comparison techniques in more detail.
7. J. W. Costerton, G. G. Geesey, and K. J. Cheng, “How Bacteria Stick,” *Scientific American*, 238.1 (1978): 86–95; J. R. Lawrence, D. R. Korber, B. D. Hyde, J. W. Costerton, and D. E. Caldwell, “Optical Sectioning of Microbial Biofilms,” *Journal of Bacteriology*, 173 (1991): 6558–6567.
8. J. W. Costerton, P. S. Stewart, and E. P. Greenberg, “Bacterial Biofilms: A Common Cause of Persistent Infections,” *Science*, 284 (1999): 1318–1322.

9. Michelle Bryner, “Why Does Soda Fizz?” *Live Science*, May 2009.
10. M. W. Eysenck, “Age Differences in Incidental Learning,” *Developmental Psychology*, 10 (1974): 936–941.
11. J. Theios, “Reaction Time Measurements in the Study of Memory Processes,” in H. Bower (ed.), *The Psychology of Learning and Motivation*, vol. 7 (New York: Academic Press, 1973), pp. 44–85.
12. R. G. Pachella, “The Interpretation of Reaction Time in Information-Processing Research,” in B. H. Kantowitz (ed.), *Human Information Processing—Tutorials in Performance and Cognition* (Hillsdale, NJ: Erlbaum, 1974), pp. 41–82.
13. G. Cobb, *Introduction to Design and Analysis of Experiments* (Emeryville, CA: Key College Publishing, 1998), adapted from p. 2.
14. D. Montgomery, *Design and Analysis of Experiments*, 7th ed. (New York: Wiley, 2009), p. 21.
15. Ibid, p. 22.

# Research Project: Testing for the Effect of Distractors

The following pages provide guided steps for conducting your own research project involving an online game. You will design your own study, collect data, analyze the results, draw conclusions, and present your results.

## Reviewing the Literature

When I (one of your authors) sat down with my 5-year-old son to teach him how to tie his shoes, I was surprised that I couldn't tell him how to do it. Even though I tie my own and my children's shoes almost every day, I had to tie the shoe myself and watch each step before I could break down the steps in order to teach my son. This is just one example of automatized behaviors, which are behaviors that can be done automatically without carefully thinking through each step in the process.\*

The Stroop effect demonstrates that automatized behaviors can interfere with other desired behaviors. In the paper assigned below, John Stroop tested the reaction time of college undergraduates in identifying colors. Students took a longer time identifying colors of ink when the ink was used to spell a different color. For example, if the word "green" was printed in blue ink, students took longer to identify the color blue because they automatically read the word "green." Even though students were told only to identify the ink color, the automatized behavior of reading interfered with the task and slowed their reaction time.

Another type of reaction time question that cognitive psychologists are interested in is the speed-accuracy tradeoff. Theios conducted a study in which a digit was shown to a subject and the time it took the subject to simply name the digit was measured.<sup>11</sup> The percentage of times a particular digit was shown (called the stimulus probability) varied from 20% to 80%. Theios found that the stimulus probability did not impact the subjects' reaction time. Pachella repeated the study by Theios; however, he measured both reaction time and accuracy. Pachella found that even though the reaction time stayed the same, the subject's accuracy changed dramatically as the percentage varied.<sup>12</sup>

In this project, you will have the opportunity to develop your own experiment using an online game. In this game, the subject is expected to place specifically shaped pegs into the appropriate holes within a short time period.

1. Read the paper by J. Stroop, "Studies of Interference in Serial Verbal Reactions," *Journal of Experimental Psychology*, 12 (1935): 643–662. Focus primarily on the first two experiments. If there are any words that you do not understand, look them up and provide a short definition for each.
2. For the second experiment in Stroop's paper, identify or answer the following:
  - a. Objective of the experiment
  - b. Any relevant background (from journals that were referenced)
  - c. Response variable(s)
  - d. Factors and levels that were tested
  - e. Variables that were held constant during the experiment
  - f. Nuisance factors (i.e., factors that are not of interest but may influence the results)
  - g. Were any interactions tested, and if so, what was observed?
  - h. What type of design was used in the experiment?
  - i. How many trials were run for the experiment?
  - j. How could you modify this experiment if you were going to create a reaction time test?

Be ready to submit your answers as well as discuss this material in class.

## Playing the Computer Game and Developing a Factorial Design

Go to the Website <http://www.pearsonhighered.com/mathstats> resources to find the Shapesplosion game, and use it to develop your own experiment. Which factors do you believe will have the most significant effect on reaction time?

\*Note that many psychologists would call this procedural knowledge instead of automatized behavior. Both are processes that can be done without conscious thought, but automatized behaviors are processes that cannot be slowed down, do not decline with age, and show no gender differences.

Submit your design at the beginning of class. Note that these games allow you to develop three new factors of your own choice. Be prepared to discuss how you addressed each of the following five points.

3. Clearly define a problem and state the objectives of your experiment. Before any experiment is conducted, it is essential that everyone involved clearly understand the objectives of the experiment, what measurements will be taken, what material is needed, and what procedures will be used.

State the null and alternative hypotheses. This is often much more difficult than it appears. First, designing an experiment often involves many people from diverse backgrounds. These people typically have different goals and use very different terminology. Second, it is important to design an experiment that is “general enough to be of scientific interest, yet specific enough that it is feasible to run within time, space, and material limitations.”<sup>13</sup>
4. Identify the response variable, factors, potential levels of each factor, and units.
  - a. Verify that the response variable provides the information needed to address the question of interest. What are the range and variability of responses you expect to observe? Is the response measurement precise enough to address the question of interest? Are you interested in the number of wins or the time it took to win?
  - b. Investigate all factors that may be of importance or potentially cause bias in the results. When the objective is identifying which factors have most influence on the response, it is usually best to keep the number of factor levels low. Even though several levels are controlled in these games, consider other factors that should be controlled. Did a subject drink a significant amount of caffeine before the game, or did the subject stay up all night studying for an exam? What can you do to account for these variables?
  - c. Once experimental factors have been identified, carefully identify a reasonable range for each factor. “In some fields there is a large body of physical theory on which to draw in explaining relationships between factors and responses. This type of non-statistical knowledge is invaluable in choosing factors, determining factor levels, deciding how many replicates to run, interpreting the results of the analysis, and so forth. Using a designed experiment is no substitute for thinking about the problem.”<sup>14</sup>
5. Identify what other factors need to be controlled during the experiment to eliminate potential biases. Identify how measurements, material, and process may involve unwanted variability. What conditions would be considered normal for this type of experiment? Are these conditions controllable? If a condition changed during the experiment, how might it impact the results? List each nuisance factor, and explain at what levels and how each will be controlled throughout the experiment, even if it is simply held constant. Will subjects be allowed a practice game before the actual experiment? How important is it to randomize the order in which the games are played?
6. Choose an experimental design.
  - a. Keep the design and analysis as simple as possible. A straightforward design is usually better than a complex design. If the design is too complicated and the data are not collected properly, even the most advanced statistical techniques may not be able to draw appropriate conclusions from the experiment.
  - b. How many trials will be run? Is the cost of replicating the experiment worth gaining a better understanding of the sample-to-sample variability? Can you completely randomize all the trials, or do you need to account for timing, subject variability, and other nuisance factors? *Important note:* If your experimental design includes multiple explanatory variables, it is essential to clearly understand the difference among completely randomized, block, and repeated measures designs. If each subject is assigned to only one condition (plays the game only once), it may be appropriate to use a completely randomized/full factorial design. If each subject is tested under several conditions (each subject plays several games), a more complex design structure, such as a block or repeated measures design, may be needed.
7. Explain how your experimental design builds on previous research.
  - a. Identify relevant background on response and explanatory variables, such as theoretical relationships, expert knowledge/experience, or previous studies.
  - b. Explain where this experiment fits into the study of the process or system. Experiments are usually iterative. While your assignment is to design, conduct, and analyze one experiment, it is important to realize that each experiment is just one step in a much larger process. Even in engineering, where experiments can be much more focused (possibly to just one machine) and variables are typically more controlled than in nature, Montgomery suggests that only 25% of your resources be used in a first experiment.

In most situations, it is unwise to design too comprehensive an experiment at the start of a study. Successful design requires knowledge of the important factors, the ranges over which these factors are varied, the appropriate number of levels for each factor, and the proper methods and units of measurement for each factor and response. Generally, we are not well-equipped to answer these questions at the beginning of the experiment, but we learn the answers as we go along. Of course there are situations where comprehensive experiments are entirely appropriate, but as a general rule, most experiments should be iterative. Consequently, we usually should not invest more than about 25 percent of the resources of experimentation (runs, budget, time, etc.) in the initial experiment. Often these first efforts are just learning experiences, and some resources must be available to accomplish the final objectives of the experiment.<sup>15</sup>

## Collecting Your Own Data

8. Prepare any questions you would like to ask cognitive psychologists or statisticians before you finalize your experimental design.
9. Write specific lab procedures that you will use while conducting the experiment. Determine who will collect the data at what time, how will you randomize the trials, how the data will be recorded, and exactly what will be measured.
10. Ensure that your group has received appropriate Institutional Review Board (IRB) approval (see the supplemental material on IRB).

## Presenting Your Own Model

11. Meet with your professor to discuss your experimental design and analysis.
12. Collect the data. In conducting the experiment, did you identify any other sources of variability that could be impacting the results? Submit lab procedures.
13. Write the research paper (see “How to Write a Scientific Paper or Poster” on the accompanying CD). Bring three copies of your research paper to class. Submit one to the professor. The other two will be randomly assigned to other students in your class to review.

## Final Revisions

Make final revisions to the research paper. Submit the first draft, other students’ comments and checklists, the data set with variable descriptions, and the final paper.

## Other Project Ideas

Several of the extended activities and end-of-chapter exercises can also be used to develop your own project ideas. In addition to the Shapesplosion game, a memory game is available at <http://www.pearsonhighered.com/mathstatsresources>.

You may want to read W. G. Hunter, “Some Ideas About Teaching and Design of Experiments, with 25 Examples of Experiments Conducted by Students,” *The American Statistician*, 31.1 (Feb. 1977): 12–17. It suggests several possible student research projects.

Here are some other questions often studied by student researchers:

- Who studies more? People involved in extracurricular activities? Males or females? Does major matter?
- Do different networks or times of day impact the length of positive or negative stories in the news? Are stories about women longer or more positive than stories about men?
- Does time of day, location, Website, or type of computer impact Internet speed?
- Do some types of classes have better attendance (or class participation) than others? Does it depend on major, class size, or level of difficulty of the course?
- Where and when is student or faculty parking available on campus?

# Block, Split-Plot, and Repeated Measures Designs: What Influences Memory?

*Do not put your faith in what statistics say until you have carefully considered what they do not say.*

—William Whyte Watt<sup>1</sup>

The preceding chapter focused on designing, analyzing, and interpreting studies based on completely randomized designs (often called full factorial designs). One advantage of the completely randomized designs discussed in Chapter 4 is that they are easily extended to multiple factors where each factor may have a different number of levels. However, there are many multivariate studies where the design structure is not as straightforward. This chapter extends the material in the previous chapter to a broader set of possible factorial design structures.

A student project on memory and several other examples are used to emphasize the following key concepts within this chapter:

- The importance of properly designing a study
- How to determine if factors are fixed or random
- How to determine if factors are crossed or nested
- The differences between completely randomized, block, split-plot, and repeated measures designs
- Model assumptions
- Details of calculating effect sizes, degrees of freedom, sums of squares,  $F$ -tests, and  $p$ -values
- The use of Hasse diagrams to visualize the design structure and determine the appropriate mathematical calculations for ANOVA
- Analysis of covariance

## 5.1 Investigation: What Influences Memory?

Two students in a statistics course, Josh and Ann, were interested in conducting an experiment to help them better understand memory. One of the first challenges in properly designing an experiment is to create a suitable research question. The specific questions that any experiment is intended to answer must be clearly identified before the experiment is carried out. The research question “What factors influence memory?” is too broad for any research, let alone a class project. Memory is a very broad topic that people have been researching for years. Josh and Ann needed to modify the general research question to something specific enough to ensure that the experiment was feasible.

Chapter 4 described how researchers often depend on nonstatistical knowledge in choosing the response variable, factors, potential levels of each factor, units, and number of trials to run. There is no need to conduct a study if the answer to the research question is already well known. However, good research questions are often developed by slightly modifying or incorporating a novel idea into previous research. Josh and Ann conducted a brief literature review and found that quite a bit of research had been conducted on *memory distracters*. A distracter is anything that can disrupt an individual’s ability to memorize (e.g., to memorize a list of words). Josh and Ann also learned that memory is measured many different ways and that some types of words tend to be more difficult to retain in short-term memory than others. For example, *concrete nouns* (tangible objects that exist physically, such as *table*, *sofa*, and *horse*) are easier to remember than more *abstract nouns* (ideas or concepts that have no physical form, such as *importance*, *fate*, and *sympathy*).

Psychologists have found many potential biases related to the type of word that is used in memory tests. So instead of creating their own lists of words, Josh and Ann decided to randomly select medium-length abstract and concrete words from an online word database. They focused their research questions on two factors: *type of word list* and *type of distracter*. Each factor was limited to only two levels: abstract or concrete words and mathematics or poetry distracters. In this experiment, each unit is a test that is randomly assigned to one of the four factor-level combinations\* listed below:

1. Abstract word list and mathematics distracter
2. Abstract word list and poetry distracter
3. Concrete word list and mathematics distracter
4. Concrete word list and poetry distracter

For each test, Josh and Ann decided to test what psychologists call *free recall* by asking each subject to read through a list of 20 words for 30 seconds, work through a distracter, and then recall as many words as he or she could. Josh and Ann needed to ensure that they had enough resources (i.e., time, materials, and knowledge of the appropriate statistical techniques) to conduct the study. Many current memory studies in psychology involve hundreds of human subjects. However, this study included only a small number of subjects (college students). To gain more observations, Josh and Ann asked each subject to take all four tests in random order, with a short break between tests.

Subjects vary dramatically in their ability to memorize words. By having each subject assigned to all four treatment combinations, the researchers can control the subject-to-subject variability and gain a better measure of the true treatment effects. However, each unit (each test) is no longer completely randomly assigned to one of the four treatment conditions. There is a new restriction that exactly four tests be assigned to each subject. This chapter focuses on experimental designs where *the random assignment of units to a treatment is restricted in order to create a more efficient study*.

Josh and Ann’s final research questions were as follows:

1. Does the type of word list (abstract or concrete) impact memory?
2. Does the type of distracter (mathematics or poetry) impact memory?
3. Does the type of distracter impact memory differently for different types of word lists?

Three additional questions were also tested, but these will not be discussed until later in the chapter:

4. Is memory of word lists affected by the subject’s academic major?
5. Does the impact of the type of distracter depend on the subject’s academic major?
6. Does the impact of the type of word list depend on the subject’s academic major?

---

\*Each explanatory variable is called a *factor*. Specific conditions within each factor are called *levels*. All studies can be classified as either observational studies or experiments. In any study, the factor-level combinations are called *conditions*. In experiments, each factor-level combination is often called a *treatment*.

**NOTE**

Two abstract word lists and two concrete word lists were needed because there would be *carry-over effects* if subjects were asked to memorize the same list of words twice. Since there was no interest in the difference between randomly selected abstract word lists, Josh and Ann did not distinguish between the two abstract lists. The same was done for distractors. Two mathematics distractors were used (count down from 262 by 3's for 30 seconds or compute the following multiplication:  $1951 \times 263$ ). If subjects were asked to count down from 262 by 3's both times they were assigned a mathematics distracter, there is a strong chance that the second time would be much less distracting. The two math distracters were assumed to have equivalent impact on memory. The same assumption was made for the two poetry distracters. Two poems by Robert Frost were selected (*The Road Less Traveled* and *Fire and Ice*). After reading a poem, the subject was asked to describe the theme of the poem.

## 5.2 Elements of a Well-Designed Experiment

In evaluating conclusions from any experiment, it is essential to realize that the experimental design and data collection procedures are just as important as (if not more important than) any statistical calculations. Details of how to properly design completely randomized experiments are provided in Chapter 4. This section briefly summarizes the key points and helps us to interpret the ideas of design in a new context. Well-designed experiments contain the following key elements:

- *Comparative groups*: The experimental units (or groups of units) should be as similar as possible.
- *Random allocation*: The treatments should be randomly assigned to units.
- *Large enough sample sizes*

### Comparative Groups

Josh and Ann realized that it was likely that each subject in their study had a very different ability to memorize words. There were also extraneous variables, such as the amount of sleep or stress level of a subject, that could influence (cause unwanted variability in) the response. In order to control for this subject-to-subject variability, each subject was asked to take all four tests (treatment combinations) and to take them on the same day. In the analysis, the four tests taken by the same person were then considered to be a group or block.

The process of arranging units into groups that are similar to one another is called **blocking**. Blocking reduces known, but unwanted, sources of variation between units. Figure 5.1 demonstrates that in block designs, units are first placed into blocks and then randomly assigned to a factor-level combination. Notice that none of the research questions asked if students had different free-recall abilities. We know that people have different memory abilities, so there is no need to test for this difference. However, accounting for subject-to-subject variability in the analysis reduces the amount of unexplained variability, allowing us to more precisely identify differences due to treatments.

#### Key Concept

Blocking can account for unwanted variability in a study. Thus, if blocking is used appropriately, the statistical analysis is likely to have greater precision in determining if there truly is a difference between treatments.

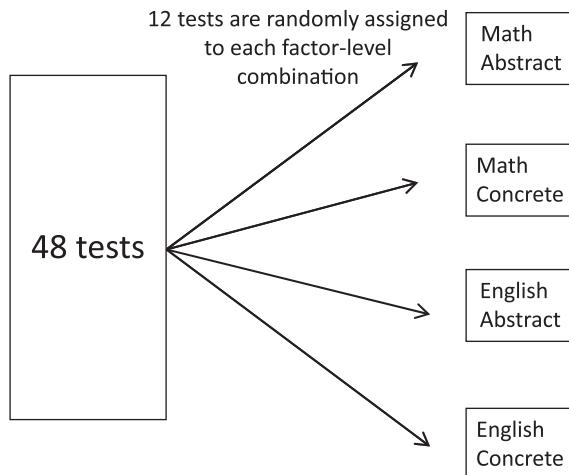
In addition to blocking, Josh and Ann restricted factors in their study that could have caused additional variability in the response variable. The test location, instructions provided to the subjects, the way words were presented, and background distractions all were carefully controlled. Their study also restricted the population of interest by testing only students from one college and by limiting the number of factors and levels.

### Random Allocation

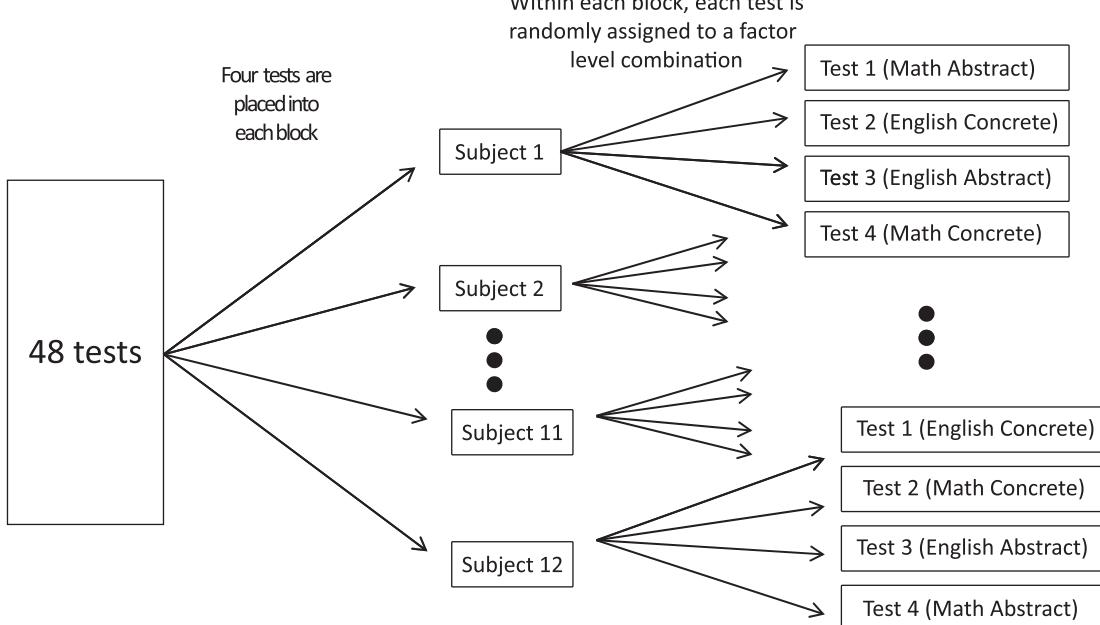
**Random allocation** is the process of randomly assigning units to treatments. In this memory study, each of the twelve subjects took four tests. Each unit (i.e., each of the 48 tests) is randomly assigned to exactly one of four treatments within each block.

The random assignment helps to reduce the likelihood of biases in the response. For example, by the fourth test, subjects may experience some fatigue or have a tendency to recall words from previous lists. Since the four treatment combinations are conducted in random order, treatment effects are not *biased* by the order in which the tests were taken.

## Completely Randomized Design



## Block Design



**Figure 5.1** Differences in randomization between completely randomized and block designs.

**Random sampling** (selecting units randomly from a larger well-defined population) is not the same as random allocation (randomly assigning treatments to units). In many studies, collecting a true random sample from a population is very difficult or impossible. If a true random sample cannot be collected, conclusions involving causation can still be made. However, the conclusions hold only for the actual population from which the sample was selected.

## Sample Size

In the memory study, 12 subjects each took 4 tests, providing 48 measurements. This may not appear to be a large enough sample size to simultaneously test the three research questions for this study. Statistics based on larger sample sizes are more reliable and increase the ability to confidently draw conclusions. However, properly designed experiments can be very effective even with small sample sizes, as we will see when we analyze the memory experiment data. Instead of attempting to calculate the sample size needed before a study is conducted, this text focuses on using the model assumptions and analysis (discussed in Chapters 2 and 4) to determine if the study is reliable.

**CAUTION**

Introductory textbooks often provide an approximate sample size for hypothesis tests. For example, a sample size of 10–15 per group will allow us to use the  $t$ -distribution to test for differences between two population means. However, even in the simple two-sample  $t$ -test, if the data are highly skewed, sample sizes greater than 30 may not be large enough to produce an accurate  $p$ -value.<sup>2</sup> The process of determining the sample size is often called power analysis. Power analysis with more than one explanatory variable often requires several simplifying assumptions. This involves specifying expected effect sizes, population variances, alpha levels, and power (the probability of properly identifying a significant difference in means when it truly exists).<sup>3</sup> While numerous sample size calculators can be found on the Internet, be very cautious about using these tools without truly understanding the many assumptions that are made in the analysis.

## 5.3 Statistical Analysis Based on the Experimental Design

Up to this point, we have discussed identifying what data to collect and how to collect the data. The rest of the chapter will focus on properly analyzing and drawing conclusions from the data. The emphasis will be on selecting the correct model and knowing what type of analysis is appropriate. To determine the appropriate analysis for any advanced experimental design, it is necessary to answer the following questions:

1. What is the randomization structure of the design? We will consider three structures that encompass the majority of designs used by researchers:
  - a. the completely randomized design, also called a factorial design or a randomized basic factorial design
  - b. the block design, also called a randomized block design
  - c. the split-plot design, also called a repeated measures design
2. Is each factor crossed or nested?
3. Is each factor fixed or random?

**NOTE**

This text follows many statistics textbooks in that we consider *repeated measures design* to be another name for *split-plot design*. In this text, we define a block design to be different from a repeated measures design. However, psychologists and social scientists often consider both block designs and split-plot designs to be repeated measures designs.

The following sections will focus on answering each of these questions for the memory study. While the terminology can cause some confusion across disciplines, in this chapter we will use only three terms for design structure: completely randomized design, block design, or split-plot design. The extended activities provide several additional examples to help you better understand these terms in multiple contexts.

**Key Concept**

In experiments with completely randomized designs, each unit is randomly assigned to one treatment. In block and split-plot designs, there are restrictions in the way units are randomly assigned to treatments.

**NOTE**

In a two-sample  $t$ -test, each unit is randomly sampled from a population (or randomly assigned to a treatment). The two-sample  $t$ -test is a special case of a completely randomized design. In a paired  $t$ -test, each subject is measured twice (e.g., each subject takes a pretest and posttest or is measured before and after a treatment is imposed). The paired  $t$ -test is a special case of a block design.

## 5.4 Three Commonly Used Design Structures

### Completely Randomized Designs

The memory example provided above was already identified as a block design. However, in order to demonstrate the differences among the three design structures, in this section the study will be treated as if it had been conducted as a completely randomized design. The four treatment combinations (two factors, each at two levels) would stay the same. The response would still be the number of words correctly recalled, and the research questions would be the same. However, in a completely randomized design, each unit (i.e., each test) would be randomly assigned to only one treatment combination with no other restrictions. In order to avoid any possible biases with subject-to-subject variability, each subject would need to be assigned to take exactly one test. Josh and Ann would have needed to have 48 subjects in order to collect 48 measurements. Since the 48 students would each be taking one test, either the student or the test could be considered the unit for this design. Here is a summary of the study that could have been collected under a completely randomized design:

- Factor 1: Wordlist with two levels (abstract or concrete)
- Factor 2: Distracter with two levels (mathematics or poetry)
- Units: each of 48 Tests (each student is assigned to one test)
- Response: each of the 48 test scores (number of words correctly recalled)

### Activity Analysis of a Completely Randomized Design

1. Treat the full Memory data set as if it had a completely randomized factorial design (*incorrectly* assume the data were collected from 48 subjects who each took one test).
  - a. Write out the three null and alternative hypotheses that will be tested, using symbols and words.
  - b. Analyze the Memory data using an ANOVA that includes the terms Wordlist, Distracter, and Wordlist\*Distracter.
  - c. Create individual value plots or side-by-side boxplots of the data. Do we have reason to question the equal variance assumption? Create a probability plot or histogram of the residuals to determine if the residuals are consistent with data from a normal distribution. What do you conclude?
  - d. Provide a  $p$ -value corresponding to each of the three hypothesis tests. Create main effects plots and an interaction plot and state your conclusions.

### Block Designs

There are many situations in which a study includes **nuisance factors**, factors that may impact the results but are not of specific interest in the study. **Extraneous variables** are any variables (known or unknown) that are not of specific interest in the study but may affect the response variable (i.e., add unwanted variability to a study). A nuisance factor can be considered as a type of extraneous variable that has been identified and incorporated into the analysis.

#### NOTE

Specific definitions tend to vary by textbook. Sometimes extraneous variables are called **confounding variables** or **lurking variables**. However, lurking and confounding variables are typically more specifically defined as variables whose effect on the response cannot be distinguished from that of an explanatory variable in the study. Some statisticians differentiate between these two by defining confounding variables as factors incorporated into the study design and lurking variables as factors not incorporated into the study.

Josh and Ann were not interested in testing whether some students performed better than other students. Large amounts of variability in memory ability between students are expected. Thus, Student is not a factor of interest, but large student-to-student variability could impact the results. Thus, it is appropriate to include the nuisance factor Student in the model and to collect the data using a block design instead of a completely randomized design.

**Blocking** is the process of grouping units based on some preexisting similarity that might impact the response of interest. Units can be sorted, reused, or subdivided to create a block. Blocking is beneficial because it can control for unwanted variability and increase the efficiency of a study.

NOTE

In this study, each subject is considered a block. Each subject is reused, since each is asked to take four tests. In a completely different study, a psychologist could have decided to give each subject a pretest, in which case subjects could be sorted into blocks based on their pretest score. Human subjects are not typically subdivided. However, a block could be a batch of dough, a plot of land, or a piece of metal that is subdivided into several pieces, each of which undergoes some type of test.

In the memory experiment, there were 48 observed test scores (`Score`) and therefore 48 experimental units (`Test`). Each of the 12 `Students` was considered a block and took four tests, representing each of the four treatment combinations in random order. The four tests were grouped by `Subject`, and the randomization of the four tests occurred within each `Subject` (the same as `Student`, here). Using a block design, the `Memory` data set is organized as follows:

- Blocks: each of the 12 `Subjects`
- Factor 1: `Wordlist` with two levels (abstract or concrete)
- Factor 2: `Distracter` with two levels (mathematics or poetry)
- Units: each of the 48 `Tests`
- Response: each of the 48 test scores

**Key Concept**

**Block designs** restrict the way units are randomized. In a completely randomized design, each unit is randomly assigned to a treatment. In a block design, the randomization occurs within each block. In this study, each of the four tests within a `Subject` block must be assigned to a different treatment. In essence, a complete block design can be thought of as a completely randomized design conducted within each block.

## Activity Analysis of a Block Design

2. Use software to conduct an ANOVA to analyze the `Memory` data set using a block design. The variables in the ANOVA should be `Student`, `Wordlist`, `Distracter`, and a `Wordlist*Distracter` interaction term. This ANOVA looks very similar to the two-way ANOVA conducted in Question 1, but now there is one additional variable in the ANOVA, `Student`.
  - a. Write the hypotheses corresponding to the three original research questions and provide a *p*-value for each hypothesis.
  - b. The goals of this study did not include determining if there was a difference between students. However, using a block design, the researchers assumed that `Student` was a nuisance factor that might explain some of the variability in their study. Based on the *p*-value for `Student` provided in the ANOVA, do you believe that student-to-student variability explains much of the variability in the data?
  - c. Create a main effects plot comparing the effects of `Student`, `Wordlist`, and `Distracter`. Explain how the *p*-values correspond to what you see in the main effects plot.
  - d. Create a probability plot or histogram of the residuals to determine if the residuals are consistent with data from a normal distribution. State your conclusions for this study. Be sure to clearly identify the population for which these conclusions hold and also state whether causation has been shown.

\* Technically, `Student` should be considered as a random factor. However, random versus fixed factors have not yet been discussed. In addition, in this particular block design, the *p*-values will be the same whether `Student` is considered fixed or random.

3. Compare the sum of squares (SS) error from Question 1 with the SS for Student and SS error in Question 2. Show a mathematical formula or simply explain how you could use the ANOVA in Question 2 to calculate SS error from Question 1.
4. Compare the SS and mean square (MS) for Wordlist, Distracter, and the Wordlist\*Distracter interaction in Questions 1 and 2. Describe any patterns that you find.
5. Explain why you might expect the MSE to be smaller in Question 2 than in Question 1.
6. Use the mean square errors to explain why the *F*-statistics and *p*-values for Wordlist in Questions 1 and 2 are not the same.



#### NOTE

Interactions involving blocks are typically not included in the ANOVA model. For example, there is no Student\*Wordlist interaction term. While interaction terms are often studied with factors of interest, the purpose of blocks is to control unwanted variability.

## Split-Plot Designs

Before the third design structure is discussed, it is important to understand the difference between replications and repeated measures. Replications occur when multiple units are assigned to each condition, while repeated measures occur when multiple measurements are taken on one unit. For example, in the completely randomized design in Question 1, there were four treatment combinations and 12 units assigned to each. Thus, there were 12 replications of each of the four treatment combinations.

In Josh and Ann's block design in Question 2, the 48 tests were treated as the units and each Student was considered a block. This text will always treat block designs as having only one type of unit (Test is considered the only type of unit in this block design). However, researchers from some disciplines would say that there are two types of units in this study, Student and Test. If Student is considered a unit, then the four measurements on each student (the four test results) are considered repeated measures. In this text, designs with two types of units are called split-plot designs.

#### Key Concept

**Replications** are multiple measurements of a treatment condition with just one measurement for each unit. To include more replications, we must include more units. **Repeated measurements** are multiple measurements per unit. To include more repeated measurements, we take more observations from the same units. Both are used in experimental design to increase the precision of estimates by increasing the number of measurements.

**Split-plot designs** have at least two sizes of units in one design and often make use of both replication and repeated measurements. Recall that Josh and Ann had additional research questions. For example, they were interested in determining if memory of word lists was related to a subject's academic major. The split-plot design can be used to incorporate these additional questions into our ANOVA.

The students used for Josh and Ann's study were not really a simple random sample selected from the college. Instead Josh and Ann used each department's list of declared majors to randomly select three English majors, three history majors, three math majors, and three computer science majors. This is a split-plot design with one whole-plot factor and two split-plot factors.

In a split-plot design, a whole-plot unit is simply a block. It is assigned to (or randomly sampled from) a condition, which is called a **whole-plot factor**. In this design, each Student was randomly sampled from Major. Then the whole-plot unit is reused or subdivided into subgroups, or **split-plot units**. The split-plot units are also assigned to a condition, or **split-plot factor**. This is the same process that occurred in the block design. Each whole-plot unit (Student) was tested four times to create four split-plot units (Tests). Within

each block, the four tests were randomly assigned to split-plot factors (*Wordlist* and *Distracter*). The split-plot design described below uses both *Student* and *Test score* as units.

Whole-plot factor: *Major* with four levels (English, history, math, and computer science)

Whole-plot unit (blocks): each of the 12 students

Split-plot factor 1: *Wordlist* with two levels (abstract and concrete)

Split-plot factor 2: *Distracter* with two levels (mathematics and poetry)

Split-plot unit: each of the 48 Tests

Response: each of the 48 test scores

### Key Concept

A split-plot design consists of a block design where the blocks are also assigned to a treatment (or sampled from distinct populations).

### NOTE

Different disciplines use different terminology for the same split-plot design. Whole-plot factors are often called **between-block factors**, while split-plot factors are often called **within-block factors**. If there is no whole-plot factor, this text will refer to the design as a *block design*. However, psychologists may call both the block and the split-plot designs *repeated measures designs* if multiple measurements are taken on one unit. In addition, if there is only one subject per block, psychologists may call the design a *within-subject design*.

### NOTE

If there is one or more whole-plot factors, this text will refer to the design as a *split-plot design*. However, many statisticians tend to call the design a *repeated measures design* if the blocks (whole-plot units) are reused or measured multiple times and a *split-plot design* if the blocks (whole-plot units) are subdivided into smaller parts. The labels used do not impact the design structure or data analysis. The designs are identical.

You can probably now appreciate that whenever an experiment is conducted, it is *essential* to ensure that enough thought is given to the structure of the experimental design before the data are collected. Unfortunately, some researchers tend to wait until all the data have been collected before contacting a statistician and asking the statistician to analyze the experimental data. If the experiment has not been designed properly, the results are often compromised, and in some cases no conclusions can be drawn, even if there are small *p*-values.

Before we analyze the memory experiment as a split-plot design, the two remaining questions listed in Section 5.3 need to be answered. In the next section, we address the question of whether factors are crossed or nested and Section 5.6 addresses how to determine whether factors are fixed or random.

## 5.5 Crossed and Nested Factors

Any two factors *A* and *B* are **crossed** if every level of *A* occurs with every level of *B*. In the memory study, *Wordlist* and *Distracter* are crossed. Every level of *Wordlist* is tested with every level of *Distracter*. In completely randomized designs, all factors of interest are crossed and there are no repeated measures.

For any two factors *A* and *B*, factor *B* is **nested within** factor *A* if levels of *B* have meaning only within specific levels of *A*. A nice visualization of nested effects is based on trees and leaves. A study may require a researcher to sample two leaves from each of several trees. There are two “levels” (leaf #1) and (leaf #2) sampled from each tree. There is no reason to average all the first leaves over all the trees and test if there is a difference between the average value corresponding to the first leaf and the

average value corresponding to the second leaf. Leaves are nested within trees because leaf #1 (i.e., “the first leaf”) has meaning only if it is known which tree it came from.

In the memory study, each Student is nested within a Major. The first mathematics student sampled is not expected to have any relation to the first English student or first computer science student sampled. So testing for differences between the averages of student #1, student #2, and student #3 across all four majors is meaningless. When two factors are nested, interaction terms cannot be calculated.

### Key Concept

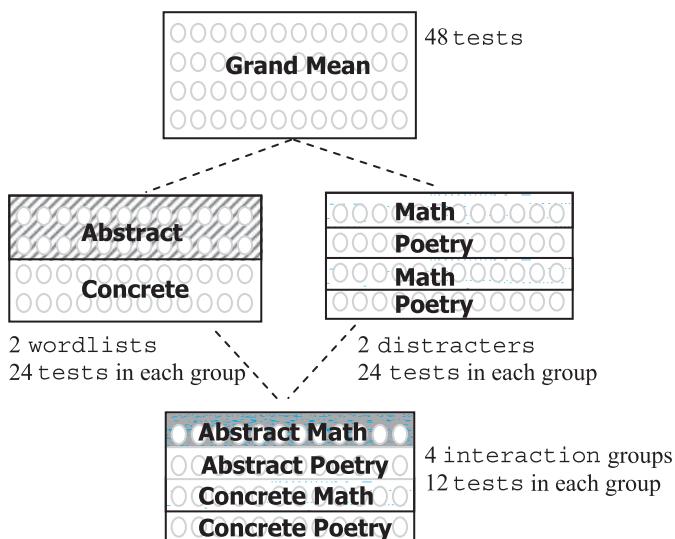
In completely randomized designs, all factors of interest are crossed. Block and split-plot designs can have either crossed or nested factors.

## Visualizing Design Structures

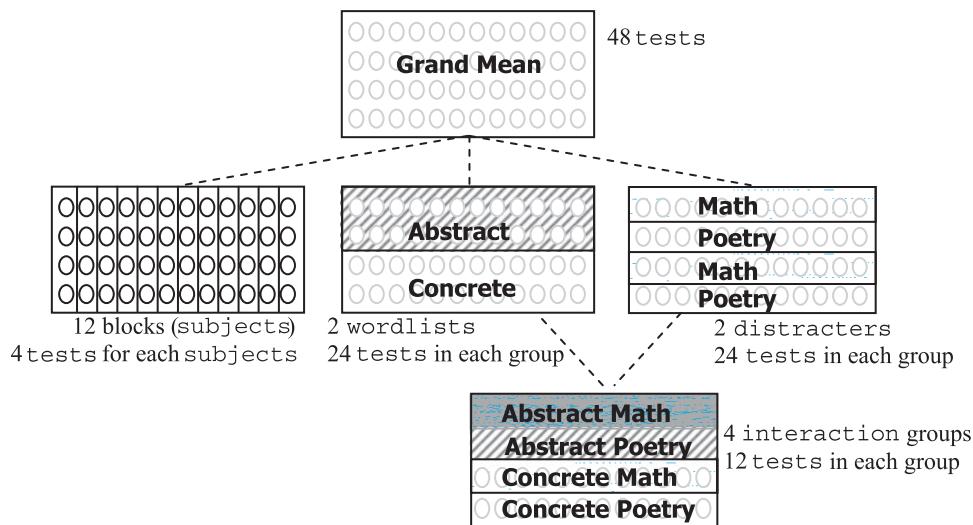
When multiple variables exist in complex models, it can often be helpful to visualize the design structure. Diagrams help us to understand the relationships between the variables and also identify all key groups (factor-level combinations) that are of interest in the study. Figure 5.2 shows that the original set of 48 units is divided based on two main factors of interest, Wordlist and Distracter. These factors are crossed, since every level of Wordlist is tested and has a meaningful interpretation with every level of Distracter.

Interaction effects can be found when two factors are crossed. Notice that there are four levels of the interaction effect (abstract math, abstract poetry, concrete math, and concrete poetry). While Wordlist and Distracter are crossed, the interaction term is nested in both Wordlist and Distracter. Each of the four levels of the interaction term occurs in only one level of Wordlist. In the same way, each of the four interaction levels occurs in only one of the Distracter levels.

Figure 5.3 shows the block design represented by Question 2. In this design, we see that the 48 units are grouped by three factors (Subject, Wordlist, and Distracter). Since each of the four interaction terms occurs with every Subject, the interaction term is crossed with Subject. The same logic can be used to show that Subject is crossed with both Wordlist and Distracter. In both the block and the split-plot designs for the memory experiment, the split-plot units (Tests) are nested within Student, Wordlist, Distracter, and the Wordlist\*Distracter interaction group.



**Figure 5.2** Design structure of a completely randomized design with two factors. Here 48 students are needed, each to take one test.



**Figure 5.3** Design structure of a block design with two crossed factors. Here 12 students are needed, each student takes all four types of tests.

## 5.6 Fixed and Random Factors

In addition to being classified as crossed or nested, each factor needs to be classified as fixed or random in order to determine the appropriate analysis for complex designs. **Fixed factors** are factors for which the levels are chosen because they are of specific interest in the study. The levels for a fixed factor would be used again by another researcher attempting to replicate the study. The factor **Wordlist** is a fixed factor. In the memory study, the type of word list (abstract or concrete) was selected because Josh and Ann were interested in testing for the effect of changing from abstract to concrete words.

**Random factors** are factors where the levels tested represent a random sample from some population of possible levels of interest. The levels of a random factor will differ for each researcher repeating an experiment. The factor **Student** is a random factor because, for example, each English student was randomly selected from a population of English majors at the college. There is no interest in determining an “English student #1” effect vs. an “English student #3” effect. Blocks and units are typically random factors.

It is not obvious whether **Major** is a random or fixed factor. If Josh and Ann specifically chose English, history, math, and computer science majors and wanted to test for differences among these, then **Major** would be considered a fixed factor. If Josh and Ann just randomly selected four majors from the college, **Major** would be considered a random factor. If it was a random factor, Josh and Ann would not attempt to determine how math majors compare to English majors, but would simply incorporate **Major** into the ANOVA to get a better feel for variability between majors. In this chapter, we will treat **Major** as fixed, but it is important to note that the only way to determine whether this factor is fixed or random is to ask the original researchers. The determination of whether a factor is fixed or random (as well as whether it is crossed or nested with other factors) often dramatically impacts the analysis. The following activity provides a demonstration of the importance of properly understanding each factor.

### Key Concept

Fixed factors have meaning only at the levels that were included in the experimental design. The same levels of that factor should be used if the experiment is repeated. If levels of factors are randomly selected, different levels will be randomly selected if the experiment is repeated. For example, if this experiment were repeated, the same distracters and types of word lists should be used, but since **Student** is a random factor, 12 new students would be randomly selected from the college.

## Activity Analysis of a Split-Plot Design

7. Is Distracter a fixed or random factor? Provide a justification for your answer.
8. Use software to conduct a split-plot design on the Memory data to test for mean differences in Score due to Major, Wordlist, Distracter, and the Wordlist\*Distracter interaction. Be sure to include the whole-plot unit Student in the model and specify that Student is a **random factor**. Write out the four null and alternative hypotheses corresponding to this study. Create a probability plot or histogram of the residuals to determine if the residuals are consistent with data from a normal distribution. Use the appropriate *p*-values, main effects plots, and interaction plot in stating your conclusions for each hypothesis.

 **NOTE**

Some data sets may show students listed as 1, 2, or 3 within each of the four majors; other data sets may show the students listed 1 through 12. In either case, the software output should still show that there are 12 students. In other words, using Student or Student2 should provide identical results.

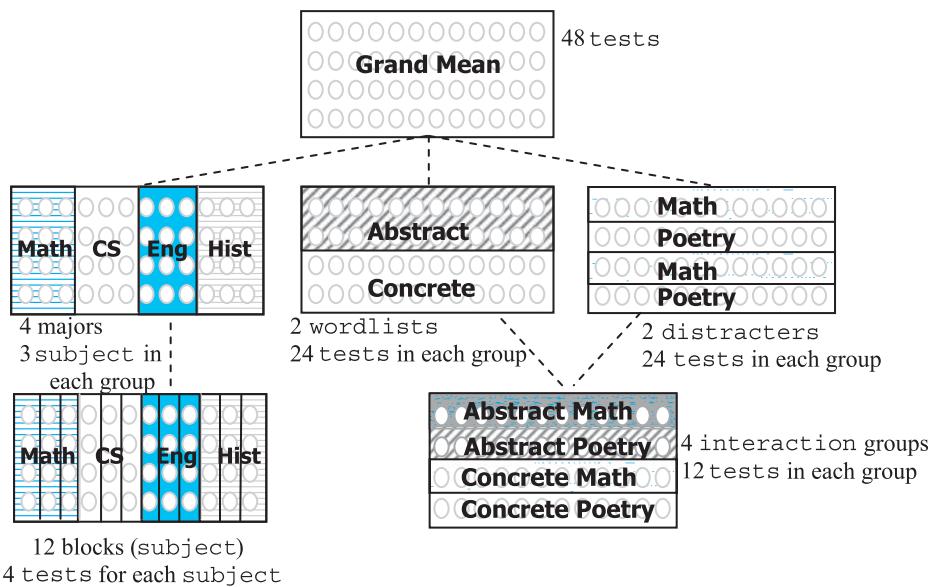
9. Compare the sums of squares (SS) in Question 2 to those in Question 8. Since the data set is still based on the 48 test scores, the total SS stays the same. List which SS stayed the same and describe how others changed. The extended activities will provide the mathematical calculations for these values.
10. Compare the *p*-values for Wordlist, Distracter, and the Wordlist\*Distracter interaction in Questions 2–8.
11. **Consequences of an Incorrect Model** Use software to conduct an analysis on the Memory data to test for mean differences in Score due to Major, Wordlist, Distracter, and Wordlist\*Distracter. *In this exercise, incorrectly assume that Student2 is a fixed factor and do not treat Student2 as nested within Major.* Submit any output that the software provides. How is the ANOVA different from what was observed in Question 8?
12. **Consequences of an Incorrect Model** Use software to conduct an analysis on the Memory data to test for mean differences in Score due to Major, Wordlist, Distracter, and Wordlist\*Distracter interaction. *In this exercise, incorrectly assume that Student 2 is a fixed factor but correctly treat Student2 as nested within Major.* Submit any output that the software provides. How is the ANOVA different from what was observed in Question 8?

 There are two separate randomizations done independently of each other in a split-plot design. The whole-plot portion by itself is simply a completely randomized design (one whole-plot factor, Major, with four levels and three whole-plot units, Students, per level). The split-plot portion (everything but the whole-plot factor) of this study is just the block design that was analyzed in Question 2. The split-plot units Test are randomly assigned to split-plot factor levels within each whole-plot unit Student. The four test scores are repeated measures for each Student. Combining the two produces the split-plot design.

The extended activities will discuss the mathematical calculations in more detail. *The key idea is that there is more than just one mean square error term of interest.* The mean square error based on the split-plot units is needed to test split-plot factors. However, to test whole-plot factors, a mean square error based on the variability of the whole-plot units will be used. Thus, Student is the appropriate unit when testing for differences in Major.

Take note that Major is not an experimental treatment. A student's academic major is simply observed. It is not reasonable to randomly assign a major to students in this study. While statements about causation can be made for Wordlist, Distracter, and the Wordlist\*Distracter interaction, be careful not to make any statements about a major *causing* a difference in the results.

Figure 5.4 displays the design structure for Question 8. The 48 units are grouped by three main factors (Major, Wordlist, and Distracter). Each subject is selected from only one Major. Thus, Subject is nested in Major, and no interaction between the 12 Subjects and Major can be calculated. However, it would be reasonable to calculate a Major\*Wordlist interaction and a Major\*Distracter interaction.



**Figure 5.4** Design structure of a split-plot design with two factors.

Figure 5.4 could be easily extended to include these two new interaction terms on the third row next to the Wordlist\*Distracter interaction term.

Recall that Josh and Ann had two final research questions that we have not yet addressed:

1. Does the impact of the type of distracter depend on the subject's academic major?
2. Does the impact of the type of word list depend on the subject's academic major?

## Activity Another Split-Plot Design

13. Is Major crossed with or nested within Wordlist?
14. Is Major crossed with or nested within Distracter?
15. Use software to analyze the Memory data using a split-plot design. Test hypotheses corresponding to mean differences in Score due to Major, Wordlist, Distracter, Wordlist\*Distracter, Major\*Wordlist, and Major\*Distracter. Be sure to specify that Student is a random factor that is nested in Major.
16. Compare the sums of squares (SS) in Question 8 to those in Question 15. Since the data set is still based on the 48 test scores, the total SS stays the same. List which SS stayed the same and describe how others changed.
17. Different designs are likely to provide different  $p$ -values even though the data are the same. Explain why the  $F$ -statistics and  $p$ -values for Wordlist, Distracter, and the Wordlist\*Distracter interaction in Question 8 are different from those in Question 15.
18. Write out the six null and alternative hypotheses corresponding to this study analyzed in Question 15. Use the appropriate  $p$ -values, main effects plots, and interaction plots as support when stating your conclusions.

## 5.7 Model Assumptions

The example used in this chapter was selected because it did not dramatically violate model assumptions or require transformations. However, the same process is used as with the simpler completely randomized designs discussed in Chapter 4. Equal variances within groups, independent and identically distributed observations, and normally distributed residuals should always be checked.

Residual plots should be checked after the model has been fit. A normal probability plot or histogram of the residuals is useful to determine if the normality assumption is violated. Residual versus fit and residual versus order plots are also useful to identify any unusual trends, skewness, or outliers. The equal variance assumptions should be checked for any error term in the model. Split-plot models assume equal variances among groups of whole-plot units and equal variances among groups of split-plot units.

## Activity Checking Model Assumptions

19. Create residual plots for the model created in Question 15. Is there evidence that the residuals are not normally distributed?
20. Any group representing a factor level or interaction that is tested with the split-plot error term (the split-plot error term is in the denominator of the  $F$ -statistic) should have equivalent variances. In addition, the variances of the differences between levels should be equivalent. Formal tests and model adjustments to account for unequal variances are beyond the scope of this text. An individual value plot (or boxplot) showing all factor-level combinations in the split-plot portion of the ANOVA can be used to roughly test for equal variances. For this memory study, create an individual value plot of each test score categorized by major, type of word list, and type of distracter. With such a small sample size in each group, it is very difficult to determine whether any of the Major, Wordlist, or Distracter factor combinations have unequal variances.
21. To test for equal variances for the whole-plot portion of the model in Question 15, calculate the average for each whole-plot unit. In our study, the whole-plot unit is Subject. Thus, calculate the average score for each of the 12 subjects (the average of the four tests for each subject). Then create an individual value plot (or boxplot) of subject average score by major. Each major will have three values, since there are three subjects for each major. While there is some evidence that the four majors have unequal variances, the small sample sizes and many group comparisons make it difficult to interpret. If the split-plot level variances are equivalent, then we can use this plot to test for equal variances in the whole-plot factors. However, if the split-plot level variances are clearly unequal, it is difficult to interpret whether the unequal variances are caused by subject variability or by submeasurements made within the split-plot.\*

The residual plots should look approximately normal. Both individual value plots do show some evidence of unequal variances. Question 21 shows that math students appear to be more consistent than computer science students. And there is one factor-level combination (English major/abstract word list/math distracter) in Question 21 with a sample variance equal to zero. However, these plots are vague simply because there are very few observations in each group. If there were equal variances in each factor-level combination, the observed data would be fairly reasonable. While there is some evidence of violations of the model assumptions, the author would conclude that there is not enough evidence to doubt the results. Thus, for these data, no transformations or more complex analysis is needed.

### MATHEMATICAL NOTE

A transformation may impact the strength of interaction terms. For example, if an interaction effect exists because the effects of two factors are multiplicative using the raw data, a log transformation may make the effects additive (and thus the interaction effect will disappear). Additive models (i.e., models with no interaction terms) are simpler, but may not address key questions of practical importance in the study. While there is no easy fix, it is important to understand the impacts of certain transformations.

\*When model assumptions are violated, it is often helpful to transform the data and then conduct a new ANOVA. Unfortunately, for this data set, a transformation was not helpful, and thus we suggest using the original data for analysis. The  $p$ -values may not be completely reliable because some model assumptions were violated. However, since three of the  $p$ -values are very small (less than 0.001), we believe that the mild lack of equal variances is unlikely to impact the conclusions for these three tests.

## 5.8 What Can We Conclude from the Memory Study?

Using only one memory experiment, this chapter explored many different statistical designs and calculated corresponding ANOVAs based on each model. The *p*-values and corresponding conclusions varied dramatically based on which design was selected for the analysis.

These examples demonstrate how easy it can be to conduct an incorrect analysis for more complex experimental designs. Simply looking at a data set does not tell us which analysis to use. We must know how the data were collected and how units were assigned to conditions. Failing to recognize the correct design structure or using the wrong experimental units often results in inaccurate *p*-values and wrong conclusions.

As with any study, conclusions should always be stated with consideration of random sampling and random allocation. If the subjects are a true random sample from these four majors, then the conclusions hold for all of these majors at this institution. However, this data set was collected toward the end of one spring semester for a final project. It would be reasonable to doubt that these results would hold for more than that particular year—or more than that semester.

Students clearly are not randomly assigned (i.e., randomly allocated) to a major. Thus, even if a significant difference between Majors is found, we cannot conclude that Major causes differences in memory abilities. The type of word list and type of distracter were randomly assigned. Thus, we can show causation for these variables if the results are significant.

If care is taken to properly design a study, ANOVA is a very powerful tool for drawing conclusions with relatively small data sets. In this memory study, proper conclusions could be drawn for six different hypothesis tests with only 48 observations. When used appropriately, block and split-plot designs make a study more efficient and help keep the variability low.

### A Closer Look

### Block and Split-Plot Designs

## 5.9 Calculating Crossed and Nested Effects

This chapter has focused on using *p*-values to determine if there is a statistically significant difference between conditions. The following two examples demonstrate how to estimate **effect sizes** (often simply called effects), which describe the amount the mean response changes as conditions change. In Chapter 4, we showed that for completely randomized designs,

- effect sizes are calculated by finding the appropriate average and subtracting the effects of all influencing factors, and
- effect sizes can be used to calculate the corresponding sum of squares.

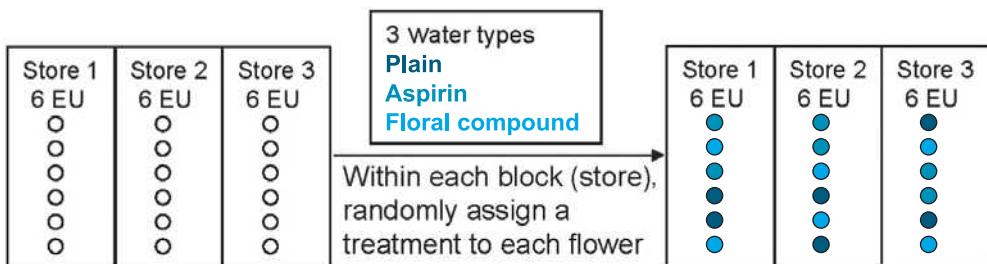
These points are still appropriate for block and split-plot designs. However, the calculations for the influencing factors depend on whether the factors are crossed or nested.

### Carnation Study

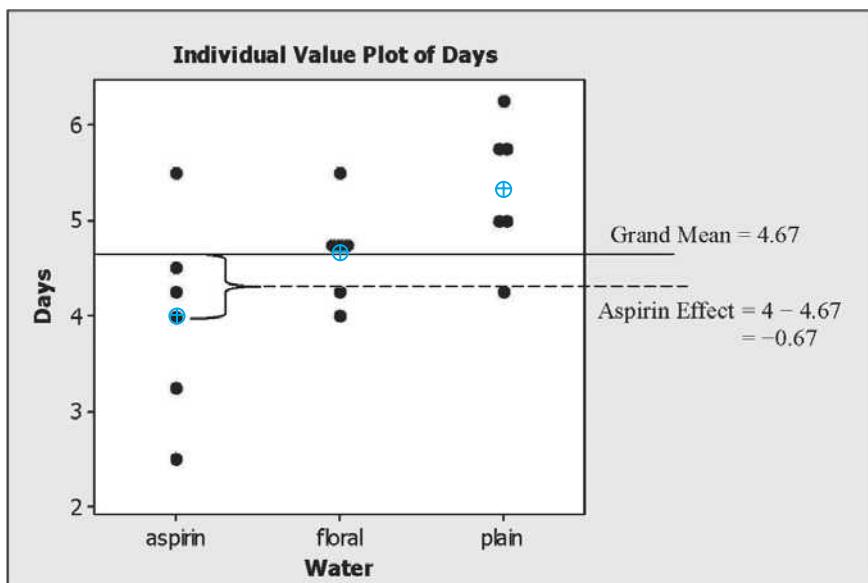
Students in an introductory statistics class tested the impact of different types of water solutions on the longevity of cut flowers. They purchased 18 white carnations and randomly assigned each flower to one of three treatments (plain water, one aspirin crushed and added to the water, and a floral compound provided by the flower shop) and then measured how many days it took until the flower wilted.

Since it is impossible to select a true random sample from all carnations sold on a particular date, the students purchased six white carnations from three different randomly selected stores in their city (for a total of 18 carnations). While not perfect, this is a very practical approach to accounting for population variability of white carnations. Store type was not a factor of interest, but it could impact the results. Thus, it is appropriate to include the nuisance factor *Store* in a block design. Figure 5.5 shows that six experimental units (EUs) are selected from each store, and then the random assignment of treatment to each unit occurs within each block (*Store*).

*Store* and *Water* are crossed factors. Since the same water solution is assigned to flowers from each store, the effects of water, aspirin, and floral compound have meaning across stores. The effect of *Aspirin* is



**Figure 5.5** Block design used in the carnation study. Random assignment of experiment units (EUs) to a water type is done within each block Store.



**Figure 5.6** Calculating the effect size of aspirin in the carnation study.

the average rating of all flowers treated with the aspirin water solution minus the grand mean (overall mean). Figure 5.6 graphically shows how the effect of aspirin is calculated.

In this study, three stores were randomly selected; thus, we are not particularly interested in comparing stores or calculating a Store effect. However, the Store 3 effect (the average rating of all flowers purchased from Store 3 minus the grand mean) is used to calculate the residuals.

Each flower (unit) is nested within one of the three stores. The first flower purchased from Store 1 is not expected to have any relation to the first flower purchased from Store 2. Thus, finding a Flower 1 effect across all three stores (finding the average effect of being the first flower chosen from each store) is not of interest to us. However, we are interested in modeling the effect of each individual flower (each unit). The effect of each individual unit is the residual value.

Effects for nested factors are calculated by finding the appropriate average and subtracting the influencing factors, but for nested factors, the factor level in which it is nested is also an influencing factor. By looking at the data, we see that a particular flower, say the 13th flower, is nested in both Store 3 and Plain water. The Flower 13 effect is simply the effect of the 13th flower after taking into account the effects of the grand mean, Store 3, and Plain water.

$$\begin{aligned}
 \text{Flower 13 effect} &= \text{Flower 13 response} - \text{Store 3 effect} - \text{Plain effect} - \text{grand mean} \\
 &= 5.75 - 0.208 - 0.667 - 4.667 \\
 &= 0.208
 \end{aligned}$$

Thus, the 13th flower lasted 0.208 day longer than expected after accounting for the fact that the flower was from Store 3 and had a plain water treatment. In other words, the residual value for this particular flower (Flower 13) is 0.208.

## Extended Activity ➔ Analyzing a Block Design

Data set: Flower

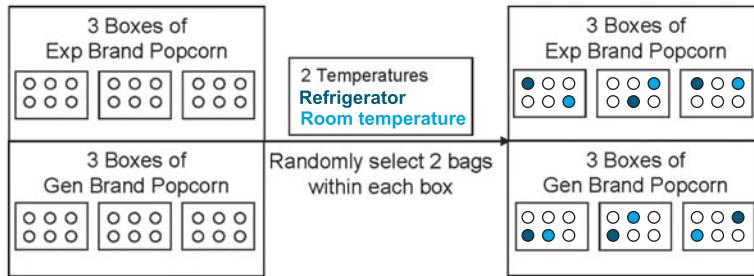
22. Identify any factors, the levels for each factor, the experimental units, and the response in the Flower data set.
23. Is Store fixed or random? Is Water fixed or random?
24. Are the two flowers assigned to each water/store condition replicates or repeated measures? Explain why.
25. State the null and alternative hypotheses for this study. Conduct an ANOVA to analyze the Flower data. Check the normality and equal variance assumptions. State your conclusions based on the *p*-values and include an appropriate graph. Section 5.9 will show that while this is a block design, the ANOVA is identical to a two-factor ANOVA with no interaction term, with one *F*-test for store effect and another for water solution effect.
26. Calculate the grand mean, the average for each Water level and the Water effects for the Flower data.

## Popcorn Brand and Storage Study

Two students in a design of experiments course wanted to test if the price and the storage location of popcorn influenced the percentage of kernels that popped.\* The students purchased three boxes of both an expensive and a generic popcorn brand (labeled Exp and Gen). Each box contained six microwavable bags of popcorn. Two bags were randomly selected from each box and stored for one week, one in the refrigerator (Frig) and the other at room temperature (Room). The bags were popped in random order according to the instructions on the box, and the percentage of popped kernels was calculated for each bag.

Since Boxes were randomly selected from each Brand population, the box-to-box variation should be measured by the experimental error (whole-plot MSE) within brands. There are three replicates (three “randomly” selected boxes) for each Brand.

To test for differences in storage temperature, Bags were randomly assigned to a treatment (Frig or Room). Figure 5.7 shows that this random assignment occurs within each Box. The popcorn Bags within a Box are repeated measures (not replicates) because the bag-to-bag variability within a Box is not representative



**Figure 5.7** The split-plot design used in the popcorn study. Three boxes of an expensive brand and three boxes of a generic brand of popcorn were selected. Then random assignment of experimental units (Bags) to a storage temperature was done within each box.

\*The students initially thought that temperature was the influencing factor, but a little research showed that the humidity of the storage location was believed to influence the percentage of popped kernels. In this study, the conditions in the refrigerator were much more humid than those in the other location.

of the population variability (the variability of all bags from a certain brand). The Bags within each Box are likely to be similar, since they probably were handled by the same person, at the same time, and at the same location. Thus, the two Bags can be considered as two measurements of the same Box.

## Extended Activity ➔ Analyzing a Split-Plot Design

Data set: Popcorn

27. Identify the whole-plot factor, the whole-plot unit, the split-plot factor, the split-plot units, and the response for the Popcorn data set. Label each factor as either fixed or random.
28. Is Box crossed with or nested within Brand? Explain.
29. Is Bag crossed with or nested within Box? Explain.
30. Is Temp crossed with or nested within Brand? Explain.
31. State three null and alternative hypotheses for this study (test for Brand, Temp, and the Brand\*Temp interaction). Conduct an ANOVA to analyze the Popcorn data. Even though storage temperature is crossed with Box (a nuisance factor), this interaction effect is of no interest and typically is not shown in an ANOVA table. Check the normality assumption and for equal variances.  
Create a main effects plot and interaction plot for Brand and Temp using the Popcorn data. Which Brand and which Temp appeared to do better? Even though the graph appears to show a difference between levels, only  $p$ -values can determine if differences are significant. In other words,  $p$ -values tell us how often we would expect effect sizes at least this large in a random sample if there really was no difference in Brand levels. State your conclusions for all three hypotheses.
32. Calculate the grand mean, the average for each Brand level, and the Brand effects for the Popcorn data. Identify these effect sizes on the main effects plot.
33. Calculate the average for each Temp level and the Temp effects for the Popcorn data. Identify these effect sizes on the main effects plot.
34. For each of the four Brand/Temp conditions, find the interaction effects with the following formula: condition (Brand/Temp) average – (Brand effect + Temp effect + grand mean). Use these values to create an interaction plot for the Popcorn data.

## Extended Activity ➔ Calculating Nested Effects

Data set: Popcorn

In the popcorn study, boxes are nested within brands. Thus, there is no overall Box 1 effect, but the effect of Box 1 depends on which brand it was selected from. Since each Box has meaning only within a Brand, there are 6 Box averages that need to be calculated. Table 5.1 shows some initial calculations of the Box effects.

35. Complete Table 5.1 to find each Box effect by calculating the Box average – (Brand effect + grand mean). Note that crossed effects always sum to zero. Nested effects (Box) also sum to zero within each appropriate factor level (Brand). For example, Box 1, Box 2, and Box 3 effects sum to zero within the expensive Brand, and Box 1, Box 2, and Box 3 effects sum to zero within the generic Brand.
36. Since Bags are the units in this study, the Bag effect is the same as a residual effect. Create a table similar to Table 5.1 to calculate the residual effect by subtracting all other effects from each Bag average (the observed percentage popped from each bag).

### Key Concept

Effect sizes are calculated by finding the appropriate average and subtracting all influencing factor effects.

**Table 5.1** Brand and Box effect calculations. Each Box effect is found by subtracting the Brand effect and grand mean from each Box average.

Brand	Box	Box Average	Brand Effect	Grand Mean	Box Effect
Exp	1	80	0.5	84	-4.5
Exp	2	86	0.5	84	1.5
Exp	3				
Gen	1				
Gen	2				
Gen	3	86.5			

## 5.10 Mathematical Calculations for ANOVA

Effects show the impact of each factor combination and identify which factors are most influential in a sample. However, a statistical hypothesis test is needed in order to determine if any of these effects are significant. Each row representing a factor of interest in the analysis of variance table corresponds to a hypothesis test to determine if there is statistical evidence that the effects are nonzero.

While effect size calculations vary depending on whether the factor is crossed or nested, the following calculations are used for all factors in all balanced designs. Certainly, we do not expect anyone to use hand calculations to conduct an ANOVA in practice. However, recognizing the relationship between effect sizes and sum of squares is helpful in conceptually understanding the logic behind the ANOVA calculations.

### Calculating Sums of Squares (SS)

**Sums of squares (SS)** are calculated by summing the squared factor effect for each observation,  $SS = \sum_{i=1}^N (\text{effect}_i)^2$ , where  $N$  is the total number of observations. For the popcorn example, there are 12 observations. Each of the six expensive-brand observations has a corresponding effect size of 0.5, and each of the six generic-brand observations has a corresponding effect size of -0.5. Thus,  $SS_{\text{Brand}}$  can be calculated as

$$\begin{aligned} SS_{\text{Brand}} &= (0.5)^2 + (0.5)^2 + (0.5)^2 + (0.5)^2 + (0.5)^2 + (0.5)^2 \\ &\quad + (-0.5)^2 + (-0.5)^2 + (-0.5)^2 + (-0.5)^2 + (-0.5)^2 + (-0.5)^2 \\ &= \sum_{i=1}^{12} (0.25) = 3 \end{aligned} \tag{5.1}$$

Table 5.2 shows the calculations for all the sums of squares in the popcorn study.

### Calculating Degrees of Freedom (df)

**Degrees of freedom (df)** are the number of free units of information. This section will provide the calculation details, but Hasse Diagrams described in Section 5.11 are very useful in visualizing how to calculate degrees of freedom. To calculate degrees of freedom in the popcorn study, let

$$a = \text{number of levels in Brand}$$

$$b = \text{number of levels in Temp}$$

$$c = \text{number of levels in Box within each Brand}$$

*For factors not nested in any other factors, the number of degrees of freedom is the number of levels minus one.* In the popcorn example, there are two levels of Brand and an effect is calculated for each level. As shown in the extended activities in Chapter 4, the restriction on the model terms requires that the effects

**Table 5.2** Calculation of each effect and sum of squares for the popcorn study.

% Popped (or Response)	Brand Effect	Temp Effect	B*T Effect	Box Effect	Bag Effect	Brand Effect Squared	Temp Effect Squared	B*T Effect Squared	Box Effect Squared	Bag Effect Squared
84	0.5	-0.33	2.83	-4.5	1.5	0.25	0.11	8.03	20.25	2.25
76	0.5	0.33	-2.83	-4.5	-1.5	0.25	0.11	8.03	20.25	2.25
86	0.5	-0.33	2.83	1.5	-2.5	0.25	0.11	8.03	2.25	6.25
86	0.5	0.33	-2.83	1.5	2.5	0.25	0.11	8.03	2.25	6.25
91	0.5	-0.33	2.83	3	1	0.25	0.11	8.03	9.00	1.00
84	0.5	0.33	-2.83	3	-1	0.25	0.11	8.03	9.00	1.00
74	-0.5	-0.33	-2.84	-3	-3.33	0.25	0.11	8.03	9.00	11.11
87	-0.5	0.33	2.84	-3	3.33	0.25	0.11	8.03	9.00	11.11
84	-0.5	-0.33	-2.84	0	3.67	0.25	0.11	8.03	0.00	13.44
83	-0.5	0.33	2.84	0	-3.67	0.25	0.11	8.03	0.00	13.44
83	-0.5	-0.33	-2.84	3	-0.33	0.25	0.11	8.03	9.00	0.11
90	-0.5	0.33	2.84	3	0.33	0.25	0.11	8.03	9.00	0.11
<b>Total</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>3.00</b>	<b>1.33</b>	<b>96.33</b>	<b>99.00</b>	<b>68.33</b>

of all levels within a factor sum to zero. Thus, knowing the effect of the generic brand automatically forces a known expensive brand effect. So there is really only one free effect size (one degree of freedom) for the Brand factor. The same holds for the two temperature levels.

$$df_{\text{Brand}} = a - 1 = 2 - 1 = 1$$

$$df_{\text{Temp}} = b - 1 = 2 - 1 = 1$$

For nested factors, restrictions in ANOVA require that all nested effects sum to zero within each level of the factor they are nested in. Box is nested in Brand. Since there are three boxes within each of the two brands, there are a total of six effects that can be calculated. However, the restriction states that the three Box effects in the expensive brand need to sum to zero. If two Box effects in the expensive brand are known, the third Box effect can be determined by subtraction. This same restriction holds for the generic brand. In general, for every Brand level, there are  $c - 1$  pieces of free information.

$$df_{\text{Box}} = a(c - 1) = 2(3 - 1) = 4$$

The degrees of freedom for an interaction of two crossed terms is the total number of interaction effects minus (the df for each term it is nested in) minus one. For the Brand\*Temp factor interaction, there are  $a \times b = 4$  effects that are calculated. The sum of these four effects is zero; thus, one of the  $a \times b = 4$  interaction effects must depend on the other effects. In addition, model assumptions in ANOVA require that effects within each treatment combination sum to zero. For example, the interaction effects within the expensive brand level sum to zero. The same is true for the generic brand level. Thus, an additional  $a - 1$  pieces of information are fixed. By the same logic, an additional  $b - 1$  pieces of information are fixed, since the interactions within each Temp level must sum to zero.

$$\begin{aligned} df_{\text{Brand*Temp}} &= \text{number of interaction effects} - \text{pieces of information already accounted for} \\ &= \text{number of interaction effects} - [df_{\text{Brand}} + df_{\text{Temp}} + 1] \\ &= ab - [(a - 1) + (b - 1) + 1] \\ &= 4 - [1 + 1 + 1] = 1 \end{aligned}$$

### ► MATHEMATICAL NOTE ▶

In many introductory textbooks, the calculation of degrees of freedom for an interaction term is written as  $(a - 1)(b - 1)$ . Note that  $ab - [(a - 1) + (b - 1) + 1] = ab - a - b + 1 = (a - 1)(b - 1)$ .

Similarly, the degrees of freedom for residuals (`Bags` in our example) also fits these restrictions.

$$\begin{aligned} df_{\text{Bag}} &= \text{number of Bags} - \text{pieces of information already accounted for} \\ &= \text{number of Bags} - [df_{\text{Box}} + df_{\text{Brand} \times \text{Temp}} + df_{\text{Brand}} + df_{\text{Temp}} + 1] \\ &= abc - [a(c - 1) + (a - 1)(b - 1) + (a - 1) + (b - 1) + 1] \\ &= 12 - [2(3 - 1) + (2 - 1)(2 - 1) + (2 - 1) + (2 - 1) + 1] = 4 \end{aligned}$$

where  $abc = 12 =$  the number of units (`Bags`).

### Key Concept

There is a very close relationship between calculating effects and calculating degrees of freedom for any factor of interest in ANOVA tables. Effects for any factor are found by calculating the appropriate average and subtracting the effects of any influencing factors. Degrees of freedom are found by calculating the number of levels and then subtracting the number of pieces of information that are already accounted for by influencing factors.

## Calculating Mean Square (MS) and F-Statistics

In Chapters 2 and 4, mean squares were calculated with the following formulas:

- **Mean square (MS)** =  $SS/df$  for each factor. MS is a measure of (between group) variability for each factor.
- **Mean square error (MSE)** =  $SS_{\text{Error}}/df_{\text{Error}}$  is equal to the pooled variance of sample units within each level.
- **F-statistic** =  $(\text{MS for each factor})/(\text{appropriate MSE})$ .

Both the carnation and the popcorn examples have units sampled from larger groups (`Flowers` within `Stores` and `Boxes` within `Brand`). In completely randomized designs and block designs (such as the carnation example), only one measurement (`Days`) is taken on each unit. Thus, there is only one measurement of unit-to-unit variability (MSE).

Split-plot designs take more than one measurement on the same unit. Three `Boxes` were randomly selected from each `Brand`, and then two measures (two `Bags`) were taken within each `Box`. Thus, the popcorn split-plot example has two sizes of units (`Box`, the whole-plot unit, and `Bag`, the split-plot unit), and each unit size is used to measure unexplained variability. Figure 5.8 provides the ANOVA for the popcorn example where the whole-plot error term is called `MSBox(Brand)` (stated `Box` nested within `Brand`) and the split-plot error term is called `MSBag` (often called `MSError`).

Recall that the *F*-statistic is a ratio of between-group variability to within-group (unit-to-unit) variability. Since each `Box` is randomly sampled from a `Brand` population, it seems reasonable that the best measure of unit-to-unit variability within `Brand` is `MSBox(Brand)` (i.e., `Boxes` are the best representation of the variation within `Brand`). Thus, the `Brand` *F*-statistic is calculated as `MSBrand/MSBox(Brand)`.

To test the effect of temperature, we took two bags that were as similar as possible (from the same box). Each `Bag` was randomly assigned to a temperature. Thus, the unit-to-unit variability within `Temp` is best represented by a measure of variability between bags. The *F*-statistic for `Temp` is `MSTemp/MSBag`. The `Brand*Temp` interaction *F*-statistic also uses `MSBag` in the denominator. Hasse diagrams, described in the next section, provide a nice set of rules for how to properly determine degrees of freedom and calculate *F*-statistics.

Source	DF	SS	MS	F	P
Brand	1	3.00	3.00	0.12	0.745
Box(Brand)	4	99.00	24.75	1.45	0.364
Temp	1	1.33	1.33	0.08	0.794
Brand*Temp	1	96.33	96.33	5.64	0.076
Error	4	68.33	17.08		
Total	11	268.00			

Figure 5.8 ANOVA table for the popcorn study.

## 5.11 Hasse Diagrams

As designs become more complex, it can become more difficult to determine which error term should be used in the denominator for every  $F$ -test. The appropriate denominator (measure of unit-to-unit variability) in the  $F$ -tests will depend on the three initial questions:

- What is the design structure?
- Is each factor crossed or nested with other factors?
- Is each factor fixed or random?

Hasse diagrams (pronounced “hahs”) are effective for determining how to properly calculate  $F$ -statistics and  $p$ -values for all balanced designs (i.e., same number of units in every condition).<sup>4</sup> A set of rules for calculating Hasse diagrams is given below. Hasse diagrams for the flower and popcorn studies are provided as examples to illustrate how these rules can be used.

### Rules for Developing Hasse Diagrams

1. Start row 1 with a node labeled  $M$  for the grand mean.
2. Put a node on row 2 for each factor that is not nested in any term. Draw an arrow from each node on row 2 to the grand mean. Place parentheses around any random factor.
3. Add a node on row 3 for any factor directly nested in row 2 (including interaction terms). Use arrows to connect the nodes in row 3 to all nodes in previous rows in which they are nested. Place parentheses around any random factor or any factor that is nested in a random factor. If an interaction term contains at least one random effect, the entire interaction is considered random.
4. On each successive row, say row  $k$ , add a node for any factor directly nested in row  $k - 1$ . (including interaction terms). Draw arrows to connect each node in row  $k$  to all the nodes in previous rows in which they are nested. Place parentheses around any random factor or any factor that is nested in a random factor.
5. When all interactions or nested factors have been exhausted, create a node for error on the bottom line. This often involves replacing the smallest unit of measurement with the word *error*. Draw arrows to nodes in the row above.
6. For each node, add a superscript that indicates the number of effects for each term (the number of interaction effects is always a product of the number of main effects).
7. For each node, add a subscript that indicates the degrees of freedom for that term. Degrees of freedom for a term are found by starting with the superscript for that particular node and subtracting out the degrees of freedom for all terms connected with arrows above it.

If the Hasse diagram is developed, finding the denominator for the appropriate  $F$ -test is typically straightforward: The denominator for testing a node for any factor (call it factor  $A$ ) is the next eligible random term below  $A$  in the Hasse diagram. If there are two or more *next eligible random terms*, use an approximate test. Approximate tests usually are a combination of existing mean square error values. Most software packages do them automatically.

Figures 5.9 and 5.10 show how to calculate Hasse diagrams for the carnation and popcorn studies. For the carnation study, there is only one error term (thus only one MSE). Error is the first random term following the Store effect and the Water solution effect. Thus, error (i.e., Flower or unit) is the denominator in both  $F$ -tests.

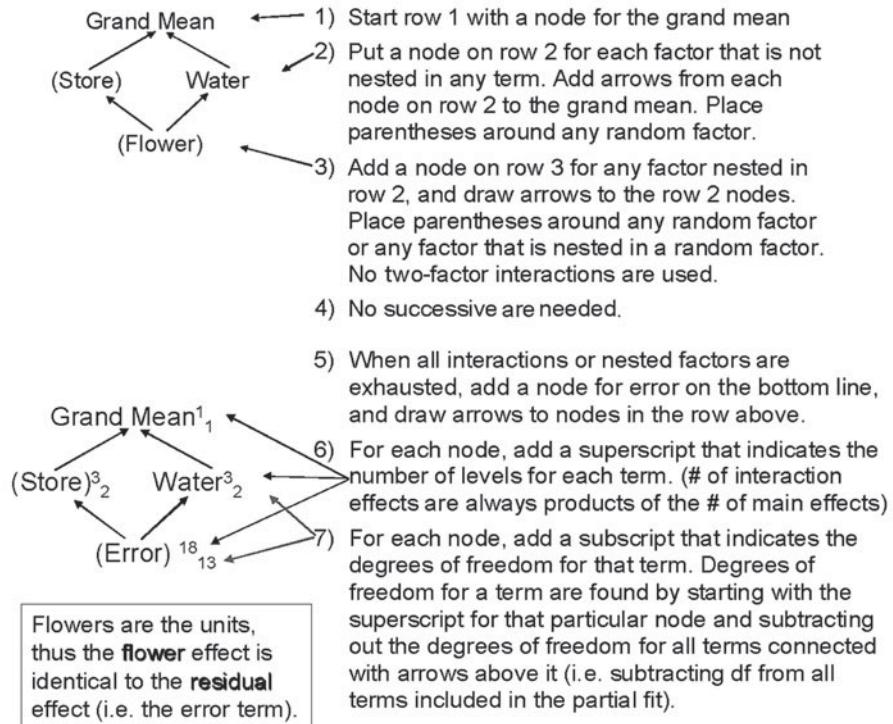


Figure 5.9 Hasse diagram for the flower study.

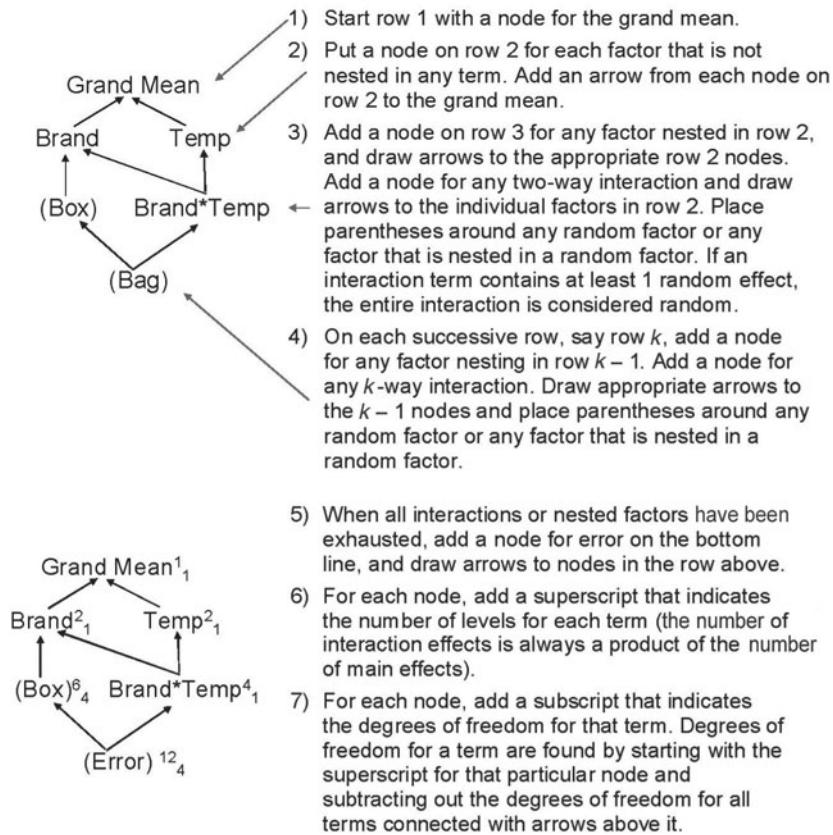


Figure 5.10 Hasse diagram for the popcorn study.

For the popcorn study, Box (whole-plot error) is the first random term below Brand and thus is used in the denominator for the Brand  $F$ -test. Bag (error or subplot error) is the first random term below Temp and Brand\*Temp; thus, error (i.e., Bag, or split-plot unit) is the denominator in both Temp and Brand\*Temp  $F$ -tests.

### Key Concept

Hasse diagrams can help to identify whether factors are crossed/nested or fixed/random. They are very beneficial in determining how to properly calculate  $F$ -statistics and  $p$ -values in more complex designs. These Hasse diagrams are appropriate only for balanced designs.

### ► MATHEMATICAL NOTE ▶

If a model includes **mixed interaction terms** (there are both fixed and random factors in an interaction term) and there are multiple terms below the mixed interaction term in the Hasse diagram, it may be necessary to determine whether the effects are **restricted** or **unrestricted**. This chapter does not include any designs this complex, but texts listed at the end of this chapter discuss restricted and unrestricted effects in more detail. In general, restricted effects sum to zero while unrestricted effects do not. In these more complex unrestricted designs, the denominator for the appropriate  $F$ -test is calculated as follows:

- As the denominator for testing a term (call it term  $A$ ), take the next eligible random term below  $A$  in the Hasse diagram.
- If there are two or more next eligible random terms, then use an approximate test. Approximate tests usually are a combination of existing mean square error values. Most software packages do them automatically.

In these more complex restricted designs, the denominator for the appropriate  $F$ -test is calculated as follows:

- As the denominator for testing a factor (call it factor  $A$ ), take the next eligible random term below  $A$  in the Hasse diagram.
- If there are two or more next eligible random terms, then use an approximate test. Approximate tests usually are a combination of existing mean square error values. Most software packages do them automatically.

## Developing Statistical Models

Hasse diagrams are useful in visualizing statistical models. Each node represents a term in the model. For the carnation example, the observed values and four nodes can be represented by the following terms:

$y_{ijk}$ : the  $k$ th observation from Store  $j$  and Water solution  $i$

$\mu$ : overall grand mean (the benchmark value)

$\alpha_i$ : effect of Water solution  $i$  ( $i = 1, 2, 3$ )

$\beta_j$ : effect of Store  $j$  ( $j = 1, 2, 3$ )

$\varepsilon_{ijk}$ : error for the  $k$ th subject ( $k = 1, 2$ ) from the  $i$ th Water solution and  $j$ th Store

The mean response in the ANOVA model is  $\mu + \alpha_i + \beta_j$ . Effects are a measure of the differences between group means. The effect  $\alpha_1$  represents the change in the response from the grand mean to the aspirin group mean (assuming that the aspirin group was labeled as the first group). The statistical model for the carnation study is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \text{ for } i = 1, 2, 3, j = 1, 2, 3, \text{ and } k = 1, 2 \text{ where } \varepsilon_{ijk} \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad (5.2)$$

For the popcorn example, the observed values and six nodes can be represented by the following terms:

$y_{ijkm}$ : the  $m$ th observation from Brand  $j$ , Box  $i$ , and Temp  $k$

$\mu$ : overall grand mean (the benchmark value)

$\alpha_i$ : effect of Brand  $i$  ( $i = 1, 2$ )

$\eta_k$ : effect of Box  $k$  ( $k = 1, 2, 3$ ) within Brand  $i$

$\beta_j$ : effect of Temp  $j$  ( $j = 1, 2$ )

$\varepsilon_{ijkm}$ : error for the  $m$ th observation from the  $i$ th Brand,  $k$ th Box, and  $j$ th Temp

The following statistical model can be used to represent the popcorn study:

$$y_{ijkm} = \mu + \alpha_i + \eta_{k(i)} + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijkm} \quad (5.3)$$

for  $i = 1, 2, 3$ ,  $j = 1, 2, 3$ , and  $k = 1, 2$   
 where  $\eta_{k(i)} \stackrel{\text{iid}}{\sim} N(0, \sigma_1^2)$ ,  $\varepsilon_{ijkm} \stackrel{\text{iid}}{\sim} N(0, \sigma_2^2)$

In the split-plot model, there are two random error terms:  $\eta_{k(i)}$ , which represents the whole-plot error (MSBox(Brand)), and  $\varepsilon_{ijkm}$ , which represents the split-plot error (MSBag).

## Extended Activity Creating Hasse Diagrams

Data set: Memory

37. Create a Hasse diagram for Josh and Ann's memory study corresponding to Question 1. Use the Hasse diagram to find the corresponding statistical model.
38. Create a Hasse diagram for Josh and Ann's memory study corresponding to Question 2. Use the Hasse diagram to find the corresponding statistical model.
39. Create a Hasse diagram for Josh and Ann's memory study corresponding to Question 8. Use the Hasse diagram to find the corresponding statistical model.
40. Create a Hasse diagram for Josh and Ann's memory study corresponding to Question 15. Use the Hasse diagram to find the corresponding statistical model.

## 5.12 Wash Your Hands: Analysis of Covariance (ANCOVA)

In the memory study at the beginning of this chapter, blocking improved the precision in our results. In other words, by accounting for subject-to-subject variability we were able to better identify true differences in the type of word list and type of distracter.

Analysis of covariance (ANCOVA) is another technique that can be used to improve the precision of the results. ANCOVA is used when there is a variable (called a covariate or concomitant variable) that is *linearly* related to the response but not related to the other factors.

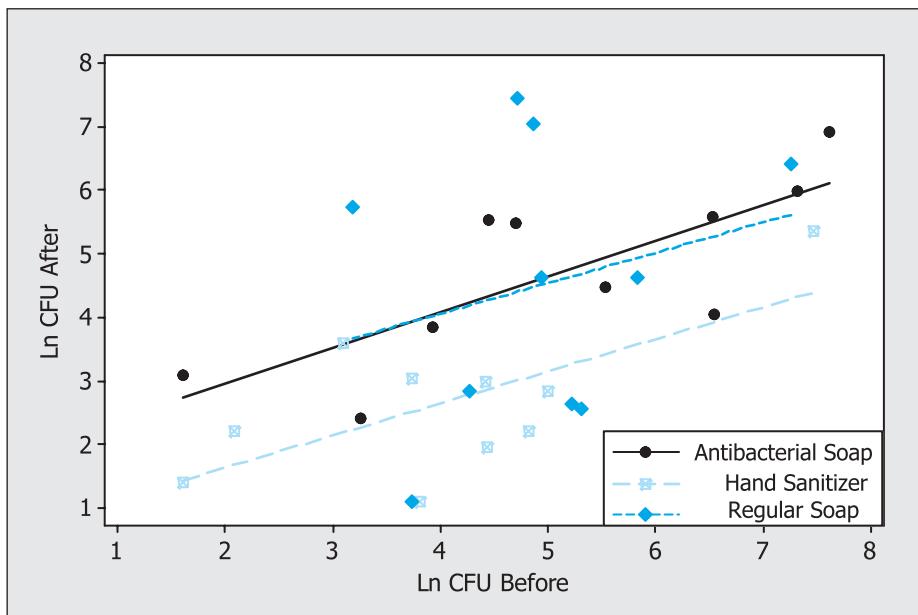
Antibacterial soaps have recently become very popular. However, some doctors fear that they are being overused. Some bacterial strains, such as *Staphylococcus aureus*, have developed resistance to these soaps, decreasing their effectiveness. If antibacterial soaps are overused, it is likely that more bacterial strains will become resistant and antibiotics will no longer be effective.

With the H1N1 outbreak in 2010, hand washing was well promoted as one of the most effective and economical methods for preventing the spread of infectious diseases. For a final project, two statistics students were interested in comparing the effectiveness of three hand-washing soaps (62% ethyl-alcohol-based hand sanitizer, 0.15% tricolsan antibacterial soap, and a soap lacking antimicrobial agents).

These researchers used sterile swabs to swab each participant's right hand before washing, first going around the fingers and then making an S-shape on the palm. The right hand was swabbed again after subjects washed their hands with one of the cleansers for 20 seconds and let their hands air dry for 3 minutes.

Swabs were then placed into a microfuge tube with 500 mL of saline and swirled for 30 seconds to knock off bacteria. Then the researchers vortexed the microfuge tube for 20 seconds and pipetted 100 mL of bacterial solution onto an L-agar plate. Plates were then put in an incubator at 30°C for 72 hours. Colony-forming units (CFUs) were then counted on each plate.

Figure 5.11 shows a linear relationship between the natural logs of the before and after CFU bacteria counts for each type of hand cleanser. Note that the three lines are almost parallel. In other words, the slopes for all three hand cleansers are very similar.



**Figure 5.11** Scatterplot of the hand-washing study.

### Key Concept

ANCOVA is effective if (1) the covariate is linearly related to the response, (2) there are similar slopes for each factor level, and (3) the covariate is not associated with the other factors. If these assumptions are not met, it is likely better to use a block design.

## Extended Activity

### Comparing Hand Cleansers

Data set: Handwash

41. Create a scatterplot of the before versus the after CFU counts. Explain why a natural log transformation is appropriate for both the before and the after CFU counts.
42. Analyze the hand soap study to compare the three hand soaps. Use the natural log of before CFU counts as the covariate and the natural log of after CFU counts as the response. Use the  $p$ -values to state your conclusions, taking into account random sampling and random allocation.
43. It would also be possible to use a block design in this study. Sort the before CFU counts into three blocks (representing low, medium, and high CFU counts). Each block should have 10 units. Analyze the data using the three blocks to test for differences in the hand-washing treatments. This is an unbalanced (i.e., unequal sample size in each group) block design with one blocking factor and one cleansing factor—no covariate should be used.

Both the ANCOVA and the block design provide similar results. The  $p$ -values are roughly equivalent and the  $R^2$  values show that both models explain a similar percentage of the overall variability.

Recall that trying several different analyses and then basing your final conclusions on the smallest  $p$ -values will tend to bias your results toward finding significant results even if there really is no significant difference between groups. While statisticians' recommendations vary, we suggest that ANCOVA is more appropriate. In general, it is best to keep data as accurate as possible. The quantitative before CFU measurements provide more specific information than the low, medium, and high blocks. In addition, Figure 5.11 shows that the ANCOVA model assumptions are met.

Calculations for the ANCOVA are left to the computer. However, it is useful to understand the basic mechanics of the ANCOVA model.

The **null hypothesis** for the ANCOVA model is that there is no treatment effect—one linear regression line is appropriate for the entire data set. The model under the null hypothesis (often called the null model) is

$$y_{ij} = \mu + \beta \times x_{ij} + \varepsilon_{ij} \quad (5.4)$$

where  $y_{ij}$  represents each observed value,  $\mu$  is the  $y$ -intercept,  $\beta$  is the slope coefficient,  $x_{ij}$  is the observed value of the covariate and  $\varepsilon_{ij}$  is the residual value.

The **alternative hypothesis** is that there is a treatment effect. In other words, the regression line is improved (sums of squared residuals are smaller) if parallel lines are created for the treatment groups. The model under the alternative hypothesis (often called the full model) is

$$y_{ij} = \mu + \alpha_i + \beta \times x_{ij} + \varepsilon_{ij} \quad (5.5)$$

where  $\alpha_i$  represents a treatment effect.

The slope  $\beta$  is assumed to be the same in both models. The hypothesis test focuses on whether the  $\alpha_i$ s (and thus the corresponding  $y$ -intercepts  $\mu + \alpha_i$ ) differ.

As in ANOVA, the  $F$ -statistic in ANCOVA is the mean square treatment over the mean square error. The mean square error is based on the sum of square residuals for the full model. The mean square treatment represents the reduction of sum of square residuals in going from the null (single line) model to the full (parallel lines) model. Thus, if the full model dramatically improves the fit, the  $F$ -statistic will be large, the corresponding  $p$ -value will be small, and we will conclude that the treatment effects  $\alpha_i$  are significantly different.

## Chapter Summary

This chapter showed how to design, analyze, and interpret block and split-plot designs. In completely randomized designs, units are randomly assigned to (or randomly selected from) a factor-level combination. These designs are an effective method for collecting, organizing, and analyzing data. However, they do not capture the unwanted variability that is caused by differing skills of each subject. Accounting for the subject-to-subject variability will reduce the unexplained variability and make it easier to identify true differences between the means.

Block and split-plot designs have some type of restriction on the randomization process that can account for subject-to-subject variability. **Blocking** is the process of grouping units based on some preexisting similarity. It is useful to include a blocking factor when units within groups (blocks) are more similar to each other than other units. When used appropriately, blocks tend to provide a more precise analysis by creating a smaller mean square error (MSE) term.

As shown in the activities, the blocking factor reduces the amount of unexplained variability and thus reduces the sum of squares error term. The MSE term is smaller only if the blocks explain enough of the variability to compensate for the loss of degrees of freedom in the error term.

While block effects can be of interest in a study, they are most often used to incorporate nuisance factors into a design in order to provide more accurate results. If blocks are preexisting conditions and not assigned to experimental units, no statement about *causation* can be made.

**Split-plot designs** consist of block designs where the blocking factor (also called the whole-plot unit) is randomly assigned to (or sampled from) a factor-level combination. These designs have two sizes of units, called whole-plot units and split-plot units.

In order to properly analyze these block or split-plot designs, it is essential to know (1) if each factor is fixed or random and (2) which factors are crossed or nested.

Two factors are **crossed** if every level of one factor can occur at every level of the second factor. Every factor-level combination has meaning. Any factor, call it factor  $A$ , is **nested** in a second factor, factor  $B$ , if the levels of factor  $A$  are different for each level of factor  $B$ . In other words, a level of factor  $A$  doesn't have meaning unless we know the level of factor  $B$ .

**Fixed factors** (often referred to as factors with fixed effects) are factors for which levels were specifically chosen because they represent something of importance to the study. The levels of **random factors** (i.e., factors with random effects) are selected at random from a larger population. The levels of random factors are

not of particular interest in the study, but are selected so that the variation between levels can be accounted for. Blocks and units are typically assumed to be random factors.

Most of the extended activities provided more examples in order to help you better understand how and when to apply block and split-plot designs to various studies. Although this chapter focused on designed experiments, the same analysis can be performed on observational studies. As in the repeated measures memory experiment, not even experiments require all factors to be experimental treatments (e.g., a student's academic major is not "assigned"). While the calculations are the same for experiments and observational studies, the conclusions that can be drawn will vary. Causation can be shown only for factors for which the treatments were imposed on the experimental units.

This chapter did not attempt to describe all types of statistical designs. But the designs discussed here are the most common designs that incorporate a factorial treatment structure. Most designs beyond the scope of this chapter still have their foundations in these three designs.<sup>5</sup> For example, most advanced courses titled "Design of Experiments" will include discussions of the following designs:

- Latin square designs are designs with two blocking factors (two nuisance factors) and one factor of interest.
- Nested designs and split-split plot designs can have three or more levels of nesting. For example, trees could be nested in fields and leaves nested in trees.
- Fractional factorial designs and balanced incomplete block designs are designs in which not every factor-level combination is tested.

The **Hasse diagram** provides a visual display of the relationships between factors for balanced (i.e., same number of units in every condition) complete designs. It provides rules for determining the appropriate linear model, degrees of freedom, ANOVA table, and *F*-tests. Thus, the Hasse diagram can be used to determine the appropriate analysis for designs more complex than the ones discussed in this chapter.

## Exercises

### E.1. Design Your Own Study 1

Assume that your university offers 16 sections of Calculus 1 in the fall semester. At your university, there are four faculty members who each teach four sections of Calculus 1. All have agreed to be involved in a test to determine if a computer-based curriculum increases core understanding of the course material. Each faculty member will teach two courses with and two courses without the computer-based curriculum, and at the end of the semester all students will take the same final exam.

- a. Identify any extraneous variables that may potentially bias your results. How will these potential biases be addressed?
- b. List each factor in the study and determine whether it is fixed or random. Also note whether it is crossed with or nested within other factors.
- c. Specify the response variable you will use.
- d. Specify the units and sample size for this study.
- e. List each factor-level combination and describe how you will use randomization in this study.
- f. List the factors and the corresponding degrees of freedom that will occur in the ANOVA table.
- g. Will this be a completely randomized, block, or split-plot design? Explain.

### E.2. Design Your Own Study 2

Assume that you have a small garden and are interested in knowing whether particular species of tomato plants will yield more tomatoes. You are also interested in knowing if adding fertilizer will increase the growth of tomatoes. You will try three species of tomatoes with and without fertilizer and measure the total weight of the tomatoes produced. Because of limited space, your garden is broken into three small plots in separate locations. Each plot has the same total area (each plot is big enough to grow 12 plants), but the three plots have differing amounts of sunlight. Write a brief outline for an experimental design.

- a. Identify any extraneous variables that may potentially bias your results. How will these potential biases be addressed?
- b. List each factor in the study and determine whether it is fixed or random. Also note whether it is crossed with or nested within other factors.

- c. Specify the units and sample size for this study.
- d. List each factor-level combination and describe how you will use randomization in this study.
- e. List the factors and the corresponding degrees of freedom that will occur in the ANOVA table.
- f. Will this be a completely randomized, block, or split-plot design? Explain.

### E.3. Tension in Tennis 1: A Block Design

Data set: *Tennis*

Students on the college tennis team were interested in knowing if string tension affected the speed and accuracy of a tennis ball. They had recently read an article in the *Journal of Sports Sciences* and decided conduct a similar study for themselves.<sup>6</sup> After warming up, five men's varsity tennis players volunteered to hit 30 serves using three different racquet tensions over three days. Day was not treated as a factor of interest in the study, but used to avoid fatigue in the players. For each serve, players aimed at a target in the back center of the service box, and accuracy was measured as the distance in inches from the center of the box to the ball strike location. A radar gun was used to measure the velocity of each player's serve. Use *Speed* as the response variable for each of the following questions:

- a. Write the null and alternative hypotheses corresponding to the original research questions.
- b. To avoid bias, all players used a Wilson K Factor KSix-One Tour 90 racquet (which has a recommended string tension of between 50 and 60 pounds).<sup>\*</sup> In addition, each player used new balls each day and racquets were restrung after each day. Identify any questions or concerns you have about potential sources of bias in this study.
- c. Use software to conduct an ANOVA to analyze the *Tennis* data set using a block design.
- d. Create a main effects plot comparing the effects of *Player* and *Racquet*. Explain how the *p*-values correspond to what you see in the main effects plot.
- e. The goals of this study did not include determining if there was a difference between the performance of *Players*. Based on the degrees of freedom, sum of squares, and *p*-value for *Player* provided in the ANOVA, do you believe that blocking was helpful in this study?
- f. Create a probability plot or histogram of the residuals to determine if the residuals are consistent with data from a normal distribution.
- g. State your conclusions for this study. Be sure to clearly identify the population for which these conclusions hold and also state whether causation has been shown.

### E.4. Tension in Tennis 2: A Block Design with Unequal Sample Sizes

Data set: *Tennis*

Repeat Exercise 3 using *Accuracy* as the response instead of *Speed*. Notice that in this study there are several missing values. When a serve hit the net, no distance measurement could be made. Statistical software will account for missing values in the analysis. Thus, the *p*-values are still appropriate (assuming model assumptions are met) even though this is no longer a balanced design.

### E.5. Music and Performance

Data set: *Music*

Three students in an introductory statistics class read an article suggesting that listening to certain types of music helps doctors perform chest compressions for cardiopulmonary resuscitation (CPR). It is suggested that optimal CPR performance is 100 beats per minute. Most doctors start out at the right pace but tend to slow down over time. Gore and Lloyd found that doctors performed much better when listening to the song *Stayin' Alive*, by the Bee Gees, which has a rhythm of 103 beats per minute.<sup>7</sup>

These students developed a research question asking if music also influences performance in other areas. They designed an experiment testing whether music tempo or length of test (1 minute or 3 minutes) influenced students' ability to type fast and accurately. Would subjects listening to *Stayin' Alive* type at a different speed than subjects listening to *Yesterday* by the Beatles?

Forty undergraduate students consented to be in the study. Each subject took four tests from the website [typingtest.com](http://typingtest.com) in random order based on two coin flips: 1 min/*Yesterday*, 1 min/*Stayin'*

---

<sup>\*</sup>A pound (lb) measures the force applied to the tennis string on the racquet.

*Alive*, 3 min/*Yesterday*, and 3 min/*Stayin' Alive*. The questions the researcher wanted to test were the effect of Song, Length, and Song\*Length on words per minute (WPM).

- a. Specify whether each factor (Subject, Song, and Length) is fixed or random and whether each factor is crossed or nested.
- b. Calculate an appropriate ANOVA for this study using WPM as the response. State your conclusions, taking into account random sampling and random allocation. Provide appropriate plots.
- c. Check the model assumptions. Create a plot to check if the data appear to be skewed or have outliers. Is there reason to doubt the equal variance assumption? Are the error terms approximately normally distributed?
- d. Draw a Hasse diagram corresponding to this study.
- e. These introductory students initially conducted a completely randomized design with only the Song, Length, and Song\*Length terms in the ANOVA model (block designs are not covered in the introductory class). They were disappointed to not see any significant results in their study. If the blocks are ignored in this study, the conclusions will be incorrect. Write a short paragraph explaining to these students why a block design (including Subjects as blocks) should be used instead of a completely randomized design. Include a main effects plot with Subject, Song, and Length (or other appropriate graphs) in your explanation.
- f. The student researchers were concerned that the order of the tests or the actual text used could influence the results. They were primarily concerned that subjects might struggle on the first test, so they decided to have each of the 40 subjects start with a 1-minute practice test. Results were not recorded for this practice test. Five distinct texts were also used for each of the tests. The order in which the five texts were given was the same for all subjects. For example, Trial 1 used the same text for all subjects.

Create a plot to see if there are any patterns in the residuals that can be explained by trial number in the study.

Note: If no pattern exists by trial number, we can have confidence that order or type of text did not bias our results. If a pattern exists by trial number, it should be noted that order or text may bias our ANOVA and conclusions. As in any study, it is dangerous to add new hypothesis tests after we have searched for patterns in our data. In addition, if a pattern existed, the design of this study would make it impossible to distinguish whether order or text was biasing the results. If the effect of order or text was important, a new study should be designed that properly accounts for these factors.

#### E.6. Adjusted Music and Performance

Data set: Music

The students in the music and performance study in Exercise 5 also recorded an adjusted WPM score (i.e., the words per minute adjusted by the student's accuracy).

- a. Calculate an appropriate ANOVA for this study using adjusted WPM as the response. State your conclusions, taking into account random sampling and random allocation. Provide appropriate plots.
- b. Assume these introductory statistics students asked you whether they should present WPM or adjusted WPM. What would you suggest? Explain why.
- c. These students also considered using Accuracy as a response variable. Explain why they should be cautious about testing several different response variables and then choosing to present the one result with the smallest  $p$ -values.

#### E.7. Baking Cookies

Data set: Cookies

Two students wanted to determine if people could taste the difference in chocolate chip cookies with varying amounts of sugar and varying amount of freshness. Nine batches were made, following the recipe on the chocolate chip bag as closely as possible except for the amount of sugar. Each batch was randomly assigned to one of three treatments: half the suggested amount of sugar, the suggested amount of sugar, and double the suggested amount of sugar.

On the day the nine batches of cookies were baked, the researchers handed out five cookies from each batch (a total of 45 cookies) to people in their dorm and asked them to rate the cookies

from 1 through 10, with 1 being inedible and 10 being the best cookie they every had. The researchers stored the rest of the cookies for a day. On the second day, the researchers handed out five more cookies from each of the original nine batches to students in their dorm and asked them to rate them from 1 through 10. The researchers did the same thing on the third day.

Split-plot designs are often used when time is a second factor. The whole-plot factor (*Sugar*) is randomly assigned to whole-plot units (*Batch*), and then these same units (*Batches*) are measured at several time points (*Day*). This is called a split plot in time, as the split plots are the time points within the units.

Note that the factor *Day* is confounded with any other effect that occurs over time. For example, suppose this study was conducted on a Saturday, Sunday, and Monday. Students may have been more stressed on Monday and unknowingly tended to give lower scores on Monday. Or more parents may have been around on the weekend and may have been more positive than students when rating the cookies.

- a. Specify whether each factor in the study is fixed or random and whether each factor is crossed or nested.
- b. Calculate an appropriate ANOVA for this study using *Taste* as the response. State your conclusions, taking into account random sampling and random allocation. Provide appropriate plots.
- c. Check the model assumptions. Create a plot to check if the data appear to be skewed or have outliers. Is there reason to doubt the equal variance assumption? Are the error terms approximately normally distributed?
- d. Draw a Hasse diagram corresponding to this study.

#### E.8. Split Plots in Agriculture

Data set: Corn

Split-plot designs were often originally used in agricultural experiments, where one factor was randomly applied to plots of land and then each plot was split up into smaller sections and a second treatment was randomly assigned to the smaller subplots. There are many reasons for using split-plot designs in agriculture. For example, a farmer can hire an airplane to spray herbicides or pesticides over large areas of land, and smaller plots within the larger plots can be used to plant rows of different species of a crop. Another example is growing strawberries in a greenhouse, where a factor, such as temperature, is applied to an entire greenhouse while another factor, such as water level, is applied to certain areas within the greenhouse.

A study was conducted to determine if different species of corn or amounts of nitrogen would impact yields.<sup>8</sup> Nitrogen is available in the soil as the result of natural and biological processes; however, farmers often add a nitrogen fertilizer to corn crops in order to increase yields. A nitrogen deficiency can considerably decrease yields. Fertilizer is expensive; thus, many farmers will choose to plant a species of corn that grows well with less nitrogen. Soil types vary with each plot of land. The amount of nitrogen can also vary from year to year, depending on the type of crop (e.g., wheat, oats, hay, soybeans, or corn) planted the previous year. Thus, the amount of nitrogen needed will be different for different plots. Researchers often conduct studies for a county or entire state, so they are primarily interested in measuring the variability between plots instead of how well a specific plot does compared to another. Determining the right species to plant as well as the right amount of fertilizer can dramatically impact a farmer's profit. In fact, a quick Google search will show that many agricultural states provide corn-nitrogen productivity calculators on the web.

Many seed companies and state-sponsored agricultural research centers have large testing areas that have been shown to be similar to other plots of land throughout their state. In a large testing area, eight 20-acre plots of land were randomly assigned to a nitrogen application rate (0, 70, 140, or 210 pounds of nitrogen per acre). Each of these 20-acre plots was also subdivided into five 4-acre subplots. Within each 20-acre plot, the subplots were assigned to be planted with one species of hybrid corn (A, B, C, D, or E). At the end of the season, the fields were harvested and the yield in bushels per acre was recorded.

- a. Specify whether each factor in the study is fixed or random and whether each factor is crossed or nested.
- b. Calculate an appropriate ANOVA for this study using *Yield* as the response. State your conclusions, taking into account random sampling and random allocation. Provide appropriate plots.

- c. Check the model assumptions. Create a plot to check if the data appear to be skewed or have outliers. Is there reason to doubt the equal variance assumption? Are the error terms approximately normally distributed?
- d. Draw a Hasse diagram corresponding to this study.
- e. Draw a picture of an aerial view of a possible 160-acre test area plot. There should be a total of 40 small plots in the final picture. Demonstrate how corn species and nitrogen rates could possibly have been randomly assigned.

#### E.9. Baking Cookies 2

Data set: Cookies2

Two students wanted to test whether ingredients (butter, Fleischmann's corn oil margarine, or unflavored Crisco), cooking time (short or long) or cookie type (chocolate chip or gingersnap) influenced taste ratings. Both main effects and interactions were of interest.

Twelve volunteers were found who were each willing to taste 12 cookies in random order (one taste for each of the two cookie types, the three ingredient types, and the two cooking times). Each volunteer ranked all 12 cookies on a scale from 1 to 10 (10 being the best).

- a. Specify whether each factor in the study is fixed or random and whether each factor is crossed or nested with all other factors.
- b. Calculate an appropriate ANOVA for this study using Rating as the response. State your conclusions, taking into account random sampling and random allocation. Provide appropriate plots.
- c. Check the model assumptions. Create a plot to check if the data appear to be skewed or have outliers. Is there reason to doubt the equal variance assumption? Are the error terms approximately normally distributed?
- d. Draw a Hasse diagram corresponding to this study.
- e. If these student researchers had asked your help in designing this study, would you have suggested having 144 people each taste one cookie and rate it? Is there a benefit to having 12 people each taste 12 cookies?

#### E.10. Paper Football

Data set: Football1

Ben, Hugh, and Alex wanted to determine if the size of the “football” made a difference in scoring accuracy in the game of paper football. (Details of the simple tabletop game can be found at <http://www.paperfootball.us>.) In addition, these students were interested in knowing if the effect of football size was dependent on a player’s experience level.

The researchers set up a goal that was 8 inches above a table and 10 inches wide. The subjects kicked/flicked the football from 42 inches away from the goal.

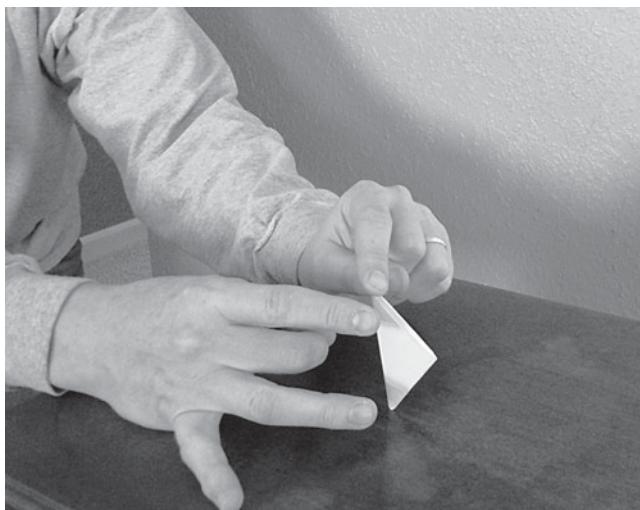


Photo courtesy of Shonda Kuiper.

There were 18 volunteers, who self-identified as experienced or inexperienced players. After a few practice kicks, each subject kicked 20 small and 20 large footballs in random order (flipping a coin before each kick) at the goal. The response for this study was the proportion of successful goals.

- Specify whether each factor in the study is fixed or random and whether each factor is crossed or nested with all other factors.
- Calculate an appropriate ANOVA for this study using Proportion as the response. State your conclusions, taking into account random sampling and random allocation. Provide appropriate plots.
- Check the model assumptions. Create a plot to check if the data appear to be skewed or have outliers. Is there reason to doubt the equal variance assumption? Are the error terms approximately normally distributed?
- Draw a Hasse diagram corresponding to this study.

### E.11. Color Distractors

Data set: Colors

Kastner et al. studied how the brain recognizes the shape and color of an object.<sup>9</sup> They found that the process of identifying the shape and color of an item is carried out not simultaneously but in steps.

Two students decided to investigate the impact of color distraction on a shape matching game called Shapesplosion. In this game, subjects are asked to match shapes by placing specifically shaped pegs into the matching hole as quickly as possible. These student researchers selected eight female science majors and eight female non-science majors at their college. Each of the 16 subjects played four games (each of the following four games in random order):

Factor 1 (game complexity): the Shapesplosion game with 15 pieces (pegs) or the game with 18 pieces

Factor 2 (color distracter): the Shapesplosion game where the peg color matches the hole color or the game where the pegs are a different color than the hole

The Time it took (in seconds) to complete each game was recorded.

- List each factor in the study and specify whether it is fixed or random and whether it is crossed or nested with all other factors.
- Calculate an appropriate ANOVA for this study using Time as the response. Check the model assumptions. Create a plot to check if the data appear to be skewed or have outliers. Is there reason to doubt the equal variance assumption? Are the error terms approximately normally distributed?
- Calculate an appropriate ANOVA for this study using log(Time) as the response. Does log(Time) better fit the model assumptions? Would you suggest using Time or log(Time) when presenting the results?
- State your conclusions for the six hypotheses that can be tested with this study. Provide appropriate plots and take into account random sampling and random allocation when stating your conclusions.
- Draw a Hasse diagram corresponding to this study.

### E.12. Color Distractors 2

Data set: Colors2

The design in Exercise 11 is sometimes called a split-plot/repeated measures (SP/RM) [1,2], where the 1 represents the number of whole-plot factors and the 2 represents the number of split-plot factors.

In fact, the previous exercise provides only half of the data. In addition to the 16 females discussed in Exercise 11, 16 males were also tested. The complete design is a split-plot/repeated measures (SP/RM) [2,2], with both gender and division as whole-plot factors.

- Draw a Hasse diagram corresponding to this study, assuming that we are interested in all possible two-way interactions (do not include three-way interactions).
- Calculate an appropriate ANOVA for this study, assuming that we are interested in testing all two-way interactions. Check the model assumptions and use an appropriate transformation if

needed. Some software packages such as Minitab, do not easily analyze a split-plot design with two whole-plot factors. Instead create one new factor (Gender and Division) with 4 levels and use this as the one whole-plot factor in your ANOVA.

- State your conclusions for this study. Provide appropriate plots and take into account random sampling and random allocation when stating your conclusions.

#### E.13. Hasse Diagrams

- Develop a general Hasse diagram for a split-plot design with the following characteristics:
  - Whole-plot factor  $A$  has  $a$  levels ( $i = 1, 2, \dots, a$ ).
  - Split-plot factor  $B$  has  $b$  levels ( $j = 1, 2, \dots, b$ ).
  - Within each whole-plot factor,  $c$  whole-plot units are selected ( $k = 1, 2, \dots, c$ ).
  - Within each  $AB$  factor-level combination,  $d$  subplot units are selected ( $m = 1, 2, \dots, d$ ).
- Write out a statistical model similar to Equation (5.2) or (5.3) for this general design.

#### E.14. Hasse Diagrams 2

- Develop a general Hasse diagram for a split-plot design with one whole-plot factor and two split-plot factors:
  - Whole-plot factor  $A$  has  $a$  levels ( $i = 1, 2, \dots, a$ ).
  - Split-plot factor  $B$  has  $b$  levels ( $j = 1, 2, \dots, b$ ).
  - Split-plot factor  $C$  has  $c$  levels ( $k = 1, 2, \dots, c$ ).
  - Within each whole-plot factor,  $d$  whole-plot units are selected ( $m = 1, 2, \dots, d$ ).
  - Within each  $ABC$  factor-level combination,  $n$  subplot units are selected.
- Write out a statistical model similar to Equation (5.2) or (5.3) for this general design.

#### E.15. Hasse Diagrams 3

- Develop a generic Hasse diagram for a split-plot design with two whole-plot factors and one split-plot factor:
  - Whole-plot factor  $A$  has  $a$  levels ( $i = 1, 2, \dots, a$ ).
  - Whole-plot factor  $B$  has  $b$  levels ( $j = 1, 2, \dots, b$ ).
  - Split-plot factor  $C$  has  $c$  levels ( $k = 1, 2, \dots, c$ ).
  - Within each  $AB$  factor-level combination,  $d$  whole-plot units are selected ( $m = 1, 2, \dots, d$ ).
  - Within each  $ABC$  factor-level combination,  $n$  subplot units are selected.
- Write out a statistical model similar to Equation (5.2) or (5.3) for this general design.

## Endnotes

---

- English professor William Whyte Watt in his book *An American Rhetoric*, 3rd ed. (New York: Rinehart and Company, 1958), p. 382.
- Tim Hesterberg, “It’s Time To Retire the ‘ $n \geq 30$ ’ Rule,” *Proceedings of the American Statistical Association, Statistical Computing Section* (CD-ROM), 2008.
- R. V. Lenth, “Some Practical Guidelines for Effective Sample Size Determination,” *The American Statistician*, 55 (2001): 187–193.
- Details on this technique can be found in P. W. Iverson and M. G. Marasinghe, “Visualizing Experimental Designs for Balanced ANOVA Models Using Lisp-Stat,” *Journal of Statistical Software*, 18 (2005): 3; O. Kempthorne, “Classifactory Data Structures and Associated Linear Models,” in G. Killianpur, P. R. Krishnaiah, and J. K. Ghosh (eds.), (New York: North Holland, 1982), 397–410; *Essays in Honor of C. R. Rao*; S. L. Lohr, “Hasse Diagrams in Statistical Consulting and Teaching,” *The American Statistician*, 49.4 (1995): 376–381; M. G. Marasinghe, and P. L. Darius, “A Structure-Based Approach for Model Determination in Experimental Designs,” *Proceedings Statistical Computing Section, American Statistical Association*, 1990, 143–150; and W. H. Taylor and H. G. Hilton, “A Structure Diagram Symbolization for Analysis of Variance,” *The American Statistician*, 35.2 (1981): 85–93.
- The following texts are useful resources for learning more about design of experiments: G. Cobb, *Introduction to Design and Analysis of Experiments*, (Emeryville, CA: Key College Publishing, 1998); D. Montgomery, *Design and*

*Analysis of Experiments*, 7th ed. (New York: Wiley, 2008); G. Oehlert, *A First Course in Design and Analysis of Experiments* (San Francisco: W. H. Freeman, 2000); S. R. Searle, *Linear Models* (New York: Wiley, 1971).

6. Rob Bower and Rod Cross, "String Tension Effects on Tennis Ball Rebound Speed and Accuracy During Playing Conditions," *Journal of Sports Sciences*, 23.7 (July 2005): 765–771.
7. Laura Gore and Julia Lloyd, "Pop Song 'Stayin' Alive' Helps People Perform Chest Compressions for CPR," *Scientific Assembly*, American College of Emergency Physicians, Oct. 23, 2008, <http://www.acep.org>, accessed 12/2/08.
8. A. Reza, *Design of Experiments for Agriculture and the Natural Sciences*, 2nd ed. (Boca Raton, FL: Chapman & Hall CRC, 2006), p. 138.
9. S. Kastner, P. De Weerd, R. Desimone, and L. G. Ungerleider, "Mechanisms of Directed Attention in the Human Extrastriate Cortex as Revealed by Functional MRI," *Science*, 282 (1998): 5386.
10. A. Paivio, "Mental Imagery in Associative Learning and Memory, *Psychological Review*, 76 (1969): 241–263.
11. H. Ebbinghaus, *Memory: A Contribution of Experimental Psychology* (New York: Columbia University Press, 1885; reprinted by Dover, 1964); C. M. MacLeod, "Forgotten but Not Gone: Savings for Pictures and Words in Long Term Memory," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14 (1988): 195–212.
12. D. L. Scarborough, "Stimulus Modality Effects of Forgetting in Short Term Memory," *Journal of Experimental Psychology*, 95 (1972): 285–289; J. Brown, "Some Tests of the Decay Theory of Immediate Memory," *Quarterly Journal of Experimental Psychology*, 10 (1958): 12–21; L. R. Peterson and M. J. Peterson, "Short Term Retention of Individual Items," *Journal of Experimental Psychology*, 58 (1959): 193–198.
13. F. L. M. Craik and R. S. Lockhart, "Levels of Processing: A Framework for Memory Research," *Journal of Verbal Learning and Verbal Behavior*, 11 (1972): 671–684; F. L. M. Craik and E. Tulving, "Depth of Processing and the Retention of Words in Episodic Memory," *Journal of Experimental Psychology: General*, 104 (1975): 268–294; J. M. Gardiner, R. I. Java, and A. Richardson-Klavehn, "How Level of Processing Really Influences Awareness in Recognition Memory," *Canadian Journal of Experimental Psychology*, 50 (1996): 112–122; R. S. Lockhart and F. I. M. Craik, "Levels of Processing: A Retrospective Comment on a Framework for Memory Research," *Canadian Journal of Psychology*, 44, (1992): 87–112.
14. G. Cobb, *Introduction to Design and Analysis of Experiments* (Emeyville, CA: Key College Publishing, 1998), adapted from p. 2.
15. D. Montgomery, *Design and Analysis of Experiments*. 7th ed. (New York: Wiley, 2009), p. 21.
16. Ibid, p. 22.

# Research Project: What Impacts Memory?

Now that you have analyzed Josh and Ann's memory study, it is time to conduct your own research project. The following pages provide guided steps for conducting your own research project involving an online memory game. You will design your own study, collect data, analyze the results, draw conclusions, and present your results.

## Reviewing the Literature

Memory is the process of retaining and recalling knowledge or experiences. Human memory is very complex and can be tested in many ways. Factors commonly used in testing memory include the following:

- *Nature of the material, such as letters, numbers, symbols, words, sentences:* Characteristics of the materials can also be altered. For example, Paivio found a difference between people's ability to recall concrete words (such as *dog, house*) and abstract words (such as *joy, ugly*).<sup>10</sup> Word length and commonness of the word can also impact recall.
- *Nature of the test:* Similar tests can have different instructions for the subjects. For example, subjects may or may not be told that the purpose of the experiment is testing memory.
- *Retention interval between when the material is presented and when it is recalled.*<sup>11</sup>
- *Rate of presentation* (how quickly the material is presented).
- *Amount of material presented.*
- *Modality:* Do people recall material better when it is presented visually, orally, or both?<sup>12</sup>
- *Study strategy:* How does each person try to learn (i.e., encode) the material? For example, does the person think about the meaning of each word he or she is trying to remember, or does he or she simply think about the appearance of the word?<sup>13</sup>

One type of memory test, serial recall, evaluates the ability of people to recall information in the specified order in which it was presented. How many items a subject can remember in order without an error, called memory span, is also studied. The Memorathon game is an example of serial recall and memory span. In this game, the subject is expected to repeat a sequence of buttons provided by an electronic device. Each time the subject successfully repeats the given sequence of buttons, the sequence gets longer. The challenge is to remember as long a sequence as possible.

The online Memorathon game, at <http://www.pearsonhighered.com/mathstatsresources>, provides the opportunity to design and conduct experiments that test which factors, such as sound, color, or speed, have the largest impact on memory.

To simply try out the game, you can go to the above site and leave all the variables blank. However, if you want to find your score in the database of results, a specific course ID and student ID will be needed.

In the following project, you will have the opportunity to develop your own experiment using an online game.

1. Read the paper by A. M. Surprenant, "Distinctiveness and Serial Position Effects in Tonal Sequences," *Perception & Psychophysics*, 63.4 (2001): 737–745. Focus only on the first of the three experiments in the paper. If there are any words that you do not understand, look them up and provide a short definition for each.
2. Using Surprenant's paper, identify or answer the following:
  - a. Objective of the experiments
  - b. Any relevant background (from journals that were referenced)
  - c. Response variable(s)
  - d. Factors and levels that were tested
  - e. Variables that were held constant during the experiment
  - f. Nuisance factors (i.e., factors that are not of interest but may influence the results)
  - g. Were any interactions tested, and if so, what was observed?
  - h. What type of design was used in the experiments?

- i. How many trials were run for each experiment?
- j. How could you modify the experiments if you were going to conduct a similar study on memory?

Be ready to submit your answers as well as discuss this material in class.

## Playing the Computer Game and Developing a Design

Go to the website <http://www.pearsonhighered.com/mathstatsresources> to find the Memorathon game, and use it to develop your own experiment. Which factors do you believe will have the most significant effect on memory?

Submit your design at the beginning of class. Note that these games allow you to develop three new factors of your own choice. Be prepared to discuss how you addressed each of the following five points.

3. **Clearly define a problem and state the objectives of your experiment.** Before any experiment is conducted, it is essential that everyone involved clearly understand the objectives of the experiment, what measurements will be taken, what material is needed, and what procedures will be used.

State the null and alternative hypotheses. This is often much more difficult than it appears. First, designing an experiment often involves many people from diverse backgrounds. These people typically have different goals and use very different terminology. Second, it is important to design an experiment that is “general enough to be of scientific interest, yet specific enough that it is feasible to run within time, space, and material limitations.”<sup>14</sup>

4. **Identify the response variable, factors, potential levels of each factor, and units.**

a. Verify that the response variable provides the information needed to address the question of interest. What are the range and variability of responses you expect to observe? Is the response measurement precise enough to address the question of interest? Are you interested in the number of wins or the time it took to win?

b. Investigate all factors that may be of importance or potentially cause bias in the results. When the objective is identifying which factors have most influence on the response, it is usually best to keep the number of factor levels low. Even though several levels are controlled in these games, consider other factors that should be controlled. Did a subject drink a significant amount of caffeine before the game, or did the subject stay up all night studying for an exam? What can you do to account for these variables?

c. Once experimental factors have been identified, carefully identify a reasonable range for each factor. “In some fields there is a large body of physical theory on which to draw in explaining relationships between factors and responses. This type of non-statistical knowledge is invaluable in choosing factors, determining factor levels, deciding how many replicates to run, interpreting the results of the analysis, and so forth. Using a designed experiment is no substitute for thinking about the problem.”<sup>15</sup>

5. Identify what other factors need to be controlled during the experiment to eliminate potential biases. Identify how measurements, material, and process may involve unwanted variability. What conditions would be considered normal for this type of experiment? Are these conditions controllable? If a condition changed during the experiment, how might it impact the results? List each nuisance factor, and explain at what levels and how each will be controlled throughout the experiment, even if it is simply held constant. Will subjects be allowed a practice game before the actual experiment? How important is it to randomize the order in which the games are played?

6. Choose an experimental design.

a. Keep the design and analysis as simple as possible. A straightforward design is usually better than a complex design. If the design is too complicated and the data are not collected properly, even the most advanced statistical techniques may not be able to draw appropriate conclusions from the experiment.

b. How many trials will be run? Is the cost of replicating the experiment worth gaining a better understanding of the sample-to-sample variability? Can you completely randomize all the trials, or do you need to account for timing, subject variability, and other nuisance factors? *Important note:* If your experimental design included multiple explanatory variables, it is essential to clearly

understand the difference among completely randomized, block, and repeated measures designs. If each subject is assigned to only one condition (plays the game only once), it may be appropriate to use a completely randomized/full factorial design. If each subject is tested under several conditions (each subject plays several games), a more complex design structure, such as a block or repeated measures design, may be needed.

7. Explain how your experimental design builds on previous research.
  - a. Identify relevant background on response and explanatory variables, such as theoretical relationships, expert knowledge/experience, or previous studies.
  - b. Explain where this experiment fits into the study of the process or system. Experiments are usually iterative. While your assignment is to design, conduct, and analyze one experiment, it is important to realize that each experiment is just one step in a much larger process. Although some researchers disagree (particularly depending on the discipline), Montgomery suggests that only 25% of your resources be used in a first experiment.

In most situations, it is unwise to design too comprehensive an experiment at the start of a study. Successful design requires knowledge of the important factors, the ranges over which these factors are varied, the appropriate number of levels for each factor, and the proper methods and units of measurement for each factor and response. Generally, we are not well-equipped to answer these questions at the beginning of the experiment, but we learn the answers as we go along.... Of course there are situations where comprehensive experiments are entirely appropriate, but as a general rule, most experiments should be iterative. Consequently, we usually should not invest more than about 25 percent of the resources of experimentation (runs, budget, time, etc.) in the initial experiment. Often these first efforts are just learning experiences, and some resources must be available to accomplish the final objectives of the experiment.<sup>16</sup>

## Collecting Your Own Data

8. Prepare any questions you would like to ask cognitive psychologists or statisticians before you finalize your experimental design.
9. Write specific lab procedures that you will use while conducting the experiment. Determine who will collect the data at what time, how the trials will be randomized, how the data will be recorded, and exactly what will be measured.
10. Ensure that your group has received appropriate Institutional Review Board (IRB) approval (see the supplemental material on IRB).

## Presenting Your Own Model

11. Meet with your professor to discuss your experimental design and analysis.
12. Collect the data. In conducting the experiment, did you identify any other sources of variability that could be impacting the results? Submit lab procedures.
13. Write the research paper (see “How to Write a Scientific Paper or Poster” on the accompanying CD). Bring three copies of your research paper to class. Submit one to the professor. The other two will be randomly assigned to other students in your class to review.

## Final Revision

Make final revisions to the research paper. Submit the first draft, other students’ comments and checklists, the data set with variable descriptions, and the final paper.

## Other Project Ideas

Several of the extended activities and end-of-chapter exercises in this chapter and in Chapter 4 can also be used to develop your own project ideas. In addition to the Memorathon game, a color distracter game is available at <http://www.cs.grinnell.edu/~kuipers/statsgames/Shapeshlosion>.

The popcorn study in the extended activities showed that the generic brand did slightly better than the expensive brand. The generic brand did have more kernels that popped; however, the size and quality of the popped kernels were much better in the expensive brand. Conduct a similar popcorn study, but use a different response variable, such as a measure of volume. You may want to adjust your model to account for other potential sources of variability, such as cooking time, power setting on the microwave, or type of microwave.

During the music and performance study in the extended activities, the researchers noticed that when students listened to “Stayin’ Alive,” they started out quickly but then slowed down dramatically, especially in the 3-minute test. When students listened to “Yesterday,” their speed stayed much more consistent throughout each test. Design and conduct a study on some activity, such as typing or exercise, to determine if listening to a fast song is helpful at first but then causes people to slow down. One way to do this would be to split each test into three subtests, such as beginning, middle, and end of the test. See whether you find evidence that subjects fatigue as the test wears on, especially when they start at a fast rate.

Take a favorite recipe and try adjusting a few ingredients. You may want to block over subjects tasting the item, or you may want to randomly assign split-plot units to other variables such as time (freshness) or storage type (stored in refrigerator, counter, or freezer).

# Categorical Data Analysis: Is a Tumor Malignant or Benign?

*It is commonly believed that anyone who tabulates numbers is a statistician.  
This is like believing that anyone who owns a scalpel is a surgeon.*

—Robert Hooke<sup>1</sup>

This chapter introduces inference techniques for data in which both the explanatory and response variables are categorical. The term *categorical data analysis* often refers to analysis of data in which the response variable is categorical. However, in this chapter we will restrict our focus to cases where both the explanatory and the response variables are categorical and where there is no natural ordering to the categories.

Most of the hypothesis tests discussed in previous chapters use the normal distribution to model the mean response. In this chapter, proportions, odds ratios, and relative risk will be used to summarize categorical data. We will start by looking at cancer cell data to determine if there is a relationship between the shape of the cell nuclei and the proportion of malignant cells. In this chapter, we will discuss how to do the following:

- Conduct a simulation study, chi-square test, and Fisher's exact test
- Calculate and properly interpret summary statistics such as relative risk and the odds ratio
- Determine which test to use based on various sampling schemes and questions of interest

## 6.1 Investigation: Is Cell Shape Associated with Malignancy?

Cancer is a disease that occurs when abnormal cells grow in the body. When DNA (a substance in every cell) is damaged, normal cells will often repair the damaged DNA. Cancer cells are cells in which the DNA is not repaired. DNA can be damaged by many things, including viruses, tobacco smoke, alcohol, and too much sunlight. Cells with damaged DNA can also be inherited. Cancer cells can continue to grow and divide and usually form tumors (a lump or mass) somewhere in the body. Cancer cells can also outlive normal cells.

It is important to note that not all tumors are cancerous. If a lump is detected, part of it can be removed surgically and a biopsy conducted to determine if the mass is *benign* or *malignant*. Benign tumors are scar tissue or abnormal growths that do not spread and are typically harmless. Malignant (or invasive) cancer cells are cells that can travel, typically through the bloodstream or lymph nodes, and begin to replace normal cells in other parts of the body. If a tumor is malignant, it is essential to remove or destroy all cancerous cells in order to keep them from spreading. If a tumor is benign, surgery is typically not needed and the harmless tumor can remain.

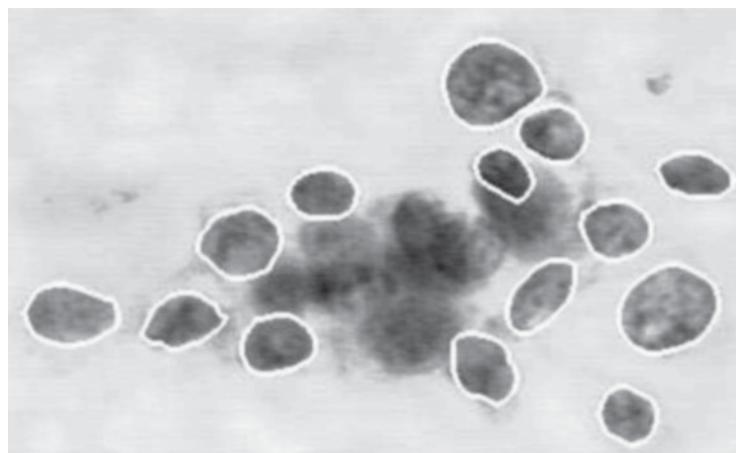
A biopsy requires surgery to remove a section (or all) of the tumor and will leave a scar. *Fine needle aspiration (FNA)* is a technique in which a small sample of the tumor is taken using a needle and visually inspected through a microscope. Since many tumors are benign, it is often preferable to have an FNA, which is less invasive and less traumatic than a biopsy.

Breast cancer is the second leading cause of cancer death among women in the United States. The American Cancer Society estimated that, in the United States in 2010, 39,840 women and 390 men died from breast cancer and 207,090 women were diagnosed with breast cancer.<sup>2</sup> This type of cancer is often detected by finding a lump (or mass) in the breast.

Wolberg and Mangasarian developed a technique to accurately diagnose breast masses using only visual characteristics of the cells within the tumor.<sup>3</sup> This system is used at University of Wisconsin hospitals to assist doctors in diagnosis of breast cancer.<sup>4</sup> Each FNA sample is placed on a slide, and characteristics of the cellular nuclei within the tumor are examined under a microscope. Several measurements, such as size, shape, and texture, are collected for each of the nuclei visible on the slide, and then an algorithm is used to determine the likelihood that a mass is benign or malignant.

In this chapter, we will focus on just a few characteristics from a relatively small data set that was collected at University of Wisconsin hospitals in Madison. We will start by determining if the shape of a cell nucleus can help us to determine whether a tumor is malignant or benign.

Typically, healthy cell nuclei have round or ellipsoid shapes. Figure 6.1 shows a sample of malignant cells that appear to have grown such that the perimeters of the cell nuclei have somewhat concave points.



**Figure 6.1** An image of malignant cells where nuclei are outlined with a curve-fitting program. Reprinted by permission. Mangasarian, Street & Wolberg, "Breast Cancer Diagnosis and Prognosis via Linear Programming," INFORMS Journal Operations Research, 43.4, 1995. © 1995, Institute for Operations Research and the Management Sciences (INFORMS).

## 6.2 Summarizing Categorical Data

Table 6.1 shows data from 37 FNA slide samples. Slides with smooth ellipsoid-shaped nuclei were classified as round, and slides with poorly shaped cell nuclei were classified as concave. A biopsy was also conducted on each of these samples to determine if each was malignant or benign.

**Table 6.1** The numbers of benign and malignant tumors for round and concave cell nuclei.

		Type		Total
		Benign	Malignant	
Shape	Round	9	7	16
	Concave	4	17	21
Total		13	24	37

In Table 6.1, both the variables are categorical. A **categorical variable** is one for which the measurement consists of categories, such as college major, political affiliation, personality type, gender, or pass/fail grade. When a variable is categorical, each subject or unit must fit in one and only one category. A **binary variable** is a special type of categorical variable that has just two possible categories.

Summarizing quantitative data such as age, weight, and income often includes calculating the mean value. However, Table 6.1 cannot be used to calculate the mean Shape or mean Type value. Since these categorical variables have no ordinal or interval meaning, it is not appropriate to focus on the mean response; rather, we focus on the proportion (or percent) of responses that fall into each category.

A table that counts the number of observations in each group, such as Table 6.1, is called a **contingency table** (or cross-tabulation table). A contingency table with two variables is called a **two-way contingency table**. Table 6.1 is also called a **2 × 2 table**, since there are two row groups and two column groups. In Table 6.1, the Shape of the cell is called the **row variable**, since each horizontal row represents one shape group. Similarly, the Type of the cell is called the **column variable**.

### NOTE

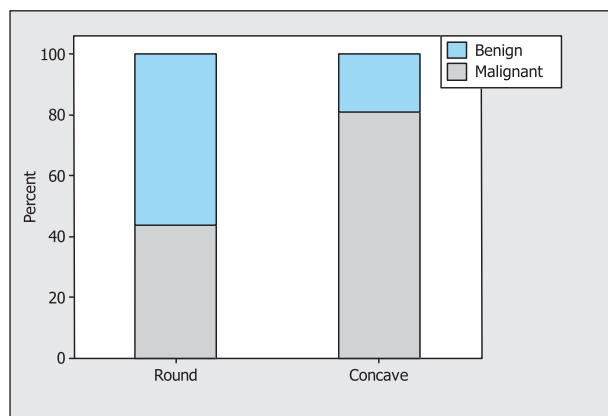
Categorical data that have a natural ordering, such as level of agreement (strongly disagree, disagree, indifferent, agree, strongly agree) or evaluation of a product (poor, fair, good, or excellent), are called **ordinal data**. Categorical data that do not have a natural ordering, such as gender or major, are called **nominal data**. Techniques for nominal data will give identical results for any ordering of the categories, whereas results based on techniques for ordinal data do depend on the ordering of the data. This chapter is restricted to examples of nominal data analysis techniques.

### Key Concept

Analysis of categorical response variables requires calculating the proportion of responses in each category rather than calculating means.

## Activity ▶ Descriptive Statistics and Graphs

1. Identify the observational units, the explanatory variable, and the response variable in the cancer cell data in Table 6.1.
2. Calculate the proportion of round cell samples that are malignant and the proportion of concave cell samples that are malignant.
3. Create a segmented bar graph using Table 6.1. Typically, the explanatory variable should be along the horizontal axis. Assuming this is a random sample from a larger population, does the graph show evidence that nucleus shape is related to the likelihood of a cell being malignant? Explain.



**Figure 6.2** Segmented bar graph of nucleus shape and malignancy.

Bar graphs are often useful in comparing two categorical variables. Figure 6.2 shows a **segmented bar graph** (also called a **stacked bar graph**) for the cancer cell data. This graph shows the conditional percentages for each nucleus shape. About 80% of the concave nuclei are malignant, whereas about 45% of the round nuclei are malignant.

## 6.3 A Simulation Study: How Likely Is It That the Observed Sample Would Occur by Chance?

Figure 6.2 shows that our sample of 37 slides indicates a relationship between the shape of the cell nuclei and malignancy. However, statistical inference is needed to draw conclusions about the entire population from which these samples were selected. In this section, we will conduct a hypothesis test to determine if the sample data provide evidence that the proportion of malignant cells is greater for concave nuclei than for round nuclei.

The hypothesis test in this example is one-sided because we have a medical reason to suspect that concave nuclei are more likely to be malignant. For this example, the null and alternative hypotheses can be written as

$$H_0: p_C = p_R \text{ vs. } H_a: p_C > p_R \quad (6.1)$$

where  $p_C$  is the true proportion of concave nuclei that are malignant and  $p_R$  is the true proportion of round nuclei that are malignant.

If the null hypothesis,  $H_0$ , is true, the two populations (of concave and round cells) have the same proportion of malignant cells and the observed difference between round and concave cells in our sample is due simply to the random sampling process. In other words, the samples just randomly happened to have more malignant cells in the concave nucleus population than in the round nucleus group.

The **p-value** for this test is the probability of obtaining a difference in sample proportions ( $\hat{p}_C - \hat{p}_R$ ) as large as or larger than the one observed in this sample when the null hypothesis is true. If the p-value is small, it is unlikely that the null hypothesis is true and we conclude that the alternative hypothesis ( $p_C > p_R$ ) is true.

One way to estimate this p-value is to simulate taking samples many times under the following three conditions:

- Assume that malignancy is unrelated to cell nucleus shape (i.e., assume that both cell nucleus shapes have the same proportion of malignant cells).
- A total of 13 benign and 24 malignant cells were observed.
- A total of 16 round cell nuclei and 21 concave cell nuclei were observed.

**Activity****Conducting a Simulation Study with Cards**

-  4. Use 37 index cards to represent this sample of 37 cancer cells. On 24 of the cards write M for “malignant,” and on 13 of the cards write B for “benign.” Shuffle the cards and randomly select 21 cards. These 21 cards can represent the concave nucleus group. How many of the 21 concave cards are also malignant?
5. Repeat the simulation process in Question 4 nine more times. Does it seem likely that 17 or more malignant cells would occur in the concave group by chance alone?

While simulations can be done by hand with cards, this process is very time consuming, as a large number of simulations are needed to get a true feel for the likelihood of an outcome (many statisticians suggest 10,000 simulations). Instead of repeating the above process 10,000 times by hand, we will use a computer program to conduct a simulation.

**Activity****Computer Simulation**

-  6. Use the technology instructions provided on the CD to conduct one simulation. In this simulation, you should have one column listing 24 malignant and 13 benign cells. Randomly select 21 rows to represent the concave nucleus shapes. Count the number of observations that fall into the concave malignant group.
7. Repeat the computer simulation process in Question 6 nine more times, each time recording the number of malignant cells you sampled in the concave nucleus group. How does this simulation compare to your index card simulation? Would you expect to get exactly the same number of samples with 17 or more malignant cells? Why or why not?
8. Use the software instructions to repeat the computer-simulated randomization process a total of 10,000 times. Create a histogram of the 10,000 simulated counts in the concave malignant group. Estimate the *p*-value by dividing the number of counts greater than or equal to 17 by 10,000.

Table 6.2 shows 10,000 simulated trials, based on Question 8. This simulation had 220 observations greater than or equal to 17, providing a *p*-value of  $P(X \geq 17) = 0.022$ , where  $X$  is the number of concave malignant cells. Thus, we can conclude that if the null hypothesis were true (the proportion of malignant cells was the same for cells with round and concave nuclei), the likelihood of finding 17 or more malignant cells out of the 21 cells with concave nuclei would be approximately 220 out of 10,000. This small *p*-value shows that the difference in our sample proportions is so large that it is unlikely to have occurred by chance. Thus, we reject the null hypothesis and conclude that cells with concave nuclei are more likely to be malignant than cells with round nuclei.

Using simulations to approximate *p*-values has many advantages. Often simple programs or macros can be written to quickly simulate thousands of samples. Computer programs can be modified to fit a variety of situations, while parametric tests with theoretical assumptions can be somewhat rigid and require large sample sizes. Simulations only provide approximate *p*-values. However, increasing the number of simulations improves the precision of the *p*-value; 10,000 simulations usually provides precise *p*-values.

**NOTE**

This simulation study is an example of a permutation test. Chapter 1 describes that *permutation hypothesis tests* are significance tests that simulate the act of randomly rearranging units into groups.

**Table 6.2** 10,000 simulated trials based on Question 8. The *p*-value is  $P(X \geq 17) = 0.022 = (2 + 21 + 197)/10,000$ , where  $X$  is the number of concave malignant cells.

Concave Malignant Cells	0	...	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Observed Number of Trials	0	...	0	1	8	118	551	1508	2483	2655	1734	722	197	21	2	0	0

**Key Concept**

Simulation studies are gaining in popularity because computer simulations can quickly and easily find accurate  $p$ -values. In addition, unlike most tests that have distributional assumptions, simulation studies do not have minimum sample size requirements and are often more accurate than distribution-based tests for studies with small sample sizes.

## 6.4 Fisher's Exact Test

While the use of simulations to determine  $p$ -values is quickly gaining in popularity, sometimes the exact  $p$ -value can be calculated based on an appropriate probability model. **Fisher's exact test** uses the **hypergeometric distribution** to calculate exact probabilities.

Just as in the simulation study, we are interested in testing  $H_0: p_C = p_R$  vs.  $H_a: p_C > p_R$ . In addition, we use the same **three assumptions** to find the  $p$ -value, the probability that 17 or more of the malignant cells occur in the concave nucleus group when  $H_0$  is true. Table 6.3 provides the same data as Table 6.1, but notation has been included to extend this test to any  $2 \times 2$  table.

In any  $2 \times 2$  table, there are a total of  $N$  observations that can be classified as either a success or a failure.  $M$  represents the **number of success**, and thus  $N - M$  is the number of failures. We are interesting in finding the probability of observing  $x$  successes in a random selection of  $n$  observations. For the cancer cell data in

**Table 6.3** The numbers of benign and malignant tumors for round and concave cell nuclei.

		Type		Total
		Benign	Malignant	
Shape	Round	9	7	16
	Concave	4	17 = $x$	21 = $n$
	Total	13	24 = $M$	37 = $N$

Table 6.3,  $x$  represents the event of interest (observing 17 concave malignant cells),  $N = 37$  represents the total number of observations,  $M = 24$  is the total number of malignant cells (number of successes), and  $n$  is the total number of observations in the concave group.

### MATHEMATICAL NOTE

In this section, we simply discuss how to conduct Fisher's exact test with statistical software. The extended activities shows that Fisher's exact test uses the hypergeometric distribution to calculate exact  $p$ -values. For any  $2 \times 2$  contingency table when there are  $N$  total observations with  $M$  total successes, the probability of observing  $x$  successes,  $P(X = x)$ , in a sample of size  $n$  is

$$\frac{\text{number of ways to select } x \text{ successes and } n - x \text{ failures}}{\text{number of ways to select } n \text{ subjects}} = \frac{\binom{M}{x} \binom{N - M}{n - x}}{\binom{N}{n}}$$

where  $\binom{M}{x} = M \text{ "choose" } x = \frac{M!}{x!(M - x)!}$ .\* Similar calculations hold for  $\binom{N - M}{n - x}$  and  $\binom{N}{n}$ .

\*For any positive integer, the notation  $n!$  is read “ $n$  factorial” and is defined as  $n! = n(n - 1)(n - 2) \cdots (3)(2)(1)$ . For example, “3 factorial” is  $3 \times 2 \times 1 = 6$  and “four factorial” is  $4! = 4 \times 3 \times 2 \times 1 = 24$ . In addition,  $0! = 1$ .

## Activity Calculating Fisher's Exact Test

9. In this cancer study, assume  $N = 37$  observations with  $M = 24$  successes. If  $n = 21$  observations are selected, use the technology instructions provided on the CD to calculate the exact probabilities  $P(X = 17)$ ,  $P(X = 18)$ ,  $P(X = 19)$ ,  $P(X = 20)$ , and  $P(X = 21)$ .
10. Assuming  $N = 37$ ,  $M = 24$  successes, and  $n = 21$ , create a histogram of the probabilities for  $x$ . Compare the probabilities in Question 9 to the probabilities in Question 8. Since Question 8 was a simulation of the hypergeometric distribution, these histograms should look very similar.
11. What is the exact  $p$ -value  $P(X \geq 17)$ ? How does this exact  $p$ -value compare to the simulated  $p$ -value?
12. There is nothing special about how a success or failure is defined. Assume we have a table of  $N = 37$  observations with  $M = 13$  successes (here a benign cell is considered a success). For a sample of size 16 (round nuclei), find  $P(X \geq 9)$ . How does this answer compare to your answer in Question 11?
13. Assume we have a table of  $N = 37$  observations with  $M = 13$  successes (here a benign cell is considered a success). For a sample of size 21 (concave nuclei), find  $P(X \leq 4)$ . How does this answer compare to your answer in Question 11?

Both the simulation in Section 6.3 and Fisher's exact test can be considered permutation tests. The simulation study provides an approximation to Fisher's exact test. Fisher's exact test provides a  $p$ -value for this problem of  $P(X \geq 17) = 0.0225$ . If there truly were no difference between the likelihood of malignancy for the two nucleus shapes (which would mean the null hypothesis  $H_0$  was true), random sampling would produce this outcome (17 or more malignant cells in the concave group) 2.25% of the time. This small probability provides evidence that  $H_0$  should be rejected.

### MATHEMATICAL NOTE

Fisher's exact test and the corresponding simulation study were derived here using a rather strong assumption about the null hypothesis. Both tests were completed under the assumption that we had observed 13 benign and 24 malignant cells and that these totals would be the same in every randomization. Statisticians call this a **conditional test of independence**. In other words, both the row and the column totals are known (fixed) before the study is conducted. However, the extended activities will show that Fisher's exact test can be used for any  $2 \times 2$  table, even when the margin totals are not fixed.<sup>5</sup>

### Key Concept

Fisher's exact test uses the hypergeometric distribution to provide exact  $p$ -values even for small sample sizes and success/failure proportions near 0% or 100%.

## 6.5 Two-Sided Hypothesis Tests

In Fisher's exact test and the simulation study, 17 or more concave malignant cells corresponded to a difference in sample proportions of  $\hat{p}_C - \hat{p}_R = 0.8095 - 0.4375 = 0.372$  or more. The  $p$ -value was the probability of calculating a sample statistic *greater than or equal to*  $\hat{p}_C - \hat{p}_R = 0.372$  given that  $p_C = p_R$ .

Before this sample was collected, the researchers had medical reasons to believe that the cells with concave nuclei (i.e., the malformed nuclei) might be more likely to be malignant. If there had been no specific reasoning to justify a one-sided hypothesis, a two-sided hypothesis test would have been more appropriate:

$$H_0: p_C = p_R \text{ vs. } H_a: p_C \neq p_R \quad (6.2)$$

The  $p$ -value for a two-sided hypothesis is the probability of calculating a sample statistic *at least as extreme as*  $\hat{p}_C - \hat{p}_R = 0.372$  given that  $p_C = p_R$ . That is, if the null hypothesis is true, what is the probability that  $\hat{p}_C - \hat{p}_R \geq 0.372$  or  $\hat{p}_C - \hat{p}_R \leq -0.372$ .

Question 14 helps to show that the difference in proportions,  $\hat{p}_C - \hat{p}_R \leq -0.372$ , corresponds to 10 or fewer concave malignant cells. Table 6.2 can then be used to calculate the approximate  $p$ -value for the two-sided hypothesis test:

$$\begin{aligned}
 p\text{-value} &= P(\hat{p}_C - \hat{p}_R \leq -0.372) + P(\hat{p}_C - \hat{p}_R \geq 0.372) \\
 \text{10 ??????} &= P(X \leq 10) + P(X \geq 17) \\
 &= \frac{1 + 8 + 118 + 197 + 21 + 2}{10,000} \\
 &= 0.0347
 \end{aligned}$$

Based on this simulation, the two-sided hypothesis test should reject the null hypothesis and conclude that  $p_C \neq p_R$ .

## Activity Calculating a Two-Sided Hypothesis Test

14. The three conditions in Section 6.3 stay the same for both the one-sided and the two-sided tests. The total number of malignant cells must be 24, the number of concave cells is 21, and the number of round cells is 16. Thus, when  $Y =$  the number of concave malignant cells,  $\hat{p}_C - \hat{p}_R = \frac{Y}{21} - \frac{24 - Y}{16}$ .
  - a. Find  $\hat{p}_C - \hat{p}_R$  when  $Y$  is 17.
  - b. Find  $\hat{p}_C - \hat{p}_R$  when  $Y = 9$ ,  $Y = 10$ , and  $Y = 11$ .

Notice that the two-sided test is not completely balanced; there is no count,  $Y$ , that exactly corresponds to  $\hat{p}_C - \hat{p}_R = -0.372$ . However, any  $Y \leq 10$  will satisfy  $\hat{p}_C - \hat{p}_R \leq -0.372$ .
15. Repeat Question 8 to estimate the  $p$ -value for the two-sided hypothesis test.
16. Use the **hypergeometric distribution** to find the  $p$ -value for the two-sided hypothesis test.

### NOTE

Two-sided tests can be somewhat cumbersome. Many statisticians suggest simply approximating a two-sided  $p$ -value by doubling the one-sided  $p$ -value. For example, the  $p$ -value for  $H_0: p_C = p_R$  vs.  $H_a: p_C \neq p_R$  is  $2 \times P(X \geq 17) = 2(0.0225) = 0.045$ .

## 6.6 The Chi-Square Test

Fisher's exact test has the advantage of providing exact  $p$ -values. Before technology was available to provide quick and easy ways to conduct Fisher's exact test and simulation studies, the chi-square test was typically used.

You may recall from previous statistics courses that the chi-square test requires that certain assumptions be met before any analysis is done. For example, large sample sizes are needed, especially if the proportion of successes in either group is close to 0% or 100%. If the sample size is large enough, the distribution of the chi-square test statistic will resemble the chi-square distribution.

### Key Concept

The steps to conduct a chi-square test are similar to those for the hypothesis tests discussed in most introductory statistics classes.\* To conduct a chi-square test you will need to do the following:

- State the null and alternative hypotheses.
- Calculate the test statistic.
- Calculate the  $p$ -value.
- Check model assumptions.
- Draw conclusions within the context of the study.

\*For  $2 \times 2$  tables, the chi-square test statistic is identical to the square of the  $Z$ -statistic when testing for equal population proportions. In addition, the  $p$ -values for the two tests will be identical.

**Step 1: State the null and alternative hypotheses:**

The null and alternative hypotheses can be written in exactly the same terms as for the two-sided permutation test:

$$H_0: p_C = p_R \text{ vs. } H_a: p_C \neq p_R$$

The null hypothesis is equivalent to stating that the distributions (of malignant cells) are the same across all groups of nucleus shapes. This test is often called a **test of homogeneity**.

**MATHEMATICAL NOTE**

Some texts suggest using the **chi-square test for homogeneity** if separate samples are selected from more than one population and using the **chi-square test for independence** when data are collected from a single sample. Even though this study is based on one sample of 37 slides, the extended activities will show that both the test of homogeneity and the test of independence are appropriate. The calculations for both tests are identical; however, the hypotheses and conclusions are different. We choose to present the test of homogeneity in this section so that the hypotheses and conclusions coincide with those for the two-sided permutation tests.

**Step 2: Calculate the test statistic:**

The chi-square test statistic is a measure of the difference between the observed counts in Table 6.1 and the expected counts that are calculated under the assumption that the null hypothesis is true (shown in Table 6.4).

In this study, 7 out of the 16 round nuclei were malignant ( $\hat{p}_R = 7/16 = 0.4375$ ) and 17 out of the 21 concave nuclei were malignant ( $\hat{p}_C = 17/21 = 0.8095$ ). Assuming that the proportion of malignant cells is the same for both nucleus shapes, the best estimate of the overall proportion of cells that are malignant is  $(7 + 17)/(16 + 21) = 24/37 = 0.64865$ .

Each expected cell count is calculated by multiplying the estimated proportion times the appropriate sample size (the total number of concave nuclei or total number of round nuclei). There are 16 round nuclei in our study, so the expected count of round malignant cell nuclei is  $16(24/37) = 10.38$ . In general, the expected counts are calculated as row total  $\times$  column total/overall total.

## Activity ◀ Calculating Expected Counts

- 17. Calculate the expected count of concave malignant cell nuclei.
- 18. Assuming the null hypothesis is true, what is the estimated proportion of benign cells in the population? If you select a sample with 16 round cells, what is the expected count of round benign cells?

Table 6.4 shows a table of expected counts under the assumption that the null hypothesis is true: The different cell nucleus shapes have the same proportion of malignant cells.

**Table 6.4** Table of expected counts for the cancer cell study.

	Benign	Malignant	Total
Round	5.62	10.38	16
Concave	7.38	13.62	21
Total	13	24	37

Where

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

For example,

$$10.38 = \frac{16 \times 24}{37} \text{ and } 5.62 = \frac{16 \times 13}{37}.$$

**NOTE**

In a chi-square test, the expected counts are calculated from the totals of the observed data. Totals in the table of observed counts (like Table 6.1) and the table of expected values (like Table 6.4) are identical, except for possible round-off error.

The chi-square test statistic is calculated to measure if the observed data are consistent with the null hypothesis. The chi-square statistic is

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \quad (6.3)$$

This statistic is the sum of each squared difference between the observed count and the expected count, weighted by the expected count. The *observed count* represents the observed cell count from Table 6.1, and the *expected count* represents the expected cell count from Table 6.4.

The chi-square statistic for the cancer cell study is

$$\begin{aligned} \chi^2 &= \frac{(9 - 5.62)^2}{5.62} + \frac{(7 - 10.38)^2}{10.38} + \frac{(4 - 7.38)^2}{7.38} + \frac{(17 - 13.62)^2}{13.62} \\ &= 2.03 + 1.10 + 1.55 + 0.84 = 5.52 \end{aligned}$$

The chi-square test statistic is always positive. If the observed and expected counts are identical, then the test statistic  $\chi^2 = 0$ . If the observed data are far from the expected data, the test statistic will be large and the null hypothesis will be rejected. The chi-square test is not used to show that one proportion is greater or less than another proportion, but to show that the proportions are **not equal**. Thus, the chi-square test is a **two-sided test**.

**NOTE**

The chi-square test can be extended to categorical data with more than two rows and more than two columns. Notice that the table of expected values and the chi-square test statistic in Equation (6.3) can be calculated with additional rows and columns.

### **Step 3: Calculate the p-value:**

The *p*-value will help us determine if a test statistic of  $\chi^2 = 5.52$  or larger is likely to occur by chance. If the null hypothesis is true, the  $\chi^2$  test statistic will follow a chi-square distribution where the degrees of freedom are calculated as (number of rows – 1) × (number of columns – 1). In the cancer cell study, there are two rows of data (round and concave) and two columns of data (benign and malignant); thus, the test statistic has  $(2 - 1) \times (2 - 1) = 1$  degree of freedom.

## Activity Conducting a Chi-Square Test

19. Use a statistical software package to conduct a chi-square test for the cancer cell data.
- Submit the computer output showing the observed and expected tables, the test statistic, and the *p*-value.
  - Assuming the data were a simple random sample from a larger population, what conclusions can you draw about the population?
  - Can you conclude that the shape of the cell nucleus *causes* a change in the likelihood that a cancer cell is malignant? Why or why not?

**Step 4: Check model assumptions:**

Just like most hypothesis tests, the chi-square test requires that each observation be independent. In addition, the chi-square test requires a large enough sample size in each of the cells to ensure that the test statistic can be accurately represented by the chi-square distribution. Some statisticians have slightly different technical assumptions for the sample size needed, but here are two general rules:

- For  $2 \times 2$  contingency tables, the sample size should be large enough that the expected count in each of the  $2 \times 2$  cells in the table is at least 5.
- For tables with more than two rows or two columns, all expected counts should be greater than 1 and the average expected count should be greater than or equal to 5.

In the cancer cell study, all the expected counts are greater than 5, so it is appropriate to use the chi-square test.

**NOTE**

In this example, some of the expected counts are just slightly greater than 5. In this case, some texts might suggest using a chi-square test with a continuity correction. However, simulation studies and Fisher's exact test can now be easily calculated with computers and provide more accurate  $p$ -values. Thus, there really is no longer a need to use a chi-square continuity correction to estimate the  $p$ -value.

**Step 5: Draw conclusions within the context of the study:**

The chi-square test provides a  $p$ -value of 0.019 for the cancer cell study. This indicates that we should reject the null hypothesis in favor of the alternative. Thus, the chi-square test leads to the same conclusion as the two-sided simulated permutation test and the two-sided Fisher's exact test: Each nucleus shape has a different proportion of malignant cells.

The  $p$ -value of 0.019 for the chi-square test is somewhat close to the result of the simulation study conducted in Section 6.3, where we found a two-sided  $p$ -value of 0.0347. Every person conducting a chi-square test on the data in Table 6.1 should get the same  $p$ -value, while each simulation study will provide a slightly different  $p$ -value. Many students mistakenly assume that the variation in the  $p$ -value for the simulation study indicates that it is less accurate than the  $p$ -value for the chi-square test. But, in fact, the simulation study is more accurate than the chi-square test. Fisher's exact test shows that the two-sided hypothesis test for the cancer cell study has an exact  $p$ -value of 0.0357. With larger sample sizes, the  $p$ -values for chi-square tests will be closer to the exact  $p$ -values.

**Activity**  **Simulating the Chi-Square Test Statistic**

20. **Degrees of Freedom** Create a  $2 \times 2$  contingency table with the same totals in the margins as in Table 6.1. Assume you counted 16 concave malignant nuclei in Question 6. Fill in the rest of the table cells. Note that you can complete the other three table counts (the concave benign, round malignant, and round benign) with just one known count value. Thus, only one table cell count is truly free—once one cell is determined, the other three are fixed. This demonstrates why  $2 \times 2$  contingency tables have only 1 degree of freedom.
21. **Degrees of Freedom** Create a  $3 \times 2$  contingency table with row totals of 25, 30, and 25 and column totals of 30 and 50. How many table cell counts are truly free (i.e., what is the smallest number of table cell counts that, when filled in, will completely specify the rest of the cells)? Describe how the formula  $(\text{number of rows} - 1) \times (\text{number of columns} - 1)$  relates to the number of free cells in a two-way contingency table of any size. It may be helpful to create a  $3 \times 3$  table or a  $3 \times 4$  table to convince yourself that this rule continues to hold.

## 6.7 What Can We Conclude from the Cancer Study?

The cancer cell data in Table 6.1 were analyzed with a simulation study (more specifically, a permutation test), a chi-square test, and Fisher's exact test. All tests suggested that the null hypothesis should be rejected, and thus we conclude that concave and round cell nuclei are associated with different proportions of malignant cells.

This study was not an experiment, since there was no random allocation of units to particular conditions. Thus, while we expect that there is an association between nucleus shape and malignancy, we cannot conclude that nucleus shape *causes* different proportions of malignancy.

The individuals who provided these 37 slide samples were not a true random sample of all North Americans, but a sample of patients who volunteered to be part of the study at the University of Wisconsin hospitals. Subjects in medical studies are rarely true random samples from the general population. We cannot be certain that these results hold for a larger population of people. However, it seems reasonable for researchers to believe that cancer cells from patients in Wisconsin are similar to cancer cells from other patients in other hospitals. Thus, we can cautiously conclude that this study provides some evidence that the different nucleus shapes are associated with different proportions of malignant cells.

### A Closer Look Contingency Tables

## 6.8 Relative Risk and the Odds Ratio

The **proportion** of malignant nuclei in this study is  $24/37 = 0.6486$ . This overall proportion of malignancy is often called the **risk** (or baseline risk) of malignancy. The **conditional proportion** is the proportion of malignant cells calculated for each cell shape. Thus, the round nuclei have a conditional proportion (of malignancy)  $= 7/16 = 0.4375$ , and the concave cells have a conditional proportion (of malignancy)  $= 17/21 = 0.8095$ .

The hypothesis tests in this chapter have focused on determining whether concave and round nuclei have the same proportion of malignancy. However, it is important to recognize that there are limitations in testing for differences in proportions. For example, assume that we have two studies. In the first study,

$$\hat{p}_1 - \hat{p}_2 = 0.52 - 0.48 = 0.04$$

In the second study,

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.01 = 0.04$$

Both studies could test whether an observed difference in proportions of 0.04 is significant. However, in the second study  $\hat{p}_1$  is five times larger than  $\hat{p}_2$ .

An alternative calculation that is commonly used for data with categorical response and explanatory variables is the relative risk. In the following calculations, we arbitrarily decided to define a malignant cell as a success.

$$\text{Relative risk} = \frac{\text{proportion of successes in group 1}}{\text{proportion of successes in group 2}} = \frac{\hat{p}_1}{\hat{p}_2} \quad (6.4)$$

In the cancer cell study, the relative risk is  $0.8095/0.4375 = 1.85$ . Thus, the risk of malignancy is 1.85 times greater for the concave group than for the round group.

The **odds** (more specifically, odds of success) can also be used to compare proportions and tend to have meaning over a broader range of potential outcomes.

$$\text{Odds} = \frac{\text{number of successes}}{\text{number of failures}} \quad (6.5)$$

The odds of malignancy in the concave group are 17 to 4, meaning that we expect 17 successes (malignant cells) for every 4 failures (benign cells). This is often stated as follows: The odds of malignancy in the concave group are  $4.25 (17 \div 4)$  to 1 ( $4 \div 4$ ). The odds of malignancy in the round group are 7 to 9. The **odds ratio** is used to compare the odds of two groups.

$$\text{Odds ratio} = \frac{\text{odds of group 1}}{\text{odds of group 2}} = \frac{\hat{\theta}_1}{\hat{\theta}_2} \quad (6.6)$$

The odds ratio in the cancer cell study is  $\frac{17/4}{7/9} = 5.5$ . Thus, the odds of malignancy are 5.5 times greater for the concave group than for the round group. When the odds ratio = 1, the odds for both groups are equal.

#### → NOTE →

In relative risk and odds ratio calculations, the group that has the lower proportion (or lower odds) is typically considered group 2 (in the denominator). That way, the relative risk and odds ratio are always numbers larger than one and easier to interpret.

#### Key Concept

In studies where the proportions are far away from 0.5, the hypothesis test for the difference in proportions may not best represent your question of interest. When both conditional proportions are small, the relative risk and odds ratio are preferable, since they take the base line rate (overall proportion of successes) into account.

## Extended Activity

### ▶ Calculating Additional Summary Statistics

Data set: Table 6.1

22. Use the data from Table 6.1 and define a benign cell as a success and round cells to be group 1. Calculate and interpret the relative risk and the odds ratio.
23. Show that the null hypothesis  $H_0: p_1 = p_2$  is mathematically equivalent to the null hypothesis  $H_0: \theta_1/\theta_2 = 1$ , where  $p$  represents the proportion successful and  $\theta$  represents the odds of success for any two groups (labeled 1 and 2).
24. **Shortcut for Calculating the Odds Ratio** Use the counts in Table 6.1 to calculate the following:

$$\frac{(\text{count of round benign})(\text{count of concave malignant})}{(\text{count of round malignant})(\text{count of concave benign})}$$

Does this calculation match any statistic in Question 22? This product of diagonals can always be used as a shortcut to calculate the odds ratio.

#### → NOTE →

The odds ratio does not depend on the choice of success or failure. In addition, the odds ratio provides identical results when the explanatory and response variables are switched.

## Cautions About Relative Risk Reduction in Medical Studies

Zocor is a drug used to lower cholesterol in order to reduce the chances of a heart attack. A five-year study was conducted to investigate the effectiveness of Zocor, using 4444 people.<sup>6</sup> People in this study were aged 35–70 and had a high risk of heart attack. In addition, all the subjects were Caucasian and 81% were males. Based on this study, television and print advertisements stated, “A clinical study among people with high cholesterol and heart disease found 41% fewer deaths from heart attack among those taking Zocor.”<sup>7</sup>

### Extended Activity

#### Comparing Relative and Absolute Risk Reduction

25. Does the advertisement claim that the study shows a 41% reduction in heart attacks among all people that use Zocor? Explain.
26. What does 41% fewer deaths mean in terms of the chances of having a heart attack?  
(Select one.)
  - a. 2222 died from heart attacks in the placebo group, and 41% fewer (1289 out of 2222) died from heart attacks in the treatment group.
  - b. 1000 died from heart attacks in the placebo group, and 41% fewer (580) died from heart attacks in the treatment group.
  - c. 100 died from heart attacks in the placebo group, and 41% fewer (58) died from heart attacks in the treatment group.
  - d. It is impossible to tell based on the quote from the advertisement.
27. The actual study found that 189 out of 2223 in the placebo group died from heart attacks and 111 out of 2221 in the treatment (Zocor) group died from heart attacks.
  - a. Use the Zocor study data to create a two-way table. Use Placebo and Treatment (Zocor) as the row variables. Use Death and Survival as the column variables.
  - b. Calculate the percentage of deaths in the placebo group and the percentage of deaths in the treatment group.
  - c. Calculate the difference between the two percentages calculated in Part B.
  - d. Calculate and interpret the relative risk of death from a heart attack.

The difference between the two percentages calculated in Question 27b,  $8.5\% - 5\% = 3.5\%$ , is called the **absolute risk reduction**. The 41% fewer deaths is actually calculated as

$$\frac{8.5\% - 5\%}{8.5\%} \approx 41\%$$

The statistic 41% is called the **relative risk reduction**. While both statistics, 3.5% and 41%, are appropriate, it is important to recognize how easily these numbers can be misunderstood. The advertisement states things so as to make the reduction in deaths due to heart attacks appear greater than it is in absolute terms: a reduction of 8.5% to 5% for the risk of death from heart attack for the next five years for a restricted sample of people with heart disease.

## 6.9 Sampling Designs

Contingency tables are widely used and easy to interpret. However, it is important to recognize that the appropriate statistical analysis cannot be determined simply by looking at a table of data; it is determined by how the data were collected. While there are numerous ways to collect data, this section will focus on three key sampling designs often used in observational studies.

1. In **cross-classification studies**, information is collected simultaneously on both variables in the study. This often occurs when all data are collected from one sample and then placed into a classification

table. The row totals and the column totals are not known prior to the data collection. The total sample size,  $N$ , may or may not be known before the data are collected.

2. In **cohort studies**, individuals (or units) who differ with respect to a certain explanatory variable are selected (or assigned to groups) and then a response variable is measured. These predetermined groups are called cohorts, and if the response variable is measured over time the design is called a **prospective design**.\* In cohort studies, the totals corresponding to the explanatory variable are known before the responses are collected.
3. In **case-control studies**, individuals (or units) are selected according to a response variable (and often called the cases and the controls). Then the individuals are classified according to some explanatory variable. Case-control studies are often retrospective studies, since historical data are typically used to collect information on the explanatory variable. In case-control studies, the totals corresponding to the response variable are known before data are collected on the explanatory variable.

Prospective studies typically provide stronger evidence of a relationship between the explanatory variable and the response variable. However, they tend to be expensive, to require a long time to gather data, and to be sensitive to attrition. When responses can be rare, such as getting lung cancer, case-control studies are preferred over cohort studies because case-control studies can ensure a large enough sample size within each group of responses. However, when data are selected based on the response variable, such as in case-control studies, studies can be particularly susceptible to difficulties with confounding.

## Extended Activity

### ▶ Retrospective Studies and the Odds Ratio

In a retrospective study of lung cancer patients in 20 London hospitals, Richard Doll identified a relationship between smoking and lung cancer.<sup>8</sup> Table 6.5 shows a partial set of data from his study. The data were collected on 60 female patients with lung cancer and 60 control females.

**Table 6.5** Data from Richard Doll's case-control study on smoking and lung cancer.

Females	Have Lung Cancer		
	Yes	No	Total
Smoker	41	28	69
Nonsmoker	19	32	51
Total	60	60	120

28. What is the proportion of females in this study who have lung cancer? What is the proportion of smokers who got lung cancer? What is the proportion of nonsmokers who got lung cancer? What is the relative risk when having lung cancer is defined as a success?
29. Notice that Doll predetermined the distribution of the response variable (this is done in case-control studies). Explain why each of the proportions in Question 28 cannot appropriately be extended to a larger population. (Hint: Is it appropriate to conclude that  $60/120 = 50\%$  of female patients in the 20 London hospitals have lung cancer? Is it appropriate to assume that a good estimate of the percentage of female nonsmoking patients in the hospitals who have lung cancer is 37.2%?)

Note that tests for equal proportions are not appropriate if the response totals are fixed prior to the study. In addition, the next section will show that since the response totals are not random, a test of independence should not be used. Fisher's exact test and a simulation study could be used. The following questions will show why it is appropriate to conduct a hypothesis test about the odds ratio.

\*Retrospective cohort studies also exist. In these designs, past (medical) records are often used to collect data. As with prospective cohort studies, the objective is to first establish groups based on an explanatory variable. However, since these are past records data on the response variable can be collected at the same time.

30. Calculate the odds of lung cancer for smokers. Calculate the odds of lung cancer for nonsmokers. Calculate and interpret the odds ratio of lung cancer. Does the odds ratio indicate a relationship between smoking and lung cancer?
31. Calculate and interpret the odds ratio of being a smoker. Does the odds ratio depend on what variable is considered the response?

### Key Concept

The statistics used to estimate an overall proportion (and risk) cannot be interpreted or generalized to a larger population in studies where the response totals (i.e., the distribution of the response variable) are fixed by the researcher. Thus, the conditional proportions and relative risk cannot be generalized to a larger population. Since the odds ratio is invariant to the choice of explanatory and response variables, it can be used when the totals of the response variable are controlled by the researcher. For  $2 \times 2$  contingency tables, Fisher's exact test can be used with any sampling design and any sample size.

## Tests for the Homogeneity of Odds

Questions 28–31 showed that when the response variable totals are fixed before the study is conducted, the proportions are not representative of a larger population. Thus, tests for the homogeneity of proportions are not appropriate. This section shows the steps in a test for the homogeneity of odds using the data in Table 6.5.

### **Step 1: State the null and alternative hypotheses:**

$$H_0: \frac{\theta_S}{\theta_N} = 1$$

$$H_a: \frac{\theta_S}{\theta_N} > 1$$

### **Step 2: Calculate the test statistic:**

We calculate the odds of cancer for the smoking group to be

$$\hat{\theta}_S = \frac{\hat{p}_S}{1 - \hat{p}_S} = 1.46$$

Similarly, the odds of cancer for the nonsmoking group are

$$\hat{\theta}_N = \frac{\hat{p}_N}{1 - \hat{p}_N} = 0.594$$

The odds ratio is  $\hat{\theta}_S/\hat{\theta}_N = 2.466$ .

It can be shown (in more advanced texts) that when sample sizes are large, the natural log of the odds ratio is approximately normal with the following standard deviation:

$$S_{LO} = \sqrt{\frac{1}{n_S(\hat{p})(1 - \hat{p})} + \frac{1}{n_N(\hat{p})(1 - \hat{p})}} = \sqrt{\frac{1}{69(0.5)(0.5)} + \frac{1}{51(0.5)(0.5)}} = 0.3693$$

where  $n_S$  and  $n_N$  are the total number of smokers and the total number of nonsmokers in the study, respectively, and  $\hat{p}$  is the overall proportion of people with lung cancer in the study. Also recall that testing whether  $\theta_S/\theta_N = 1$  is equivalent to testing whether  $\ln(\theta_S/\theta_N) = 0$ .

Thus, a test statistic is calculated as

$$Z = \frac{\ln\left(\frac{\hat{\theta}_S}{\hat{\theta}_N}\right) - 0}{S_{LO}} = \frac{0.90266}{0.3693} = 2.44$$

**Step 3: Calculate the p-value:**

$$P(Z > 2.44) = 0.0073$$

**Step 4: Check model assumptions:**

It is reasonable to assume that each patient in the study is independent of the others. A general rule is that as long as the sample size in each cell is greater than or equal to 5, the normality assumption is appropriate.

**Step 5: Draw conclusions within the context of the study:**

In this study, the small *p*-value leads us to reject the null hypothesis and conclude that the odds ratio is greater than one. In other words, if the population odds of cancer were identical for both smokers and nonsmokers, a sample odds ratio as large or larger than 2.466 was very unlikely to occur by random chance. This study is not an experiment; thus, it is inappropriate to conclude that smoking causes larger odds of cancer. This is not a true random sample of patients in the 20 London hospitals. However, it is reasonable to cautiously expect that these patients are representative of all hospital patients in the 20 London hospitals.

## 6.10 Comparing Tests of Homogeneity and Independence

The calculations involved in the tests of homogeneity and independence are identical; however the type of question asked will impact the conclusions that can be drawn.

The **test of homogeneity** is used to determine if proportions are equal across two or more populations. The hypothesis test related to the cancer cell study

$$H_0: p_L = p_H$$

$$H_a: p_L \neq p_H$$

fits this situation. Tests of homogeneity are appropriate whenever one of the variables is clearly defined as the response variable. As shown in the previous section, a test of homogeneity of proportions is not acceptable when the response totals are known in advance. However, a test of the homogeneity of odds is appropriate even if the response totals are fixed.

The **test of independence** is used to determine if two random variables within a population are independent. It is not necessary to determine which variable is the explanatory variable and which is the response variable. If any of the marginal (i.e., row or column) totals are known (fixed) in advance, at least one of the variables is not random, and thus the test of independence is not appropriate. The null and alternative hypotheses can be written as

$H_0$ : the row and column variables are independent (more specifically,  $H_0$ : nucleus shape and malignancy are independent)

$H_a$ : the two variables are not independent

**NOTE**

Tests of homogeneity and independence are not restricted to chi-square tests. For example, the two-proportion *z*-test is also a test of homogeneity.

In both tests, the null hypothesis is that there is no relationship between the row variable and the column variable. Tests of independence are essentially testing if the probability of a success for one variable depends on the value of a second variable. They are appropriate whenever both the row and the column variables are random (e.g., the column totals are not predetermined by the researcher), as in cross-classification studies.

Tests of homogeneity are appropriate whenever it is clear that one variable should be treated as the response and the other as the explanatory variable. Thus, tests of homogeneity can be used with all three study designs listed earlier. However, case-control study designs (where the response totals are fixed) are appropriate only for testing homogeneity of odds (not homogeneity of proportions). Table 6.6 summarizes the appropriate hypothesis tests for each sampling design. Specific examples of each type of sampling design are provided in the end-of-chapter exercises.

**Table 6.6** Sampling designs and appropriate hypothesis tests.

	Sampling Designs				
Marginal Totals	Cross-Classification	Cohort	Case-Control	Hypergeometric Experiment	Randomized Experiment
<b>Total N Fixed</b>	Either is OK	Yes	Yes	Yes	Yes
<b>Response Totals Fixed</b>	No	No	Yes	Yes	No
<b>Explanatory Totals Fixed</b>	No	Yes	No	Yes	Yes
Hypothesis Tests	Cross-Classification	Cohort	Case-Control	Hypergeometric Experiment	Randomized Experiment
<b>Test of Independence</b>	Yes	No	No	No	No
<b>Test of Homogeneity of Proportions</b>	Yes	Yes	No	No	Yes
<b>Test of Homogeneity of Odds</b>	Yes	Yes	Yes	No	Yes
<b>Fisher's Exact Test</b>	Yes	Yes	Yes	Yes	Yes

In the cancer cells study, we concluded that different cell nucleus shapes are associated with different proportions of malignant cells. In this study, the researchers had a clear explanatory variable (`shape`) and response variable (`Type`). Thus, the test of homogeneity is appropriate. However, a test of independence is also appropriate, since each subject was selected from one larger population: patients with suspicious tumors at the University of Wisconsin hospitals. After the patients were selected, the biopsy and FNA were conducted and each observed slide was classified as either malignant or benign and either round or concave.

#### Key Concept

Although chi-square tests homogeneity and independence involve exactly the same mathematical calculations, the conclusions can be different. The choice of test is based on (1) how the sampling was conducted and (2) whether there are clearly defined explanatory and response variables. A test of homogeneity is appropriate whenever the response variable is clearly defined from the beginning of the study. If one sample is selected from one population and each observation is grouped into categories for both variables (i.e., neither explanatory nor response totals are known in advance), the test of independence is appropriate.

While the chi-square test of independence is frequently used, the conclusions from this test are not very informative. The test does not prove that two variables are independent, but only identifies that we have significant evidence that two variables are dependent. No measurement of the degree of independence (or indication of whether this level of dependence is of practical importance) is given.

## 6.11 Chi-Square Goodness-of-Fit Tests

Many statistical techniques are based on specific model assumptions. For example, many procedures discussed throughout this text for calculating hypothesis tests and confidence intervals assume that the error terms follow a normal distribution. If the model assumptions are violated, the tests should not be considered reliable.

Goodness-of-fit tests are used to determine how well our observed data “fit” the model assumptions. In other words, we want to determine how “close” the observed values are to those values that we would expect under a specific (theoretical) distribution.

Before a goodness-of-fit test can be used, we need some prior knowledge (or benchmark values) about a theoretical model. The chi-square goodness-of-fit test is applied to data that are placed into groups. Sample data are placed into classes (observed groups), and a theoretical model is used to calculate the expected number of observations in each group. In this section, we will use goodness-of-fit tests to determine if the observed counts (or proportions) are consistent with hypothetical counts (or proportions).

### Key Concept

In a chi-square goodness-of-fit test with  $n$  observations:

- Each of the  $n$  observations fits into exactly one of  $G$  categories.
- The null hypothesis specifies the expected probability ( $p_i$ ) for each cell.
- Expected counts are calculated as  $n \times p_i$ .
- The test statistic is  $\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$ .
- If each of the expected cell counts is at least 5, the  $p$ -value is calculated from the chi-square distribution with  $G - 1$  degrees of freedom.

## Extended Activity

### Playing a Dice Game

Your friend offers to play a simple game with you. He has a six-sided die and will roll it 30 times. Each time he rolls a 5 or 6, you will pay him \$5. Each time he rolls a 1, 2, 3, or 4, he will pay you \$5. The results of your friend's 30 rolls for this game are shown in Table 6.7.

**Table 6.7** Observed results of 30 rolls of a die.

	1	2	3	4	5	6
Observed	2	5	3	3	8	9

If you had agreed to play this game, you would have owed your friend some money. Table 6.7 is called a one-way table with six cells. We can consider these results a random sample from all possible rolls of this die, and we can assume that each roll is independent. If the die was fair, we would assume (our null hypothesis would be) that each of the six outcomes was equally likely.

32. From visual inspection of the results, do you have any reason to suspect the die wasn't fair? Explain.
33. In this example, there are  $G = 6$  groups (cells). Assuming the die is fair, the probability that a roll results in a 1 is  $p_1 = 1/6$ , the probability that a roll results in a 2 is  $p_2 = 1/6$ , etc. In goodness-of-fit tests, the expected values are found from some hypothesized distribution that determines the probability for each of the  $G$  groups.

In this case, we assume  $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$ . The expected count for each of the 6 groups can be found from  $n \times p_i$ . Assuming all outcomes are equally likely, fill out the expected counts in Table 6.8.

**Table 6.8** Observed and expected results of 30 rolls of a die.

	1	2	3	4	5	6
Observed	2	5	3	3	8	9
Expected						

34. Conduct a chi-square test to determine if there is a significant difference between what was observed and what was expected.
- State the null and alternative hypotheses in words.
  - Calculate the chi-square statistic using Equation (6.3).
  - In goodness-of-fit tests such as this, where all parameters are well defined, the *degrees of freedom is equal to the number of cells minus 1*. In this example, if we know how many 1s, 2s, 3s, 4s, and 5s were rolled, the number of 6s is fixed. Calculate the *p*-value for this study.
  - Are there enough observations to assume the chi-square distribution is appropriate?
  - Do you have evidence to believe your friend used an unfair die? Clearly state your conclusions.



### CAUTION

The chi-square goodness-of-fit test is applied to grouped data (i.e., data put into classes). Continuous data can easily be placed into groups. However, it is important to recognize that the value of the chi-square test statistic depends on how the data are grouped.

## Chapter Summary

This chapter focused on analyzing and drawing conclusions from categorical data. Fisher's exact test, simulation studies, and chi-square tests were used to analyze various data sets. Like most hypothesis tests, each of these tests is valid only if care is taken to ensure that the appropriate assumptions are met. All tests assume that the observations are independent.

**Fisher's exact test** is appropriate for any sample size and any sampling design involving  $2 \times 2$  contingency tables. **Simulation studies** are used to approximate Fisher's exact test. While simulation studies are approximate, conducting 10,000 iterations will typically provide precise *p*-values.

Simulation studies have the advantage of being very adaptable. For example, they can easily be modified to any number of categories within a study. (The research project at the end of the chapter provides an example of a simulation study where the independent observation assumption is violated.)

When sample sizes are large enough, the chi-square test statistic will follow the chi-square distribution. In order for the *p*-values from **chi-square tests** to be appropriate, the following conditions should be satisfied:

- For  $2 \times 2$  contingency tables, the sample size should be large enough that the expected count in each cell of the table is at least 5.
- For tables with more than two rows or two columns, all expected counts should be greater than 1 and the average expected count should be greater than or equal to 5.

Chi-square tests are always two-sided hypothesis tests. However, they can be easily extended to contingency tables with more than two rows and two columns. If there are more than two rows and two columns, the chi-square test does not specifically test the significance of each cell. However, the end-of-chapter exercises show that it is reasonable to look at each cell's contribution to the chi-square statistic.

Three key sampling designs were described in this chapter.

- In a **cross-classified design**, information is collected simultaneously on both variables in the study. The row totals and the column totals are not known prior to the data collection.
- In **cohort studies**, the totals corresponding to the explanatory variable are known before the responses are collected.
- In **case-control studies**, the totals corresponding to the response variable are known before data are collected on the explanatory variable.

Testing the **equivalence of two proportions** ( $H_0: p_1 = p_2$ ) is mathematically equivalent to testing whether the **odds ratio** is equal to one ( $H_0: \theta_1/\theta_2 = 1$ ), where  $p$  represents the proportion successful and  $\theta$  represents the odds of success for any two groups (labeled 1 and 2). However, when the proportions of interest are very small, it may be more appropriate to look at the odds ratio than at the proportions. Tests for equal proportions are not appropriate for case-control studies, because the fixed response totals cause the sample proportions to not accurately represent the population.

Chi-square tests for homogeneity and independence involve exactly the same mathematical calculations; however, the hypotheses and conclusions are different. A **test of homogeneity of proportions** is appropriate whenever the explanatory and response variables are clearly defined from the beginning of the study. While the explanatory totals may be known in advance, the response totals cannot be fixed before the data are collected. **Tests of independence** are appropriate if one sample is selected from one population, and thus neither explanatory nor response totals are known in advance (i.e., both explanatory and response totals can be considered random).

This chapter was limited to studies involving only one categorical explanatory variable and one categorical response variable. Additional tests such as the Mantel-Haenszel procedure can be used to test for equal proportions or odds with two explanatory variables. The following chapters discuss more general procedures, using logistic and Poisson regression to analyze data with a categorical response variable and multiple categorical or quantitative explanatory variables.

## Exercises

---

### E.1. Cancer and Smoking: Fisher's Exact Test and Simulation Studies

Data set: Table 6.5

Answer the following questions for the data displayed in Table 6.5.

- Was either the explanatory (row) or the response (column) variable fixed before the study was conducted?
- Is this an example of an experiment or an observational study?
- Is this a cross-classification, cohort, or case-control study? Explain.
- Create a segmented bar chart for the data.
- Create a simulation study to test the one-sided hypothesis that smokers are more likely to have lung cancer. Provide a  $p$ -value and state your conclusions.
- Use Fisher's exact test to test the one-sided hypothesis that smokers are more likely to have lung cancer. Provide a  $p$ -value and state your conclusions.

### E.2. Statistics Enrollment for Humanities and Science Majors

Data set: Table 6.9

A study was conducted by introductory statistics students to see whether majors in the sciences are as likely to take a statistics course as majors in the humanities. They sampled 50 seniors from among humanities majors and another 50 seniors from among science majors and asked if they had taken a statistics course or were scheduled to take one before they graduated. Their data are provided in Table 6.9.

**Table 6.9** Results from the statistics enrollment study.

	Yes	No
Humanities Major	23	27
Science Major	32	18

- Was either the explanatory (row) or the response (column) variable fixed before the study was conducted?
- Is this an example of an experiment or an observational study?
- Is this a cross-classification, cohort, or case-control study?
- Create a segmented bar chart for the data.
- Create a simulation study to test the one-sided hypothesis that science majors are more likely to take a statistics course. Provide a  $p$ -value. Assuming that students took care to collect a simple random sample of seniors from each division, what conclusions can be drawn?
- Use Fisher's exact test to test the one-sided hypothesis that science majors are more likely to take a statistics course. Provide a  $p$ -value. Assuming that students took care to collect a simple random sample of seniors from each division, what conclusions can be drawn?

### E.3. Statistics Enrollment for Humanities and Science Majors 2

Data set: Table 6.9

- Use Exercise 2 to determine if a chi-square test of homogeneity or a test of independence is appropriate for the data in Table 6.9. Give an explanation for your answer.
- Use the data in Table 6.9 to determine if the same proportion of humanities and science majors take a statistics course. Assuming that students took care to collect a simple random sample of seniors from each division, what conclusions can be drawn?
- Use a simulation study to determine if a difference in proportions at least as large as the one in the sample data is likely to occur by chance (using a two-sided hypothesis test). Assuming that students took care to collect a simple random sample of seniors from each division, what conclusions can be drawn?

### E.4. A Lady Tasting Tea

Data set: Table 6.10

At a summer tea party in the late 1920s, a lady claimed that she could determine by taste if tea had been poured into the milk or if milk had been poured into the tea. Sir Ronald Fisher, well known for his work in design of experiments, discussed how to test this lady's claim. The lady was asked to taste several cups in random order and identify each as tea first or milk first. Although the actual results were not published, rumor has it that the lady correctly identified every single cup.<sup>9</sup>

Let's assume that the lady was given 14 cups of tea in random order. She was told that half the cups had tea first and the other half had milk first. Table 6.10 provides hypothetical results.

**Table 6.10** Results of the hypothetical tea-tasting study.

	Lady Guesses Tea First	Lady Guesses Milk First	Total
Tea First	7	0	7
Milk First	0	7	7
Total	7	7	14

Note that the study was designed with exactly seven tea first and seven milk first cups; thus, the row totals are fixed. In addition, since the lady was told this, it is reasonable to set her answers to have exactly seven tea first guesses and seven milk first guesses. Thus, the column totals are also fixed.

- Use Table 6.6 to determine the name of this type of study design.
- Which type of statistical tests can be used with this study design?
- Conduct a hypothesis test to determine if the lady could correctly identify the order in which items were poured into a cup. Submit a *p*-value and clearly state your conclusions.
- Conduct a hypothesis test to determine if the lady could correctly identify the order in which items were poured into a cup. However, instead of using Table 6.10, assume she made four mistakes: She guessed tea first for two milk first cups, and she guessed milk first for two tea first cups. Based on the new data, submit a *p*-value and clearly state your conclusions.

### E.5. Donner Party

Data set: Table 6.11

Members of the Donner party attempted a new route between Fort Bridger, Wyoming, and the Humboldt River, Nevada. This new route took much longer than expected, and the entire party was trapped in the Sierra Nevada in the winter of 1846–1847. By the time they were rescued in the spring, many members had died. Table 6.11 lists the gender and survival status of all adults who were trapped during this trip.

- Plot the data with a segmented bar graph. Describe the overall patterns seen in the data.
- Is this an example of an experiment or an observational study?
- Is this a cross-classification, cohort, or case-control study?
- Create a simulation study to test the one-sided hypothesis that females were more likely to survive. Provide a *p*-value and clearly state your conclusions.

**Table 6.11** Survival status of Donner party members.

	Survived	Died	Total
Male	23	30	53
Female	25	9	34
Total	48	39	87

- e. Use Fisher's exact test to test the one-sided hypothesis that females were more likely to survive. Provide a  $p$ -value and clearly state your conclusions.
- f. D. K. Grayson stated, "The differential fate of the members of the Donner Party lends strong support to the argument that females are better able than males to withstand conditions marked by famine and extreme cold."<sup>10</sup> Does the small  $p$ -value lead us to conclude that females have stronger survival abilities than males? Can you suggest any other reasonable explanation?

#### E.6. Statistics Enrollment: Testing for Homogeneity of Odds

Data set: Table 6.9

- a. Calculate and interpret the odds ratio for the data in Table 6.9.
- b. Use the data from Table 6.9 and define taking a statistics course (yes) as a success. Conduct a hypothesis test for the homogeneity of odds. State the 1-sided null and alternative hypotheses.
- c. Calculate the test statistic (the  $Z$  statistic).
- d. Provide the  $p$ -value and state your conclusions within the context of the study.

#### E.7. Cancer Cells: Testing for Homogeneity of Odds

Data set: Table 6.1

Use the data from Table 6.1 and define a benign cell as a success. Conduct a hypothesis test for the homogeneity of odds.

- a. State the null and alternative hypotheses.
- b. Calculate the odds ratio and the test statistic (the  $Z$  statistic).
- c. Provide the  $p$ -value and state your conclusions within the context of the study.

#### E.8. Playing a Dice Game

In the extended activity in Section 6.10, you conducted a goodness-of-fit test to determine if the die was fair. Now conduct a simulation study to determine if there is a significant difference between what was observed and what was expected. Simulate rolling 30 dice 10,000 times. Count the number of times you have 17 or more 5s and 6s in the 30 rolls. Calculate a  $p$ -value and explain why it is not surprising that the  $p$ -value is very different than what was observed in Question 34.

#### E.9. Comparing Treatments for Drug Addiction: A $3 \times 2$ Contingency Table

Data set: Table 6.12

Many people with cocaine addictions also suffer from mood disorders like depression or a manic-depressive disorder. To aid in the process of trying to break a patient's addiction to cocaine, a common antidepressant known as lithium is often given to patients in drug treatment centers. A randomized, comparative experiment was performed to determine whether another commonly prescribed antidepressant, disipramine, could be used to reduce the likelihood of patients' using cocaine again or whether antidepressant therapy helped at all.<sup>11</sup> The 72 patients selected for study over a 3-year period were randomly assigned to one of three treatments: disipramine, lithium, or a placebo. After the end of the 6-week treatment period, these patients were interviewed to determine if they had been successful in stopping their cocaine habit. A success was defined as having stopped using cocaine for 3 weeks or more during the 6-week treatment period. The data from the study are shown in Table 6.12.

The null and alternative hypothesis can be written as

$$H_0: p_D = p_L = p_C$$

$H_a$ : at least one treatment group has a different proportion of successes than the others

**Table 6.12** Results from a study of the effect of treatment with different antidepressant drugs (or a placebo) on the success of cocaine addicts in remaining drug-free for at least 3 weeks during a 6-week treatment period.

	Success	Failure	Total
<b>Disipramine (D)</b>	14	10	24
<b>Lithium (L)</b>	6	18	24
<b>Placebo/Control (C)</b>	4	20	24
<b>Total</b>	24	48	72

where  $p_D$ ,  $p_L$ , and  $p_C$  represent the true proportions of successes in the D, L, and C groups, respectively.

- Was either the explanatory (row) or the response (column) variable fixed before the study was conducted?
  - Is this an example of an experiment or an observational study?
  - Use Part A to determine if a chi-square test of homogeneity or a test of independence is appropriate for this study.
  - Create a segmented bar chart for the data.
  - Conduct a chi-square test to determine if this difference between success rates is likely to occur by chance. Calculate a table of expected counts assuming that the null hypothesis is true. (If the treatment group makes no difference in the success or failure of the patient in remaining free from cocaine, we would expect the success rates of all three groups to be the same.)
  - Are the sample sizes large enough to assume that the chi-square distribution is appropriate? Explain.
  - What are the test statistic, the degrees of freedom, and the  $p$ -value?
  - Can we conclude that for this group of 72 people the treatment did cause a difference in success rates?
  - These 72 people were not randomly selected from a larger population. Can we conclude that these results hold for a larger group of people?
- E.10. **Friday Night Collisions: Drawing Conclusions from a Goodness-of-Fit Test**

Data set: Table 6.13

A local organization conducted a traffic study to determine if traffic collisions were more likely on Friday and Saturday nights than on other nights. The city council was considering a proposal that bars in a specific district be closed earlier in an effort to reduce late-night accidents in the area. The assumption was that a larger number of traffic accidents would correspond to the heaviest bar traffic (which occurred on the weekends). Table 6.13 shows data collected on traffic accidents that had occurred in that district over the past two years.

**Table 6.13** Results from the collision study.

	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
<b>Observed Collisions</b>	5	7	3	11	8	16	12

- Conduct a chi-square test to determine if the proportion of accidents is the same each day of the week. Write the null and alternative hypotheses and clearly state your conclusions. Do we have evidence to show that weekends are more likely to have accidents than other days?
- Which day contributes most to the chi-square statistic? In other words, of the seven days, which day has the largest  $(\text{observed} - \text{expected})^2/\text{expected}$  value? Note that taking the square root of this value has some similarities to calculating standardized residuals.

- c. Another way to conduct the study would be to start with the null hypothesis that weekends are twice as likely to have accidents as weekdays. Thus, the null hypothesis states that  $1/9$  of the accidents occur on each weekday,  $2/9$  of the accidents occur on Friday, and  $2/9$  of the accidents occur on Saturday. Conduct a chi-square test and clearly state your conclusions. Do we have evidence to show that weekends are more likely to have accidents than other days?

Note that there are limitations to what can be concluded from a chi-square test. Failing to reject the null hypothesis is not the same as proving the null hypothesis is true. While we can prove that the probability of an accident in the district is not the same for each day, we can't state which days are more likely to have accidents. Similarly, we can prove  $H_a$ : weekends are not twice as likely to have accidents as weekdays, but failing to reject  $H_0$  does not prove that weekends are twice as likely to have accidents as weekdays.

### E.11. Zocor Again

Data set: Table 6.14

The extended activities in Section 6.7 discussed a clinical study on Zocor, a drug used in lowering cholesterol, to determine if there was a difference between the treatment (Zocor) and the placebo in reducing the number of deaths from heart attacks. Table 6.14 shows the original data from this study.

**Table 6.14** Results of the Zocor study.

	Died	Survived	Total
Placebo	189	2034	2223
Treatment	111	2110	2221
Total	300	4144	4444

- Conduct a chi-square test to test the hypothesis that the proportion of deaths is different depending on the drug (treatment or placebo). Write the null and alternative hypotheses and clearly state your conclusions.
- The extended activities showed that the percentage of deaths from heart attacks only changed from 8.5% to 5% over a five-year time period. Is this difference of 3.5% too small to be of practical importance? Why or why not?

### E.12. The Pill Scare: Understanding Relative Risk Reduction

Data set: Table 6.15

In October 1995, the United Kingdom Committee on Safety of Medicines (CSM) issued a warning to 190,000 general practitioners, pharmacists, and directors of public health about oral contraceptive pills containing gestodene or desogestrel. The warning, based on three unpublished epidemiological research studies, stated, "It is well known that the pill may rarely produce thrombosis (blood clots) involving veins of the legs. New evidence has become available indicating that the chance of thrombosis occurring in a vein increases about two-fold for some types of pills compared to others."<sup>12</sup> Table 6.15 provides data from one of the studies.

**Table 6.15** Impact of third generation contraceptive pills on venous thrombosis.

	Venous Thrombosis		Total
	Yes	No	
Third Generation Pill (contains gestodene or desogestrel)	127	249	376
Second Generation Pill (does not contain gestodene or desogestrel)	132	402	534
Total	259	651	910

Since the occurrence of venous thrombosis is very rare (1 in 7000 for people using the second generation pill),<sup>13</sup> 259 subjects were selected who had thrombosis and 651 similar subjects (from hospitals and community) who did not have thrombosis. Then these subjects were classified by the type of contraceptive they used.

- a. Was either the explanatory (row) or the response (column) variable fixed before the study was conducted?
- b. Is this an example of an experiment or an observational study?
- c. Is this a cross-classification, cohort, or case-control study?
- d. Create a segmented bar chart for the data.
- e. Use a two-sided hypothesis and Fisher's exact test to determine if the type of contraceptive impacts the likelihood of thrombosis. Do you expect that researchers took care to collect a simple random sample of subjects? What conclusions can be drawn?

The warning contained no numerical information other than the fact that the chance of blood clots was likely to double when birth control pills contained gestodene or desogestrel. This warning was widely publicized throughout the press, and evidence suggests that, as a result of this warning, many women ceased contraception altogether. Evidence shows a strong association between the warning and an increase in the number of unintended pregnancies and abortions (especially in women younger than 20 years old). This resulted in an estimated increase in cost of £21 million for maternity care and £4 to £6 million for abortion provision.<sup>14</sup>

- f. Remember that the actual occurrence of venous thrombosis is only 1 in 7000. If third generation pills double the chances of venous thrombosis, the likelihood of occurrence is still only 2 in 7000. Explain the difference between absolute risk reduction and relative risk reduction in this study.
- g. Death from venous thrombosis related to third generation pills is estimated to be 1 in 11 million, much lower than the probability of death resulting from pregnancy.<sup>15</sup> In 2005, the lifetime risk of maternal death in developed countries was 1 in 7300.<sup>16</sup> The CSM warning did suggest that patients see a doctor before altering their contraceptives; however, it appears that many women simply stopped taking any contraceptives. Write a brief statement to the press, general practitioners, pharmacists, and directors of public health about this study.

#### E.13. Baby Weights: Goodness-of-Fit Tests with Quantitative Data

Data set: BabyWeight

The file BabyWeight provides newborn weights for a simple random sample of 135 infants born in the United States in 1995. Can we conclude that the population from which this sample came is normally distributed?

- a. Find the 10th, 20th, 30th, . . . percentiles of the standard normal distribution. For example, the 10th percentile of the standard normal distribution is  $-1.28155$ .
- b. Standardize the sample data (subtract the sample mean and divide by the sample standard deviation). Place the sample data into categories based on the standard normal percentile groups, and report the number of data values in each category. For example, in the first group you would count the number of standardized sample observations that are less than or equal to  $-1.28155$  to get the observed number of data points for that group. How many data points would you expect in each group?
- c. Conduct a chi-square goodness-of-fit test and clearly state your conclusions.

With this goodness-of-fit test, we are testing the form of the distribution instead of particular parameter values. Before doing many types of hypothesis tests or calculating confidence intervals, it is appropriate to determine if the data fit certain model assumptions, such as a certain distributional form. Often this process includes determining if the data were sampled from a normal distribution, as in this question. The steps in conducting a goodness-of-fit tests are as follows:

- Group the observed responses ( $y$ ) into  $k$  arbitrarily chosen classes. In this exercise, we used evenly spaced percentiles of the normal distribution to create classes, but other classes (and more or fewer classes) could also be used. Classes must be distinct so that each observation can fall into only one class. If the result of the test is very sensitive to the choice of classes, then you cannot have much confidence in your conclusions.

- Calculate the expected frequencies based on some theoretical assumption about the distribution of the population (the random variable  $Y$ ).
- Calculate the chi-square statistic. If the observed frequencies are very different from the expected frequencies, the test statistic will be large and we can reject the theoretical model.

Note that we can never prove that the population actually follows a normal (or any other theoretical) model. We can only prove that a proposed model is wrong or fail to reject the model. Care should be taken not to state that we have proven that any population actually follows a particular distribution.

#### E.14. Simulating the Chi-Square Distribution

Data set: Table 6.1

- Calculate the chi-square test statistic for the simulated table in Question 20. Since the table cell counts (at least for the concave malignant group) were assumed to be randomly generated, this test statistic can be thought of as one random sample drawn from a chi-square distribution with 1 degree of freedom.
- Use software to generate 10,000 possible counts for the concave malignant group, using the cancer cell data. Calculate the chi-square test statistic for each of the 10,000 simulations. Create a histogram of the 10,000 chi-square test statistics. Describe the shape of this simulated chi-square distribution with 1 degree of freedom. Since 37 observations is still a fairly small sample size, the simulated distribution does not look completely like the theoretical chi-square distribution. Larger sample sizes will cause the simulated distribution to look much more like the true chi-square distribution.
- Use statistical software to randomly generate 10,000 values from a chi-square distribution with 1 degree of freedom. Compare this distribution to the one generated in Part B. Do they look the same?
- Repeat Part B for a larger sample size. Assume that there are 160 round and 210 concave nuclei. In addition, assume that there are 240 malignant and 130 benign cells. As in Part B, generate 10,000 possible counts for the concave malignant group for this new table. Calculate the chi-square test statistic for each of the 10,000 simulations. Create a histogram of the 10,000 chi-square test statistics. Describe the shape of this simulated chi-square distribution with 1 degree of freedom.

The theoretical chi-square distribution is continuous. As shown in Part B and C, when the number of counts in each cell is small, the test statistic does not accurately follow a chi-square distribution. While the chi-square test works best with large sample sizes, most statisticians agree that the cell counts in Part B are sufficient to be modeled by the chi-square distribution.

## Endnotes

---

1. R. Hooke (1918–2003), *How to Tell the Liars from the Statisticians* (New York: Marcel-Dekker, 1983).
2. American Cancer Society, *Cancer Facts and Figures 2008*, Atlanta, GA, [http://www.cancer.org/acs/groups/conten...](http://www.cancer.org/acs/groups/content/@epidemiologysurveillance/documents/document/acspc026210.pdf?) accessed 2/25/11.
3. W. Wolberg and O. Mangasarian, “Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology,” *Proceedings of the National Academy of Sciences of the United States of America*, 87.23 (Dec. 1990): 9193–9196.
4. N. Street, W. Wolberg, and O. Mangasarian, “Nuclear Feature Extraction for Breast Tumor Diagnosis,” *International Symposium on Electronic Imaging: Science and Technology*, 1905 (1993): 861–870.
5. For a more complete discussion of sampling schemes and appropriate tests see F. Ramsey and D. Schafer, *The Statistical Sleuth* (Pacific Grove, CA: Duxbury, 2002) or A. Agresti, “A Survey of Exact Inference for Contingency Tables,” *Statistical Science*, 7. 1 (1992): 131–177.
6. T. R. Pedersen et al., “Randomized Trial of Cholesterol Lowering in 4444 Patients with Coronary Heart Disease: The Scandinavian Simvastatin Survival Study (4S),” *Lancet*, 344 (1994): 1383–1389.
7. S. Woloshin, L. Schwartz, and H. G. Welch, *Know Your Chances: Understanding Health Statistics* (Berkeley: University of California Press, 2008), p. 42.

8. R. Doll and A. B. Hill, "Smoking and Carcinoma of the Lung: Preliminary Report," *British Medical Journal*, 2.4682 (1950): 739–748.
9. D. Salsburg, *The Lady Tasting Tea* (New York: Freeman, 2001).
10. "Donner Party Deaths: A Demographic Assessment," *Journal of Anthropological Research*, 46 (1990): 223–242.
11. D. M. Barnes, "Breaking the Cycle of Addiction," *Science*, 241 (1988): 1029–1030.
12. A. Furdei, "The Public Health Implications of the 1995 'Pill Scare,'" *Human Reproduction Update*, 5.6 (1999): 621–626.
13. G. Gigerenzer et al., "Helping Doctors and Patients Make Sense of Health Statistics," *Psychological Science in the Public Interest*, Oct. 8, 2008.
14. A. Furdei, "The Public Health Implications of the 1995 'Pill Scare,'" *Human Reproduction Update*, 5.6 (1999): 621–626.
15. Ibid.
16. *Maternal Mortality in 2005*, [http://www.who.int/whosis/mme\\_2005.pdf](http://www.who.int/whosis/mme_2005.pdf), accessed 11/3/10.
17. N. J. Gotelli and G. R. Graves, *Null Models in Ecology* (Washington, DC: Smithsonian Institution Press, 1996). Gotelli and Graves define their concept this way: "[A] null model is a pattern-generating model that is based on randomization of ecological data or random sampling from a known or imagined distribution. The null model is designed with respect to some ecological or evolutionary process of interest. Certain elements of the data are held constant, and others allowed to vary stochastically to create assemblage patterns. The randomization is designed to produce a pattern that would be expected in the absence of a particular ecological mechanism" (pp. 3–4). Gotelli and Graves motivate the scientist's interest in null models, stating that if the data are consistent with a properly constructed null model, we can infer that the biological mechanism is not operating, but if the data are inconsistent with the null model, "this provides some positive evidence in favor of the mechanism" (p. 6).

# Research Project: Infant Handling in Female Baboons

## Reviewing the Literature

Adult female primates interact with the infants of other females. Observation of these interactions has spawned several descriptive terms to reflect the range of behavior: auntling, babysitting, play-mothering, allomothering, and kidnapping. The term *infant handling* is a neutral and inclusive term for all such interactions, which include, but are not limited to, pulling, hitting, holding, and carrying infants. The research on such behavior spans four decades, and the social, functional, and evolutionary understanding of infant handling continues to be the subject of study and debate.

Primateologist Vicki Bentley-Condit of Grinnell College studied interactions between female and infant yellow baboons (*Papio cynocephalus cynocephalus*) at the Tana River National Primate Reserve in Kenya. She collected the data for her study by observing baboons in 20-minute focal samples over an 11-month period in 1991–1992. Her subjects included 23 female baboons, 11 of which were mothers with infants (no mother with more than one offspring). Bentley-Condit observed and recorded interactions between females and infants, excluding interactions between a mother and her own offspring.

1. Read the paper by V. K. Bentley-Condit, T. L. Moore, and E. O. Smith, “Analysis of Infant Handling and the Effects of Female Rank Among Tana River Adult Female Yellow Baboons Using Permutation/Randomization Tests,” *American Journal of Primatology*, 55 (2001): 117–130. If there are any words that you do not understand, look them up and provide a short definition for each.

Identify the following and be ready to discuss this material in class:

- a. Objective of the study
- b. Any relevant background (from journals that were referenced)
- c. Response variable(s)
- d. Explanatory variables and levels that were tested
- e. Variables that were held constant during the study
- f. Nuisance factors (i.e., factors that are not of interest but may influence the results)

## Developing a Research Hypothesis

Bentley-Condit computed a “dominance hierarchy score” for each female, using a calculation based on aggressive and submissive interactions between female dyads in the troop. Scores can theoretically range between –22 and +22. Figure 6.3 shows these scores, which are broken into clusters: high ranks (code = 1) are above 10; mid ranks (code = 2) are between 0 and 10; low ranks (code = 3) are below 0.

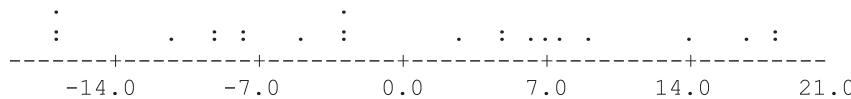


Figure 6.3 Dominance hierarchy (female rank) scores for 23 female handlers.

One objective of Bentley-Condit’s study was to see if female rank impacted the pattern or success rate of infant-handling interactions. Her research hypothesis was that female handlers would be disinclined to handle infants above their own rank, and she established the following research hypothesis for investigation\*:

**Research Hypothesis 1 (RH1):** Females will tend to handle the infants of females that are ranked the same as or lower than themselves.

Table 6.16 gives the total number of interactions for each female–infant pair *over the entire study period*. For example, the value of 13 in the (2, 1) cell (row = 2, column = 1) represents 13 interactions by handler

\*We call this Research Hypothesis 1, as we will introduce other possible research hypotheses later.

**Table 6.16** Data matrix giving the total number of interactions of female handlers (columns) and infants (rows). Boldface numbers give ranks of handlers or infants, with 1 being a high ranking, 2 being a mid ranking, and 3 being a low ranking. Each handler and infant has a two-letter ID; infants' IDs are separated from their mothers' IDs with a slash. Horizontal and vertical lines separate the rank categories.

		Handler's Names and Ranks																						
Infants/Mothers Names and Mother's Ranks		KM <b>1</b>	KN <b>1</b>	NQ <b>1</b>	PO <b>1</b>	HQ <b>2</b>	LL <b>2</b>	NY <b>2</b>	PS <b>2</b>	SK <b>2</b>	ST <b>2</b>	WK <b>2</b>	AL <b>3</b>	CO <b>3</b>	DD <b>3</b>	LS <b>3</b>	LY <b>3</b>	MH <b>3</b>	ML <b>3</b>	MM <b>3</b>	PA <b>3</b>	PH <b>3</b>	PT <b>3</b>	RS <b>3</b>
KG/KM	<b>1</b>	0	0	4	1	1	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	2	1
HZ/HQ	<b>2</b>	13	23	7	5	0	2	1	1	5	6	18	1	6	3	0	1	4	1	0	9	0	10	1
LC/LL	<b>2</b>	4	0	1	4	3	0	2	1	1	5	3	1	0	0	1	0	2	1	1	1	0	1	6
NK/NY	<b>2</b>	12	4	10	5	9	1	0	2	3	11	7	8	6	3	1	0	2	1	1	5	3	2	3
PZ/PS	<b>2</b>	1	3	4	1	0	0	0	0	0	0	2	0	2	0	0	0	3	0	1	1	0	3	0
CY/CO	<b>3</b>	2	2	7	3	1	1	2	0	3	12	16	3	0	2	0	0	2	0	0	1	0	0	2
LZ/LS	<b>3</b>	1	0	3	2	1	1	0	0	2	0	5	2	2	2	0	1	9	2	0	0	0	3	2
MQ/ML	<b>3</b>	0	1	5	2	2	4	2	2	2	4	5	7	5	2	1	1	7	0	4	4	1	0	2
MW/MH	<b>3</b>	3	0	7	4	2	3	0	5	2	8	13	7	14	2	0	0	0	4	0	8	0	13	6
MX/MM	<b>3</b>	2	3	4	5	0	0	0	0	0	5	2	9	3	1	0	0	2	0	0	1	2	2	3
PK/PH	<b>3</b>	2	0	6	4	3	4	1	0	0	15	10	8	5	1	0	3	1	1	6	3	0	7	5

KM of infant HZ over the observational period of the study. Handler KM was ranked 1 (high), and infant HZ had a mother, HQ, which was ranked 2 (mid). Note that the dominance is ranked as high > mid > low, which numerically translates to rank 1 > rank 2 > rank 3 in Table 6.16.

Part (a) of Table 6.17 provides a summary of the 678 interactions shown in Table 6.16. In Table 6.17, each interaction is classified only by the rank of the infant (meaning the infant's mother) and the rank of the female handler. For example, the 5 count in the (1, 2) cell in Part (a) is the sum total of all interactions of a high-ranked infant with a mid-ranked female; that is,  $5 = 1 + 0 + 0 + 0 + 3 + 1 + 0$ , reading across the counts corresponding to mid-ranked handlers and the one high-ranked infant in Table 6.16. If the research hypothesis were true, we would expect counts to be relatively high in cells where the rank of the handler was at least as high as that of the infant and relatively low in the other cells.

**Table 6.17** Interactions by infant rank and handler rank appear in (a); column percentages are in (b). Part (c) gives the expected values from a chi-square test for independence, and Part (d) gives the adjusted, standardized Pearson residuals:  $(\text{observed} - \text{expected}) / \sqrt{\text{expected}}$ .

		Handler's Rank					Handler's Rank		
		Hi	Mid	Low	Infant	Hi	Mid	Low	
Infant	Hi	5	5	3	Infant	2.9%	2.2%	1.1%	
	Mi	97	83	95		57.1%	36.7%	33.9%	
	Lo	68	138	184		40.0%	61.1%	65.0%	
		(a)					(b)		
		Handler's Rank					Handler's Rank		
		Hi	Mid	Low	Infant	Hi	Mid	Low	
Infant	Hi	3.26	4.33	5.41	Infant	0.96	0.32	-1.04	
	Mi	68.95	91.67	114.38		3.38	-0.90	-1.81	
	Lo	97.79	130.00	162.21		-3.01	0.70	1.71	
		(c)					(d)		

Note that there are a few instances where mid- or low-ranked handlers interact with a high-ranked infant. This disobeys the research hypothesis, so the question before us is whether the data exhibit a statistical tendency toward the research hypothesis. *That is, allowing for variability, is the general pattern in favor of the research hypothesis (RH1) strong enough to rule out chance as a plausible explanation for the pattern?*

This last question is the question that Professor of Anthropology Bentley-Condit brought to Professor of Statistics Moore after she herself had observed patterns, in a descriptive fashion, which suggested support for the research hypothesis. In this project, we will work through these steps to investigate the research hypothesis in light of the data:

- We will interpret Table 6.17 descriptively in light of RH1.
  - We will then apply the concept of permutation tests to the data to confirm or deny our descriptive analysis.
- In the process of pairing the descriptive with the inferential analysis, we will learn these lessons:
- Permutation tests provide an inferential tool in situations where standard methods do not exist.
  - Permutation tests are easy to compute in these new situations, given the availability of inexpensive and fast computing power.
  - Permutation tests are more flexible than standard methods, providing more choice in the test statistic.

## Descriptive Analysis of the Data

Part (b) of Table 6.17 uses simple percentages to describe the pattern in the data. You should verify with a calculator that the counts in Part (a) are consistent with the percentages in Part (b). In describing a pattern in a two-way table, it is almost always helpful to first identify the explanatory and response variables. In observational studies, like the one we have here, this distinction can be a bit arbitrary, but it is still a useful distinction to make. Here we treat handler rank as the explanatory variable ( $X$ ) and infant rank as the response variable ( $Y$ ).

Notice how Part (b) of Table 6.17 appears to support the research hypothesis. In the top row, we see that the highest probability of handling a high-ranked infant is associated with the high-ranked handler (2.9% versus 2.2% and 1.1%). Similarly, mid-ranked infants are more likely to be handled by high- or mid-ranked females than by low-ranked females.

When data are placed in a two-way contingency table, it is often reasonable to consider conducting a chi-square test (for independence). Part (c) of Table 6.17 contains the expected counts, assuming that the rows and column variables are independent.

Part (d) of Table 6.17 highlights the extent to which the expected counts in Part (c) differ from the observed counts in Part (a), as well as the direction of these differences. Cells in Part (d) have positive values when the observed count is greater than the expected count and negative values when the observed count is less than the expected count. Moreover, the values in Part (d) can be interpreted just as we interpret Z-scores. For example, the 1.71 in the (3, 3) cell says that the observed number of interactions of low-ranked handlers and low-ranked infants is much higher than one would expect if handler rank and infant rank were independent, to the tune of almost 2 standard deviations. This value gives support to the research hypothesis, as do other entries in Part (d).

2. Verify that the calculations in Part (c) of Table 6.17 are the expected counts, assuming that the row and column variable are independent.
3. Give a succinct interpretation of each cell in Part (d) in terms of infant handling in the context of handler vs. mother status.
4. Read the baboon data into a computer software package. Then re-create Table 6.16 and Part (a) of Table 6.17.
5. While there is a pattern that seems to support the research hypothesis, a significance test is needed to determine if the observed pattern can plausibly be attributed to random chance. Perform a chi-square test for independence on Part (a) of Table 6.17. State the  $p$ -value.

## Permutation Tests

The chi-square test in Question 5 is not appropriate, since the chi-square test assumes that each observation is independent. In other words, the chi-square test would require that the 678 observations enter the table one at a time. For example, consider the 97 high-on-mid interactions. The chi-square test assumes that if handler rank is independent of infant rank, then these 97 interactions could just as likely have been distributed into other cells of the  $2 \times 2$  table. But this ignores the “clustering” of the counts by individual baboons. For example, part of those 97 interactions are the 13 interactions of KM on HZ/HQ, interactions that could, under the null hypothesis, have ended up in other cells of the  $2 \times 2$  table, but they would have entered those other cells as a cluster: all 13 in the cell or not. In a sense, we have two levels of observation: the handler–infant pairs and the individual instances of infant handling. While the expected values (and corresponding residuals in Part (d) of Table 6.17) are fine for informal descriptive analysis, the classical chi-square test is problematic because the clustering of observations from handler–infant pairs occurring in each cell violates a key assumption.

The beauty of the permutation test is that it is adaptable enough to respect the clustering of the counts by handler–infant pairs. For our permutation test, the null hypothesis and corresponding null model are as follows:

$H_0$ : handler rank and infant rank are independent

*Null Model:* For the data in Table 6.16, the dominance ranks can be viewed as meaningless labels attached at random to handlers and infants. Thus, data sets produced by permuting these ranks in all possible ways that respect infant–mother pairs are equally likely.

In the null model, the elements held constant are the counts. The elements that are allowed to vary stochastically (randomly) to create “assemblage patterns” are the female and infant ranks. The assemblage patterns generated become new, hypothetical data sets from which we can construct the sampling distribution of a permutation test statistic. To complete the picture, we need some test statistic that reflects the level of agreement between the data and the null model. For now, denote the test statistic by the letter  $C$ . Here then is the formal process that becomes our test:

- Assign ranks at random to infants and handlers, using the rank distributions of the data set. That is, randomly assign ranks so that the 11 infants are assigned, in this case, 1 to high, 4 to mid, and 6 to low. In addition, randomly assign the 23 handlers so that there are 4 high, 7 mid, and 12 low (with the restriction that the 11 handlers that are the mothers must be assigned to the same rank as their infants). This assignment leads to the original data table but with permuted ranks.

- Re-form the  $3 \times 3$  table of interaction counts (as in Part (a) of Table 6.17) based on these new handler and mother ranks.
- Compute the value of a test statistic,  $C$ , for this new table.

This three-step process defines a sampling distribution for  $C$  under  $H_0$ . The  $p$ -value of a particular data set is defined to be  $P(C \geq C_D)$ , where  $C_D$  represents the value of the test statistic observed in the original data.

To compute the sampling distribution exactly would require a complete enumeration of all possible permutations. There are 42,688,800 such permutations, a prohibitively large number to enumerate, so we instead approximate the sampling distribution by randomly generating permutations of the row and column ranks according to the three-step process outlined above. *We can then determine an empirical p-value as the proportion of permutations that result in  $C \geq C_D$ .*

In the following discussion, we will use a test statistic that reflects RH1. We denote this test statistic by LTE (less than or equal to), so we replace  $C$  with LTE in the above three-step process. Here is the definition of LTE:

$$\begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = a + d + g - b + e + h - c - f + i = \text{LTE}$$

The  $\cdot$  symbol is meant to represent the familiar dot product from vector algebra. For the  $3 \times 3$  table in Part (a) of Table 6.17, the calculation of LTE is thus

$$\begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 5 & 5 & 3 \\ 97 & 83 & 95 \\ 68 & 138 & 184 \end{bmatrix} = 5 + 97 + 68 - 5 + 83 + 138 - 3 - 95 + 184 = 472$$

Notice that the LTE statistic adds counts that support RH1 and subtracts counts that run counter to RH1. Hence, large LTE values lend evidence for RH1, while small values lend evidence against RH1. The question is, Is 472 a large value? Is it large enough to be considered “statistically significant”?

The permutation test (steps 1–3 above) permutes the infant–mother ranks at random many times, recomputing the  $3 \times 3$  table and the LTE statistic many times, to ascertain the answer to this question.

6. Use software instructions to calculate the LTE statistic for the observed data.
7. Table 6.18 gives the original data set, but with one random rearrangement (permutation) of the infant and handler ranks. Answer the questions that follow to complete one iteration of the permutation test.
  - a. The random permutation should not change the number of elements in each group (i.e., the frequency in each rank needs to stay the same). Count the frequency of ranks 1, 2, and 3 for the infants. Is the distribution the same as in the original data set?
  - b. In similar fashion, is the distribution of ranks for the handlers the same as in the original data set?
  - c. Now, verify that any handler that is also a mother has the same rank as her infant. (This makes this permutation a valid permutation of ranks, as described in the first step of the randomization test above.)
  - d. Complete step 2 for this table by hand. That is, re-form the table into a  $3 \times 3$  table.
  - e. Finally, compute the LTE statistic for this table by hand.
  - f. Repeat Parts D and E using statistical software.
8. Use the software instructions provided to run 10,000 iterations of the permutation test.
9. Does the chi-square test, which we argue above is inappropriate for these data, suggest about the same level of significance as the permutation test? Try to reason why this would make sense in light of the way in which the data are structured.

## A New Research Hypothesis

Within the period of analyzing the data, Bentley-Condit and colleagues recognized the possibility of a second research hypothesis, which in the paper is described as “immediately lower,” rather than “lower,” the latter being our RH1. Specifically, we define Research Hypothesis 2 this way:

**Table 6.18** The original data table with a random permutation of the infant–mother ranks and the handler ranks.

Infant's/Mother's Name and Mother's Rank	Handler's Name and Rank																						
	KM 1	KN 3	NQ 1	PO 3	HQ 3	LL 3	NY 2	PS 2	SK 3	ST 1	WK 3	AL 2	CO 2	DD 3	LS 3	LY 1	MH 3	ML 3	MM 3	PA 2	PH 2	PT 3	RS 2
KG/KM 1	0	0	4	1	1	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	2	1
HZ/HQ 3	13	23	7	5	0	2	1	1	5	6	18	1	6	3	0	1	4	1	0	9	0	10	1
LC/LL 3	4	0	1	4	3	0	2	1	1	5	3	1	0	0	1	0	2	1	1	1	0	1	6
NK/NY 2	12	4	10	5	9	1	0	2	3	11	7	8	6	3	1	0	2	1	1	5	3	2	3
PZ/PS 2	1	3	4	1	0	0	0	0	0	0	2	0	2	0	0	0	3	0	1	1	0	3	0
CY/CO 2	2	2	7	3	1	1	2	0	3	12	16	3	0	2	0	0	2	0	0	1	0	0	2
LZ/LS 3	1	0	3	2	1	1	0	0	2	0	5	2	2	2	0	1	9	2	0	0	0	3	2
MQ/ML 3	0	1	5	2	2	4	2	2	2	4	5	7	5	2	1	1	7	0	4	4	1	0	2
MW/MH 3	3	0	7	4	2	3	0	5	2	8	13	7	14	2	0	0	0	4	0	8	0	13	6
MX/MM 3	2	3	4	5	0	0	0	0	0	5	2	9	3	1	0	0	2	0	0	1	2	2	3
PK/PH 2	2	0	6	4	3	4	1	0	0	15	10	8	5	1	0	3	1	1	6	3	0	7	5

**Research Hypothesis 2 (RH2):** Females will tend to handle the infants of females that are ranked immediately below them in the three-tiered ranking system (or ranked the same, for a three-ranked handler).

The rationale for RH2 is given in the paper.

10. Below we define a new test statistic, LT. Explain why this test statistic “makes sense” when the research hypothesis of interest is RH2. Compute the observed value of LT for Table 6.16.

$$\begin{bmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = -a + d + g - b - e + h - c - f + i = \text{LT}$$

11. Perform a permutation test with the LT statistic. Submit no more than a two-page summary of your analysis, including appropriate conclusions. Assume you are working with a primatologist who has only an introductory statistics background. In this summary, explain how the permutation test is conducted and why a permutation test is a better approach than a chi-square test for this study.

While this is an observational study, rejecting the null hypothesis cannot lead to a firm causal interpretation. Still, the rejection of the null hypothesis in this context is considered by many scientists to be a useful part of the analysis and a worthwhile next step beyond identifying an observed pattern from the descriptive analysis.<sup>17</sup>

# Logistic Regression: The Space Shuttle Challenger

*The best thing about being a statistician is that you get to play in everyone's backyard.*

—John Tukey<sup>1</sup>

There are many investigations where a researcher is interested in developing a regression model when the response variable is dichotomous (has only two categories). Dichotomous responses can be represented with binary data (data with values of only zero or one). Logistic regression is used to examine the relationship between one or more explanatory variables and a binary response variable.

In this chapter, we will look at several studies, including O-ring failure data provided by the National Aeronautics and Space Administration (NASA) after the space shuttle Challenger disaster, in order to introduce the following logistic regression techniques:

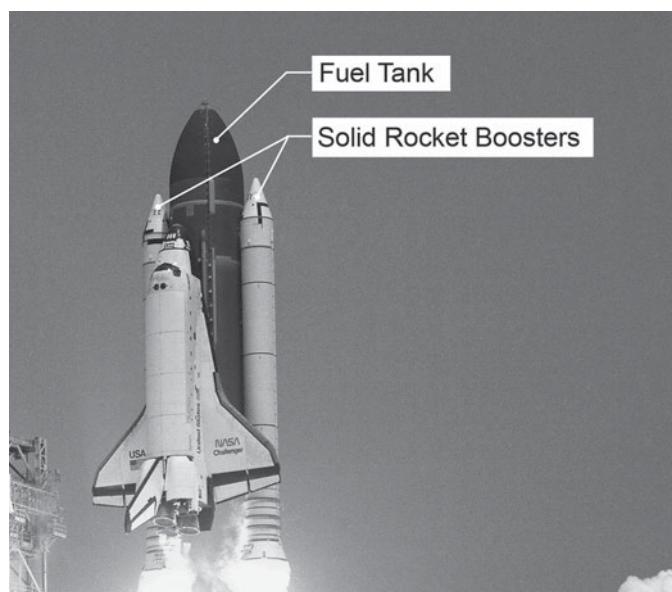
- Calculating and interpreting the logistic regression model
- Using the Wald statistic and likelihood ratio tests to determine the significance of individual explanatory variables
- Calculating the log-odds function and maximum likelihood estimates
- Conducting goodness-of-fit tests to evaluate model appropriateness
- Assessing regression model performance by looking at a classification table, showing correct and incorrect classification of the response variable
- Extending logistic regression to cases with multiple explanatory variables

## 7.1 Investigation: Did Temperature Influence the Likelihood of an O-Ring Failure?

On January 28, 1986, the NASA space shuttle program launched its 25th shuttle flight from Kennedy Space Center in Florida. Seventy-three seconds into the flight, the external fuel tank collapsed and spilled liquid oxygen and hydrogen. These chemicals ignited, destroying the shuttle and killing all seven crew members on board. Reports to President Reagan and videos of the event are available at the Kennedy Space Center website.\*

Investigations showed that an O-ring seal in the right solid rocket booster failed to isolate the fuel supply. Figure 7.1 shows the space shuttle Challenger just after ignition with the fuel tank and two 149.16-foot-long solid rocket boosters. Figure 7.2 shows a diagram of a solid rocket booster. Because of its size, the rocket boosters were built and shipped in separate sections. A forward, center and aft field joint connected the sections. Two O-rings (one primary and one secondary), which resemble giant rubber bands 0.28 inch thick but 37 feet in diameter, were used to seal the field joints between each of the sections.

An O-ring seal was used to stop the gases inside the solid rocket booster from escaping. However, the cold outside air temperature caused the O-rings to become brittle and fail to seal properly. Gases at 5800°F escaped and burned a hole through the side of the rocket booster.



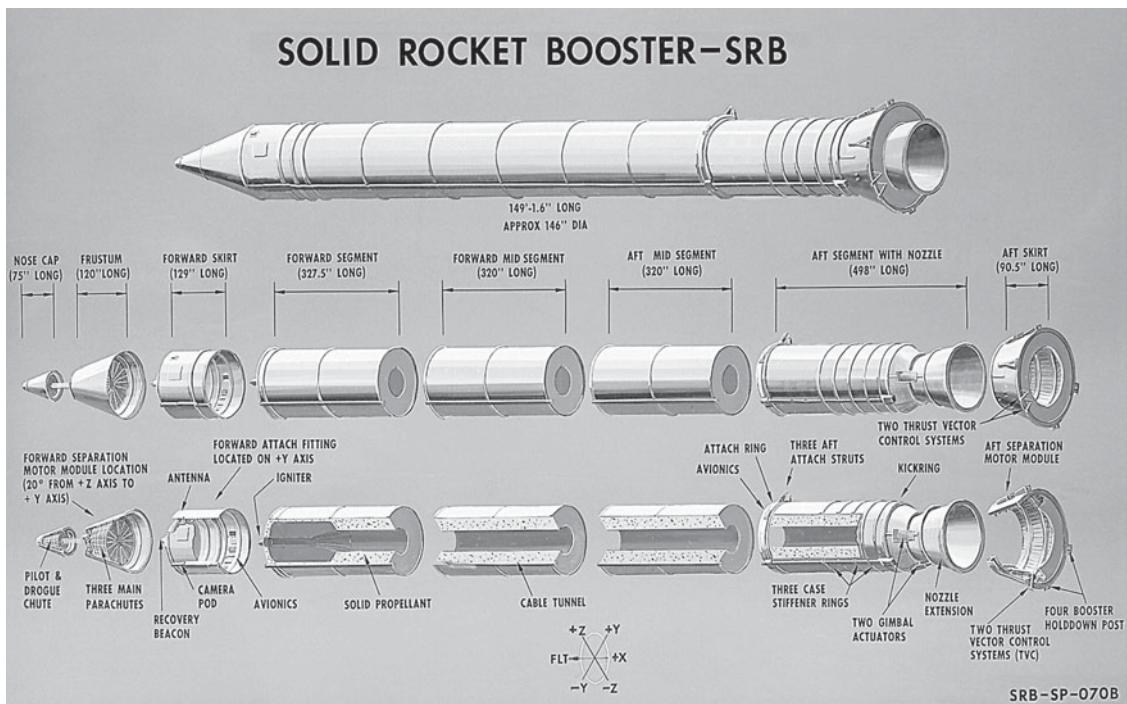
**Figure 7.1** Picture of the space shuttle Challenger just after ignition. Each solid rocket booster had six O-rings, two at each field joint. The O-rings at the right aft field joint failed. (Photo courtesy of NASA)

*The Report of the Presidential Commission on the Space Shuttle Challenger Accident*, also known as the Rogers' Commission Report, states:

“O-ring resiliency is directly related to its temperature. . . . A warm O-ring that has been compressed will return to its original shape much quicker than will a cold O-ring when compression is relieved. . . . A compressed O-ring at 75 degrees Fahrenheit is five times more responsive in returning to its uncompressed shape than a cold O-ring at 30 degrees Fahrenheit. . . . At the cold launch temperature experienced, the O-ring would be very slow in returning to its normal rounded shape. . . . It would remain in its compressed position in the O-ring channel and not provide a space between itself and the upstream channel wall. Thus, it is probable the O-ring would not . . . seal the gap in time to preclude joint failure due to blow-by and erosion from hot combustion gases. . . . Of 21 launches with ambient

---

\*Take a few minutes to view the BBC video using this link: [http://news.bbc.co.uk/onthisday/hi/dates/stories/january/28/newsid\\_2506000/2506161.stm](http://news.bbc.co.uk/onthisday/hi/dates/stories/january/28/newsid_2506000/2506161.stm).



**Figure 7.2** Diagram of a solid rocket booster. (Illustration courtesy of the Report to the President by the Presidential Commission on the Space Shuttle Challenger Accident/NASA)

temperatures of 61 degrees Fahrenheit or greater, only four showed signs of O-ring thermal distress: i.e., erosion or blow-by and soot. Each of the launches below 61 degrees Fahrenheit resulted in one or more O-rings showing signs of thermal distress.”<sup>2</sup>

A lamentable aspect of this disaster is that the problem with the O-rings was already understood by some engineers prior to the Challenger launch. In February 1984, the Marshall Configuration Control Board sent a memo about the O-ring erosion that occurred on STS 41-B (the 10th space shuttle flight and the 4th mission for the Challenger shuttle). Messages continued to increase in intensity, as evidenced by a 1985 internal memo from Thiokol Corporation, the company that designed the O-ring. Employees from Thiokol wrote the following to their Vice President of Engineering: “This letter is written to ensure that management is fully aware of the seriousness of the current O-Ring erosion problem in the SRM joints from an engineering standpoint.”<sup>3</sup>

With the temperature on January 28, 1986, expected to be 31°F, Thiokol Corporation recommended against the Challenger launch. However, this flight was getting significant publicity because a high school teacher, Christa McAuliffe, was on the flight. The flight had already been delayed several times, and there was no quick solution to the O-ring concern. The engineers were overruled, and the decision was made to go ahead with the launch. The eventual presidential investigation stated,

“The decision to launch the Challenger was flawed. Those who made that decision were unaware of the recent history of problems concerning the O-rings and the joint and were unaware of the initial written recommendation of the contractor advising against the launch at temperatures below 53 degrees Fahrenheit and the continuing opposition of the engineers at Thiokol after the management reversed its position. They did not have a clear understanding of Rockwell’s concern that it was not safe to launch because of ice on the pad. If the decision makers had known all of the facts, it is highly unlikely that they would have decided to launch 51-L on January 28, 1986.”<sup>4</sup>

It seems that even though some engineers did comprehend the severity of the problem, they were unable to properly communicate the results. Prior to the ill-fated Challenger flight, the solid rocket boosters for 24 shuttle launches were recovered and inspected for damage. Even though O-ring damage was present in some of the flights, the O-rings were not damaged enough to allow any gas to escape. Since damage was very minimal, all 24 prior flights were considered a success by NASA.

**Table 7.1** O-ring damage on 24 space shuttle launches.

Flight Number	Date	Ambient Temperature (°F)	Successful Launch
1	4/12/1981	66	1
2	11/12/1981	70	0
3	3/22/1982	69	1
4	6/27/1982	80	*
5	11/11/1982	68	1
6	4/4/1983	67	1
7	6/18/1983	72	1
8	8/30/1983	73	1
9	11/28/1983	70	1
10	2/3/1984	57	0
11	4/6/1984	63	0
12	8/30/1984	70	0
13	10/5/1984	78	1
14	11/8/1984	67	1
15	1/24/1985	53	0
16	4/12/1985	67	1
17	4/29/1985	75	1
18	6/17/1985	70	1
19	7/29/1985	81	1
20	8/27/1985	76	1
21	10/3/1985	79	1
22	10/30/1985	75	0
23	11/26/1985	76	1
24	1/12/1986	58	0

\* Flight 4 is a missing data point because the rockets were lost at sea.

Table 7.1 shows the temperature at the time of each launch and whether any damage was visible in any of the O-rings. In this chapter, we will define a successful launch as one with no evidence of any O-ring damage. In Table 7.1, Successful Launch is a categorical variable, with 0 representing a launch where O-ring damage occurred and 1 indicating a successful launch with no O-ring damage. Throughout the rest of this investigation, the relatively small data set in Table 7.1 will be used to demonstrate techniques that can be used to determine if the likelihood of O-ring damage is related to temperature.

## Activity Describing the Data

1. Based on the description of the Challenger disaster O-ring concerns, identify which variable in the Shuttle data set in Table 7.1 should be the explanatory variable and which should be the response variable.
2. Imagine you were an engineer working for Thiokol Corporation prior to January 1986. Create a few graphs of the data in Table 7.1. Is it obvious that temperature is related to the success of the O-rings? Submit any charts or graphs you have created that show a potential relationship between temperature and O-ring damage.

In this chapter, we will develop a regression model using a binary response variable, *Successful Launch*. For the space shuttle data set,  $y = 1$  represents a successful flight with no O-ring damage and  $y = 0$  represents a flight with some O-ring damage. Binary response data occur in many fields; for example, we may want to know

- whether a disease is present or absent
- whether or not a person is a good credit risk for a loan
- whether or not a high school student should be admitted to a particular college
- whether or not an individual is involved in substance abuse

The next section describes why the least squares regression model is not appropriate when the response is binary. **Logistic regression** is used to examine the relationship between one or more explanatory variables and a binary response variable. Like other regression models, logistic regression models often have explanatory variables that are quantitative, but they can be categorical as well.

## 7.2 Review of the Least Squares Regression Model

In Chapters 2 and 3, you saw that the ordinary least squares regression model has the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, 3, \dots, n \quad (7.1)$$

where  $n$  is the number of observations,  $y_i$  is the  $i$ th value of a *continuous response variable*,  $Y$ ,  $\beta_0$  and  $\beta_1$  are regression coefficients,  $x_i$  is the  $i$ th value of the explanatory variable, and  $\varepsilon_i$  represents normally distributed errors with a constant variance. Equation (7.2) states that the mean response (the expected response at each particular  $x_i$ ) is equal to the *linear predictor*  $\beta_0 + \beta_1 x_i$  for each observed value  $x_i$ .

$$E(Y_i | x_i) = \beta_0 + \beta_1 x_i \quad \text{for } i = 1, 2, 3, \dots, n \quad (7.2)$$

where  $\beta_0$  and  $\beta_1$  are parameters that can be estimated with sample data. In addition to assuming that the regression model has a linear predictor, we assume that the error terms in the least squares regression model are independent and follow the normal distribution with a zero mean and a fixed standard deviation:

$$\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad \text{for } i = 1, 2, 3, \dots, n \quad (7.3)$$

Equation (7.3) states that each **independent and identically distributed** error term follows a normal probability distribution that is centered at zero and has a constant variance.

### NOTE

When there is only one explanatory variable, as in Equation (7.1), ordinary least squares regression is often called simple linear regression. As shown in Chapter 3, the model is called least squares regression because the line minimizes the sum of the squared residuals (the difference between an observed value and the expected response). Least squares estimates for  $\beta_0$  and  $\beta_1$  (represented as  $b_0 = \hat{\beta}_0$  and  $b_1 = \hat{\beta}_1$ ) can be calculated even when the normality and equal variance assumptions are violated. However, these assumptions about the error terms are needed to conduct hypothesis tests and construct confidence intervals for  $\beta_0$  and  $\beta_1$ .

### Activity ▶ Building a Least Squares Regression Model

3. Use the data in Table 7.1 to create a scatterplot with a least squares regression line for the space shuttle data. Calculate the predicted response values ( $\hat{y} = b_0 + b_1 x$ ) when the temperature is 60°F and when the temperature is 85°F.

Logistic regression models have many similarities to ordinary least squares regression models. However, the regression line created in Question 3 does not appropriately predict the response. The following section demonstrates important differences in regression models when the response variable is binary.

## 7.3 The Logistic Regression Model

When the response variable is binary, the response is typically defined as a probability of success, instead of 0 or 1. For example, in Question 3, when the temperature is 60°F, the least squares regression line estimates that the probability of a successful launch is 0.338. The expected response at each particular  $x_i$  is defined as

$$\begin{aligned}\pi_i &= P(Y_i = 1) = \text{probability that a launch has no O-ring damage at temperature } x_i \\ &= E(Y_i | x_i) \\ &= \beta_0 + \beta_1 x_i \quad \text{for } i = 1, 2, 3, \dots, n\end{aligned}\tag{7.4}$$

While the linear model ( $\beta_0 + \beta_1 x_i$ ) in Equation (7.4) is simple, it is not appropriate to use, since probabilities must be between 0 and 1. For example, with a temperature value  $x_i = 50$ , the least squares regression model in Question 3 would predict a probability of  $-0.036$ . In order to restrict the predictions to values between 0 and 1, an S-shaped function called the log-odds function will be used.

Logistic regression uses the following model to fit an S-shaped relationship between  $\pi$  and  $x$ :

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i\tag{7.5}$$

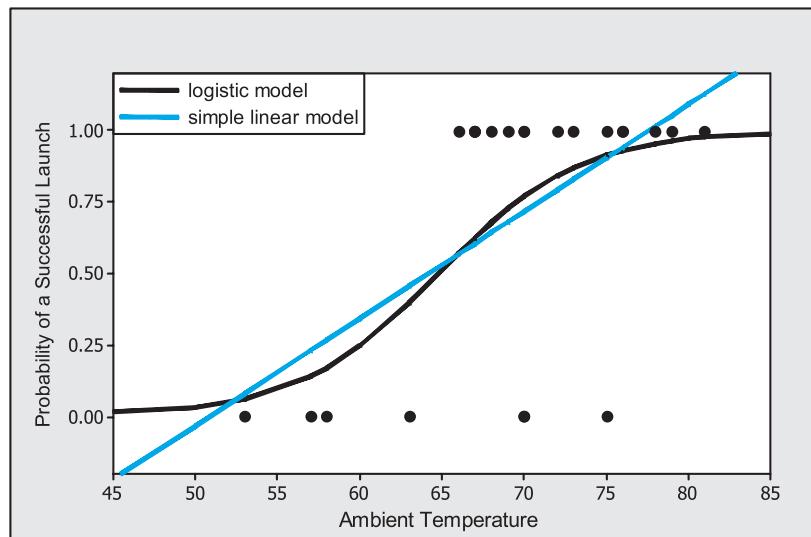
where  $\ln$  represents the natural log,  $\beta_0$  and  $\beta_1$  are regression parameters, and  $\pi_i$  is the probability of a successful launch for a given temperature ( $x_i$ ). The ratio  $\pi/(1 - \pi)$  is called the **odds**, the probability of success over the probability of failure. Thus, the function  $\ln[\pi/(1 - \pi)]$  is called the **log-odds** of  $\pi$  or the **logistic or logit transformation** of  $\pi$ .\* Figure 7.3 shows both the least squares regression model and the logistic regression model for the space shuttle data.

### MATHEMATICAL NOTE

In Chapter 6, the *odds* of an outcome are defined as  $\pi/(1 - \pi)$ , the probability of a success (no O-ring damage) over the probability of a failure (O-ring damage). For example, if a computer randomly selects a day of the week, the odds of selecting Saturday (Saturday is considered a success) are 1 to 6, since

$$\text{odds} = \frac{\pi}{1 - \pi} = \frac{\frac{1}{7}}{\left(1 - \left(\frac{1}{7}\right)\right)} = \frac{1}{6}.$$

Similarly, the odds are 6 to 1 against Saturday being selected (any day but Saturday is a success).



**Figure 7.3** Space shuttle data with a simple linear regression model and a logistic regression model.

\*Throughout this chapter, we will use terms such as *log-odds* or *log-likelihood*, but we actually use natural logs ( $\ln$ ) in our calculations.

Equation (7.5) can be solved for  $\pi_i$  to show that

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \quad (7.6)$$

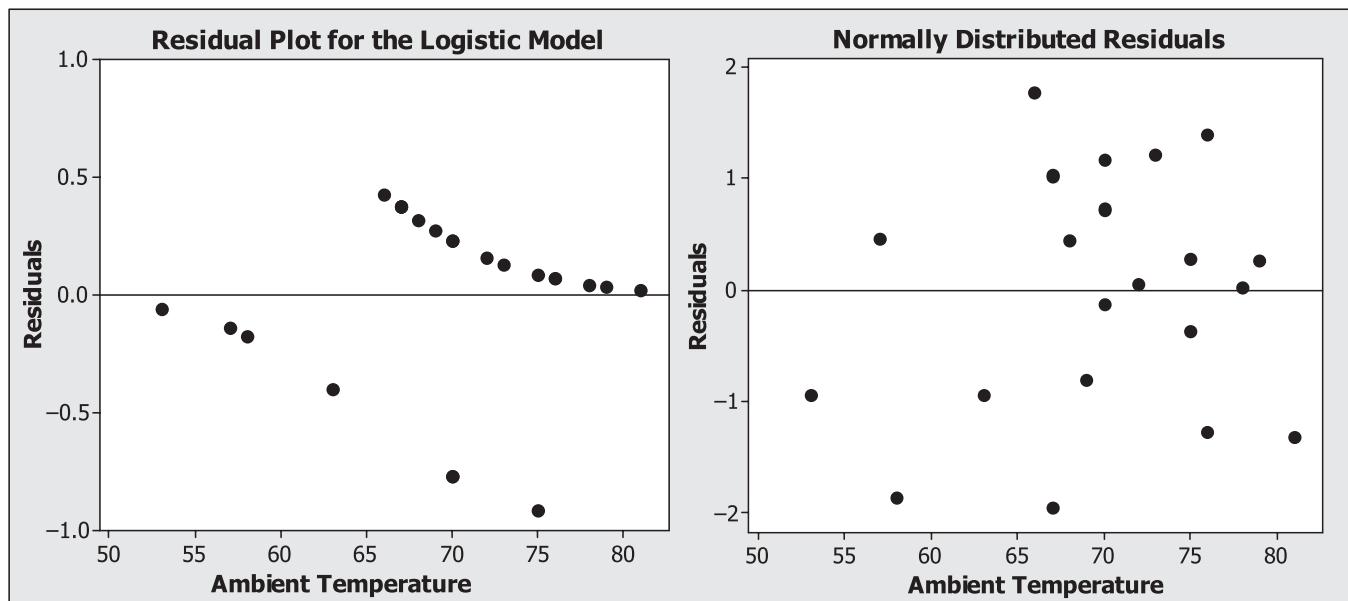
#### MATHEMATICAL NOTE

Binary logistic regression assumes that for each  $x_i$  value, the response variable ( $Y_i$ ) follows a **Bernoulli distribution** (described in the extended activities). This means we assume that (1) each  $Y_i$  is **independent**, (2) each response falls into **exactly one of two categories represented by either a zero or a one**, and (3) for each  $x_i$ ,  $P(Y_i = 1) = \pi_i$  and  $P(Y_i = 0) = 1 - \pi_i$  (more specifically written as  $P(Y_i = 1|x_i) = \pi_i$  and  $P(Y_i = 0|x_i) = 1 - \pi_i$ ). This third assumption states that for any given explanatory variable (a specific temperature value), the **probability of success (no O-ring failures)** is constant.

It is possible to use least squares regression techniques to estimate  $\beta_0$  and  $\beta_1$  in logistic regression models. However, the assumptions needed for hypothesis tests and confidence intervals using the ordinary least squares regression model are not met. Specifically, even after the log-odds transformation, Figure 7.3 demonstrates that the residuals are not normally distributed and the variability of the residuals depends on the explanatory variable.

Residuals (observed values minus expected values) are used to estimate the error terms. Visual inspection of residual plots is often used to check for normality. Recall from previous work in regression that if the residuals are normally distributed, the scatterplot of the residuals versus the explanatory variable should resemble a randomly scattered oval of points. For example, it should resemble the random scatter you would see if you happened to drop 23 coins (one for each residual value).

In logistic regression, the residuals are  $y_i - \hat{\pi}_i$ . If the observed response  $y_i = 0$ , then the residual value is  $-\hat{\pi}_i$ . If the observed response  $y_i = 1$ , then the residual value is  $1 - \hat{\pi}_i$ . This leads to the two curves shown in Figure 7.4. When the temperature is low in the space shuttle data (around 55°F, as seen



**Figure 7.4** A scatterplot of the residuals from the space shuttle logistic regression model and a sample of what a scatterplot of normally distributed residuals might look like.

in Figure 7.3), the observed responses tend to be zero and the predicted responses ( $\hat{\pi}_i$ s) are small positive numbers. Thus, the residual values ( $-\hat{\pi}_i$ s) are negative and close to zero. When the temperature is high, the observed responses tend to be one, the predicted responses are close to one, and the residual values are positive and close to zero.

### Key Concept

When the response in a regression model is binomial,  $\pi_i = P(Y_i = 1)$  = probability of a success (a launch has no O-ring damage at temperature  $x_i$ ). In simple linear regression models with a binomial response,

$$y_i = \pi_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for } i = 1, 2, 3, \dots, n \quad (7.7)$$

With the logit transformation, logistic regression models with a binomial response have the following form:

$$y_i = \pi_i + \varepsilon_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} + \varepsilon_i \quad \text{for } i = 1, 2, 3, \dots, n \quad (7.8)$$

While the logit transformation results in a nice S-shaped curve, the error terms in Equations (7.7) and (7.8) are not constant and are not normally distributed. Thus, hypothesis tests and confidence intervals cannot be calculated using least squares regression.

## 7.4 The Logistic Regression Model Using Maximum Likelihood Estimates

Logistic regression is a special case of what is known as a generalized linear model. Generalized linear models expand linear regression models to cases where the normal assumptions do not hold. All generalized linear models have three components:

- **A linear predictor:** In the shuttle example, there is only one explanatory variable,  $x$  = Temperature, so the linear predictor in Equation (7.5) is  $\beta_0 + \beta_1 x$ . However, just as in other multiple regression models, the linear predictor can include many explanatory variables, including indicator variables and interaction terms as well as other transformed variables.
- **A random component:** Each error term is assumed to be independent. However, in generalized linear models, the error terms are not required to follow a normal distribution. In addition, generalized linear models do not require that the variability of the error terms be constant.
- **A link function:** A link function is a function that fits the expected response value to a linear predictor. In Equation (7.5), the link function is the log-odds function,  $\ln[\pi/(1 - \pi)]$ . The link function depends on the distribution of the response variable. In logistic regression, the response variable is binary. Other link functions for binary response data do exist, but the log-odds function is the most common because it is the easiest to interpret.

Generalized linear models can also be used when the response variable follows other distributions. For example,  $y$  may follow a Poisson or gamma distribution. Textbooks on generalized linear models derive link functions for each of these types of response variables. In least squares regression where the response has a normal distribution, as in Equation (7.1), the link function is simply the identity function. In other words, the response needs no transformation in simple linear regression models.

Clearly the logistic regression model in Figure 7.3 is nonlinear. So it may seem somewhat surprising to consider logistic regression as a generalized linear model. The reason we still call this model linear is that the link function, the log-odds transformation, is modeled with a linear predictor,  $\beta_0 + \beta_1 x$ .

Instead of using least squares estimates, generalized linear models use the method of maximum likelihood to estimate the coefficients  $\beta_0$  and  $\beta_1$ . The extended activities provide more detail on calculating maximum likelihood estimates in logistic regression. In the space shuttle example, we will simply use a computer software package to find maximum likelihood estimates of  $\beta_0$  and  $\beta_1$ .

**NOTE**

In least squares regression, we often transform the response variable ( $Y$ ) so that the data fit model assumptions. In addition to linearizing data, transforming  $Y$  impacts the variability and the distribution of the error terms. In generalized linear models, the link function transforms the expected response ( $\pi$ ) to fit a linear predictor. For those who have had calculus, link functions are also differentiable and invertible.

### Activity ◀ Using Software to Calculate Maximum Likelihood Estimates

4. Solve Equation (7.5) for  $\pi_i$  to show that Equation (7.6) is true.
5. Use Equation (7.6) to create six graphs. In each graph, plot the explanatory variable ( $x$ ) versus the expected probability of success ( $\pi$ ) using  $\beta_0 = -10$  and  $-5$  and  $\beta_1 = 0.5, 1, \text{ and } 1.5$ . Repeat the process for  $\beta_0 = 10$  and  $5$  and  $\beta_1 = -0.5, -1, \text{ and } -1.5$ .
  - a. Do not submit the graphs, but explain the impact of changing  $\beta_0$  and  $\beta_1$ .
  - b. For all of these graphs, what value of  $\pi$  appears to have the steepest slope?
6. Use statistical software to calculate the maximum likelihood estimates of  $\beta_0$  and  $\beta_1$ . Compare the maximum likelihood estimates to the least squares estimates in Question 3.

Figure 7.3 shows a logistic regression model using maximum likelihood estimates of  $\beta_0$  and  $\beta_1$ . Using Equation (7.6) and the maximum likelihood estimates from Question 6, we can estimate the probability that a launch has no O-ring damage at temperature  $x_i$ :

$$\hat{\pi}_i = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}} = \frac{e^{-15.043 + 0.232x_i}}{1 + e^{-15.043 + 0.232x_i}} \quad \text{for } i = 1, 2, 3, \dots, n \quad (7.9)$$

Notice that  $\pi$  in Equation (7.6) has been replaced by  $\hat{\pi}$  in Equation (7.9) because the parameters in the linear regression model ( $\beta_0$  and  $\beta_1$ ) have been estimated with our sample data;  $b_0 = \hat{\beta}_0 = -15.043$  and  $b_1 = \hat{\beta}_1 = 0.232$ .

### Activity ◀ Estimating the Probability of Success with Maximum Likelihood Estimates

7. Use Equation (7.9) to predict the probability that a launch has no O-ring damage when the temperature is 31°F, 50°F, and 75°F.

At this point, it seems reasonable to question why the O-rings were not considered a higher risk at the time of the 1986 Challenger launch. After all, the odds of a successful launch (no O-ring damage) at the expected temperature of 31°F are about 1 to 2555 and the predicted odds change dramatically based on temperature. It is important to recognize that the previous launches did not result in the same disaster as the Challenger launch because the O-rings showed only “minor” damage. This wasn’t enough for gas to escape—only an indicator that the O-rings might not be as resilient as expected.

**CAUTION**

Estimating a value for a temperature of 31°F is extrapolating beyond our data set. Just as in least squares regression, caution should be used when making predictions outside the range of explanatory variables that are available.

## 7.5 Interpreting the Logistic Regression Model

Interpretation of logistic regression models is often done in terms of the **odds of success** (odds of a launch with no O-ring damage). When the temperature is 59°F, the odds of a successful launch with no O-ring damage are  $\hat{\pi}/(1 - \hat{\pi}) = 0.2066/(1 - 0.2066) = 0.2605 \approx 0.25 = 1/4$ . Thus, at 59°F, we state that the odds of a successful launch are about 1 to 4. When the temperature is 60°F, the odds of a successful

launch are  $\hat{\pi}/(1 - \hat{\pi}) = 0.3285 \approx 0.333 \approx 1/3$ . At 60°F, we state that the odds of a successful launch are about 1 to 3.

The slope is not as easy to interpret for a logistic regression model as for a simple linear regression model. While ordinary least squares regression focuses on  $b_1$ , logistic regression measures the change in the odds of success by the term  $e^{b_1}$ , which is called the **odds ratio**. If we increase  $x_i$  by 1 unit in a logistic regression model, the predicted odds that  $y = 1$  (i.e., the launch will not have any O-ring damage) will be multiplied by  $e^{b_1}$ . For example, when the temperature changes from 59° to 60°, the odds increase by a multiplicative factor of  $e^{b_1} = e^{0.232} = 1.2613$ . In other words,

$$\begin{aligned}\text{odds of success at } 59^\circ\text{F} \times e^{b_1} &= 0.2605(1.2613) \\ &= 0.3285 \\ &= \text{odds of success at } 60^\circ\text{F}\end{aligned}$$

For any temperature value  $x_i$ , this relationship can also be stated as

$$\text{odds ratio} = e^{b_1} = \frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)} \quad (7.10)$$

#### ► MATHEMATICAL NOTE ▼

Taking the exponent of Equation (7.5), we can write the odds of success as

$$\text{odds} = \left( \frac{\pi_i}{1 - \pi_i} \right) = e^{\beta_0 + \beta_1 x_i} = e^{\beta_0} (e^{\beta_1})^{x_i} \quad (7.11)$$

Thus, as  $x_i$  increases by 1,

$$e^{\beta_0} (e^{\beta_1})^{x_i+1} = e^{\beta_0} (e^{\beta_1})^{x_i} (e^{\beta_1}) \quad (7.12)$$

#### Key Concept

The slope in a logistic regression model is typically described in terms of the odds ratio  $e^{b_1}$ . If we increase  $x_i$  by 1 unit, the predicted odds will be multiplied by  $e^{b_1}$ . In our example, if the temperature increases by one degree, we increase the odds of a successful launch by  $e^{b_1} = e^{0.232} = 1.2613$  times. Similarly, if we decrease  $x_i$  by 1 unit, the predicted odds will be multiplied by  $e^{-b_1} = 1/e^{b_1}$ .

### Activity ◀ Interpreting a Logistic Regression Model

8. Calculate the odds of a launch with no O-ring damage when the temperature is 60°F and when the temperature is 70°F.
9. When  $x_i$  increases by 10, state in terms of  $e^{b_1}$  how much you would expect the odds to change.
10. The difference between the odds of success at 60°F and 59°F is about  $0.3285 - 0.2605 = 0.068$ . Would you expect the difference between the odds at 52°F and 51°F to also be about 0.068? Explain why or why not.
11. Create a plot of two logistic regression models. Plot temperature versus the estimated probability using maximum likelihood estimates from Question 6, and plot temperature versus the estimated probability using the least squares estimates from Question 3.

Thus far, we have developed a model to estimate the odds of a successful launch with no O-ring failures. However, we have not yet discussed the variability of the estimates or how confident we can be of the results. In the next section, we will discuss two hypothesis tests that can be used to determine if the odds of a successful launch are related to temperature. In other words, can we conclude that the logistic regression coefficient  $\beta_1$  is not equal to zero?

**Key Concept**

The probability, the odds, and the log-odds are three closely related calculations. Even though any of the three could be used to express the concepts of interest, the log-odds are often used to estimate the coefficients, while interpretation of logistic regression models typically relies on expected probabilities and odds because they are easier to interpret.

## 7.6 Inference for the Logistic Regression Model

### Assumptions for Logistic Regression Models

Inference for logistic regression uses statistical theory that is based on limits as the sample size approaches infinity. The techniques, based on what is called asymptotic theory, work well with large sample sizes, but are only approximate with outliers or small sample sizes. It is common for logistic regression models to be developed for data sets of any size, but savvy statisticians will always use caution when interpreting the results for data sets with small sample sizes (such as the space shuttle example).

### The Wald Statistic

**Wald's test** is often used to test the significance of logistic regression coefficients. Just as in least squares regression, we set up a hypothesis test to determine if there is a relationship between the explanatory and response variables:

$$H_0: \beta_1 = 0 \text{ vs. } H_a: \beta_1 \neq 0$$

Wald's test is similar to the one-sample Z-test seen in introductory statistics courses. The Wald statistic is calculated as

$$Z = \frac{b_1 - 0}{\text{se}(b_1)} = \frac{0.232}{0.108} = 2.14 \quad (7.13)$$

**NOTE**

Some texts use a chi-square statistic instead of the Z-statistic given in Equation (7.13). Most probability textbooks explain that the square of the test statistic in Equation (7.13) follows a chi-square distribution with 1 degree of freedom. Both techniques provide identical  $p$ -values.

where  $b_1$  is the maximum likelihood estimate of  $\beta_1$  and  $\text{se}(b_1)$  is the standard error of  $b_1$ . The maximum likelihood estimate,  $b_1$ , is asymptotically normally distributed (i.e.,  $b_1$  is normally distributed when the sample size is large). Thus, the Z-statistic in Equation (7.13) will follow a standard normal distribution when the null hypothesis is true and the sample size is large. In this model, we see that the estimated slope coefficient  $b_1 = 0.232$  has a  $p$ -value of  $P(|Z| \geq 2.14) = 0.032$ .

Wald confidence intervals can also be created. In logistic regression, the confidence interval is often discussed in terms of the odds ratio. For example, a 95% confidence interval for  $\beta_1$  is given as

$$\begin{aligned} (e^{(b_1 - Z^* \text{se}(b_1))}, e^{(b_1 + Z^* \text{se}(b_1))}) &= (e^{0.232 - 1.96(0.108)}, e^{0.232 + 1.96(0.108)}) \\ &= (e^{0.02}, e^{0.44}) \\ &= (1.02, 1.56) \end{aligned} \quad (7.14)$$

where  $1.96 = Z^*$  represents a value corresponding to a 95% confidence interval for a normal distribution with mean of 0 and standard deviation of 1. When  $\beta_1 = 0$ , and thus the odds ratio  $e^{\beta_1} = 1$ , the odds of success for temperature  $x_i$  are the same as the odds of success for any other temperature. Thus,  $e^{\beta_1} = 1$  tells us that there is no association between the explanatory variable and the response.

**Key Concept**

When a 95% Wald confidence interval for the odds ratio does not contain 1, we reject the null hypothesis  $H_0: \beta_1 = 0$  (using an  $\alpha$ -level of 0.05) and conclude that the odds of success do depend on the explanatory variable  $x_i$ . If the interval does contain 1, we fail to reject  $H_0: \beta_1 = 0$ .

The Minitab output in Figure 7.5 shows Wald's test and the corresponding confidence interval for the odds ratio. In the space shuttle example, the 95% confidence interval does not include  $e^{\beta_1} = 1$ ; thus, we can reject the null hypotheses and conclude that the odds of a successful launch do depend on the temperature. Even though computer software provided a small  $p$ -value and a confidence interval that does not include 1, it is important to note that there are only 23 observations in this study. While Wald's test is reasonable with very large sample sizes, with smaller sample sizes it is known to have a tendency to result in a type II error—failing to reject the null hypothesis when it should be rejected.<sup>5</sup>

We can also calculate the odds ratio of a successful launch between 60°F and 70°F. When  $x_i$  increases by 10°F, the odds are multiplied by  $(e^{\beta_1})^{10} = 1.2613^{10} = 10.19$ . Thus, we have approximately 10 times higher odds of a successful launch when the temperature is 70°F than when it is 60°F. A 95% confidence interval for the odds ratio of a successful launch between 60°F and 70°F can be given by  $(1.02^{10}, 1.56^{10}) = (1.22, 85.40)$ . This 95% confidence interval has a very wide range; the odds of success at 70°F could be just slightly larger than the odds at 60°F or 85 times as large as the odds at 60°F. This large range suggests that our estimate of the odds ratio, 10.19, is highly variable. More data are needed to better understand the true odds ratio.

### Activity ▶ Wald Confidence Intervals and Hypothesis Tests

12. Calculate the odds ratio of a successful launch between 31°F and 60°F. Provide a confidence interval for this odds ratio and interpret your results.
13. The coefficients in Equation (7.9) were calculated when a successful launch was given a value of 1. Conduct a logistic regression analysis where 1 indicates an O-ring failure and 0 represents a successful launch.
  - a. Explain any relationships between the model shown in Equation (7.9) and this new model.
  - b. How did the regression coefficients change?
  - c. How did the odds ratio change?
  - d. Create a 95% Wald confidence interval for the new odds ratio and interpret the results.

#### Binary Logistic Regression: O-Ring Failures versus Ambient Temperature

Link Function: Logit

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% Lower	CI Upper
Constant	-15.0429	7.37862	-2.04	0.041			
Ambient Temperature (F)	0.232163	0.108236	2.14	0.032	1.26	1.02	1.56
Log - Likelihood = -10.158							
Test that all slopes are zero: G = 7.952, DF = 1, P - Value = 0.005							

Figure 7.5 Minitab output for the space shuttle study.

\*Recall that  $Z$  in Equation (7.13) is a statistic calculated from the sample data and  $Z^*$  in Equation (7.11) is called a critical value.  $Z^*$  represents a value based on a desired level of confidence.  $Z^*$  is known before the data are collected, while  $Z$  is based on the sample data.

## The Likelihood Ratio Test

The **likelihood ratio test (LRT)** is derived by calculating the difference between the adequacy of the full and restricted log-likelihood models. A **full model** (sometimes called an **unrestricted model**) includes all parameters under consideration in the model. In this example, there are only two parameters,  $\beta_0$  and  $\beta_1$ , but the full model could include more parameters if more explanatory variables were in the model. The **restricted model** (also called a **reduced model**) is a model with fewer terms than the full model. In our example, only  $\beta_0$  is in the restricted model (no explanatory variables are in this model). When no explanatory variables are in the restricted model, the restricted model is also called a **null model**. If the full model has a *significantly better fit* (the expected values are closer to the observed values) than the restricted model, we reject the null hypothesis  $H_0: \beta_1 = 0$  and conclude that  $H_a: \beta_1 \neq 0$ .

The log-likelihood (restricted) function, described in the extended activities, is a measure of the fit of the model that includes only the intercept:

$$\text{Restricted Model: } \pi_i = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

For the restricted model, the null hypothesis is true and  $\pi_i$  is constant for any  $x$ -value. The log-likelihood (full) function measures the fit of the model that includes all of the parameters of interest:

$$\text{Full Model: } \pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

When the null hypothesis  $H_0: \beta_1 = 0$  is true, it can be shown that

$$G = 2 \times \text{log-likelihood (full model)} - 2 \times \text{log-likelihood (restricted model)} \sim \chi^2 \quad (7.15)$$

The **G-statistic** measures the difference between the fits of the restricted and full models. In essence, we are measuring how much better the fit is when an explanatory variable (temperature) is added to the logistic model. If the *p*-value corresponding to the G-statistic is small, the difference in fits is so large that it is unlikely to occur by chance, and thus we conclude that  $H_a: \beta_1 \neq 0$  (the fit of the full model is *significantly* better than that of the restricted model).

Degrees of freedom for the LRT equal the number of parameters in the full model minus the number of parameters in the restricted model. In our case, this is  $2 - 1 = 1$ . Different software packages will present this test in slightly different ways. In the Minitab output in Figure 7.5, the log-likelihood of the full model (-10.158) and the G-statistic (7.952) are provided.

Other statistics packages may not give the G-statistic, but they will give enough information so that the LRT can be calculated. The R output shown in Figure 7.6 gives the **null deviance** [ $K - 2 \times \text{log-likelihood (restricted model)}$ ] and the **residual deviance** [ $K - 2 \times \text{log-likelihood (full model)}$ ], where  $K$  is a constant value.

Note that

$$\begin{aligned} G &= 2 \times \text{log-likelihood (full model)} - 2 \times \text{log-likelihood (restricted model)} \\ &= \text{null (restricted model) deviance} - \text{residual (full model) deviance} \\ &= 7.952 \end{aligned}$$

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-15.0429	7.3786	-2.039	0.0415*
Temperature	0.2322	0.1082	2.145	0.0320*
Null deviance: 28.267 on 22 degrees of freedom				
Residual deviance: 20.315 on 21 degrees of freedom				

Figure 7.6 R-output for the space shuttle study.

## Activity ▶ The Likelihood Ratio Test

14. Use statistical software to calculate the LRT for the space shuttle data. Submit the  $p$ -value and state your conclusions.

The  $G$ -statistic is relatively large, indicating that we have some evidence to reject  $H_0: \beta_1 = 0$  and conclude that temperature is related to the odds of a successful launch with no O-ring damage. When sample sizes are large and there is only one explanatory variable in the model, the  $p$ -values will be approximately the same for the LRT test and Wald's test. In the space shuttle example, the LRT and Wald's test have somewhat different  $p$ -values.

The likelihood ratio test is more reliable and is often preferred over Wald's test for small sample sizes.<sup>6</sup> However, unlike the LRT, Wald's test can have one-sided alternative hypothesis tests as well as nonzero hypothesized values. It is difficult to determine the actual sample size needed for Wald's test or the LRT test to perform well. Some statisticians suggest a minimum sample size of 100 observations.<sup>7</sup> Thus, it is best to label each  $p$ -value as approximate when using these tests with a small sample size.

### Key Concept

With a large sample size, Wald's test and the likelihood ratio test provide accurate tests for  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$ . While the LRT test tends to be more reliable with smaller sample sizes, use caution when interpreting the results for data sets with **small sample sizes** (such as the space shuttle example).

## 7.7 What Can We Conclude from the Space Shuttle Study?

The space shuttle example is an observational study, since the launches were not “randomly assigned” to the temperature groups, so we cannot conclude solely from this data set that low temperatures *caused* O-ring damage. Wald's test and the likelihood ratio test both provided some evidence that the odds of a successful launch are related to temperature. However, a sample size of 23 is not large enough for us to be confident that the  $p$ -values are reliable. The logistic regression model provides some indication that the probability of a successful launch is related to temperature. Other information, such as scientists understanding that cold temperatures cause O-rings to be more brittle, also strengthens the conclusion.

### A Closer Look Logistic Regression

## 7.8 Logistic Regression with Multiple Explanatory Variables

Wolberg and Mangasarian developed a technique to accurately diagnose breast masses using only visual characteristics of the cells within the tumor.<sup>8</sup> A sample is placed on a slide, and characteristics of the cellular nuclei within the tumor, such as size, shape, and texture, are examined under a microscope to determine whether the cancer cells are benign or malignant. Benign tumors are scar tissue or abnormal growths that do not spread and are typically harmless. Malignant (or invasive) cancer cells are cells that can travel, typically through the bloodstream or lymph nodes, and begin to replace normal cells in other parts of the body. If a tumor is malignant, it is essential to remove or destroy all cancerous cells in order to keep them from spreading. If a tumor is benign, surgery is typically not needed and the harmless tumor can remain.

In Chapter 6, we used contingency tables with only two variables, cell shape and type, to better understand how to analyze two categorical variables. This section will describe the process of variable selection in logistic regression, using the radius and the concavity of cell nuclei to estimate the probability that a tumor is malignant. In this data set, radius is actually the average radius (in micrometers,  $\mu\text{m}$ ) of all visible cell

nuclei from a slide, but we will refer to this variable simply as the cell radius for the tumor. The concavity of the cell nuclei is an indicator of whether the visible cell nuclei from the sample have the nice round shape of typical healthy cells or whether cells appear to have grown in such a way that the perimeters of the cell nuclei tend to have concave points.

## Extended Activity

### Estimating the Probability of Malignancy in Cancer Cells

Data set: Cancer2

15. Create a logistic regression model using Radius and Concave as explanatory variables to estimate the probability that a mass is malignant.
  - a. Using Radius as the first explanatory variable,  $x_1$ , and Concave as the second explanatory variable,  $x_2$ , submit the logistic regression model. In other words, find the coefficients for the model
$$y_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}} + \varepsilon_i \quad \text{for } i = 1, 2, \dots, n$$
  - b. Submit the likelihood ratio test results, including the log-likelihood (or deviance) values.
  - c. Concave = 0 represents round cells and Concave = 1 represents concave cells. Calculate the event probability when Radius = 4 and the cells are concave. Also calculate the event probability when Radius = 4 and the cells are not concave.
16. Create a logistic regression model using only Radius as an explanatory variable to estimate the probability that a mass is malignant.
  - a. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values.
  - b. Calculate the event probability when Radius = 4.

When there are multiple explanatory variables in a logistic regression model, such as in the model created in Question 15, the likelihood ratio test compares the full model to the null model, which excludes the radius and concavity terms. Thus, the null hypothesis is that the coefficient corresponding to each of the explanatory variables is zero. In Question 15, the LRT is testing

$$H_0: \beta_1 = \beta_2 = 0 \text{ vs. } H_a: \text{at least one of the coefficients is not zero}$$

The  $G$ -statistic for this hypothesis test is 527.42 with  $3 - 1 = 2$  degrees of freedom (the number of parameters in the full model minus the number of parameters in the restricted model) and a corresponding  $p$ -value  $< 0.001$ . Thus, we can reject  $H_0$  and conclude that at least one of the explanatory variables is significantly related to the probability that the cells are malignant.

The coefficients in multiple logistic regression models are discussed in terms of the odds of success. Any coefficient ( $\beta_j$ ) indicates how the response will change corresponding to the  $j$ th explanatory variable, conditional on all other explanatory variables in the model. When the  $j$ th explanatory variable is increased by one unit, the odds of success will be multiplied by  $e^{\beta_j}$ . When an explanatory variable ( $x_j$ ) is binary, as Concave is,  $e^{\beta_j}$  represents the odds ratio between the two groups. However, just as in ordinary least squares regression with multiple explanatory variables, these coefficients are conditional on the other terms in the model.

## Extended Activity

### Interpreting Odds and Model Coefficients

Data set: Cancer2

17. Using the logistic model from Question 15, use the odds ratio to interpret the coefficient for concavity.
18. Create a logistic regression model using Radius and Concave as explanatory variables and considering benign cells a success (1 instead of 0). Use the odds ratio to interpret the coefficient for concavity. Compare this interpretation to the one in Question 17.

## 7.9 The Drop-in-Deviance Test

Figure 7.7 provides the expected probabilities for the model created in Question 15. The probability of a cell being malignant appears to depend on Radius. In addition, it appears that concavity is an important variable in the model, since concave cells tend to have a higher estimated probability of being malignant. In Chapter 3, **variable selection** is described as a process of determining which explanatory variables should be included in a regression model. Ideally, we would like the simplest model (i.e., the model with the fewest terms) that best explains the response (i.e., the model that has the smallest residuals). In this example, we want to determine if the model in Question 16 can estimate the probability of malignancy just as accurately as the slightly more complex model in Question 15. If the models have similar abilities to estimate the probability of malignancy, then we will prefer the simpler model.

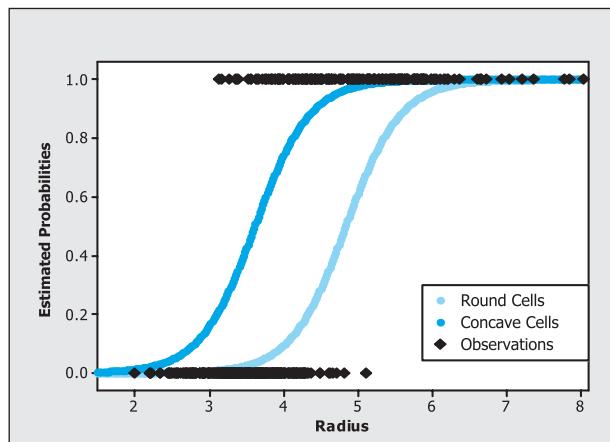
The logic for determining whether additional terms should be in a logistic regression model is essentially the same as for the LRT discussed in connection with the space shuttle study. The log-likelihood (or deviance) can be used to measure how well any model fits the data. If a full model with two explanatory variables,  $x_1$  and  $x_2$ , has a much better fit than a restricted model with just one explanatory variable,  $x_1$ , we can conclude that  $x_2$  is significant and should be included in the model.

If the LRT shows no significant difference between the full and the restricted model, the coefficient of the second variable,  $x_2$ , can be set to zero and we can conclude that including the additional variable in the model does not improve our ability to estimate the probability of success (in this case, success represents a malignant cell).

In the cancer study, we can compare the full log-likelihood (or deviance) from the two-term model in Question 15 to the restricted log-likelihood (or deviance) from the one-term model in Question 16.

$$\begin{aligned}
 G &= 2 \times \text{log-likelihood (full model)} - 2 \times \text{log-likelihood (restricted model)} \\
 &= 2(-112.008) - 2(-165.005) \\
 &= \text{null (restricted model) deviance} - \text{residual (full model) deviance} \\
 &= 105.994
 \end{aligned} \tag{7.16}$$

Just as in the LRT described in connection with the shuttle example, degrees of freedom are calculated as the number of parameters in the full model minus the number of parameters in the restricted model. Thus, we can test whether Concave ( $x_2$ ) should be included in the model by finding the  $p$ -value, which is the percentage of the  $\chi^2$  distribution with  $3 - 2 = 1$  degree of freedom that exceeds  $G$ . The  $p$ -value corresponding to Equation (7.16) is less than 0.0001. We have strong evidence that the explanatory variable, Concave, is important in the model when Radius is already included. Thus, the logistic regression model in Question 15 is preferred over the model in Question 16.



**Figure 7.7** A scatterplot of the observed data and estimated probabilities for both round cells (Concave = 0) and concave cells (Concave = 1).

When the LRT is used for variable selection, it is often called the **change-in-deviance test** or **drop-in-deviance test**. This test is valid only when the restricted model is nested within the full model. A restricted model is **nested** in a full model when every explanatory variable in the restricted model is also in the full model.

### Key Concept

To use the drop-in-deviance test to determine if  $x_j$  should be included in a model:

1. Calculate the deviance (or  $-2 \times \log\text{-likelihood}$ ) for the full (i.e., unrestricted) model with all variables of interest.
2. Calculate the deviance (or  $-2 \times \log\text{-likelihood}$ ) for the reduced (i.e., restricted) model (e.g., the model including all the variables in step 1 except for  $x_j$ ).
3. Calculate  $G = 2 \times \log\text{-likelihood (full model)} - 2 \times \log\text{-likelihood (restricted model)} = \text{deviance (restricted model)} - \text{deviance (full model)}$ .
4. Calculate the degrees of freedom, the number of parameters in the full model minus the number of parameters in the restricted model.
5. Find the  $p$ -value, the percentage of the  $\chi^2$  distribution that exceeds  $G$ .
6. If the  $p$ -value is small, reject  $H_0: \beta_j = 0$  and conclude that  $x_j$  should be included in the model. If the  $p$ -value is large, the explanatory variable,  $x_j$ , can be eliminated from the model.

### ► MATHEMATICAL NOTE

The drop-in-deviance can also be used to simultaneously test multiple variables. For example, let's assume that there were four shape measurements (round, slightly concave, moderately concave, and severely concave). These four levels would be used to create three indicator variables. If we are interested in testing whether Shape is important, we should test all three coefficients simultaneously. The steps in testing three coefficients simultaneously are identical to the steps listed above except that instead of just testing for  $x_j$ , we are simultaneously testing for  $x_2$ ,  $x_3$ , and  $x_4$ . Then the null hypothesis would be  $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ , where each of these coefficients corresponds to one Shape indicator variable in the full model.

Drop-in-deviance tests are often used in combination with stepwise regression techniques in order to identify the best model for prediction. In the end-of-chapter exercises, backward elimination procedures analogous to those used in multiple least squares regression will be used to find an appropriate model. The procedure starts with all explanatory variables of interest in the model. Variables that do not appear to be significant (or do not significantly reduce the size of the residuals) are removed in a stepwise process. The process is continued until all variables in the model are significant (or the model consists of only variables that are important in reducing the size of the residuals).

Just as in least squares regression, there is often no “best” multivariate logistic regression model. Different stepwise procedures, sampling variability, the desired balance between the accuracy and the simplicity of the model, and choice of terms tested all can influence the final model that is selected. While there should be careful justification for selecting a final model, caution should be used before stating that any selected model is “the best.”

### ► CAUTION

Recall that stepwise techniques are useful for developing models if the goal is to estimate or predict a response. However, they are not appropriate if the goal of developing a regression model is theory testing. Stepwise procedures involve multiple testing on the same variables. This leads to unreliable  $p$ -values when testing the coefficients of individual explanatory variables. Thus, it is inappropriate to use stepwise procedures to find a good model and then test each coefficient on your final model to determine if the individual explanatory variables are significant.

**NOTE**

Some software packages have programs that will automatically suggest a model for logistic regression (just as is done for stepwise procedures in least squares regression). While this chapter focuses only on the drop-in-deviance test, there are other techniques that can be used in variable selection.

$$\text{Akaike's Information Criterion (AIC)} = -2 \times \log\text{-likelihood} + 2 \times p$$

$$\text{Bayesian Information Criterion (BIC)} = -2 \times \log\text{-likelihood} + p \times \ln(n)$$

where  $p$  is the number of estimated parameters (the number of explanatory variables plus 1) and  $n$  is the sample size. Both the AIC and the BIC adjust for the number of parameters in the model and are more likely to select models with fewer variables than the drop-in-deviance test. Both techniques suggest choosing a model with the smallest AIC or BIC value.

## 7.10 Measures of Association

When sample sizes are small, a model may have a strong association (a clear pattern is visible) but not have enough evidence to show that the independent variable is significant. Conversely, if there were thousands of observations in a data set, a hypothesis test might conclude that an independent variable was significant even though there was only a very weak association. Thus, researchers typically report both significance tests and a measure of association when discussing results. While there is no widely accepted equivalent to  $R^2$  in logistic regression, this section will describe calculations that can be used to measure the strength of association.

To measure the strength of association in the space shuttle logistic regression model, pair each observed success with every observed failure. In the shuttle example, there are 16 successes and 7 failures; thus, there are  $16 \times 7 = 112$  pairs. For each pair, use the logistic regression model to estimate the probability of success for both the observed success and the observed failure. If the observation corresponding to a success has a higher estimated probability, the pair is called a **concordant pair**. If the observation corresponding to a success has a lower estimated probability, the pair is called a **discordant pair**. **Tied pairs** occur when the observed success has the same estimated probability as the observed failure.

- To find **Somers' D**, take the number of concordant pairs minus the number of discordant pairs and then divide by the total number of pairs.
- To find **Goodman-Kruskal gamma** (also called Goodman and Kruskal's gamma), take the number of concordant pairs minus the number of discordant pairs and then divide by the total number of pairs *excluding ties*.
- To find **Kendall's tau-a**, take the number of concordant pairs minus the number of discordant pairs and then divide by the total number of pairs of observations including pairs with the same response value.

If all possible pairs were concordant, then Somers'  $D$  would equal 1. If the model had no predictive power, we would expect half the pairs to be concordant and half to be discordant. This would correspond to Somers'  $D = 0$ . Thus, a value of 0 for Somers'  $D$  (as well as Goodman and Kruskal's gamma) indicates no effect of the explanatory variable on the response variable.

### Extended Activity

#### Measures of Association

Data set: `Shuttle`

19. The first and eleventh launches form a pair, since at 63°F there was an O-ring failure and at 66°F there was a success (no O-ring failure). This is a concordant pair, since the probability of success is higher when there was an observed success. Estimate the probability of success for each temperature.
20. Calculate the expected probabilities of the first (66°F) and 22nd (75°F) observations. Is this a concordant or discordant pair?
21. Identify two launches in the space shuttle data that represent a tied pair.
22. Various statistical software packages tend to provide different measures of association. Use statistical software to calculate the Goodman-Kruskal gamma, Somers'  $D$ , or Kendall's tau-a for the space shuttle data.

Minitab output for the space shuttle data is provided in Figure 7.8. The output shows that 75.9% of the pairs were concordant, while 19.6% of the pairs were discordant. This provides some evidence of association between temperature and probability of a successful launch.

$$\text{Somers' } D = (85 - 22)/112 = 0.56$$

$$\text{Goodman-Kruskal gamma} = (85 - 22)/(112 - 5) = 0.59$$

$$\text{Kendall's Tau-a} = (85 - 22)/(253) = 0.25$$

where  $253 = 23 \times 22/2$ , the total number of pairs

Somers'  $D$  and the Goodman-Kruskal gamma are very close to each other because there are very few tied pairs. There are no  $p$ -values corresponding to these measures, but they are useful for comparing different models with different explanatory variables or comparing models based on different link functions.

Measures of Association (Between the Response Variable and Predicted Probabilities)				
Pairs	Number	Percent	Summary Measures	
Concordant	85	75.9	Somers' D	0.56
Discordant	22	19.6	Goodman-Kruskal Gamma	0.59
Ties	5	4.5	Kendall's Tau-a	0.25
Total	112	100.0		

Figure 7.8 Minitab output showing measures of association for the space shuttle data.

## 7.11 Review of Means and Variances of Binary and Binomial Data

If you have worked with discrete probability models, you will recognize that binary data follow a Bernoulli distribution if the following conditions are true:

- Each observation,  $y_i$ , is independent.
- Each  $y_i$  falls into exactly one of two categories represented by either a zero or a one.
- The probability of success,  $P(Y_i = 1) = \pi_i$ , is constant for each observation.

The Bernoulli distribution can be displayed as in Table 7.2.

Table 7.2 is often represented with the following mathematical function to model the Bernoulli distribution:

$$P(Y = k) = \pi^k(1 - \pi)^{1-k} \quad \text{for } k = 0, 1 \quad (7.17)$$

The expected value (mean) of  $y$  is the average outcome:

$$E(Y) = 0 \times P(Y = 0) + 1 \times P(Y = 1) = 0 \times (1 - \pi) + 1 \times \pi = \pi \quad (7.18)$$

Note that each outcome is not equally likely. Thus, each outcome is weighted by its probability. The variance of  $y$  is the average value of the squared deviation of the observed value,  $y$ , and the expected value,  $\pi$ :

$$\text{Var}(Y) = E[(Y - \pi)^2] = (0 - \pi)^2 \times P(Y = 0) + (1 - \pi)^2 \times P(Y = 1) = \pi(1 - \pi) \quad (7.19)$$

Table 7.2 The Bernoulli model.

Value of $y$	0	1
Probability	$1 - \pi$	$\pi$

In the above equations,  $\pi$  is the same for every observational unit and the results for each observational unit are independent of each other. However, in regression we focus on the expected value of  $y$  given  $x$ ,  $E(Y|x_i)$ . This represents the expectation that the probability of success depends on an explanatory variable. In the space shuttle example, the expected probability of a successful launch will change depending on the temperature. For any given temperature value,  $x_i$ , the probability of success is constant,  $\pi_i$ , and the observations are independent. Thus, for a particular  $x_i$ , the corresponding mean and variance are

$$E(Y|x_i) = \pi_i \quad \text{and} \quad \text{Var}(Y|x_i) = \pi_i(1 - \pi_i) \quad (7.20)$$

Thus, in logistic regression with Bernoulli response variables, the variance of  $Y$  will depend on  $x$ . This violates the key assumption of constant variance in least squares regression models.

Logistic regression is also appropriate when the response is a count of the number of successes. A count follows a binomial distribution if the following conditions are true:

- There are  $n_i$  independent observations at a given level of  $x$  ( $x_i$ ).
- $\pi_i = P(Y_i = 1|x_i)$  is the probability of success, and this probability is constant for any given  $x_i$  ( $0 \leq \pi_i \leq 1$ ).
- Each response has only two possible outcomes. However, instead of recording a 0 or 1 value for each outcome, we typically record  $y$  as the count of successes (or proportion of successes) at a particular  $x_i$  value.

Many introductory textbooks show that if the data follow a binomial distribution, the probability that there are  $k$  successes in  $n_i$  independent observations is

$$P(Y = k|x_i) = \binom{n_i}{k} \pi_i^k (1 - \pi_i)^{n_i - k} \text{ for } k = 1, 2, \dots, n_i \quad (7.21)$$

where  $\binom{n_i}{k} = \frac{n_i!}{k!(n_i - k)!}$  is called the binomial coefficient and is read as “ $n_i$  choose  $k$ .”

The binomial coefficient counts the number of ways in which  $k$  successes can occur in  $n_i$  observations. For any integer  $k$ ,  $k!$  (read “ $k$  factorial”) is calculated as

$$k! = k(k - 1)(k - 2) \cdots (3)(2)(1) \quad \text{where } 0! = 1$$

Using a technique similar to that for the Bernoulli distribution, it can be shown that the conditional mean (probability of success) and variance of binomial response variables are

$$E(Y|x_i) = n_i \times \pi_i \quad \text{and} \quad \text{Var}(Y|x_i) = n_i \times \pi_i(1 - \pi_i) \quad (7.22)$$

## Extended Activity Understanding the Binomial Distribution

23. Assume that for a particular temperature  $x_i = 70^\circ\text{F}$  the true probability of success is  $\pi_i = 0.75$ . If there are four launches made at  $x_i = 70^\circ\text{F}$ , use Equation (7.21) to find the probability that all four launches are successful,  $P(Y_i = 4|x_i = 70)$ . Also find the probability that one of the four launches is successful,  $P(Y_i = 1|x_i = 70)$ .
24. When there are four observations ( $n_i = 4$ ) and  $\pi_i = 0.75$ , use the basic formula for calculating means

to find  $E(Y_i|x_i) = \mu_{x_i} = \sum_{k=0}^4 k \times P(Y_i = k|x_i)$ .

25. When  $n_i = 4$  and  $\pi_i = 0.75$ , find  $\text{Var}(Y_i|x_i)$ .

## 7.12 Calculating Logistic Regression Models for Binomial Counts

In the previous examples,  $y$  was considered a binary random variable (either 0 or 1). In this example, logistic regression will be used when the response is a count of the number of successes (i.e., the response is binomial). Table 7.3 shows the cancer cell data with the Radius variable grouped into five levels. Clearly grouping is not necessary for logistic regression, but this grouping is done to demonstrate how to conduct logistic regression when the response is binomial.

**Table 7.3** Cancer cell data classified into groups based on the size of the radius.

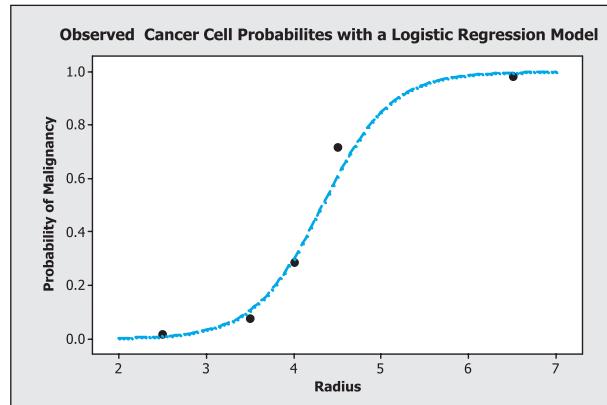
$i$	Median radius ( $x_i$ )	Benign ( $n_i - y_i$ )	Malignant ( $y_i$ )	Total ( $n_i$ )	Proportion Malignant
1	2.5	113	2	115	0.017
2	3.5	140	12	152	0.079
3	4	85	34	119	0.286
4	4.5	17	43	60	0.717
5	6.5	2	121	123	0.984
<b>Total</b>		<b>357</b>	<b>212</b>	<b>569</b>	

## Extended Activity ➔ Binomial Logistic Regression

Data set: Table 7.3

26. Create a logistic regression model based on Equation (7.8) to predict the probability of a malignant cell from the grouped radius data in Table 7.3.
- What are the maximum likelihood estimates  $b_0$  and  $b_1$ ?
  - Substitute  $b_0$  and  $b_1$  into Equation (7.9) to estimate the probability of malignancy when the radius is 4.5.
  - Use Wald's test and the  $G$ -statistic to determine if this model provides evidence that the probability of malignancy is related to cell radius.

Figure 7.9 plots the observed percentages of malignant cells and the corresponding logistic regression model. Notice that the observed and expected probabilities are fairly close. However, the observed percentage of malignant cells was higher than expected when the cell radius was 4.5.



**Figure 7.9** A logistic regression model, with  $\hat{\pi}_i$  estimated from Equation (7.8) using maximum likelihood estimates, plotted with the observed probability of malignancy for the grouped data in Table 7.3.

## 7.13 Calculating Residuals for Logistic Models with Binomial Counts

When the response variable follows a binomial distribution as in Table 7.3, Pearson residuals and deviance residuals are often calculated. These residuals are calculated to evaluate how well a model fits the data.

Using our estimate  $\hat{\pi}_i = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}}$  for  $\pi_i$ , we find the **Pearson residual** by taking the number of observed successes ( $y_i$ ) minus the estimated number of expected successes ( $n_i \times \hat{\pi}_i$ ) and then dividing by the estimated standard deviation:

$$\text{Pearson residual for the } i^{\text{th}} \text{ radius value} = \frac{y_i - n_i \times \hat{\pi}_i}{\sqrt{n_i \times \hat{\pi}_i \times (1 - \hat{\pi}_i)}} \quad (7.23)$$

Since the variance is not constant, each residual is weighted by its own estimated standard deviation.

#### NOTE

We can calculate the Pearson residual using count data because we know that for any individual radius,  $x_i$ , the response variable,  $y_i$ , follows a binomial distribution with an estimated mean  $n_i \times \hat{\pi}_i$  and variance  $n_i \times \hat{\pi}_i \times (1 - \hat{\pi}_i)$ .

### Extended Activity

#### Evaluating Residuals in Binomial Logistic Regression

Data set: Table 7.3

27. Create a logistic regression model to predict the probability of a malignant cell from the grouped radius data in Table 7.3.
  - a. Use software to calculate the Pearson residuals.
  - b. Create a histogram and/or a normal probability plot of these residuals. Are the residuals normally distributed?
  - c. Create a scatterplot of the Pearson residuals versus radius. How accurate are the estimates when radius is 2.5 and when radius is 4.5?
28. For each observation, the deviance residual measures the change in deviance from the full to the null model. Each deviance residual is the square root of the deviance goodness-of-fit statistic for each cell (or distinct radius value). The deviance residual is

$$D_i \pm \sqrt{2 \times \left[ y_i \times \ln\left(\frac{y_i}{n_i \hat{\pi}_i}\right) + (n_i - y_i) \times \ln\left(\frac{n_i - y_i}{n_i - n_i \hat{\pi}_i}\right) \right]} \quad (7.24)$$

where the sign is positive if the observed  $y_i$  is higher than the estimated mean and negative if the observed  $y_i$  is less than the estimated mean. Repeat Question 27 but use the deviance residual instead of the Pearson residual.

### 7.14 Assessing the Fit of a Logistic Regression Model with Binomial Counts

A goodness-of-fit test measures how well a model fits the observed data. We will discuss three goodness-of-fit tests for logistic regression based on the chi-square distribution. As discussed in Chapter 6, a chi-square goodness-of-fit test is an asymptotic test that measures the accumulated distance between observed values and expected values (i.e., values predicted by our model). In each of the three tests, the null and alternative hypotheses are

$H_0$ : the logistic regression model provides an adequate fit to the data

$H_a$ : the model does not adequately fit the data

If the predicted values in the model are relatively close to the observed data values (i.e., the model is a good fit for the data), then the test statistic will be small and the  $p$ -value will be large. If the test statistic is large, it suggests “lack of fit”;  $H_0$  should be rejected and other models should be tried to better fit the data.

**Key Concept**

Goodness-of-fit tests assess the overall fit of a logistic regression model. If  $H_0$  is rejected, the model is not a good fit. Failing to reject  $H_0$  does not guarantee that the model is a "best fit" or even a good fit, but rather that we simply don't have enough evidence to prove that it's a poor fit.

Test 1: The **Pearson chi-square goodness-of-fit test** is the traditional chi-square goodness-of-fit test seen in many introductory statistics courses. Degrees of freedom are calculated as the number of groups (classes) minus the number of parameters being estimated. In our case, groups are classified by median radius; we have five groups and are estimating two parameters ( $\beta_0$  and  $\beta_1$ ), so there are  $5 - 2 = 3$  degrees of freedom.

Test 2: The **deviance goodness-of-fit test** (also called the likelihood ratio chi-square test) is based on the sum of squared deviance residuals. The test statistic follows a chi-square probability distribution where the degrees of freedom are calculated as the number of groups minus the number of parameters being estimated. The Pearson test statistic and the deviance test statistic tend to be similar. For determining a model, the deviance goodness-of-fit test is preferred over the Pearson test.<sup>9</sup>

**NOTE**

The sum of squared Pearson residuals is equal to the Pearson chi-square test statistic. When there is one explanatory variable, the likelihood ratio test is equivalent to the deviance goodness-of-fit test. The residual deviance and the degrees of freedom given in R are identical to those given in the deviance goodness-of-fit test statistic in Minitab.

Test 3: Hosmer and Lemeshow developed a test similar to the Pearson chi-square goodness-of-fit test. The key distinction is that the groups are formed differently. While the Pearson test uses the *explanatory variable* to form groups, the **Hosmer-Lemeshow test** uses the *predicted values* to sort the data and form groups (the default is 10 groups). In Table 7.3, we predetermined the five groups, so the Pearson and the Hosmer-Lemeshow tests are identical in this example. However, when the explanatory variable is continuous (not grouped), the Hosmer-Lemeshow test is more reliable than the Pearson chi-square goodness-of-fit test.

**Extended Activity****Calculating Residuals by Hand**

Data set: Cancercells

29. Calculate the Pearson chi-square goodness-of-fit test by hand for the Cancercells data.
  - a. Complete Table 7.4 by using the logistic regression model from Question 26 to calculate the expected (predicted) cell counts.

**Table 7.4** Observed and expected values using the logistic regression model for the Cancercells data, where  $\hat{\pi}_i$  is estimated from the logistic regression model [ $\hat{\pi}_i = e^{b_0 + b_1 x_i} / (1 + e^{b_0 + b_1 x_i})$ ].

i	Median radius	Observed			Expected		
		Benign	Malignant	Total	Benign	Malignant	Total
1	2.5	113	2	115	115(1 - $\hat{\pi}_1$ ) = 113.967	115( $\hat{\pi}_1$ ) = 115(0.00898) = 1.033	115
2	3.5	140	12	152		152 $\hat{\pi}_2$ = 16.13	152
3	4	85	34	119			119
4	4.5	17	43	60			60
5	6.5	2	121	123			123
	<b>Total</b>	<b>357</b>	<b>212</b>	<b>569</b>	<b>357</b>	<b>212</b>	<b>569</b>

- b. Calculate the Pearson chi-square test statistic:

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \quad (7.25)$$

using the observed count and expected count from Table 7.4. Note that there are 10 observed and 10 expected cells.

- c. Find the  $p$ -value by using software or looking up the statistic in a chi-square probability distribution table.
- d. Interpret the results and clearly state how your conclusions are impacted by random sampling and random allocation to treatments in the original study design. Remember that a small  $p$ -value provides evidence that the model is *not* a good fit.
30. Calculate the deviance chi-square goodness-of-fit test by hand for the `Cancercells` data given in Table 7.3. Expected counts, degrees of freedom, and  $p$ -values are found the same way as in the Pearson test. The only difference is in the calculation of the test statistic:

$$\text{Deviance chi-square test statistic} = D^2 = 2 \times \sum \left[ \text{observed count} \times \ln \left( \frac{\text{observed count}}{\text{expected count}} \right) \right] \quad (7.26)$$



Describe differences (if any) between your conclusions here and those from the Pearson test in Question 29.

Figure 7.10 shows the Minitab output for the logistic regression model fit and the three goodness-of-fit tests.

Note that all three goodness-of-fit tests show small  $p$ -values, indicating that we can reject  $H_0$  and conclude that the model is *not a good fit*. In addition, the cell counts shown in Question 29 reveal that the sample size is large enough for us to believe the tests are reliable. Thus, other models should be tried. The logistic regression model shown in Figure 7.10 looks reasonable, but there are many possible explanations as to why the data are not considered a good fit:

- The groups based on median radius may not be appropriate. For example, radii of 1.51 and 2.49 were considered part of the same group. We could create more groups with smaller class sizes. Clearly we lose information whenever data are grouped into categories, so using the original data is likely the best option.
- Additional explanatory variables may need to be included in the model. Just as in ordinary regression, it may be appropriate to include interaction terms or transformations (such as the square, cube, or log) of the explanatory variable(s).
- The binomial model may not appropriately model the response variable or a transformation other than the logit transformation may need to be tried. *Notice that the logit model assumes symmetry; the curves in the S are the same shape and symmetric around the midpoint of the data.*

Binary Logistic Regression: Malignant, Benign versus median radius							
Link Function: Logit							
Logistic Regression Table							
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% Lower	CI Upper
Constant	-11.1359	1.14458	-9.73	0.000			
median radius	2.57286	0.284750	9.04	0.000	13.10	7.50	22.90
Log - Likelihood = -173.258							
Test that all slopes are zero: G = 404.924, DF = 1, P - Value = 0.000							
Goodness-of-Fit Tests							
Method	Chi-Square	DF	P				
Pearson	10.3559	3	0.016				
Deviance	8.0215	3	0.046				
Hosmer-Lemeshow	10.3559	3	0.016				

Figure 7.10 Minitab output for the `Cancercells` study.

- A few outliers may be significantly influencing the results. For example, when the median radius is 4.5, the observed probability is not very close to the expected value. This point greatly contributes to the large chi-square statistics in Questions 29 and 30.

Since chi-square tests are based on asymptotic theory, the  $p$ -values are not reliable unless there is a large enough sample size. A general rule for analyzing a  $2 \times 2$  table of counts is that the expected count for each cell must be at least 5. For larger tables, the expected counts for all cells must be at least 1, and the average of expected cell counts should be greater than 5.

When the data are **sparse** (the observed, and therefore the expected, cell counts are too small), chi-square tests may tend to fail to reject the null hypothesis which states that the model is a good fit for the data.<sup>10</sup> In other words, these tests may not have good power for detecting particular types of lack of fit. The Hosmer-Lemeshow test is designed to correct for this problem (when there are continuous explanatory variables) by grouping the data. Hosmer and Lemeshow suggest that you have a sample size of at least 400 before using their test.<sup>11</sup>

## Extended Activity

### Goodness-of-Fit Tests for Continuous Explanatory Variables

Data set: Cancer2

31. Use computer software to create a logistic regression model to predict the probability of a malignant cell using the continuous variable `Radius` as the explanatory variable. Conduct the three goodness-of-fit tests for this model.

The goodness-of-fit tests in Question 31 look very different from those calculated with the grouped data in Figure 7.10. Remember that for goodness-of-fit tests, the degrees of freedom are calculated as the number of groups minus the number of parameters being estimated. There are two parameters estimated, and the stated 454 degrees of freedom for the Pearson and deviance tests indicate that 456 groups were formed (based on each distinct radius value given in the data). Since there are 569 observations in this data set, the Pearson and deviance goodness-of-fit tests do not meet the sample size requirements for a reliable test (most of the cells have only one observation).

In Question 31, the Pearson and deviance goodness-of-fit tests fail to reject  $H_0$ : the logistic regression model provides an adequate fit to the data. However, the violation of the sample size requirement makes these tests inappropriate to use.

The Hosmer-Lemeshow test in Question 31 has only 8 degrees of freedom, since the default is to form 10 groups based on the fitted values. To form the groups, statistical software sorts the estimated probabilities and then attempts to create 10 groups of equal size. The observed and expected values are given in the software output. Notice that there are still two cells with expected values less than 1. Recent studies have shown that the Hosmer-Lemeshow test is somewhat sensitive to the way groups are formed, and other more specific tests have been developed.<sup>12</sup> However, since the  $p$ -value is much bigger than  $\alpha = 0.05$ , there does not appear to be strong evidence that the model is not a good fit.

Hosmer and Lemeshow state that slight violations of the sample size requirements are acceptable.<sup>13</sup> They suggest that if there is a sample size concern, you simply group a few columns. For example, in Question 31, grouping columns 1 and 2 and grouping columns 9 and 10 would satisfy the sample size requirements.

## Extended Activity

### Goodness-of-Fit Tests for Continuous Explanatory Variables

Data set: Cancer2

32. Conduct the Hosmer-Lemeshow goodness-of-fit test again, but this time adjust the group size (or number of groups) so that the sample size requirement is not violated. Did the  $p$ -value of this new Hosmer-Lemeshow test change your conclusions?

#### Key Concept

The Pearson, deviance, and Hosmer-Lemeshow tests assess model fit with chi-square tests. When one or more explanatory variables are continuous (which is often the case), the deviance and Pearson chi-square tests are not useful because there are not enough observations in each observed and expected cell (the number of distinct values of the explanatory variable is nearly equal to the number of observations). The Hosmer-Lemeshow test can be used with continuous explanatory variables because it uses the predicted values to group the data.

## 7.15 Diagnostic Plots

Just as in least squares regression, residual plots are useful in understanding logistic models. Goodness-of-fit tests are useful, but, as stated earlier, they tend to fail to reject the null hypothesis even when the model is not appropriate. When these chi-square tests fail to conclude that a model is inappropriate, residual plots can be used to verify that the model fit is appropriate. In addition, if the goodness-of-fit tests do reject the null hypothesis (conclude that the model is not appropriate), residual plots can help identify where the issues of model fit occur.

Large residual values are useful in identifying observations that are not explained well by the model. In addition to residuals, several scatterplots can be used to identify outliers and influential observations. Before creating these plots, we will need to define a few additional terms.

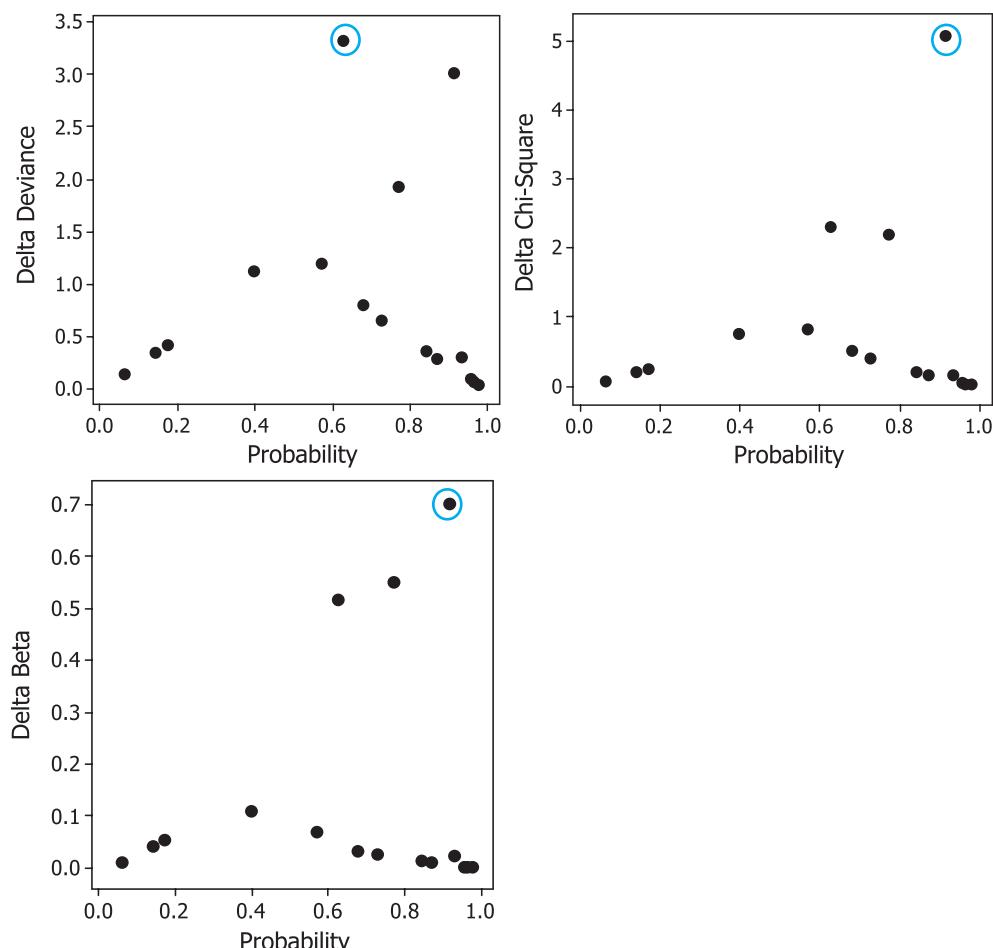
A **covariate pattern** is a set of all observations with identical explanatory variables. In the space shuttle example, there are four observations where the temperature is 70°F. These four observations form a covariate pattern. In the Cancer2 data set, a covariate pattern is a group of observations with both the same `Radius` value and the same `Concave` value.

**Delta chi-square** ( $\Delta\chi^2$ ) is a measure of the change in the Pearson goodness-of-fit statistic ( $\chi^2$ ) when a particular observation (or covariate pattern) is eliminated. In other words, for a particular  $x_i$  value, delta chi-square is  $\Delta\chi_i^2 = \chi^2 - \chi_{(i)}^2$ , where  $\chi_{(i)}^2$  is the Pearson goodness-of-fit statistic with the  $i$ th observation (or covariate pattern) eliminated.

**Delta deviance** ( $\Delta D^2$ ) is a measure of the change in the deviance goodness-of-fit statistic ( $D^2$ ) when a particular observation (or covariate pattern) is eliminated. In other words, for a particular  $x_i$  value, delta deviance is  $\Delta D_i^2 = D^2 - D_{(i)}^2$ , where  $D_{(i)}^2$  is the deviance goodness-of-fit statistic with the  $i$ th observation (or covariate pattern) eliminated.

**Delta beta** measures the difference in the regression coefficient when a particular observation (or covariate pattern) is removed.

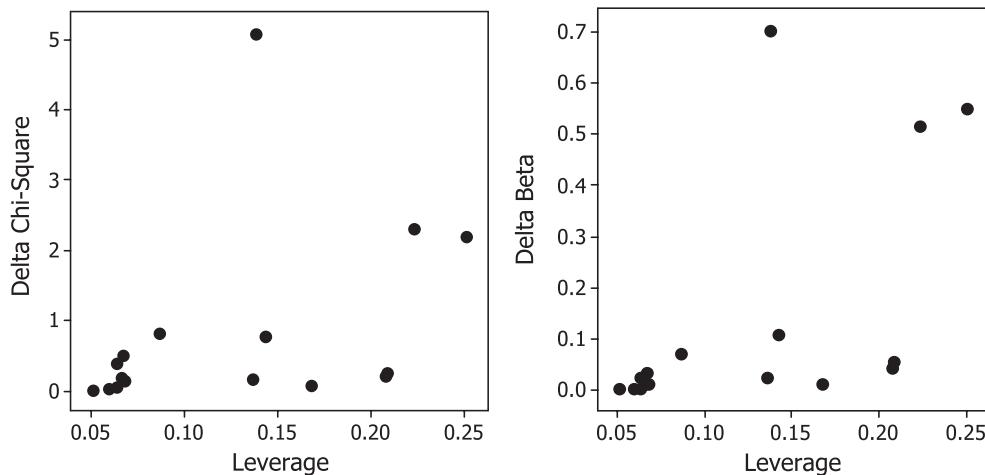
Figure 7.11 shows the delta chi-square, delta deviance, and delta beta values plotted against the expected probabilities ( $\hat{\pi}_i$ ). The determination as to whether or not an observation (covariate pattern) is an outlier or



**Figure 7.11** Scatterplots of delta deviance, delta chi-square, and delta beta values versus the expected probabilities from the space shuttle data. Circled values represent launches at 70°F.

overly influential is somewhat subjective. Outliers typically appear as extreme values in the upper corners of the scatterplot. As a rough estimate,  $\Delta\chi^2$  or  $\Delta D^2$  greater than 4 and delta beta values greater than 1 may be considered unusual observations. With large sample sizes,  $\Delta\chi^2$  and  $\Delta D^2$  approximately follow the chi-square distribution, and the 95th percentile of the chi-square distribution with 1 degree of freedom equals 3.84.<sup>14</sup> Figure 7.11 has one covariate pattern that has a fairly large  $\Delta\chi^2$  and  $\Delta D^2$  values. It corresponds to the two launches that occurred at 75°F. Notice in Figure 7.11 that a failure at 75°F appears to be somewhat unusual. In this example, the  $\Delta\chi^2$  and  $\Delta D^2$  values are not extreme enough to be of major concern (all values are close to or less than 4).

Figure 7.12 shows the delta chi-square and delta beta values plotted against the leverage values. Leverages are values between 0 and 1 that depend only on the explanatory variables (not the response). Large values indicate that the observation (covariate pattern) has extreme values and may have a large influence on regression coefficients.



**Figure 7.12** Scatterplots of delta chi-square and delta beta values versus leverage from the space shuttle data. The extreme value along the y-axis represents the launches at 75°F. There do not appear to be any extreme leverage values.

### Key Concept

The contribution of a single observation depends on both its residual and leverage. The delta chi-square and delta deviance values can be used to detect observations that have a strong influence on the goodness-of-fit statistics. A large delta beta value indicates a covariate pattern with large leverage and/or large residual values.

## Extended Activity

### Identifying Outliers and Influential Observations

Data set: Cancer2

33. Run a logistic regression model with both explanatory variables Radius and Concave.
  - a. Create histograms of the standardized Pearson residuals and deviance residuals.
  - b. Create scatterplots of delta deviance, delta chi-square, and delta beta (standardized) versus the expected probabilities.
  - c. Create scatterplots of delta deviance, delta chi-square, and delta beta (standardized) versus the leverage.
  - d. Identify any observations (covariate patterns) that appear to be outliers or influential observations.

## 7.16 Maximum Likelihood Estimation in Logistic Regression\*

Maximum likelihood estimation is a fairly complex topic. The goal of this section is simply to provide an example of how to calculate maximum likelihood estimates with binomial data. After completing this section, you should have a better understanding of how the LRT and the deviance test are calculated. However, as stated earlier, maximum likelihood estimates are very computationally intensive and are best left to computer algorithms.

### MATHEMATICAL NOTE

When the assumptions about the residuals in the least squares regression model described in Section 7.2 are satisfied, the maximum likelihood estimates of  $\beta_0$  and  $\beta_1$  are identical to the least squares estimates.

### Maximum Likelihood Estimator for Binary Data

Consider a set of independent binary responses  $y_1, y_2, \dots, y_n$ . Since each observed response is independent and follows the Bernoulli distribution shown in Equation (7.17), the probability of a particular outcome can be found as

$$\begin{aligned} P(Y_1 = k_1, Y_2 = k_2, \dots, Y_n = k_n) &= P(Y_1 = k_1)P(Y_2 = k_2) \cdots P(Y_n = k_n) \\ &= \pi^{k_1}(1 - \pi)^{1-k_1}\pi^{k_2}(1 - \pi)^{1-k_2} \cdots \pi^{k_n}(1 - \pi)^{1-k_n} \\ &= \pi^{\sum_{i=1}^n k_i} (1 - \pi)^{\sum_{i=1}^n (1-k_i)} \end{aligned} \quad (7.27)$$

where  $k_1, k_2, \dots, k_n$  represent a particular observed series of 0 or 1 outcomes and  $\pi$  is a probability,  $0 \leq \pi \leq 1$ . Once  $k_1, k_2, \dots, k_n$  have been observed, they are fixed values. **Maximum likelihood estimates** are functions of sample data that are derived by finding the value of  $\pi$  that maximizes the likelihood function. For a given observed data set  $y_1, y_2, \dots, y_n$ , when Equation (7.27) is a function of  $\pi$ , it is called the **likelihood function** and denoted  $L(\pi)$ .

### NOTE

Equation (7.27) is often called a joint probability function when considered as a function of the data. However, when the data are assumed to be fixed and Equation (7.27) is considered a function of  $\pi$ , it is called a likelihood function.

The **maximum likelihood estimate** is the value of  $\pi = P(Y = 1)$  that maximizes Equation (7.27). For simplicity, it is common to find the value of  $\pi$  that maximizes the log of the likelihood function. Recall that the value of  $\pi$  that maximizes the likelihood function,  $L(\pi)$ , will also maximize the log-likelihood function,  $\ln L(\pi)$ .

$$\begin{aligned} \ln L(\pi) &= \ln\left(\pi^{\sum_{i=1}^n k_i} (1 - \pi)^{\sum_{i=1}^n (1-k_i)}\right) \\ &= \sum_{i=1}^n k_i \ln(\pi) + (n - \sum_{i=1}^n k_i) \ln(1 - \pi) \end{aligned} \quad (7.28)$$

### Key Concept

The principle of maximum likelihood estimation is to choose a value of  $\pi$  such that the observed data set is most likely to occur (i.e., the likelihood function is maximized).

To find the maximum value of  $\ln L(\pi)$ , we take the derivative of  $\ln L(\pi)$  and set the first derivative equal to 0:

$$\frac{d[\ln L(\pi)]}{d\pi} = \sum_{i=1}^n k_i \frac{1}{\pi} + (n - \sum_{i=1}^n k_i) \frac{-1}{(1 - \pi)} = 0 \quad (7.29)$$

\*Calculus required.

Then we solve the following equivalent equation in terms of  $\pi$ :

$$(1 - \pi) \sum_{i=1}^n k_i - \pi(n - \sum_{i=1}^n k_i) = 0 \quad (7.30)$$

This provides the maximum likelihood estimator of  $\pi = P(Y = 1)$ :

$$\hat{\pi} = \frac{\sum_{i=1}^n k_i}{n} \quad (7.31)$$

In this example, the maximum likelihood estimator is the same as our well-known frequentist approach to estimating  $\pi = P(Y = 1)$ . However, this is not always the case.

#### MATHEMATICAL NOTE

The second derivative of  $\ln L(\pi)$  can also be calculated to show that the function is concave down. Thus,  $\hat{\pi}$  is a local maximum and not a local minimum. Likelihood functions with just one parameter typically have only one critical value, which is the maximum value. When more than one parameter is involved, more care needs to be taken to ensure that the critical values are actually maximum likelihood estimates.

### Extended Activity Calculating Maximum Likelihood Estimates

34. For binomial response data that follow the binomial distribution, the log-likelihood function for *one observation* ( $y_1$ ) is

$$\begin{aligned} \ln[P(Y_1 = k)] &= \ln \left[ \binom{n_1}{k} \pi_1^k (1 - \pi_1)^{n_1 - k} \right] \\ &= C + k \ln(\pi_1) + (n_1 - k) \ln(1 - \pi_1) \end{aligned} \quad (7.32)$$

Where  $C$  is a constant that does not influence  $\pi_1$ , we can drop  $C$  from the above equation without any impact on the maximum likelihood estimate. Assume that you observed  $y_1 = 5$  malignant cells out of a sample of  $n_1 = 12$ . Substituting these values into Equation (7.32) and dropping  $C$  simplifies the log-likelihood function to  $\ln [P(Y_1 = 5)] = 5 \ln(\pi_1) + 7 \ln(1 - \pi_1)$ .

- Plot the log-likelihood function for several values of  $\pi_1$  between 0 and 1. Use the plot to estimate the value of  $\pi_1$  that will maximize the log-likelihood function.
- If you have had calculus, set the derivative of the log-likelihood function to 0 and solve for  $\pi_1$ . What is the maximum likelihood estimate for  $\pi_1$ ?

### Maximum Likelihood Estimator for Logistic Regression Models

The previous example did not address cases in logistic regression where the observations depend on one or more explanatory variables. To use maximum likelihood estimation in logistic regression, from Equation (7.6) we see that

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Replacing this value for  $\pi_i$  in Equation (7.28) gives

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^n k_i \ln \left( \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) + (n - \sum_{i=1}^n k_i) \ln \left[ 1 - \left( \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \right] \quad (7.33)$$

To find the maximum likelihood estimates of  $\beta_0$  and  $\beta_1$ , take the derivative of Equation (7.33) with respect to  $\beta_0$  and respect to  $\beta_1$ . This will provide two equations and two unknowns. However, the two equations will not be linear and cannot be solved directly.

### ► MATHEMATICAL NOTE ▾

Finding estimates for  $\beta_0$  and  $\beta_1$  that maximize the log-likelihood function is an iterative technique that quickly becomes complex even for small data sets and is not typically done by hand. Iterative techniques start with initial estimates of  $\beta_0$  and  $\beta_1$ . An iterative technique such as the Newton-Raphson method repeatedly provides new estimates for  $\beta_0$  or  $\beta_1$  that increase the log-likelihood until the log-likelihood does not notably change.<sup>15</sup>

## Chapter Summary

Throughout this chapter, we have discussed how to conduct logistic regression for binary response data. When the observed response variable is binary,  $y_i = 1$  typically represents a success (or outcome of interest) and  $y_i = 0$  represents a failure. In the logistic regression model, the estimated response is typically defined as the **log-odds** of the probability of a success:

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_i \quad \text{for } i = 1, 2, \dots, n$$

where  $\pi_i = E(Y_i|x_i) = P(Y_i = 1)$ . Solving the above equation for  $\pi_i$  shows that the **probability of success** can be calculated with the following equation:

$$\hat{\pi}_i = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}}$$

where  $b_0$  and  $b_1$  are **maximum likelihood estimates** of the regression coefficients. While the logit transformation can create a nice S-shaped curve, the model assumptions for least squares regression are violated. *Thus, hypothesis tests and confidence intervals should not be calculated using least squares regression.* For hypothesis testing, both logistic and least squares regression assume a linear predictor and independent observations. In addition, outliers and highly correlated explanatory variables can influence hypothesis test results in both logistic and least squares regression. However, logistic regression does not assume that the error terms are normally distributed or have equal variances. Logistic regression hypothesis tests are based on asymptotic tests and require large sample sizes.

**Wald's test** and the **likelihood ratio test** can be used to test the significance of individual explanatory variables. If Wald's test and the likelihood ratio test provide different results, the results of the likelihood ratio test should be used. However, the likelihood ratio test can only test whether the slope is equal to zero ( $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$ ) while Wald's test allows us to test for any value of the slope and also allows for one-sided hypothesis tests and confidence intervals.

Goodness-of-fit tests, such as the **Pearson chi-square test**, the **deviance test**, and the **Hosmer-Lemeshow test**, can be used to assess how well the model fits the data. Only the Hosmer-Lemeshow test should be used when an explanatory variable is continuous. Even though there is no widely accepted equivalent to  $R^2$  in logistic regression, **Somers' D**, **Goodman-Kruskal gamma**, and **Kendall's tau-a** are used to measure the strength of association by means of a classification table showing correct and incorrect classifications of the response variable.

The slope in logistic regression is typically described using the **odds ratio**,  $e^{b_1}$ . If we increase the explanatory variable by one unit, the predicted odds will be multiplied by  $e^{b_1}$ . If the explanatory variable is also categorical,  $e^{b_1}$  is interpreted as the odds ratio between the two groups.

**Variable selection** is a process of determining which explanatory variables should be included in a regression model. We prefer to select the model with the fewest number of terms that still best explains the response. The **drop-in-deviance test** compares the log-likelihood (or deviance) of a full model (a model with several terms) and a restricted model (a model with a smaller subset of the terms from the full model). If the test shows no significant difference between the full and the restricted model, we conclude that the additional terms in the full model can be eliminated. The drop-in-deviance test can be used sequentially until all terms that do not impact the deviance have been removed from the model.

## Exercises

---

### E.1. Bird Nest Study

Data set: `Birdnest`

The file `Birdnest` contains data for 99 species of North American passerine birds. Passerine are “perching birds” and include many families of familiar small birds (e.g., sparrows and warblers), as well as some larger species like crows and ravens, but do not include hawks, owls, water fowl, wading birds, and woodpeckers. One hypothesis of interest was about the relationship of body size to type of nest. Body size was measured as average length of the species. Although nests come in a variety of types (see the `Nesttype` variable), in this data set nest type was categorized into either closed or open. “Closed” refers to nests with only a small opening to the outside, such as the tree cavity nest of many woodpeckers or the pendant-style nest of an oriole. “Open” nests include the cup-shaped nest of the American robin. (Note: `closed? = 1` for closed nests; `closed? = 0` for open nests.)

- a. Create a logistic regression model using `bird length (Length)` to estimate the probability that a bird species has a closed nest type. Interpret the model in terms of the odds ratio.
- b. Use the Wald statistic to create a 95% confidence interval for the odds ratio.
- c. Test  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  using both Wald’s test and the likelihood ratio test. State your conclusion based on these tests.
- d. Conduct the Hosmer-Lemeshow test to assess how well the model fits the data.
- e. Explain why the Pearson chi-square test is not reliable for this example.
- f. Report and interpret Somers’  $D$ , Goodman-Kruskal gamma, or Kendall’s tau-a for this logistic regression model.

### E.2. Survival of the Donner Party: Logistic Regression and Chi-Square Tests

Data set: `Donner`

In 1846, a group of 87 people (called the Donner Party) were heading west from Springfield, Illinois, for California.<sup>16</sup> The leaders attempted a new route through the Sierra Nevada and were stranded there throughout the winter. The harsh weather conditions and lack of food resulted in the death of many people within the group. Social scientists have used the data to study the theory that females are better able than men to survive harsh conditions.

- a. Create a logistic regression model using `Gender` to estimate the probability of `Survival`.
- b. Interpret the model in terms of the odds ratio. Use the Wald statistic to create a 95% confidence interval for the odds ratio.
- c. Calculate and interpret the likelihood ratio test.
- d. Test  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  using both Wald’s test and the likelihood ratio test. State your conclusion based on these tests.
- e. Report and interpret Somers’  $D$ , Goodman-Kruskal gamma, or Kendall’s tau-a for this logistic regression model.
- f. Create a two-way contingency table using `Gender` and `Survival` as row and column variables. Conduct a chi-square test for equal proportions (e.g., is the proportion of survival the same for males and females?). In addition, use this two-way table to create the odds ratio. How does the analysis of the two-way table compare to the logistic regression analysis?
- g. D. K. Grayson states, “The differential fate of the members of the Donner Party lends strong support to the argument that females are better able than males to withstand conditions marked by famine and extreme cold.” While there is some evidence that `Gender` is associated with `Survival`, explain why the data cannot be used to show that being female *causes* a higher probability of survival.

### E.3. Drug Treatment and Criminal Conviction: Logistic Regression and Two-Way Tables

Data set: `Convict`

A study was conducted to determine if a relationship existed between criminal conviction and years of education.<sup>17</sup> Sixty people who had taken part in a drug rehabilitation program were classified by years of education. Each person was categorized as having a “short” education (15 years or less) or

a “long” education (more than 15 years); also recorded was whether or not they had a post-treatment conviction.

- Create a logistic regression model using years of education to estimate the probability of conviction. Interpret the model in terms of the odds ratio.
- Interpret the results of Wald’s test and the LRT.
- Conduct Fisher’s exact test and a chi-square test of independence (discussed in Chapter 6) using the `Convict` data. How do these tests compare to the logistic regression model?
- Report and interpret Somers’  $D$ , Goodman-Kruskal gamma, or Kendall’s tau-a for this logistic regression model.

#### E.4. Severe Idiopathic Respiratory Distress

Data set: `SIRDS`

A study was conducted to determine if a relationship existed between infant birthweight and the likelihood of that infant surviving severe idiopathic respiratory distress syndrome (SIRDS).<sup>18</sup> Data from 50 infants who displayed this syndrome are in the `SIRDS` file.

- Create a logistic regression model using birthweight to estimate the probability of survival. Interpret the model in terms of the odds ratio.
- Use the Wald statistic to create a 90% confidence interval for the odds ratio.
- Test  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  using both Wald’s test and the likelihood ratio test. State your conclusion based on these tests.
- Conduct the Hosmer-Lemeshow test to assess how well the model fits the data. Ensure that you have appropriate sample sizes in each group.
- Explain why the Pearson and deviance goodness-of-fit tests are not appropriate.
- Report and interpret Somers’  $D$ , Goodman-Kruskal gamma, or Kendall’s tau-a for this logistic regression model.
- Create histograms of the standardized Pearson residuals and deviance residuals.
- Create scatterplots of delta deviance, delta chi-square, and delta beta (standardized) versus the expected probabilities.
- Create scatterplots of delta deviance, delta chi-square, and delta beta (standardized) versus the leverage.
- Identify any observations (covariate patterns) that appear to be outliers or influential observations.

#### E.5. Tattoos: The Drop-in-Deviance Test

Data set: `Tattoos`

Lunn and McNeil show a data set comparing two methods of surgical tattoo removal.<sup>19</sup> In addition to which method was used, the depth of the tattoo and the patient’s gender were recorded. In this data set, Removal = 1 represents a successful removal and Removal = 0 represents a poor removal.

- Create a logistic regression model using `Method`, `Gender`, and `Depth` to estimate the probability that a tattoo was removed successfully. Based on Wald’s test, which variables appear to be significant? Based on the likelihood ratio test ( $H_0: \beta_1 = \beta_2 = \beta_3 = 0$  versus  $H_a: \text{at least one of the coefficients is not zero}$ ), can we conclude that at least one of these terms is significant?
- Create a logistic regression model using only `Gender` and `Depth` to estimate the probability that a tattoo was removed successfully. Based on Wald’s test, which variables appear to be significant? Based on the likelihood ratio test ( $H_0: \beta_1 = \beta_2 = 0$  versus  $H_a: \text{at least one of the coefficients is not zero}$ ), can we conclude that at least one of these terms is significant?
- Use the drop-in-deviance test to compare the models in Parts A and B. Can you conclude that `Method` is related to the probability of success?
- Construct a model using only `Method` to predict the probability of a successful removal. What do Wald’s test and likelihood ratio test reveal?
- Explain why the drop-in-deviance test in Part C is better than the approach in Part D for determining if `Method` is related to the probability of a successful tattoo removal.

#### E.6. Lung Cancer and Birdkeeping: The Drop-in-Deviance Test

Data set: `Birdkeeping`

Holst, Kromhout, and Brand conducted a case-control study to determine if keeping birds increased the chances of getting lung cancer.<sup>20</sup> It is believed that people who keep birds may

inhale more allergens and dust particles and thus may be more likely to get lung cancer. They collected data on 49 people with lung cancer and 98 people with similar demographics who did not have lung cancer.

Gender: 1 represents females

Status: 1 represents high economic status

Age: owner's age

Smoked: number of years the person has been smoking

Cigarettes: number of cigarettes smoked per day

Bird: 1 represents owning a bird

LungCancer: 1 represents having lung cancer

The first five variables (Gender, Status, Age, Smoked, and Cigarettes) are known to impact the likelihood of having lung cancer and are assumed to be important in modeling the likelihood of having lung cancer. We are primarily interested in determining if the odds of lung cancer change based on whether an individual is a bird keeper.

- a. Create a logistic regression model using all six explanatory variables to estimate the probability of having lung cancer. Based on Wald's test, which variables appear to be significant? Based on the likelihood ratio test, can we conclude that at least one of these terms is significant?
- b. Create a logistic regression model using the first five variables (Gender, Status, Age, Smoked, and Cigarettes) to estimate the probability of having lung cancer.
- c. Use the drop-in-deviance test to compare the models in Parts A and B. Can you conclude that birdkeeping is related to the probability of having lung cancer?
- d. Construct a model using only the variable Bird to estimate the probability of having lung cancer. What do Wald's test and the likelihood ratio test reveal?
- e. Explain why the drop-in-deviance test in Part C is better than the approach in Part D for determining if birdkeeping is related to the probability of having lung cancer.

#### E.7. Surviving Third-Degree Burns

Data set: Burns

Fan, Heckman, and Wand analyzed third-degree burn data from the University of Southern California General Hospital Burn Center.<sup>21</sup> In the Burns data set, 435 patients (adults ages 18–85) were grouped according to the size of the third-degree burns on their body. The explanatory variable is listed as the midpoint of set intervals:  $\ln(\text{area in square centimeters} + 1)$ . The response in this data set is whether or not the patient survived (1 represents a survival).

- a. Create a logistic regression model using area to estimate the probability of survival.
- b. Calculate the observed and expected probabilities. Plot both of these probabilities against the median area.
- c. Interpret the model in terms of the odds ratio. Use the Wald statistic to create a 95% confidence interval for the odds ratio.
- d. Test  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  using both Wald's test and the likelihood ratio test. State your conclusion based on these tests.
- e. Conduct the Pearson, deviance, and Hosmer-Lemeshow goodness-of-fit tests to assess how well the model fits the data. Interpret the results.
- f. Report and interpret Somers'  $D$ , Goodman-Kruskal gamma, or Kendall's tau-a for this logistic regression model.
- g. What conclusions can you draw from this study?

#### E.8. Multiple Explanatory Variables in the Bird Nest Study

Data set: Birdnest

The Birdnest data discussed in Exercise 1 contain many characteristics of North American passersines ("perching birds").

- a. Create a logistic regression model using bird length (Length) and average number of eggs (No. eggs) to estimate the probability that a bird species has a closed nest type.

- b. Use the drop-in-deviance test to determine if the model with Length and No. eggs should be used instead of the model with only Length as an explanatory variable (as shown in Exercise 1).
- c. Use the drop-in-deviance test to create and interpret a best model to estimate the probability of the nest being closed (closed = 1). Start with a full model that includes body length in centimeters (Length), average number of eggs in a clutch (No. eggs), egg color (Color), incubation time of the eggs in days (Incubation), and number of days the chicks stay in the nest (Nestling).

#### E.9. Survival of the Donner Party: Multiple Explanatory Variables

Data set: Donner

- a. Create a logistic regression model using Gender and Age to estimate the probability of survival. Create a plot of the estimated probability of survival using Age as the explanatory variable and grouping the data by Gender. Use the plot and the model to interpret the coefficients in terms of the odds ratios.
- b. Create and interpret a logistic regression model using Gender, Age, and Gender\*Age to estimate the probability of survival. Create a plot of the estimated probability of survival using Age as the explanatory variable and grouping the data by Gender.
- c. Explain any key differences between the plots created in Parts A and B. Discuss how adding the interaction term Gender\*Age impacts the model.
- d. Assuming that the model in Part B is your final model, use the Hosmer-Lemeshow test to assess the model goodness-of-fit.

#### E.10. Variable Selection Techniques and Multicollinearity

Data set: Cancer2

- a. Create a logistic regression model using Radius, Concave, Radius\*Radius, and Radius\*Concave as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values.
- b. Even though in Part A Wald's test shows the highest  $p$ -value for Radius, it is typically best to attempt to keep the simplest terms in the model. Generally, keeping simpler terms in the model makes the model easier to interpret.\* Thus, we suggest as a first attempt keeping Radius in the model and eliminating the variable with the next highest  $p$ -value. Create a logistic regression model using Radius, Concave, and Radius\*Concave as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test to determine if Radius\*Radius should be included in the model.
- c. Use a scatterplot to compare Radius to Radius\*Radius and calculate the correlation between these two terms. Are these variables highly correlated?
- d. Chapter 3 discusses **multicollinearity** (highly correlated explanatory variables). Explain whether you believe Radius is important in the logistic regression model. Why is the  $p$ -value for Radius so large in Part A but very small in Part B?
- e. Create a logistic regression model using Radius and Concave as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test to determine if Radius\*Concave should be included in the model.
- f. Create a logistic regression model using only Concave as an explanatory variable to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test to determine if Radius should be included in the model.
- g. Submit a final model and provide a justification for choosing that model.

---

\*When several variables can potentially be in a model, many texts suggest conducting backward elimination on only the primary variables of interest. After backward elimination on these terms is complete, include appropriate interaction terms and terms for curvature and again conduct backward elimination on these more complex terms.

### E.11. And the Winner Is . . . : Variable Selection Techniques

Data set: Oscars

In 2009, three Grinnell students (Allie Greenberg, Hannah Lytle, and Phillip Brogdon) conducted an analysis to estimate the probability of winning the Academy Award for Best Picture. The Academy Awards, or “Oscars,” are given annually to honor high achievement in the film industry. The Academy consists of over 6000 members who nominate their colleagues and vote to decide on the winners of this prestigious award.

In their analysis, these students included all films nominated for Best Picture from 1979 to 2008. Winning Best Picture was considered the response (1 represents a win), and the explanatory variables included whether or not the picture won any of 17 other awards that were given out during that year’s ceremony.

- a. Create a logistic regression model using all 17 explanatory variables. Which variables appear to be most significant?
- b. Create and compare multiple logistic regression models. Submit the model with the fewest number of terms that best estimates the probability of winning the Best Picture award.
- c. Academy Award for Best Picture in 2009 went to *Hurt Locker*. Use your final model in Part B to predict the likelihood that *Hurt Locker* would win the Best Picture award. *Avatar* and *The Blind Side* were also nominated. Use your final model to estimate the probability that each of these movies would win Best Picture.

### E.12. And the Winner Is . . . 2: Variable Selection Techniques

Data set: Oscars2

While the previous information is interesting, there are some limitations in its use. The winners of the 17 other awards are known only a few minutes before the Best Picture award is announced. Instead of using other Academy Awards to predict the probability of Best Picture, it may be more useful to use other award ceremonies to predict the likelihood of winning Best Picture. Hannah, Allie, and Phillip also collected the following data:

Number of Golden Globe nominations

Number of Golden Globe wins

Number of Screen Actors Guild (SAG) nominations

Number of SAG wins

- a. Create a logistic regression model using all four explanatory variables. Which variables appear to be most significant?
- b. Using only the Oscars2 data set, submit the model with the fewest number of terms that best estimates the probability of winning the Best Picture award.
- c. Compare your model in Part B of Exercise 11 to the one from Part B above. Explain which model is better.
- d. Combine the Oscars and Oscars2 data sets to include a total of 21 explanatory variables. Create the model with only a few variables that best estimates the probability of winning the Best Picture award. Is this new model better than the one created in Part B of Exercise 11?

## Endnotes

---

1. D. Leonhardt, “John Tukey, 85, Statistician, Coined the Word ‘Software’,” New York Times Archives on the Web, 7/28/2000, stat.bell-labs.com/who/tukey/nytimes.html. John Tukey (1915–2000) had a formal background in chemistry and mathematics. Conducting data analysis during World War II peaked his interest in statistics, and he became one of the most influential statisticians of the 20th century.
2. *Report of the Presidential Commission on the Space Shuttle Challenger Accident, in Compliance with Executive Order 12546 of February 3, 1986*, <http://science.ksc.nasa.gov/shuttle/missions/51-1/docs/rogers-commission/table-of-contents.html>.

3. *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, <http://history.nasa.gov/rogersrep/v4part7.htm>.
4. *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, <http://history.nasa.gov/rogersrep/v1ch5.htm>.
5. See S. Menard, *Applied Logistic Regression Analysis*, 2nd ed. (Thousand Oaks, CA: Sage Publications, 2002).
6. See A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed. (New York: Wiley, 2007) or D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. (New York: Wiley, 2007).
7. J. S. Long, *Regression Models for Categorical and Limited Dependent Variables* (Thousand Oaks, CA: Sage Publications, 1997). Other texts will show that sometimes exact logistic regression models can be created in data sets with only a small number of observations for binary outcomes.
8. W. Wolberg and O. Mangasarian, “Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology,” *Proceedings of the National Academy of Sciences of the United States of America*, 87. 23 (Dec. 1990): 9193–9196.
9. S. Menard, *Applied Logistic Regression Analysis*, 2nd ed. (Thousand Oaks, CA: Sage Publications, 2002).
10. A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed. (New York: Wiley, 2007).
11. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. (New York: Wiley, 2000).
12. D. W. Hosmer, T. Hosmer, S. Le Cessie, S. Lemeshow, “A Comparison of Goodness-of-fit Tests for the Logistic Regression Model,” *Stat Med*, 16 (1997): 965–980.
13. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression* (New York: Wiley, 1989), p. 143.
14. Ibid.
15. For more details, see D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 4th ed. (Hoboken, NJ: Wiley, 2006), Appendix C.14.1.
16. D. K. Grayson, “Donner Party Deaths: A Demographic Assessment,” *Journal of Anthropological Research*, 46, 3 (1990). Also see <http://www.xmission.com/~octa/DonnerParty/Roster.htm>.
17. S. Wilson and B. Mandelbrote, “Drug Rehabilitation and Criminality,” *British Journal of Criminology*, 18 (1978): 381–386.
18. P. K. van Vliet, and J. M. Gupta, “Sodium Bicarbonate in Idiopathic Respiratory Distress Syndrome,” *Archives of Disease in Childhood*, 48 (1973): 249–255.
19. Data are modified from A. D. Lunn and D. R. McNeil, *The SPIDA Manual* (Sydney: Statistical Computing Laboratory, 1988).
20. P. A. Holst, D. Kromhout, and R. Brand, “For Debate: Pet Birds as an Independent Risk Factor for Lung Cancer,” *British Medical Journal*, 297.6659 (1988): 6659 1319–1321.
21. J. Fan, N. E. Heckman, and M. P. Wand, “Local Polynomial Kernel Regression for Generalised Linear Models and Quasi-Likelihood Functions,” *Journal of the American Statistical Association*, 90 (1995): 47 141–150.
22. “Young Americans Say Alcohol, Marijuana, Cigarettes, and Lottery Tickets Are Easily Accessible,” Annenberg Public Policy Center of the University of Pennsylvania, [http://www.appcpenn.org/Downloads/Adolescent\\_Risk\\_Tobacco/risk\\_report.pdf](http://www.appcpenn.org/Downloads/Adolescent_Risk_Tobacco/risk_report.pdf), accessed 11/30/07.
23. G. C. Homans, “Social Behavior as Exchange,” in Peter Kivisto, ed., *Social Theory: Roots and Branches*, 2nd ed. (Los Angeles: Roxbury, 2003), pp. 295–304.
24. Ibid, p. 297.
25. 2005 Iowa Youth Survey Report, [http://www.iowayouthsurvey.org/images/2005\\_county\\_reports/79.Poweshiek.pdf](http://www.iowayouthsurvey.org/images/2005_county_reports/79.Poweshiek.pdf), accessed 9/9/07.
26. D. L. Franko, D. Thompson, S. G. Affenito, B. A. Barton, and R. H. Striegel-Moore, “What Mediates the Relationship Between Family Meals and Adolescent Health issues?” *Health Psychology*, 27.2 (2008): S109–S117.
27. Ibid, p. 3.
28. Ibid, p. 5.

# Research Project: Substance Abuse Among Youth

Now that you have analyzed studies using logistic regression, it is time to conduct your own research project. The following pages provide guided steps for conducting your own research project involving data collected by the Iowa Youth Survey on teen substance abuse. In this project, you will have the opportunity to use the likelihood ratio test to create a logistic regression model with multiple explanatory variables. The course website, the homework, and the extended activities also provide several additional examples that can be used to develop your own project ideas.

## Reviewing the Literature

Understanding why junior high and high school students engage in behavior that is harmful to their health is the key to providing effective preventive programming. Young adults perceive their popular peers as those most likely to engage in risky behaviors, and strong social codes compel them to associate “cool” with drugs and alcohol.<sup>22</sup> Social theorist George Homans explains that this counterproductive behavior is based on individuals’ making choices to maintain a group’s equilibrium.<sup>23</sup> If the majority of the members of a group emit certain behaviors in similar frequencies, then a person stabilizes his behavior to keep himself in the best of circumstances, rationally or not.<sup>24</sup> While not all teens abuse substances, if the perception exists that being “cool” means drinking alcohol and using drugs, then a strong desire to fit in with a peer group can result in substance abuse.

In the 2004 and 2005 National Survey on Drug Use and Health (<https://nsduhweb.rti.org>), 9.42% to 10.03% of youth ages 12 to 17 in Iowa reported illicit drug use in the month preceding the survey. Usage as low as 8.31% and as high as 14.44% was reported, placing Iowa in the second lowest tier for youth drug usage. The same surveys indicate that Iowa has a high binge-drinking rate among youth ages 12 to 17. The report places Iowa in the highest tier for percentage of youth reporting binge-drinking, at a rate of 12.70% to 15.55%. Grouped with Iowa are Kansas, Minnesota, Missouri, Montana, Nebraska, North Dakota, South Dakota, Wisconsin, and Wyoming, making up almost the entirety of the Midwest region of the United States.

Distributed every three years to every county in Iowa since 1975, the Iowa Youth Survey (IYS) includes questions about students’ behaviors and attitudes/beliefs, as well as their perceptions of their peer, family, school, and neighborhood/community environments.<sup>25</sup> In 2005, the IYS was distributed to 6th, 8th, and 11th graders ranging in age from 10 to 18. At the state level, by age 16, 65% of 11th graders reported having had their first alcoholic drink and 20 percent of 11th graders reported having consumed an alcoholic beverage on 1 to 2 days of the last 30. Also reported by 11th graders was that by age 16, 29% at the state level had tried marijuana.

Psychological and sociological research indicates that family life is an important predictor of present and future alcohol use. Franko, Thompson, Affenito, Barton, and Striegel-Moore studied the relationship between family meals and substance abuse among female adolescents.<sup>26</sup> They found a negative relationship between frequency of family meals and smoking. This negative relationship was not observed for alcohol use. In the paper assigned below, Eisenberg, Olson, Neumark-Sztainer, Story, and Bearinger report on the association between alcohol use and frequency of family meals in a diverse adolescent sample from Minneapolis/St. Paul area in Minnesota. These researchers found that frequency of family meals was inversely associated with alcohol use even after controlling for family connectedness.

In this research project, we will use a random sample of the results of the 2005 Iowa Youth Survey to determine if there is a relationship between a healthy home life and youth substance abuse.

1. Read the paper M. E. Eisenberg, R. E. Olson, D. Neumark-Sztainer, M. Story, and L. H. Bearinger, “Correlations Between Family Meals and Psychosocial Well-Being Among Adolescents,” *Archives of Pediatrics and Adolescent Medicine*, 158 (2004): 792–796. If there are any words that you do not understand, look them up and provide a short definition for each. Identify the following and be ready to discuss this material in class:
  - a. Objective of the study
  - b. Any relevant background (from journals that were referenced)
  - c. Response variable(s)
  - d. Potential explanatory variables and levels that were tested
  - e. Variables that were held constant during the study
  - f. Nuisance factors (i.e., factors that are not of interest but may influence the results)

## Exploring the Data

Because the IYS is dependent on the student's ability to read and answer the questions honestly, each questionnaire was submitted to 26 validity checks assessing the possibility of inconsistent responses, improbable responses, and patterned responses. Of the 98,246 surveys distributed statewide, only 142 failed four or more of the validity checks.<sup>27</sup> While not every school district participates and county reports are more subject to sampling error than the comprehensive state reports, the IYS can still reliably be used to assess areas for improvement as well as those working well in a community.<sup>28</sup> The data set `IYSdata` contains a random sample of the 2005 responses. Below is a list of the variables provided in the data set:

- In what grade are you in school?
- What is your current age?
- Are you a male or female?
- Do you live on a farm, in a small town, a small city or a large city?
- How many different times has your family moved to a different home or apartment in the last 2 years?
- On the average during the school year, how many hours a week do you spend at church or synagogue worship service, programs, or activities?
- On the average during the school year, how many hours a week do you spend doing school assignments?
- During the last 30 days, on how many days did you have 5 or more drinks of alcohol (glasses of wine, liquor, mixed drinks) in a row, that is within a couple of hours?
- In the last 30 days, how many times have you driven a car or other motor vehicle after using any amount of alcohol or other drugs?
- In the past 30 days, on how many days have you smoked cigarettes?
- In the past 30 days, on how many days have you smoked cigars?
- In the past 30 days, on how many days have you had at least one drink of alcohol (glass, bottle or can of beer; glass of wine, liquor or mixed drink)?
- In the past 30 days, on how many days have you used marijuana (pot, grass, hash, bud, weed)?
- In the past 30 days, on how many days have you sniffed glue or breathed the contents of gases or sprays in order to get high?
- In the past 30 days, on how many days have you used methamphetamines (crank, ice)?
- In the past 30 days, on how many days have you used cocaine (coke, rock, crack)?
- In the past 30 days, on how many days have you used amphetamines other than methamphetamines (like stimulants, uppers, speed)?
- In the past 30 days, on how many days have you used prescription medications that were not prescribed for you by your doctor?
- In the past 30 days, on how many days have you used over the counter medications different from the directions?
- I believe that working hard now will make my life successful in the future.
- In my school, if I got in trouble at school for breaking a rule, at least one of my parents/guardians would support the school's disciplinary action.
- My teachers care about me.
- In my home there are clear rules about what I can and cannot do.
- I have a happy home.
- There are people living in my home who have a serious alcohol or drug problem.
- I feel very close to at least one of my parents/guardians.
- I can talk about the things that bother me or I don't understand with someone in my home.
- I can get help and support when I need it from someone in my home.
- A parent/guardian knows where I am and who I am with, especially in the evening and on weekends.
- A parent/guardian checks to make sure I have done the things I am supposed to do (school homework, household chores, get home on time, etc.).

A parent/guardian generally finds out if I have done something wrong, and then punishes me.

When I am doing a good job, someone in my home lets me know about it.

Someone in my home helps me with my schoolwork.

At least one of my parents/guardians goes to school activities that I am involved in.

How wrong would your parents/guardians feel it would be for you to drink beer, wine or hard liquor (for example vodka, whiskey, gin) without their permission?

How wrong would your parents/guardians feel it would be for you to smoke cigarettes?

How wrong would your parents/guardians feel it would be for you to smoke marijuana?

How wrong would your parents/guardians feel it would be for you to use any illegal drugs other than alcohol, cigarettes or marijuana?

2. Which of the variables in the list would be the most appropriate response variable to represent substance abuse? Note that in this project the response variables will be treated as *categorical*, where a 1 represents that the student surveyed had used the substance in the last 30 days and a 0 represents that the student had not used the substance in the last 30 days.
3. Identify which explanatory variables you would expect to influence substance abuse among youth.
4. Identify any missing factors or factors that are assumed to be controlled within the model. What conditions would be considered normal for this type of study? Are these conditions controllable? If this condition changed during the study, how might it impact the results?
5. Use the response from Question 2, the explanatory variables from Question 3, and the data provided to create a preliminary model. This model initially should not include interaction or squared terms. You may want to talk to a sociologist to discuss questions that have arisen about model assumptions and variable selection.
6. Use variable selection techniques to simplify your model in Question 5. After the model has been simplified, you may want to consider including appropriate interaction or squared terms.

## Presenting Your Own Model

7. Meet with your professor to discuss your model assumptions and analysis. Include a concise discussion of what the Wald statistics, the drop-in-deviance test, chi-square tests, and measures of association tell you about your model. Be careful to determine whether each test is reliable (e.g., you may need to adjust group sizes in the Hosmer-Lemeshow test).
8. Conduct a final analysis of your data and use the discussions of prior work to write a 5- to 7-page research paper describing your analysis and discussing the results (see “How to Write a Scientific Paper or Poster” on the accompanying CD). Bring three copies of your research paper to class. Submit one to the professor. The other two will be randomly assigned to other students in your class to review.

## Final Revisions

Make final revisions to the research paper. Submit the first draft, other students' comments and checklists, and the final paper.

## Other Project Ideas

Several of the extended activities can also be used to develop your own project ideas. There are many places where data are publicly available.

- Use the website <http://devdata.worldbank.org/data-query> to collect data from the World Bank.
- Information on a variety of sports can be found at <http://cbs.sportsline.com>, <http://sportsillustrated.cnn.com>, <http://www.nfl.com/stats/team>, <http://www.ncaa.org>, and <http://www.baseball-reference.com>.

- Information from many federal agencies can be found at <http://www.fedstats.gov>, the Bureau of Labor Statistics (<http://www.bls.gov/cpi>), the Behavioral Risk Factor Surveillance System (<http://www.cdc.gov/brfss>), the National Center for Health Statistics (<http://www.cdc.gov/nchs>), and the U.S. Census Bureau (<http://www.census.gov>).
- College and university information can be found at <http://www.collegeboard.com>, <http://www.act.org>, and <http://www.clas.ufl.edu/au>.
- The National Center for Educational Statistics website is <http://nces.ed.gov>.
- Information about movies produced each year can be found at <http://www.the-numbers.com>, <http://www.boxofficemojo.com>, <http://www.imdb.com>, and <http://www.rottentomatoes.com>.

# Poisson Log-Linear Regression: Detecting Cancer Clusters

8

*If you can't solve a problem, then there is an easier problem you can solve: find it.*

—George Polya<sup>1</sup>

**C**hapter 7 discussed regression models in which the response variable was binary (it took on only one of two possible values). This chapter focuses on building regression models when the response (dependent) variable is a count. Just like least squares regression and logistic regression, Poisson regression can have one or more explanatory variables and the variables can be either categorical or quantitative. The activities in this chapter are based on a recent study to detect cancer clusters in a northeastern region of the United States and will be used to emphasize the following key concepts:

- Using the binomial and Poisson distributions to model count data
- Calculating and interpreting the Poisson regression model
- Using Wald's test to determine the significance of individual explanatory variables
- Conducting deviance goodness-of-fit tests to evaluate model appropriateness
- Using residual plots to assess regression model performance
- Extending Poisson regression to cases with more than one explanatory variable

The extended activities provide several additional examples and mathematical details. Finally, the research project provides the opportunity to build your own Poisson model to investigate whether several variables can be used to model the count of grand slams that are hit in regular-season professional baseball games.

## 8.1 Investigation: Are Cancer Rates Higher for People Living near a Toxic Waste Area?

Suppose toxic waste is discovered in soil samples in the neighborhood where you grew up. A company that has long since gone out of business has been identified as the apparent source of the waste. Even more worrisome is that you have been hearing that a number of your childhood friends have been diagnosed with cancer. You plan to investigate just how many cancer cases have been identified among your neighbors and determine whether this represents an unusually large number of cases. This is a problem that is playing out all over the country and the world as sources of toxic waste are discovered.

Statistically, this is a problem of identifying clusters. A **cluster** is identified when the number of cases in an area exceeds what we would expect to occur by chance. That is, the number of cases is so large that we believe the rate in the area is actually much higher than typical (i.e., we reject the null hypothesis that the cancer rate in this area is similar to the rate in other areas). Here are a couple of examples that have received a lot of media attention.

One of the few films with statisticians in key roles, *A Civil Action*<sup>2</sup> tells the true story of a cluster investigation. A series of studies and a large-scale lawsuit were initiated after the observation of several childhood leukemia cases in Woburn, Massachusetts. Interestingly, two water wells in the town, one contaminated and one not, provided a setting in which to examine the effect(s) of well-water contamination on the number of leukemia cases. A book by Jonathan Harr<sup>3</sup> and the later movie served to underscore the societal ramifications and importance of cluster identification for the average citizen.

A 2002 report from the National Cancer Institute discussed the results of a large case-control study initiated in 1993 called “Breast Cancer and the Environment on Long Island,” which examined the relationship of certain environmental contaminants to the rate of breast cancer. This case-control study examined and compared the histories of breast cancer patients (cases) and those not diagnosed with the disease (controls). This long-term study involved significant public monies and generated considerable concern.<sup>4</sup> Additional findings reported in 2003 indicated a possible association between exposure to electromagnetic fields and increased risk for breast cancer.

The variable of interest in both studies may appear to be a count of the occurrences of a rare outcome (e.g., a leukemia or breast cancer diagnosis). However, in such studies, any analysis or comparisons made using the data should take into consideration the rate of occurrences. The **rate**, also called an **incidence rate**, is a count of events divided by some measure of that unit’s exposure.

**Exposure** is usually some measure of time, space, unit of matter, or group size. In the cancer studies, exposure is expressed in terms of person-years. The number of **person-years** is the total number of years all people have been exposed (total number of people  $\times$  number of years at risk). Thus, the rate (the number of newly diagnosed cancers divided by person-years) depends on both the number of people in the area and the length of time each person lived in the area. Since rates can be very small, the National Cancer Institute typically expresses the **cancer incidence rate** as the cancer rate per 100,000 person-years (cancer rate  $\times$  100,000).

In this chapter, we will discuss an investigation of an alleged cancer cluster in Randolph, Massachusetts, described by Day, Ware, Wartenburg, and Zelen.<sup>5</sup> We will show how the Poisson probability distribution can often be used to describe the pattern of possible outcomes for rare events. When the response (dependent) variable can be modeled with a Poisson distribution, Poisson regression can be used to estimate the rate (or count) of occurrence as a function of explanatory variables. More specifically, we will use Poisson regression to model the rate of cancers as a function of age and location to compare cancer rates for Randolph and a nearby neighborhood.

### NOTE

Poisson regression is often expressed in terms of counts. This is straightforward when the level of exposure is constant for every observed response. However, when the amount of exposure varies for each response, counts in the model will depend on exposure levels.

## 8.2 Comparing Count Data for Groups

In the spring of 1984, 67 cases of cancer were reported among people living in 325 homes in the Bartlett-Green Acres neighborhood in the town of Randolph. Neighbors thought that this seemed like a high number of cancer cases, and they asked scientists at nearby Harvard University School of Public Health to advise

them about their cancer risk. That is, the residents wanted to know if the 67 cases in Randolph would be considered a cancer cluster.

## Activity ➔ Calculating the Cancer Rate per Person-Year

1. If the 325 homes in Randolph have an average of 3.5 people in each home (1138 people), then, based on the 67 reported cases, what proportion of people reported a cancer diagnosis?
2. If the 325 homes in Randolph have an average of 3.5 people in each home (1138 people) and if we assume people have lived in the community for an average of 25 years each, then what is the total number of person-years at risk in the community?
3. Given your answer to Question 2 and the fact that 67 cancer cases were reported, what is the observed cancer rate (number of cases per person-year) in Randolph? What is the cancer incidence rate?

Of course, we expect the number of cancer cases to vary even among neighborhoods with the same person-years of exposure and the same underlying environmental conditions. As an initial analysis, we will use a statistical model for count data to decide whether or not the count of 67 cases in Randolph is so rare that chance variation is not a reasonable explanation (i.e., calculate a *p*-value). After all, even cancer clusters, although rare, can occur just by chance.

In fact, if you have studied any probability, you already know that even rare events eventually will happen if given enough opportunities to occur. For example, you are not likely to be *the* winner of a weekly state lottery where the chance of winning is just one in one million. However, if five million people play the lottery each week for 20 weeks, then we expect at least one person to be a winner. The chance that *some* player will win is very high.

We will approach the problem of deciding whether the 67 cases in Randolph actually constitute a cancer cluster as statisticians would—using probability models. Could this community actually be like similar communities in terms of its exposure to cancer-causing agents and this large number of observed cases just be due to chance variation? Or is the number of cases in Randolph so large that it is difficult to argue that chance variation is a reasonable explanation?

In order to decide, we will first need to have some sense of the cancer rate in other communities. We will proceed by assuming that the rate of cancer in this Randolph neighborhood is the same as the nationwide incidence rate (the null hypothesis). From 1973 to 1977, the average cancer incidence rate nationally was 326 cases per 100,000 people in a year, or  $326/100,000 = 0.00326$  cases per person-year.

Since we know the exposure in Randolph from Question 2, we can find the expected count of cancer cases (i.e., average number of cases expected) when the null hypothesis is true using the relationship

$$\text{expected count} = \text{exposure} \times \text{rate} \quad (8.1)$$

## Activity ➔ Comparing Cancer Rates per Person-Year

4. Using a national cancer incidence rate of 326 cases per 100,000 people in a year and the person-years of exposure for Randolph in Question 2, calculate the expected count of cancer cases. Based on Equation (8.1), does the observed value of 67 cases out of 325 homes seem unusually high to you?
5. Day and colleagues actually made the assumption that residents lived in Randolph for an average of 12.5 years each. Using this assumption, does the observed value of 67 cases out of 325 homes seem unusually high to you?

By comparing your answers to Questions 4 and 5, you can appreciate the importance of estimating the exposure accurately. If we assume a 25-year exposure per person, we expect 92–93 cases, and the observed 67 cases is actually less than expected. However, if we assume a 12.5-year exposure period per person, we expect only 46–47 cases. Then the observed count of 67 cases is higher than the national rate, but we don't yet know whether it's unusually high.

**Key Concept**

Whenever count data for groups are compared, it is important to consider the background exposure (e.g., the group size, the length of time exposed, and/or the distance from the alleged site) in order to fairly compare rates.

## 8.3 Building Models for Count Data

We have computed expected counts of cancer cases, but these estimates only tell us the average number of cases for neighborhoods with the same person-years of exposure. In order to determine how unusual 67 cases is, we need a statistical model. We want to determine if the occurrence of 67 (or more) cases is likely to occur by chance. That is, we want to calculate a  $p$ -value.

We will begin with a simple model for the possible numbers of cases that could have been observed, and then we can use either probability calculations or a computer simulation to determine a  $p$ -value. We will make the following simplifying assumptions for the probability model:

- A: People obtain cancer diagnoses independently of one another.
- B: Each person randomly selected in a given year has a probability of 0.00326 of being diagnosed with cancer that year.

### Activity ➔ Evaluating and Building Two Simple Models

6. In what ways might model assumption A be false?
7. Discuss how assumption B might be violated if the ages of Randolph residents were quite different from those of residents of the rest of the country.
8. In what other way might model assumption B be false?
9. We will use model assumptions A and B to simulate cancer counts in 10,000 “neighborhoods,” each with  $325 \times 3.5 = 1138$  persons who have lived in the neighborhood an average of 25 years and experience a cancer incidence rate of 326 cases per 100,000 people in a year. Using the technology instructions provided on the CD, perform the simulation of cancer counts in 10,000 neighborhoods as described above. Draw a histogram of the cancer counts for the 10,000 simulated neighborhoods and assess graphically and numerically how unusual it is to observe at least 67 cases of cancer. Estimate a  $p$ -value for the test that the rate of cases is higher for the Randolph neighborhood than for the rest of the nation.
10. Repeat Question 9 under the assumption that exposure is just 12.5 years per person on average.

Your histograms in Questions 9 and 10 illustrate possible cancer counts assuming the rate in Randolph is equal to the national rate (based on person-years that would be appropriate for the Bartlett–Green Acres area in Randolph). In Question 10, when exposure averages 12.5 years per person, you can see from the histogram that communities with 67 cases are possible but very unlikely.

**Key Concept**

Using two simplifying assumptions (independence of cases and constant probability of a case when the exposure is the same), we can construct computer simulations to estimate the probability distribution of the possible case counts for a certain exposure level.

The scientists at Harvard started with an initial analysis very much like this one and decided that the alleged cluster of cancers in Randolph deserved a more detailed investigation. They decided to consider the 20-year period from 1964 to 1984 as the period of exposure for study. By examining census data and interviewing residents, Day and colleagues estimated that the number of persons at risk during the period from 1964 to 1984 was 2314 and that the mean duration of residence in the Bartlett–Green Acres neighborhood

during the period from 1964 to 1984 was about 11.5 years. Thus, there were  $2314 \times 11.5 = 26,661$  person-years of exposure and we would expect  $26,661 \times 0.00326 = 86.75$  cases—much higher than the observed 67 cases.

At this point, it appears that there was nothing out of the ordinary occurring in Randolph in terms of the incidence rate of cancers. This is a common outcome after an expert investigation of an alleged cancer cluster. In fact, there is debate over how much funding and resources should be allotted to investigate alleged disease clusters.<sup>6</sup> Just as in any study, it is important to consider additional variables that could be influencing the results. In the following sections, we develop more advanced methods that incorporate additional variables into the model.

## 8.4 The Binomial Model for Count Data

You might recognize that the model used to perform the computer simulations is a **binomial probability model**, described in Chapter 7. For example, in Question 10, you determined, for each simulated neighborhood, the count of cases ( $Y$ ) from among  $n = 1138 \times 12.5 = 14,425$  person-years, assuming that each person-year independently has the same chance ( $p = 0.00326$ ) to be a cancer case.

The probability distribution for the possible values of a binomial random variable  $Y$  with  $n$  trials and probability of success  $p$  is well known:

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \quad \text{for } y = 0, 1, 2, \dots, n \quad (8.2)$$

While statistical software packages will complete the calculations of the binomial probabilities for us, Chapter 7 shows that the mean and variance for the binomial probability model are

$$\text{mean of } Y = E(Y) = np \quad \text{and} \quad \text{Var}(Y) = np(1 - p) \quad (8.3)$$

In particular, note that the mean of  $Y$  is the expected count, so it is no surprise that there is a connection between Equations (8.3) and (8.1).

$$\text{expected count} = E(Y) = np = \text{exposure} \times \text{rate} \quad (8.4)$$

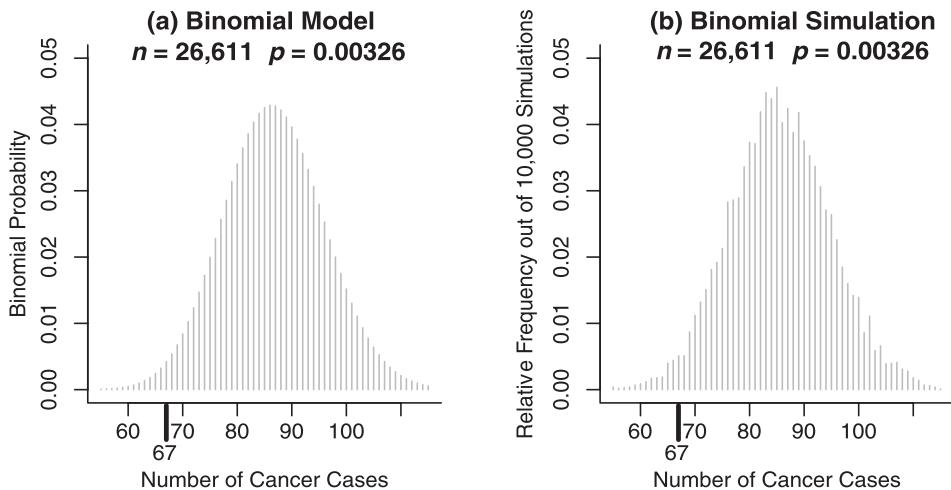
The binomial distribution in Equation (8.2) can be used directly to calculate the  $p$ -value without a simulation. Figure 8.1 gives two histograms. The first is the binomial probability distribution for counts of cancer cases from neighborhoods like Bartlett–Green Acres ( $n = 2314 \times 11.5 = 26,611$  and  $p = 0.00326$ ). The second histogram is based on a simulation of 10,000 hypothetical neighborhoods (similar to those in Questions 9 and 10 but using  $n = 26,611$  and  $p = 0.00326$ ).

From Equation (8.3), the expected value for this study is  $E(Y) = np = 26,611 \times 0.00326 = 86.75$  cases and the standard deviation is  $\text{sd}(Y) = \sqrt{\text{Var}(Y)} = \sqrt{np(1 - p)} = \sqrt{86.75(0.99674)} = 9.3$  cases. Given that the histograms in Figure 8.1 are symmetric and mound-shaped, this suggests that about 95% of the numbers of cases generated according to the national incidence rate in neighborhoods with 26,611 person-years of exposure should be between 68 and 105 (the mean plus or minus about 2 standard deviations). This is consistent with the simulation results and the probability histogram of the binomial probability distribution shown in Figure 8.1.

Figure 8.2 illustrates the shape, center, and spread of the binomial probability distribution for different choices of  $n$  and  $p$ . Note that the distribution is not always symmetric. Questions 11 and 12 address various shapes of the binomial distribution for different values of  $n$  and  $p$ .

### Activity Understanding the Binomial Distribution

- 11. Examine Figure 8.2. When is the binomial probability model most symmetric? When exposure  $n$  is large? When exposure  $n$  is small? When rate  $p$  is large? When rate  $p$  is small?
- 12. Examine Figure 8.2. When is the binomial probability model most right skewed (a long right tail)? When exposure  $n$  is large? When exposure  $n$  is small? When rate  $p$  is large? When rate  $p$  is small?



**Figure 8.1** The binomial model for Randolph cancer cases using  $n = 2314 \times 11.5 = 26,611$  and  $p = 0.00326$ : (a) a binomial probability histogram and (b) an empirical histogram from a simulation of 10,000 hypothetical neighborhoods using the binomial model.

## 8.5 The Poisson Model for Count Data

Questions 4 and 5 illustrate that time can be a critical component of exposure. Length of residence was certainly relevant for assessing the claim of a cancer cluster for persons living in the Bartlett–Green Acres neighborhood of Randolph.

The binomial model assumes that there are a fixed number of units  $n$ , and we count the number of successes out of those  $n$  units. However, in this example we are counting the number of occurrences based on some interval. Such intervals (such as time, space, or volume of matter) lie on a continuum and so might not lead to an integer value for the exposure. We already saw this in Question 5, where neither the number of persons ( $325 \times 3.5 = 1137.5$ ) nor the average years of residence (12.5) was an integer. In fact, we should have calculated the exposure as  $1137.5 \times 12.5 = 14,218.75$  person-years instead of  $1138 \times 12.5 = 14,225$  person-years.

Instead of using the binomial model for counts (where exposure values are limited to integers), we will consider the **Poisson probability model** for counts (which allows for continuous exposure levels). The formula for calculating Poisson probabilities is given in Equation (8.5), but we will again let the computer do the calculations.

$$P(Y = y) = \frac{e^{-t\theta}(t\theta)^y}{y!} \quad \text{for } y = 0, 1, 2, 3, \dots \quad (8.5)$$

where  $\theta$  is the cancer rate (i.e., the number of occurrences per unit of exposure, previously defined as  $p$ ) and  $t$  is the amount of exposure (i.e., the total number of person-years, previously defined as  $n$ ).

### CAUTION

Some texts have  $\lambda$  expressed as  $\lambda = t\theta$ . In addition, some researchers call  $\lambda$  the cancer rate (for the entire exposed population). It is important to recognize in this text the cancer rate is always in reference to  $\theta$ , not  $\lambda$ .

Like the binomial model, the Poisson model has an expected value of counts that is related to exposure and rate via Equation (8.1) as follows:

$$\text{mean of } Y = E(Y) = \lambda = t\theta \quad \text{and} \quad \text{sd}(Y) = \sqrt{\lambda} = \sqrt{t\theta} \quad (8.6)$$

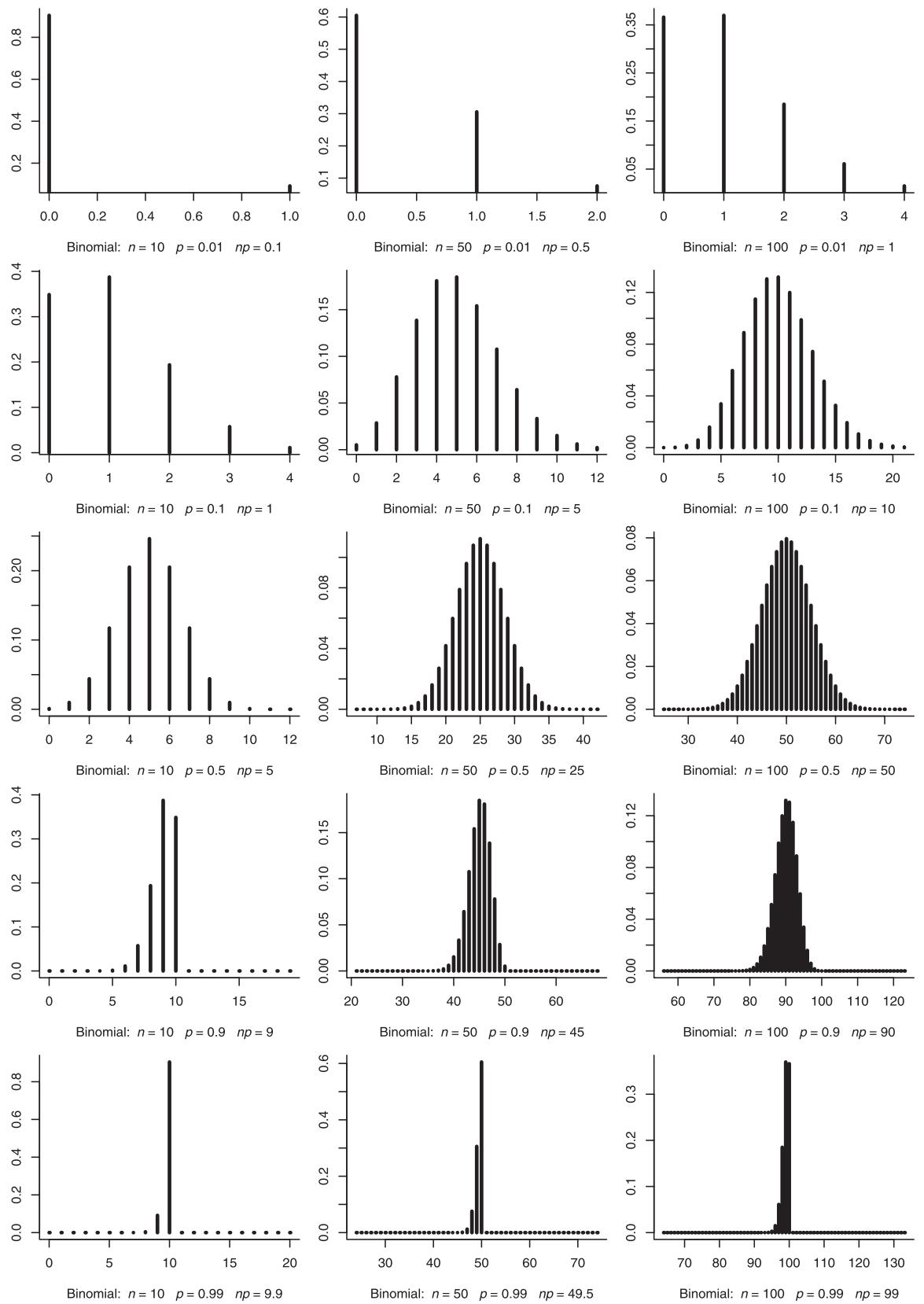


Figure 8.2 Examples of binomial probability distributions for various values of  $n$  and  $p$ .

Equation (8.6) shows that the Poisson model assumes that the variance is equal to the mean. Figure 8.3 displays several examples of Poisson probability distributions using the same values for exposure and rate as for the examples of binomial distributions in Figure 8.2. We let  $t = n$  and  $\theta = p$  to make a fair comparison between the distributions in the two figures. Question 13 addresses the similarity between the binomial and Poisson distributions for different values of  $n$  and  $p$ .

## Activity ◀ Comparing the Binomial and Poisson Distributions

13. Compare Figures 8.2 and 8.3. When is the Poisson model most like the binomial model? When exposure  $n$  is large? When exposure  $n$  is small? When rate  $p$  is large? When rate  $p$  is small?
14. Compare the formulas for the standard deviation for the binomial and Poisson models shown in Equations (8.3) and (8.6). Consider exposure  $= n = t$  and rate  $= p = \theta$  as you make your comparison.  
When is the standard deviation for the Poisson model most like the binomial standard deviation? When exposure  $n$  is large? When exposure  $n$  is small? When rate  $p$  is large? When rate  $p$  is small? Are your answers here consistent with your answers to Question 13?

The further  $p$  is from 0.5, the more the binomial probability distribution is skewed—either to the right when  $p$  is small (nearer to zero) or to the left when  $p$  is large (nearer to one). The Poisson distribution is right skewed but gets more symmetric as  $\lambda = t\theta$  grows. So the binomial probabilities are most like Poisson probabilities when  $p$  is small.

### ► MATHEMATICAL NOTE ▼

When  $p$  is small, it can be shown that the limit of the binomial model in Equation (8.2) is equal to the Poisson model in Equation (8.5) as  $n$  goes to infinity. A general rule of thumb is that the Poisson model provides a good approximation for the binomial model when  $p < 0.02$  and  $np > 5$ . For the binomial analysis of the cancer counts for the Bartlett–Green Acres neighborhood,  $n = 26,611$  person-years,  $p = 0.00326$ , and  $np = 86.75$ , so the Poisson and binomial distributions would provide similar models for these data.

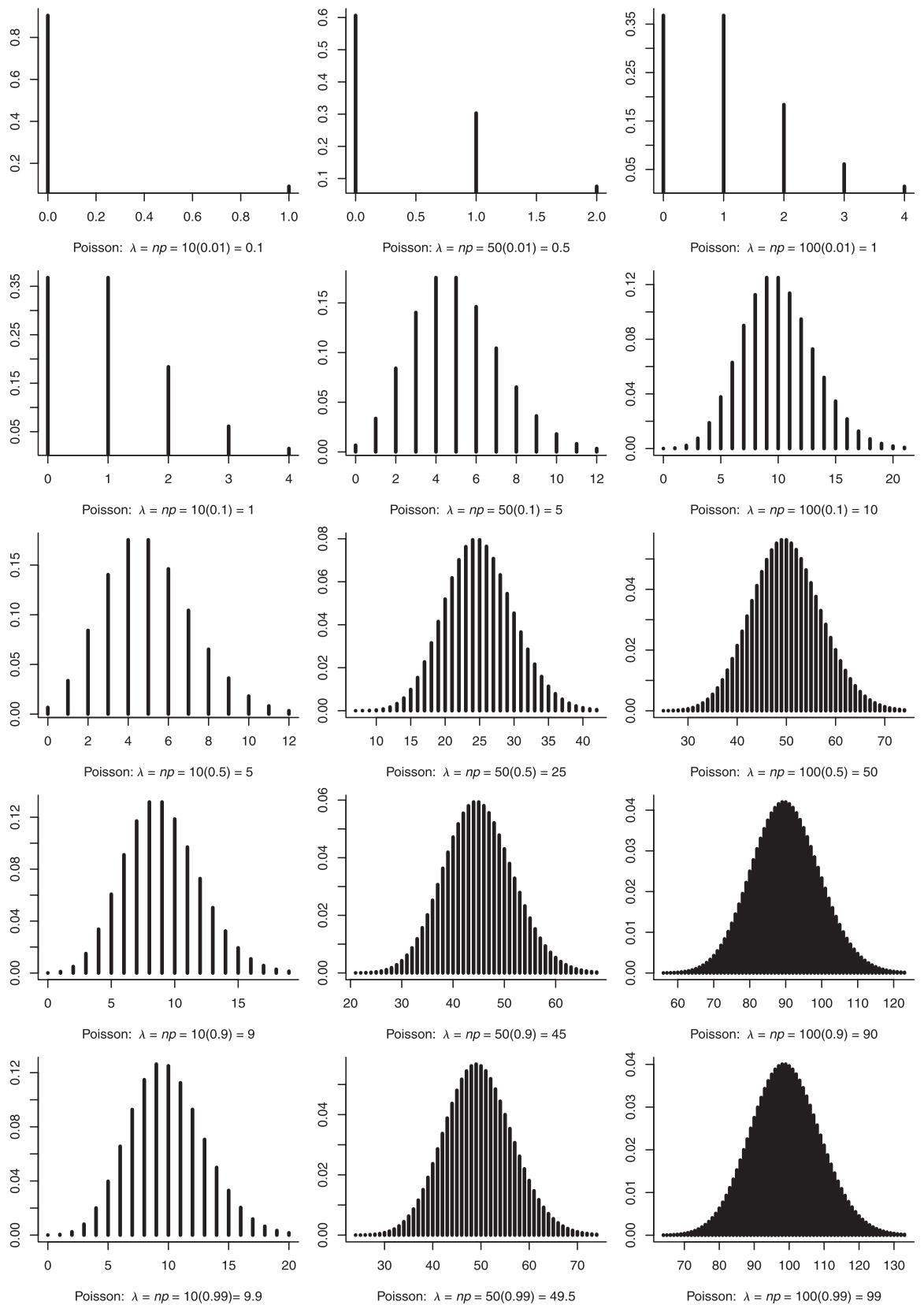
Note that with the Poisson model we still assume that the diagnostic status of each person-year is generated independently of the others, implying that people obtain cancer diagnoses independently of one another and from year to year. So your thoughts about the validity of model assumptions in Questions 6 and 7 are still relevant.

### Key Concept

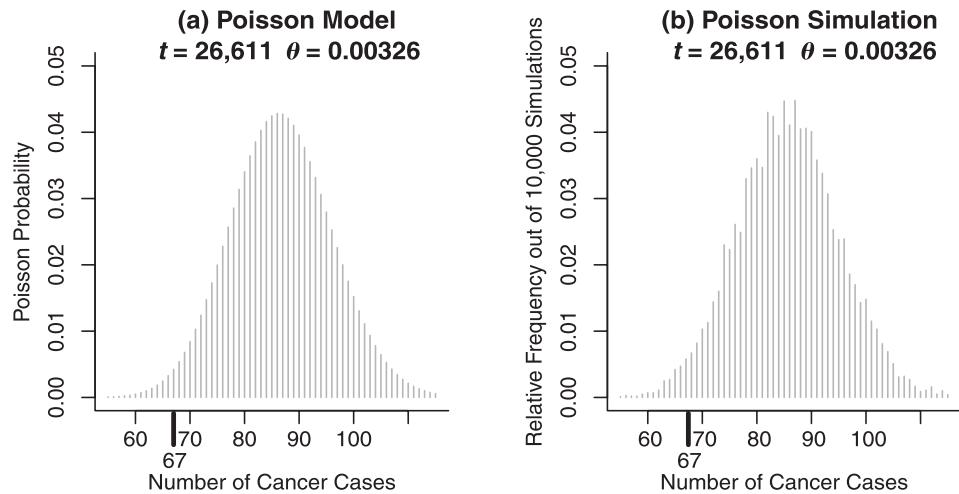
Using two simplifying assumptions (independence of cases and constant probability of a case when the exposure is the same), the binomial and Poisson models provide descriptions of the probability distribution of the possible case counts for a certain exposure level.

Figure 8.4 gives the Poisson probability distribution for counts of cancer cases from neighborhoods like Bartlett–Green Acres (exposure  $t = 2314 \times 11.5 = 26,611$  person-years and rate  $\theta = 0.00326$ ). Alongside the Poisson probability histogram is an empirical histogram from a simulation of 10,000 hypothetical neighborhoods. You can see that the histograms in Figure 8.4 are almost identical to those in Figure 8.1.

Note that unlike in the binomial model, there is no limit on the number of occurrences  $y = 0, 1, 2, 3, \dots$  during an exposure level. However, Figure 8.3 shows that when  $\lambda = t\theta$  is large, the normal distribution can be used to model counts. Thus, in regression models with large counts of occurrences, the response variable will tend to be normally distributed and least squares regression (described in Chapter 3) can often be used. However, when the number of occurrences is low ( $\lambda = t\theta$  is small), the Poisson distribution is skewed to the right and Poisson regression models tend to be more appropriate.



**Figure 8.3** Examples of Poisson probability distributions for various  $\lambda = np = t\theta$ .



**Figure 8.4** The Poisson model for Randolph cancer cases using  $t = 26,611$  and  $\theta = \text{rate} = 0.00326$ : (a) a Poisson probability histogram and (b) an empirical histogram from a simulation of 10,000 hypothetical neighborhoods using the Poisson model.

#### Key Concept

The Poisson model can be used when the number of occurrences is small and when the variance is equivalent to the mean.

## 8.6 Adding a Covariate to the Poisson Count Model

In the previous statistical models in this chapter, every individual in the neighborhood was assumed to have an identical probability (incidence rate) of getting cancer. Incorporating an explanatory variable, often called a **covariate**, allows us to more accurately estimate the response. For example, we know that age influences a person's likelihood of getting cancer. For younger people (ages 0–30), the rate of cancer is probably fairly constant, while for older people (ages 50–80), the cancer rate probably rises fairly steeply from the youngest to the oldest ages in this group. So the cancer rate is not rising at a constant rate, but probably grows exponentially with age. We will express this relationship mathematically as  $\theta = e^{\beta_0 + \beta_1 x}$ , where  $\theta$  is the cancer rate and  $x$  is age (in years).

Figure 8.5 is a picture of what this model might look like: a Poisson count model for cancer, where the cancer rate is allowed to grow exponentially with age. As age increases, so does the mean for the Poisson count model. In addition, the probability distribution of counts becomes more variable, but more symmetric. This is consistent with Equation (8.6) and the shape of the Poisson probability distribution for various means shown in Figure 8.3.

At each age, the Poisson probability distribution with mean  $\lambda = t\theta = te^{\beta_0 + \beta_1 x}$  is shown in Figure 8.5. This illustrates that different groups of individuals (even at the same age, with the same exposure  $t$  and the same rate  $\theta$ ) will vary in the number of individuals with cancer according to the Poisson model for counts.

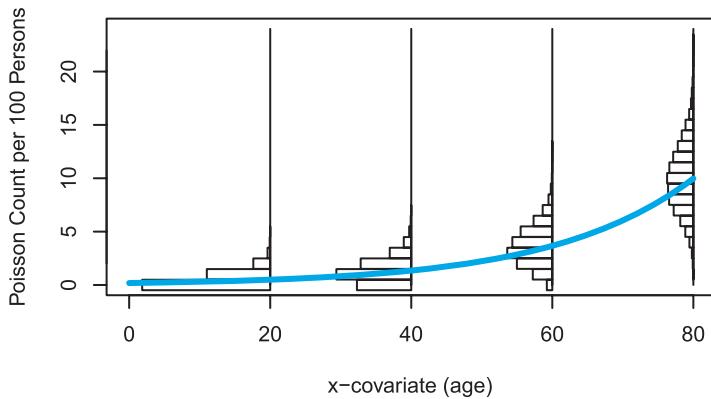
Statisticians express the exponential relationship between the expected cancer count,  $E(Y) = \lambda$ , and the covariate  $x = \text{age}$  depicted in Figure 8.5 as<sup>\*</sup>

$$\log[E(Y)] = \log(\lambda) = \log(t \times \theta) = \log(t \times e^{\beta_0 + \beta_1 x}) = \log(t) + \beta_0 + \beta_1 x \quad (8.7)$$

For Poisson regression, we have explicitly expressed the mean count  $\lambda$  in terms of both the rate ( $\theta$ ) and the exposure ( $t$ ), and you see that there is an additional term in the model in Equation (8.7) that is called an **offset**:

$$\log[E(Y)] = \log(\lambda) = \log(t) + \beta_0 + \beta_1 x = \text{offset} + \beta_0 + \beta_1 x \quad (8.8)$$

\*In order to simplify Equation (8.7), we used two basic properties of the logarithm function. For any two positive numbers  $a$  and  $b$ ,  $\log(ab) = \log(a) + \log(b)$  and  $\log(e^a) = a$ .



**Figure 8.5** Depiction of a Poisson count model for cancer cases where the cancer rate ( $\theta$ ) is allowed to grow exponentially with age ( $x$ ) according to the rule  $\theta = e^{\beta_0 + \beta_1 x} = e^{-1.5 + 0.05x}$  (represented by the thick, solid curve). The solid curve shows the relationship between the mean count and age:  $E(Y) = \lambda = t\theta = te^{\beta_0 + \beta_1 x}$  ( $t = 1$  here for simplicity).

Poisson regression uses an **offset term** to allow for situations where not every observed count has the same level of exposure. The offset in Equation (8.8) is the logarithm of the exposure level. Our basic model for counts shown in Equation (8.8) can be rewritten as

$$\log(\lambda/t) = \beta_0 + \beta_1 x \quad (8.9)$$

Thus, rather than simply modeling counts, the offset term allows us to model rates, such as counts per person-year.

Poisson regression is an example of a **generalized linear model (GLM)**. Generalized linear models, introduced in Chapter 7, extend the linear regression models to cases where the normal assumptions do not hold.

Generalized linear models all have the same basic form:

$$g[E(Y)] = g(\mu) = \beta_0 + \beta_1 x \quad (8.10)$$

where  $E(Y) = \mu$  is the mean for the probability model for the data and  $g$  is the **link function**. The link function is either flat, always increasing, or always decreasing (i.e., a monotone function); it cannot be increasing for some values of  $\mu$  and decreasing for others. In Chapter 7, the most common link function for binary response data was given as  $\ln[\mu/(1 - \mu)]$ . In Poisson regression, the most common link function is  $\log(\mu)$ .

For Equation (8.7), the mean is  $\mu = \lambda = t\theta$  and the link function is  $g(\mu) = \log(\mu)$ . The mean count  $\lambda$  can never be negative, and the logarithm function is always increasing for positive values. The GLM in Equation (8.7) is called a **Poisson log-linear regression model** and is sometimes simply called a **Poisson regression model**.

The form of the generalized linear model in Equation (8.7) may look somewhat familiar to you if you have some experience with simple linear regression models with a log-transformed response variable of the form

$$\log(Y) = \beta_0 + \beta_1 x + \varepsilon$$

or, equivalently, when the errors  $\varepsilon$  have mean zero,

$$E[\log(Y)] = \beta_0 + \beta_1 x \quad (8.11)$$

There is a key difference between Equation (8.11) and the generalized linear model given in Equation (8.7). Equation (8.11) is based on the log transformation of the observed responses and models the mean of  $\log(Y)$  as a function of the covariate,  $x$ . The GLM in Equation (8.7) uses the logarithm as the link function and models the *logarithm* of the mean of  $Y$  as a function of the covariate.

**NOTE**

Least squares regression models described in Chapter 3 have the additional assumption that the variance of the error terms is constant:  $\text{Var}(\varepsilon) = \sigma^2$  and does not depend on  $x$ . We know from Equation (8.6) that an interesting feature of a Poisson count model is that its expected value,  $E(Y) = t\theta = \lambda$ , is also its variance:  $\text{Var}(Y) = t\theta = \lambda$ . Thus, as the mean count changes, the variation in the observed counts will change as well. You can see this in Figure 8.5, where the distribution of cancer cases has more spread as age increases (and thus,  $\lambda$  is larger). This is a direct violation of the assumption of constant variance for simple linear regression modeling.

**Key Concept**

A generalized linear model (GLM) describes the relationship between the mean of a nonnormally distributed response variable and one or more covariates. A link function relates the mean response to a linear combination of the covariates. In modeling rates, the offset is used to incorporate different exposure levels for the observations.

## Fitting the Poisson Regression Model to the Cancer Count Data

We want to use a Poisson log-linear model to compare the Bartlett–Green Acres (BGA) age-specific cancer rates to some standard. We will use the Connecticut Tumor Registry (CTR) age-specific rates to make the comparison.\* The data appear in Table 8.1. The four age groups are the groups reported in the CTR. Day and colleagues provided the Randolph data by age group after carefully checking each reported case through in-person interviews (they found only 49 valid cases of adult cancers, rather than the 67 cases originally reported by the neighborhood). Here, we extract the number of cases for each age group for the years 1980–1984 for Connecticut using the reported rates and population data.<sup>7</sup> Note that the CTR case counts are close approximations, not exact, since we reproduced the case counts from the rounded case rates that were reported.<sup>†</sup>

We will fit our first Poisson regression model to the Randolph data in the top half of Table 8.1 and ignore the Connecticut Tumor Registry data in the bottom half of the table. To estimate the regression coefficients ( $\beta_0$  and  $\beta_1$ ) in the Poisson regression model in Equation (8.7), we will need three important quantities as inputs: the cancer counts for each group, the exposure (for the offset term in the model), and covariate information on median age for individuals in each group. When observed data are used to calculate  $b_0$  and  $b_1$ , Equations (8.7) and (8.9) can be written in terms of the estimated count  $\hat{\lambda}$ ,

$$\log(\hat{\lambda}) = \log(t) + b_0 + b_1x \quad (8.12)$$

or re-expressed in terms of the logarithm of the estimated rate  $\hat{\lambda}/t$ ,

$$\log(\hat{\lambda}/t) = b_0 + b_1x \quad (8.13)$$

### Activity Fitting a Poisson Regression Model with One Covariate

15. Use the technology instructions provided on the accompanying CD and the data from only the BGA location to fit a Poisson regression model with the cancer count for each group as the response ( $Y$ ) and median age as the covariate ( $x$ )—i.e., find the values of  $b_0$  and  $b_1$ . Don’t forget to include the logarithm of exposure as an offset when you fit the model.
16. In Question 15, you found that the estimated Poisson regression model was  $\log(\hat{\lambda}/t) = -8.67 + 0.049(\text{age})$ , which can be re-expressed in terms of the estimated cancer rate:

$$\hat{\lambda}/t = e^{-8.67+0.049(\text{age})} \quad (8.14)$$

\*The Connecticut Tumor Registry was established in 1941, making it the longest-running tumor registry available in the United States.

<sup>†</sup>We use counts since most statistical computing software uses the counts, population sizes, and covariate information as inputs and for fitting the log-linear Poisson model to data.

Then for the BGA location, we can estimate the cancer rate,  $(\hat{\lambda}/t)$ , for a given value of median age.

- Estimate the mean number of cancer cases in BGA when the median age is 62. Use this model to estimate the cancer rate in BGA when the median age is 62. Find the estimated cancer incidence rate (recall that the cancer incidence rate is equal to the rate  $\times 100,000$ ). For median age equal to 62 years,  $t = 1682$  for the BGA data.
- Calculate the difference between the observed cancer incidence rate from Table 8.1 and your estimated cancer incidence rate. How do they compare?

**Table 8.1** Cancer cases and cancer incidence rates (CIR) by age group for the Bartlett-Green Acres neighborhood of Randolph (BGA), 1964–1984, and similar data from the Connecticut Tumor Registry (CTR), 1980–1984.

Location	Cases	Person-Years	CIR	Age Group	Median Age
BGA	4	14,489	27.6	17–43	30.0
BGA	25	5,431	460.3	44–59	51.5
BGA	8	1,682	475.7	60–64	62.0
BGA	12	3,398	353.2	65–79	72.0
<b>BGA Overall</b>	<b>49</b>	<b>25,000</b>	<b>19.6</b>		
CTR	752	839,007.5	89.6	17–43	30.0
CTR	2,422	517,613	467.9	44–59	51.5
CTR	1,612	156,670	1028.9	60–64	62.0
CTR	5,465	285,798	1912.2	65–79	72.0
<b>CTR Overall</b>	<b>10,251</b>	<b>1,799,088.5</b>	<b>569.8</b>		

## 8.7 Interpreting Poisson Regression Model Parameters

Our Poisson model for the cancer rate as a function of (median) age is  $\log(\lambda/t) = \beta_0 + \beta_1 x$ . This model can be expressed as

$$\lambda/t = e^{\beta_0 + \beta_1 x} \quad (8.15)$$

Thus,  $e^{\beta_0} = \lambda/t$  is the cancer rate when  $x = 0$ , and  $e^{\beta_1}$  represents the multiplicative change in the cancer rate associated with each unit increase in  $x$ .

To get a feel for how to interpret  $e^{\beta_1}$ , suppose we have a simple covariate for age ( $x = 0$  for young people and  $x = 1$  for older people). Let  $Y$  be the counts (the number of cancer cases in each age group). Based on our model in Equation (8.15), the rate of cancer for young people is

$$\lambda/t = e^{\beta_0 + \beta_1(0)} = e^{\beta_0}$$

and the rate for older people is

$$\lambda/t = e^{\beta_0 + \beta_1(1)} = e^{\beta_0} \times e^{\beta_1}$$

That is, in this model, the exponentiated coefficient for age has a multiplicative effect. For example, if  $e^{\beta_1} = 4$ , then we would say that the cancer rate for older people is four times as high as the cancer rate for younger people.

The estimate for  $\beta_1$  that you calculated for the model in Question 15 is for an age covariate ( $x$ ) that is considered a quantitative, continuous time variable (not just a zero-one variable, as in our example of old versus young people above).

Then for a  $c$ -unit increase in the value of the covariate, there is a multiplicative increase of  $e^{c\beta_1}$  to the rate  $\lambda/t$ . For example, in the Poisson log-linear model of BGA cancer rates as a function of age, you should have found that  $\beta_1$  was estimated to be  $b_1 = 0.049$ . Thus, the cancer rate for the group with median age equal to 62 is estimated to be  $e^{62 \times 0.049} / e^{51.5 \times 0.049} = e^{10.5 \times 0.049} = 1.67$  times as high as the rate for the group with median age of 51.5 years.

**Key Concept**

Because the Poisson regression model involves a logarithmic transformation of the mean rate  $\lambda/t$ , an increase in the covariate by 1 will increase the mean rate by a multiple of  $e^{\beta_1}$ . When the covariate increases by  $c$  units, the mean rate will be multiplied by  $e^{c\beta_1}$ .

**Activity**  *Interpreting Parameter Estimates*

17. Fit a Poisson regression model to the CTR data with median age as the covariate. How does the parameter estimate  $b_1$  compare to the parameter estimate found for the Poisson regression model for the BGA data in Question 15? Which model shows the logarithm of cancer rates growing faster with age?
18. Use the previous question to estimate the cancer rate at age 62 and at age 72 from the Poisson log-linear model for the CTR data. Compute the estimated cancer incidence rates as well. Calculate the difference between the observed cancer incidence rates from Table 8.1 and the estimated cancer incidence rates that you calculated. Are your answers consistent with the observed values in Table 8.1?
19. What is the estimated multiplicative increase in the CTR cancer incidence rate associated with any 10-year increase in age?
20. Use the full set of cancer counts in Table 8.1 (both BGA and CTR locations) to fit a Poisson regression model to the cancer rate data with median age as the covariate. Note that this model ignores the location variable—that is, it treats the effect of (median) age on the cancer rate the same way, regardless of whether the counts are from BGA or CTR. Do you think this makes sense in light of the estimated coefficients you found in Questions 15 and 17?
21. Now use the combined counts (for both BGA and CTR locations) to fit a Poisson regression model to the cancer rate data with `location` as the covariate. Note that for this problem a dummy variable (dummy variables are discussed in Chapters 3 and 7) is used to code the location variable. Let `location` value = 0 correspond to BGA and `location` value = 1 correspond to CTR. How many times higher is the cancer rate for the CTR location compared to BGA?

**8.8 Poisson Regression Models with More Than One Covariate**

The key question in this investigation is whether the cancer rates are higher in Randolph than in other locations. In order to address this question, a more complex model with two covariates (both `age` and `location`) is needed. As shown in Chapters 3 and 7, including more covariates in a regression model can account for more variation in the response. In this study, the two covariates, `age` and `location`, can be included to form the Poisson regression model

$$\log(\lambda) = \log(t) + \beta_0 + \beta_1(\text{age}) + \beta_2(\text{location})$$

which can be re-expressed in terms of the rate,  $\lambda/t$ :

$$\lambda/t = e^{\beta_0 + \beta_1(\text{age}) + \beta_2(\text{location})} \quad (8.16)$$

Then, for example, we can interpret the term  $e^{\beta_1}$  as the multiplicative effect on the rate  $\lambda/t$  corresponding to a 1-year increase in age, after adjusting for the location. The  $e^{\beta_2}$  term would correspond to the multiplicative effect on the rate of changing locations from BGA to CTR (assuming that `location` = 0 is BGA and `location` = 1 is CTR), after adjusting for age.

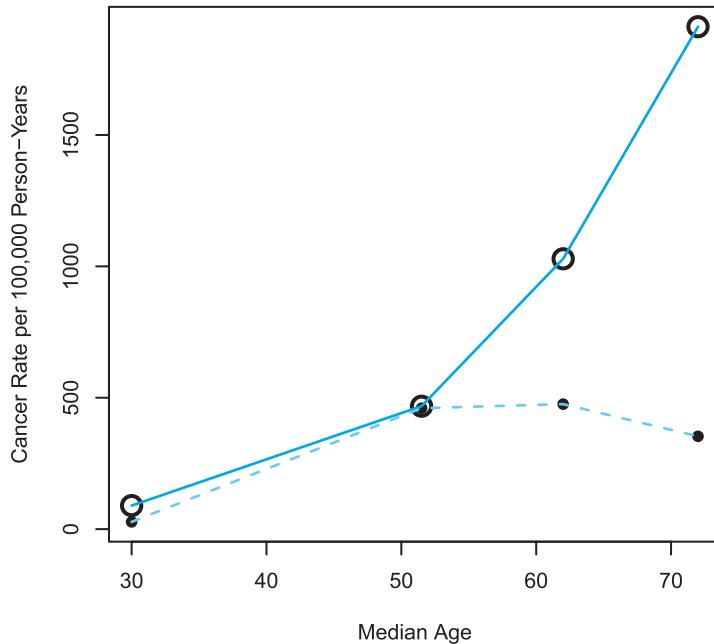
**Activity**  *Poisson Regression Models with More Than One Covariate*

22. Fit the Poisson regression model with two covariates: median age and `location`. Interpret the terms  $e^{b_1}$  and  $e^{b_2}$ , where  $b_1$  and  $b_2$  are estimates of the model coefficients corresponding to `age` and `location`, respectively.

23. After adjusting for median age, how many times higher is the rate for the BGA location than the CTR location?

Based on your answers to Questions 20 and 21, you probably noticed that the estimated cancer rates appear to differ by location and by median age. So it may be necessary to include an additional term in the model that was fit in Question 22.

Whenever possible, we should try to plot the log of the counts versus the values of the quantitative covariate. If the pattern is approximately linear, then the Poisson regression model may be appropriate for the data. Let's examine a graph of the cancer incidence rates by age for the BGA and CTR locations. Figure 8.6 illustrates how the cancer incidence rates change with age for the two locations.



**Figure 8.6** Plot of cancer rates per 100,000 person-years versus median age at two locations: the Bartlett–Green Acres neighborhood in Randolph (rates denoted by •, connected by dashed lines) and the Connecticut Tumor Registry (rates denoted by ○, connected by solid lines).

### Activity ▶ Comparing Rates by Age and Location

24. Examine Figure 8.6 and describe the relationship between age and the cancer incidence rate for the CTR data. Is the relationship linear?
25. Create a graph of the logarithm of the cancer incidence rates versus the median of the age group by location. Describe the relationship between age and cancer rate separately for the CTR data and the BGA data. Are the relationships linear?
26. Can you use the same general description to describe how the rates change with age regardless of location?

### Including an Interaction Term in a Poisson Regression Model

It is useful to create a plot of the logarithm of the observed counts versus the covariate(s) and look for a general linear pattern to determine if a log-linear model is sensible. In Question 25, you should have seen that the logarithm of the cancer rate was linearly related to median age for the CTR data, but not for the BGA

data. Still, the number of cases in two of the age groups is small for BGA data (one group even has less than 5 cases), so we can expect a great deal of variability from a line to occur just by chance.

Note that the rates change differently with age for the BGA and CTR locations in Figure 8.6 and in your graph for Question 25. We say that there is a potential **interaction** between age and location in terms of their relationship to the cancer rate. Inclusion of an interaction term in the model allows the multiplicative effect of age on cancer rate to differ for the BGA and CTR locations.

In order to examine the interaction between age and location in terms of their relationship to the cancer rate, we will use a single Poisson regression model that includes both `age` and `location` as covariates. (`location` will be an indicator or dummy variable coded as 0 or 1; see Chapter 3 for a discussion of indicator variables.) Further, we will add a term to capture the **interaction effect**, if any, between `age` and `location`. The interaction model for the cancer rate is as follows:

$$\lambda/t = e^{\beta_0 + \beta_1(\text{age}) + \beta_2(\text{location}) + \beta_3(\text{age} * \text{location})} \quad (8.17)$$

## Activity ▶ Fitting an Interaction Term

27. Fit a Poisson regression model using the cancer count data from Table 8.1. Include the covariates `median age` and `location` and an interaction term for `age` by `location`.
28. Use the estimated coefficients from the interaction model found in Question 27 to write an expression for CTR cancer rates (not the logarithm of the rates) as a function of `age`. Repeat this process for the BGA location. The `location` variable will take the value 0 or 1, with `location BGA` assigned to be 0 and `location CTR` assigned to be 1. Note that when the `location` value is 0, the interaction term value (which is the product of the `location` and `median age` variables) will also be 0.
29. Use the results of Question 27 to estimate the multiplicative change in the cancer rate for a 10-year increase in `median age` when the `location` is CTR. Repeat this exercise for the BGA location.

## 8.9 Inference for Poisson Regression Models

In this section, we will formally assess the significance of the covariates in the Poisson regression model. The techniques in this section are very similar to those discussed in Chapter 7. Table 8.2 shows some of the output typically reported by computer software when fitting a Poisson regression model to the CTR cancer rates with `age` as the explanatory variable ( $x$ ). You should have obtained similar estimates and output from Question 15.

**Table 8.2** Results from a Poisson regression of cancer rates on median age for data from the Connecticut Tumor Registry.

Coefficient	Estimate	Std. Error	Z-value
Intercept ( $b_0$ )	-9.0762	0.0496	-182.9
Median Age ( $b_1$ )	0.0714	0.0008	91.8

Null deviance (restricted model): 11,402.834 on 3 degrees of freedom

Residual deviance (full model): 19.038 on 2 degrees of freedom

## Wald's Test

Table 8.2 reports the information needed to conduct **Wald's test**: the intercept and slope estimates, their standard errors, and corresponding **Wald statistics**. Wald's test can be used to determine if there is a significant linear relationship between the covariate(s) and the logarithm of the cancer rate (adjusting for other covariates in the model if there are multiple explanatory variables).

The Wald statistic for testing  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  is

$$Z = \frac{\text{estimate}}{\text{standard error}} = \frac{b_1}{\text{se}(b_1)} = \frac{0.0714}{0.0008} = 91.8$$

When the sample size is large, the Wald statistic has an approximate standard normal distribution. Hence, the Wald statistic of 91.8 corresponds to a very small  $p$ -value, and so we reject  $H_0: \beta_1 = 0$  to conclude that age has a significant linear relationship with the logarithm of cancer counts.

#### ► MATHEMATICAL NOTE ▼

There is no clear rule for a “large” sample size for Wald’s test; however, Wald’s test and likelihood ratio test (discussed below) have similar large sample properties. One suggestion<sup>8</sup> for the large sample condition to be met for the likelihood ratio test is to have the majority of the predicted Poisson means,  $\hat{\lambda}_i$ , greater than 5.

## The Likelihood Ratio Test

In addition to Wald’s test, we can use the **likelihood ratio test**, first introduced in Chapter 7, to determine if at least one covariate is significant in the model. The test compares the **full model**, which includes all the parameters under consideration, to the **restricted model**, which contains a subset of the parameters in the full model. In our current example with the CTR data, the full model contains the parameters  $\beta_0$  and  $\beta_1$ , while the restricted model contains the parameter  $\beta_0$ .

The LRT statistic,  $G$ , is given by

$$G = \text{residual deviance of restricted model} - \text{residual deviance of full model} \quad (8.18)$$

When the sample size is large,  $G$  has an approximate **chi-square distribution** with degrees of freedom equal to the number of parameters in the full model minus the number of parameters in the restricted model.<sup>9</sup> When the restricted model does not contain any covariates, it is often referred to as the **null model**.

Using the information displayed in Table 8.2, we can see that the LRT statistic for the CTR data example is given by  $G = 11,402.834 - 19.038 = 11,383.8$ . The degree of freedom is  $2 - 1 = 1$ , so the  $p$ -value is quite small ( $<0.0001$ ). Hence, the LRT also rejects the null hypothesis that  $\beta_1 = 0$ , and we conclude that age is significantly related to the logarithm of the cancer rate for the CTR location.

Wald’s test and the LRT will typically yield the same results, especially when the sample size is large. When the sample size is small, the LRT is more reliable and should be used. For the CTR data, the predicted Poisson means are 817.6, 2342.1, 1500.5, and 5590.8, so the large sample condition appears satisfied, and we should treat the results of either test as valid.

#### Key Concept

Wald’s test and the likelihood ratio test can be used to determine the significance of a covariate in a Poisson regression model if the sample size is large enough.

## Activity ► Assessing the Significance of Model Covariates

30. Examine the output for the interaction model estimated in Question 27. Is there significant evidence that the linear relationship between age and the logarithm of cancer rate depends on the location? Use Wald’s test to address this question.
31. Perform the LRT to determine if the relationship between age and the logarithm of cancer rate significantly depends on the location. Note that you will need to fit two models (one with and one without the interaction term) to obtain the residual deviances for the full and restricted models. Is your conclusion identical to that found in Question 30? Do you think the LRT is valid to use for the data? Why or why not?

Examination of the  $p$ -value associated with the interaction term helps us decide if the observed difference in the age trends at the two locations is something we could expect to have occurred simply by chance or if the observed difference in the trends is difficult to explain as chance variation (i.e., the  $p$ -value is small) and is suggestive of a significant difference in the trends.

You should find that the  $p$ -value for Wald's test for the model coefficient of the interaction term is about 0.012, while the  $p$ -value for the LRT is about 0.016. Both tests suggest that there is a statistically significant difference in the rate trends by age for the two locations. Thus, the effect of age on the logarithm of counts depends on which location you are considering.

#### ► MATHEMATICAL NOTE ▶

Caution should be used when conducting multiple Wald's tests. In Chapter 3, we showed that  $p$ -values are unreliable when multiple tests are conducted. We also showed in Chapter 3 that tests on individual coefficients can be unreliable when explanatory variables in the model are highly correlated. In cases where the significance of several covariates is being assessed, the LRT is preferred.

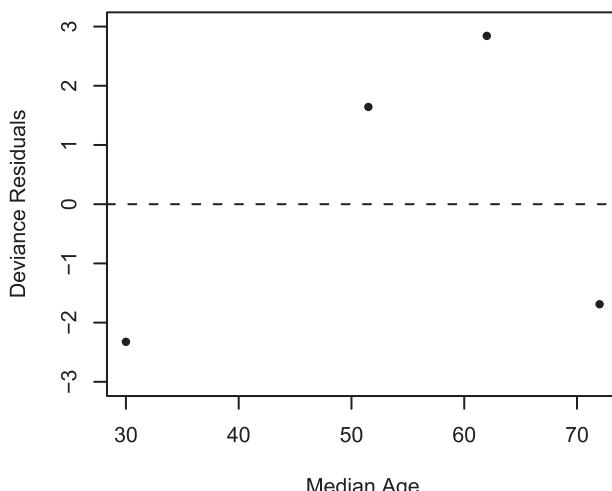
## 8.10 Assessing the Fit of the Poisson Regression Model

We will close this investigation with a look at some diagnostic tools used to assess the adequacy of a generalized linear model like the Poisson regression model: the **deviance residuals** and the **deviance statistic**.

Just as in least squares regression, described in Chapter 3, residual plots are useful for investigating whether the Poisson regression model is appropriate. A plot of the deviance residuals versus values of the covariate can be used to assess if there is lack of fit for individual observations, as well as indicate the potential form of the relationship between the covariate and the logarithm of the rate. The deviance residual (first described in Chapter 7) for the Poisson regression model is defined as

$$D_i = \pm \sqrt{2y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - 2(y_i - \hat{\lambda}_i)} \quad (8.19)$$

where the sign of the residual is positive if the observed count  $y_i$  is greater than the estimated mean count  $\hat{\lambda}_i$ . Figure 8.7 displays the deviance residuals after fitting a Poisson regression model to the CTR data, as you did in Question 15. Most computer packages have an option to report the deviance residuals and create this plot.



**Figure 8.7** Plot of the deviance residuals from a Poisson log-linear regression of cancer rate on median age, using the Connecticut Tumor Registry data as reported in the bottom half of Table 8.1.

Deviance residuals greater than 2 in absolute value indicate possible lack of fit. In Figure 8.7, two observations have deviance residuals greater than 2 in magnitude, which may indicate a poor-fitting model. We will address this potential problem in Questions 32–34. The sum of the squared deviance residuals is equal to the deviance statistic  $D^2$ , given by

$$D^2 = 2 \sum y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) \quad (8.20)$$

The deviance measures how much the model with the covariate(s) deviates from the **saturated model**—i.e., the model that simply uses the observed counts as the estimates for  $\lambda$  at each age observed in the data. This model will fit the data better than less complex models that attempt to fit a smooth relationship between  $E(Y) = \lambda$  and age, as the Poisson regression model does; however, the saturated model is not as parsimonious (i.e., it is not as simple to describe) as a GLM model function:  $\log(\lambda) = \log(t) + \beta_0 + \beta_1 x$ . A large deviance is “bad” in the sense that it suggests that the model did not describe the patterns of variability in the data well.

#### ► MATHEMATICAL NOTE ▾

The deviance statistic is also identical to the residual deviance commonly reported in the Poisson regression output. For example, see the “Residual deviance” in Table 8.2.

When the sample size is large such that the fitted Poisson means  $\hat{\lambda}_i$  mostly exceed 5, the deviance statistic has an approximate chi-square distribution with degrees of freedom equal to the number of observed counts  $n$  minus the number of parameters  $p$ .<sup>10</sup> Recall from Chapter 7 that the null hypothesis for this deviance test is that the model is a good fit. Thus, a small  $p$ -value (rejecting the null hypothesis) leads us to conclude that the model is not a good fit. A large  $p$ -value indicates that we have no evidence to suspect that the model is wrong. Just as in any hypothesis test, this does not mean that the model is correct. Perhaps we just do not have enough data to detect patterns that deviate from the model.

In the example used in Table 8.2 for the CTR data, we are modeling the logarithm of the cancer incidence rate as a linear function of age. There are  $p = 2$  parameters that are estimated, the intercept ( $\beta_0$ ) and slope ( $\beta_1$ ) of the linear function, and we observed  $n = 4$  counts for the CTR data. So the degrees of freedom for the deviance test are equal to  $n - p = 4 - 2 = 2$ . The deviance statistic reported in Table 8.2 is 19.038, which results in a small  $p$ -value (approximately 0) based on a chi-square distribution with 2 degrees of freedom. Hence, the deviance is so large that we reject the null hypothesis that the model fits the data well.\*

When the  $p$ -value is small, we conclude that the model is not a good fit. However, there are various reasons why the model may not fit well:

- The Poisson distribution may be the wrong model for the counts.
- The relationship between the logarithm mean count and age may not be linear or age alone may not be sufficient to describe the relationship (i.e., additional explanatory variables may be needed).
- There may be outliers in the data set.

#### Key Concept

The deviance statistic for a generalized linear model is a measure of how well the model summarizes the relationship between the covariate and the mean response compared to a saturated model (the observed response at each level of the covariate is used as the estimate). A small deviance implies that the Poisson regression model summarizes the data almost as well as the saturated model. If the deviance is small, we prefer the Poisson regression model because it is simpler to describe and allows us to make predictions for the mean response even at covariate levels not observed in the data.

\*The mean of a chi-square distribution is equal to the degrees of freedom. Thus, some statisticians suggest a general rule that rejects the null hypothesis if the deviance is much larger than the degrees of freedom.

## Activity Assessing the Model

32. When there are only four data points, we should be cautious about drawing any definite conclusions; however, Figure 8.7 does provide an indication of a pattern. Examine the residual plot in Figure 8.7 and comment on the appropriateness of modeling the cancer incidence rate as a linear function of age. Is there another relationship (not linear) between the incidence rate and age that you think might make more sense now that you have seen this residual plot?
33. Given your answer to Question 32, fit a different Poisson log-linear model to the CTR data that you believe may more accurately reflect the relationship between cancer incidence rates and age. Note that the models  $\log(\lambda) = \log(t) + \beta_0 + \beta_1 x^2$  and  $\log(\lambda) = \log(t) + \beta_0 + \beta_1 x + \beta_2 x^2$  are both still generalized linear models. The functions are linear in terms of the model coefficients, yet allow a nonlinear relationship between  $x$  and  $\log(\lambda)$ . What does the deviance statistic suggest about the fit of your new model?
34. Return to the model that you fit to the combined data (including an interaction term between age and location) in Question 27.
  - a. Assess the validity of the model using Wald's tests, the deviance statistic, and the deviance residuals.
  - b. Use the LRT to determine if the quadratic term  $(\text{age})^2$  should be incorporated into the model.

## 8.11 What Can We Conclude from the Cancer Rate Study?

The activities of this chapter have focused on fitting the Poisson log-linear model to the Bartlett–Green Acres cancer data and the Connecticut Tumor Registry data. The logarithm of the (mean) cancer rate was modeled as a function of the location, the (median) age of the subject, and an interaction term.

Wald's test shows that the interaction term ( $p$ -value = 0.012) is significant. Thus, there is evidence that the relationship between median age and the logarithm of cancer counts depends on location. There is no evidence that the cancer rate in Randolph (Bartlett-Green Acres) is significantly different from the cancer rates provided by the Connecticut Tumor Registry data. However, the deviance was 19.038, indicating that the model did not fit the data adequately.

The Poisson model was used to estimate the cancer rate of age and location. For the CTR data, when age increases by 10 years, we expect the cancer rate to be two times as large. For the BGA locations, an increase in age of 10 years has less impact on the cancer rate.

Plots of the deviance residuals indicate that a quadratic term may improve our Poisson regression model. The data do not represent a random sample from a larger population, but specifically compare two groups. While the Poisson regression model is relatively effective for the data shown in Table 8.1, it should not be expected to accurately estimate cancer rates in other locations or other time frames. In addition, this investigation is an observational study; thus, even if location were significant, we could not conclude that location *caused* different rates.

### A Closer Look

### Poisson Regression

## 8.12 Estimation Methods for Generalized Linear Models

You may be familiar with the method of **least squares estimation** used for simple linear regression, as in Equation (8.9). The basic idea is to make the errors, or **residuals**, as small as possible by finding estimates  $b_0$  and  $b_1$  for model parameters  $\beta_0$  and  $\beta_1$  that minimize the sum of the squares of the residuals for data collected on variables  $x$  and  $y$  from  $n$  individuals:

$$\begin{aligned} \text{sum of squared residuals} &= \sum_{i=1}^n (\text{residual}_i)^2 \\ &= \sum_{i=1}^n [\log(y_i) - (\log(t) + b_0 + b_1 x_i)]^2 \end{aligned} \quad (8.21)$$

There are at least two problems with using this approach for the log-linear Poisson regression model in Equation (8.7). First, as discussed above in our comparison of regression and generalized linear modeling, the emphasis of the least squares method would be on finding estimates that fit a linear function to the logarithm of the individual responses:  $\log(y_i)$ . Our actual goal is fitting a linear model to the logarithm of the mean response:  $\log[E(Y)] = \log(\lambda)$ .

Second, the least squares method treats each residual the same way, but from Figure 8.5 we can see that, depending on the value of the covariate ( $x$ ), some residuals will naturally be quite large (when the Poisson mean is large) and the distribution of residuals will be right skewed (when the Poisson mean is small). This is due to the fact that the Poisson probability distribution is right skewed (see Figure 8.3 for a reminder). There is a method called *weighted least squares* that might help, but a better solution is to use a method that gives good estimates for lots of generalized linear modeling situations—not just for the Poisson regression model specifically.

The method commonly used by statistical computing packages to find parameter estimates for generalized linear models is called **maximum likelihood estimation**. The goal of this estimation method is to find estimates  $b_0$  and  $b_1$  for  $\beta_0$  and  $\beta_1$  in Equation (8.7) such that the observed data are most likely to have occurred if  $\beta_0 = b_0$  and  $\beta_1 = b_1$  and less likely to have occurred if  $\beta_0$  and  $\beta_1$  are any other two numbers.

## Activity ▶ Maximum Likelihood Estimation

35. Here is a simple exercise that might give you a better sense of the idea of maximum likelihood estimation. Suppose you observe that, among all the students taking quantitative classes with you now and in recent semesters or quarters, about 30% use a Macintosh computer and 70% use a Windows PC. Let's consider possible values for the true population parameter  $p$  = the proportion of students on campus recently taking quantitative courses who use a Macintosh computer.

Do you think that your experience of observing 30% Macintosh users is more likely to occur if  $p = 80\%$  or if  $p = 40\%$ ? Do you think observing 30% Macintosh users is more likely to occur if  $p = 40\%$  or if  $p = 35\%$ ? What about if  $p = 35\%$  or if  $p = 32\%$ ? Do you think observing 30% Macintosh users is more likely to occur if  $p = 20\%$  or if  $p = 25\%$ ? What about if  $p = 25\%$  or if  $p = 28\%$ ?

You should be seeing a pattern now. Observing 30% Macintosh users is most likely if indeed  $p = 30\%$ . We say that “given the observation that 30% of students were Macintosh users, the maximum likelihood estimate for the true proportion  $p$  is  $\hat{p} = 30\%$ .”

Many statistical models are more complex than the example in Question 35. In particular, for the Poisson GLM in Equation (8.7), we would need to use some methods from calculus to derive expressions for  $b_0$  and  $b_1$  and simple expressions for a solution are not always available. Usually, statistical computing software proceeds as follows. First, initial estimates  $b_0$  and  $b_1$  are determined from the data—perhaps using the least squares method as a starting point. Successive choices for  $b_0$  and  $b_1$  are made from there by an algorithm\* that ensures that the successive estimates make the data more likely to have occurred, using the new estimates for the parameters. Eventually, the next choices for  $b_0$  and  $b_1$  will be quite similar to those at the last step and the algorithm is said to have “converged” to the maximum likelihood estimates for  $\beta_0$  and  $\beta_1$ . This is much like the process of narrowing your way down to the “best” estimate that we led you through in Question 35.

## 8.13 Do No-Smoking-at-Work Policies Keep Smoking at Home?

Several regular smokers were selected from a larger Minnesota survey about smoking habits. All reported smoking about the same number of cigarettes per day in recent months. Each was asked to report on the number of cigarettes smoked the day before (January 26) during the same 2-hour period. They were also asked whether they had been at work or at home or elsewhere during that 2-hour period.

\*We will let the computer handle the task of calculating maximum likelihood estimates (MLEs), but if you are curious, common algorithms to compute MLEs are the Newton-Raphson and Fisher scoring methods.

Six of the smokers were either at work or at home during the 2-hour period. Data on the number of cigarettes smoked by these 6 smokers in the 2-hour period and whether at home or at the office (which requires smokers to go outside in Minnesota in January!) appear in Table 8.3.

**Table 8.3** Data on six smokers for one 2-hour period. The response variable is the number of cigarettes smoked during the period, and the explanatory variable is 0 if the person was at home or 1 if the person was at work.

<b>Number of Cigarettes</b>	3	0	0	1	2	1
<b>Location</b>	0	1	1	1	0	0

## Extended Activity

### Smoking Habits

Data set: Smoking

36. Are the observed responses (3, 0, 0, 1, 2, and 1) counts or rates? What are the units of these values?
37. What is the amount of exposure related to each of the observed responses? Note that since exposure is identically equal to one (i.e., one 2-hour period) for each individual, no offset will be needed in the generalized linear model, as in Equations (8.7) and (8.10).
38. Why do you think it was important to choose subjects who “all reported smoking about the same number of cigarettes per day in recent months”?
39. Hypothesize the direction in which you expect the number of cigarettes smoked to differ between work and home.
40. Calculate the average number of cigarettes smoked during a 2-hour period at each location. Does the observed difference support your hypothesis in Question 39?
41. Write down the Poisson regression model equation relating the logarithm of the mean count of cigarettes smoked to the covariate location. Then write two separate equations for the model: one for each of the possible locations (location = 0 if the person is at home, and location = 1 if the person is at work). Set each of these two equations equal to the logarithm of the corresponding average that you found in Question 40. Solve the resulting system of two equations and two unknowns ( $\beta_0$  and  $\beta_1$ ) for estimates of the intercept and slope ( $b_0$  and  $b_1$ ).
42. Interpret the exponentiated slope estimate in the context of these data.
43. If the Poisson model is correct, then according to the facts about the mean and standard deviation of the Poisson probability distribution given in Equation (8.6), we should find that the observed average number of cigarettes smoked is similar in size to the observed sample variance. Calculate the sample variance in the number of cigarettes smoked during a 2-hour period at each location and compare the results with the averages you already calculated in Question 40.
44. Use the computer to fit a Poisson regression model to the data with  $Y$  = number of cigarettes smoked during the 2-hour period as the response variable and location as the explanatory variable.
45. How are the estimates in Question 44 related to the estimates that you calculated in Question 41?
46. According to the computer output, is the difference in the expected number of cigarettes smoked between locations statistically significant? Do you think that the test is valid to use for these data? Why or why not?
47. Perform a randomization test similar to those discussed in Chapters 1 and 6 to determine if there is a statistically significant difference in the smoking rate depending on location.

### Another Method for Assessing Model Fit: Pearson Residuals and Statistic

In Section 8.10, we looked at the deviance statistic from a Poisson regression and examined a graph of the deviance residuals in Figure 8.7. **Pearson residuals** can also be used to test whether a model fits the data well. Pearson residuals are fairly easy to understand compared to deviance residuals. The idea is this. A basic residual is

$$\text{observed count} - \text{estimated count from the model} = y_i - \hat{\lambda}_i = y_i - e^{b_0 + b_1 x_i}$$

We often look at standardized residuals by dividing each residual by an estimate of the standard deviation of the observation. According to Equation (8.6), a feature of a Poisson count model is that its expected value,  $E(Y) = \lambda$ , is also its variance:  $\text{Var}(Y) = \lambda$ . So the standard deviation of a count  $Y_i$  is  $\sqrt{\lambda_i}$ . The standardized, or Pearson, residual is thus

$$r_i = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \quad (8.22)$$

where  $\hat{\lambda}_i$  is the predicted count for individual  $i$ . The sum of squared Pearson residuals is called the Pearson chi-square statistic:

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \quad (8.23)$$

Just as for the deviance statistic, when the amount of data is large, the Pearson chi-square statistic can be compared to a chi-square probability distribution with  $n - p$  degrees of freedom to assess whether the model fit is reasonable.<sup>11</sup>

## Extended Activity Calculating Pearson Residuals

Data set: Smoking

- 48. Using your earlier calculations of Poisson regression parameter estimates for the Minneapolis smokers data, calculate the six Pearson residuals and sketch a graph of the residuals versus the estimates for the average cigarette count for the two locations.
- 49. Calculate the Pearson chi-square statistic and compare it to the deviance calculated by the computer in Question 44. Are these two values similar, as expected?

Note that since all the predicted mean counts are less than 5, we should not attempt to calculate a  $p$ -value for the Pearson chi-square (or deviance) statistic.

## 8.14 Is the Number of Species on Archipelago Islands Related to Island Area, Elevation, and Neighboring Islands?

The 30 islands in the Galápagos archipelago have long been studied by botanists, zoologists, and biologists to learn about species survival and the process of natural selection in an almost experimental setting. The islands are essentially uninhabited by humans, and all experience the same surrounding climate. Yet some species of birds, plants, and mammals thrive on only a few or even just one of the islands. In addition, some islands have a wide variety of species, while others are not nearly as biodiverse.

Data on plant species on the Galápagos islands have been collected more than once. The most recent extensive data were reported by Johnson and Raven.<sup>12</sup> We will consider two response variables: island total observed species count and island endemic species count. There are five potential covariates that we expect to be related to species count: island area ( $\text{km}^2$ ) and elevation (meters), the distance (km) from the island to its nearest neighbor (adjacent island) and to the largest island ( $\text{km}^2$ ) in the archipelago (Santa Cruz), and the area of the adjacent island ( $\text{km}^2$ ).

If we are interested in testing the hypothesis that several covariates significantly contribute to the Poisson regression model's ability to predict the mean count, then we can use the likelihood ratio test discussed in Section 8.9. For the Galápagos islands data, suppose we want to test whether at least one of the covariates  $\log(\text{area})$  or  $\log(\text{elevation})$  is significant. That is, we want to test

$$H_0: \beta_1 = \beta_2 = 0 \text{ vs. } H_a: \text{at least one coefficient is not zero} \quad (8.24)$$

Then the LRT statistic is given by  $G = 3511 - 646.2 = 2864.8$  with  $3 - 1 = 2$  degrees of freedom. The corresponding  $p$ -value  $<0.0001$ , so we reject the null hypothesis and conclude that at least one of the covariates is significantly related to the logarithm of the mean total species count.

## Extended Activity

### Modeling the Number of Plant Species in the Galápagos

Data set: gala

50. Using the Galápagos islands data (`gala`), create plots of the logarithm of the observed counts of total plant species,  $\log(\text{species})$  versus each of the five potential covariates: `area`, `elevation`, `nearest`, `scruz`, and `adjacent`. Does the Poisson regression model seem a good choice for these data? Why or why not? If so, which covariates do you think will be most helpful in a model attempting to explain the variability in island species counts?
51. Repeat the previous question using the logarithm of each of the covariates in place of the original covariate value. Caution: Always check for any zero values before using a logarithm transformation. For these data, the island of Santa Cruz is 0 km from itself. A quick fix is to set the value at a very small nonzero number (e.g., 0.1 km).
52. Fit a Poisson regression model to the total species count with covariates  $\log(\text{area})$  and  $\log(\text{elevation})$ . Confirm the results of the LRT discussed above to determine if at least one of the two covariates is significant.
53. Perform the LRT to determine if the addition of `nearest`, `scruz`, and `adjacent` to the restricted model fit in Question 52 significantly improves the model's predictive ability.
54. Examine the computer output for the full model from Question 53. Are there any covariates that *do not* appear significant in the model for species count, but *did* appear to be strongly related to the species count in your graphs from Questions 50 and 51?

### Multicollinearity in Poisson Regression

If you have some experience with multivariate regression, you may recognize the odd result observed in Question 54 as a problem of **multicollinearity**—some of the covariates are highly correlated with each other (so we may need just one in the model, not both).

## Extended Activity

### Examining Multicollinearity

Data set: gala

55. Calculate the sample correlation between all the pairs of covariates and comment on whether or not multicollinearity could be a source of the apparent paradox between the statistical significance of the covariate and the appearance of the graph.
56. Given the computer output from Question 53, are there any covariates that appear significant in the model for species count but *did not* appear to be strongly related to the species count in your graphs from Questions 50 and 51?
57. Report the deviance statistic from the computer output. Calculate the appropriate model degrees of freedom for the deviance statistic. Compare the deviance statistic to a chi-square probability distribution with the same degrees of freedom. If the observed deviance is large, then the model deviates significantly from the saturated model. We would then conclude that our simpler, generalized linear model does not explain the variability in species count well.

In the model constructed in Question 53, you should have found that, island area appears significantly related to species number, while island elevation does not. However, your graphs in Questions 50 and 51 should have indicated that both of these variables on the log scale were strongly related to species count. The explanation is that the correlation between the logarithms of these two covariates is greater than 0.90. They are highly correlated variables. This makes sense. Only a large island can have a very high peak.

You should have also found in Question 53 that the area of the nearest island (`adjacent`) and the distance from Santa Cruz island (`scruz`) are statistically significant covariates in the Poisson log-linear model for the mean species count. However, your graphs in Questions 50 and 51 should have indicated that neither of these two variables was strongly related to species count.

So we have an apparent paradox. The output from the computer software implies that three of the five covariates are significant:  $\log(\text{area})$  distance from nearest island and distance from Santa Cruz island. In fact,

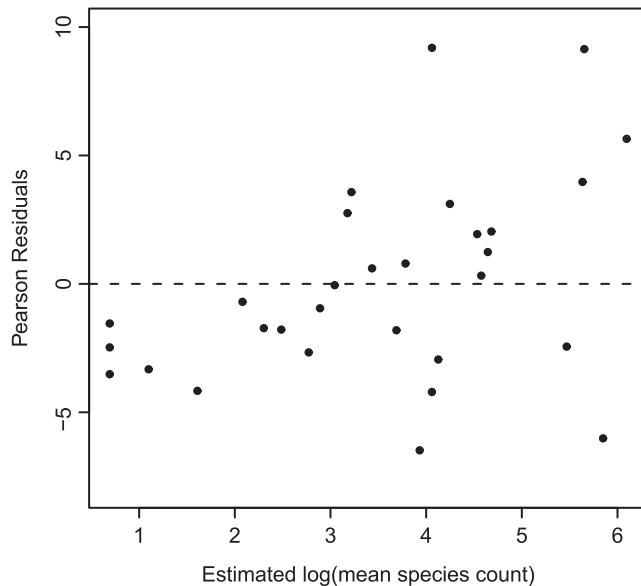
the largest of the  $p$ -values for testing  $H_0: \beta_i = 0$  for each of these three variables is still less than 0.0001. However, the graphs indicate that, of these three variables, only  $\log(\text{area})$  is strongly related to species count.

The deviance statistic provides more evidence that something is not quite right. If the model fits so well (three out of five covariates are statistically significant), then why is the deviance (427.48) so large ( $p$ -value  $<0.0001$  on  $n - p = 30 - 6 = 24$  degrees of freedom)? After all, the log-linear model for the mean seems reasonable when we look at the plots of the logarithm of the observed counts versus the logarithm of each of the covariates, as in Question 51.

Perhaps the model does not fit well in some other way. Let's look more closely at the variance. That is, let's examine the residuals. Perhaps there are a couple of outliers contributing to the large deviance. Perhaps the residuals are larger than the model would suggest (the Pearson residuals should be approximately normal with mean zero and variance one when we have a lot of data). Here, we have 30 observed counts. None are zero and only four of the 30 are less than 5. Figure 8.8 shows the Pearson residuals for the generalized linear model

$$\begin{aligned}\log(\lambda) = & \beta_0 + \beta_1 \log(\text{area}) + \beta_2 \log(\text{elevation}) \\ & + \beta_3 \text{nearest} + \beta_4 \text{scruz} + \beta_5 \text{adjacent}\end{aligned}$$

where  $\lambda = E(Y) = \text{mean species count}$  and  $Y$  follows a Poisson probability distribution.



**Figure 8.8** Pearson residuals from a Poisson regression of the mean species count for the 30 Galápagos islands. Five covariates were used in the model: island area and elevation, nearest island area, and distance from nearest island and from Santa Cruz island.

## Poisson Variance Is Larger Than the Mean: Overdispersion

From Figure 8.8, we can observe that some Pearson residuals are quite large (near 8 or 10) and many are moderately large (around 5). This implies that there is more variability in the data than the Poisson model assumes. Most importantly, this means that the standard errors reported for the Wald statistics are likely underestimated. That is, the Wald statistics reported for testing the significance of a covariate in the model are likely too large. So the reported  $p$ -values are too small. We should not make decisions about the significance of covariates from these  $p$ -values.

In this situation, we say that the Poisson model is **overdispersed** (has more spread in the data than the model implies). Remember that for a Poisson probability model,  $E(Y) = \lambda$  and  $\text{Var}(Y) = \lambda$ . One method for handling the problem of overdispersion is to modify this model somewhat and let  $\text{Var}(Y) = \phi\lambda$ , where  $\phi$  is

called the **overdispersion parameter**. Presumably,  $\phi > 1$  for the Galápagos islands species count model, since the variability observed is larger than a simple Poisson model implies.

There are two estimates for  $\phi$  that are in common use: the deviance statistic over the model degrees of freedom and the sum of the squared Pearson residuals (the Pearson chi-square statistic) over the degrees of freedom. Here are formulas for the common estimates of the dispersion parameter:

$$\hat{\phi} = \frac{\text{deviance statistic}}{n - p} \quad \text{or} \quad \hat{\phi} = \frac{\chi^2}{n - p} \quad \text{where } \chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i} \quad (8.25)$$

is the Pearson chi-square statistic introduced in Equation (8.23). Then the Wald statistic for testing  $H_0: \beta_1 = 0$  versus  $H_a: \beta_1 \neq 0$  is adjusted to be

$$Z = \frac{\text{estimate}}{\text{standard error}} = \frac{b_1}{\sqrt{\hat{\phi}} \text{ se}(b_1)} \quad (8.26)$$

Since  $\hat{\phi}$  will be larger than one for an overdispersed model, the Wald statistic will be adjusted down and the  $p$ -value will be larger. Most statistical computing software will make this adjustment for you if you instruct it to fit the overdispersed Poisson log-linear regression model. Theoretically, the  $t$ -distribution is most appropriate when testing coefficients in overdispersed Poisson models and is often used in statistical software; however, the normal approximation (and hence the Wald test) are typically adequate.<sup>13</sup>

## Extended Activity

### Overdispersion

Data set: gala

58. Fit an overdispersed Poisson log-linear regression model to the Galápagos island species count data, using the total species count as the response. Use all five covariates (with area and elevation on the log scale) and compare the adjusted  $p$ -values with those you obtained in Question 53. Which variables are no longer significant in the model?
59. Remove insignificant covariates from the model in Question 58 using backward elimination. For example, remove the “least significant” (largest  $p$ -value) covariate first and refit the model. Then remove the next “least significant” covariate from that model. Repeat until there are only significant variables in the model. Use the results to estimate the species count for an island in a different but very similar archipelago. This island has area 2 km<sup>2</sup> and elevation 200 m; it is 3 km from the nearest other island in the archipelago and 20 km from the largest island in the archipelago, and the area of the adjacent island is 15 km<sup>2</sup>.

## Chapter Summary

This chapter has focused primarily on fitting a Poisson log-linear model, also referred to as a Poisson regression model, to count data. The Poisson model is an example of a **generalized linear model** that uses the **link function**  $g(\lambda) = \log(\lambda)$  to model the relationship between the mean response count,  $E(Y) = \lambda$ , and the covariate,  $x$ . This relationship can be expressed as

$$\log[E(Y)] = \log(\lambda) = \log(t) + \beta_0 + \beta_1 x$$

where  $\log(t)$  is known as the **offset** and is used to account for the exposure level, if present. **Exposure** is a measure related to the number of subjects in a group or the duration of time that subjects have been at risk. The offset term will be needed in the model if the exposure level is not equal to one for each subject in the model. This is to allow for fair comparison of the counts.

The parameters of the Poisson model are estimated using **maximum likelihood estimation** and computational algorithms that are implemented in statistical software packages. To test whether a covariate is significant, we can use the **Wald statistic**, defined as

$$Z = \frac{b_1}{\text{se}(b_1)}$$

This statistic follows a standard normal distribution when the sample size is large.

To assess the overall fit of the Poisson model, we can use the **deviance statistic** or the **Pearson chi-square statistic**, given, respectively, by

$$D^2 = 2 \sum_{i=1}^n y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) \quad \text{and} \quad \chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}$$

which follow approximate chi-square distributions with  $n - p$  degrees of freedom when the sample size ( $n$ ) is large and  $p$  is the number of parameters in the model (including  $\beta_0$ ). A small  $p$ -value corresponding to either statistic indicates that the Poisson regression model does not fit the data well. The deviance and Pearson chi-square statistics are appropriate when each of the predicted Poisson mean counts,  $\hat{\lambda}_i$ , is greater than 5.

Some problems that may occur with fitting the Poisson model to count data include **multicollinearity** and **overdispersion**. Multicollinearity is a problem that occurs when the covariates are highly correlated with each other. If multicollinearity is present, then the regression coefficients and their corresponding  $p$ -values can be unreliable. Overdispersion occurs when there is more variability in the response data than the Poisson model assumes. If overdispersion is present, then the standard errors of the estimated coefficients will be underestimated, resulting in tests that are biased toward finding significant covariates. A plot of the Pearson residuals versus the estimated log counts can usually indicate whether overdispersion is present. To account for overdispersion in the model, the **dispersion parameter**,  $\phi$ , can be estimated using

$$\hat{\phi} = \frac{\text{deviance statistic}}{n - p} \quad \text{or} \quad \hat{\phi} = \frac{\chi^2}{n - p}$$

where  $\chi^2$  is the Pearson chi-square statistic in Equation (8.23). The estimated dispersion parameter can be incorporated into the expression for the Wald statistic, and adjusted tests can be conducted to assess the significance of the covariates.

## Exercises

---

- E.1. In 1996, the Bureau of Consumer Protection of the Federal Trade Commission (FTC) conducted a study to investigate the accuracy of electronic checkout scanners. They inspected the pricing accuracy of scanners on 17,928 items in 294 department, discount, drug, food, and other retail stores, and discovered that 4.82% of the items scanned for amounts that differed from the advertised or posted price. Let's assume that 4.82% represents the true error rate.
- Briefly explain why the binomial probability model is appropriate for the counts of mis-scanned items in a random selection of scanned items.
  - Find the probability that in the next 1000 randomly selected items that are scanned for purchase, exactly 15 are scanned incorrectly for price.
  - Is it reasonable to use the Poisson model to approximate the probability found in Part B?
  - Regardless of your answer to Part C, use the Poisson probability to find the approximate probability that exactly 15 are scanned incorrectly for price. Is the Poisson approximation close to the exact binomial probability?
- E.2. Suppose your friend types very accurately, making a typographical error in only 1 out of every 2500 words.
- In a 10,000-word essay, what is the probability that he makes exactly 3 errors? Use the binomial model to answer this question.
  - Briefly explain why it is appropriate to answer the above question using the Poisson approximation.
  - Answer Part A using the Poisson probability model. How do your two answers compare?
- E.3. **Cancer Counts Revisited**

Data set: BGA\_CTR

Consider the cancer count data for the Bartlett–Green Acres neighborhood. Between 1964 and 1984, the number of persons at risk was 2314 and the mean duration of residence was 11.5 years, yielding

26,611 for the number of person-years of exposure. Assuming that the national cancer incidence rate is 326 cases per 100,000 people in a year, answer the following:

- Using the binomial model, find the exact probability that at least 67 cancer cases are observed.
- Since  $p = 0.00326 < 0.02$  and  $np > 5$ , the Poisson model will provide a good approximation to the binomial model. Using the Poisson model, compute the approximate probability that at least 67 cancer cases are observed. Compare your answer to that from Part A.
- Do your answers appear consistent with the information provided in Figures 8.1 and 8.4?

#### E.4. Death by Horse Kicks

Data set: `horsekick`

The form for the distribution shown in Equation (8.5) was discovered and published by Siméon-Denis Poisson (1781–1840). One of the first applications of the Poisson distribution was in the 19th century, to model the number of deaths by horse or mule kicks in 10 corps of the Prussian army.<sup>14</sup> Preece, Ross, and Kirby have a nice description of the data in their paper, which uses Poisson regression.<sup>15</sup> In fact, there were 14 corps, but 4 are typically deleted.

The `horsekick` Data set contains the total number of deaths by horse kicks per year for 10 corps of the Prussian army over a 20-year time period. Perform a Poisson regression analysis using deaths per corp as the response and year as the covariate, and address the following questions.

- Conduct Wald's test to determine if there is significant evidence that the number of deaths by horse kick depends on time.
- Use the deviance and Pearson statistics to investigate goodness of fit of the model.
- Examine the Pearson residuals to check for overdispersion.

#### E.5. AP Exams

Data set: `APexams`

Are high-achieving high school students more inclined to take Advanced Placement exams? In a recent student survey, the co-author of this book asked students in his statistics courses to report the number of Advanced Placement (AP) exams they had taken in high school. In addition, information on their self-reported cumulative high school grade point average (`HSGPA`) and gender was recorded. Assume that students attended high school for four years.

- What is the amount of exposure related to each of the observed responses? Is the exposure for any of the subjects different from the exposure for the others?
- Is it necessary to include an offset for these data in the generalized linear model, as in Equations (8.7) and (8.12)? Why or why not?
- Fit a Poisson regression model to the data with  $Y = \text{number of AP exams taken while attending high school}$  as the response and `HSGPA` as the explanatory variable.
- Based on the computer output, interpret the estimated coefficient corresponding to `HSGPA`.
- What is the estimated multiplicative increase in the rate of taking AP exams for any half-unit increase in `HSGPA`?
- Investigate the adequacy of the model, as well as any departures from the model assumptions, such as zero counts and overdispersion. Should the results of Wald's test be trusted? Why or why not?

#### E.6. Is the Number of Hurricanes Increasing?

Data set: `hurricanes`

With all the discussion about climate change in the past couple of decades, it is no surprise that the occurrence of several intense hurricanes over the past several years has garnered some attention. You may remember that in 2005 the U.S. Gulf coast was hit by hurricanes Emily, Katrina, and Rita. All three were category five on the Saffir-Simpson scale. That is the highest intensity on this five-level scale.

The Data set `hurricanes` contains 59 lines, one for each year from 1950 through 2008. For each year, there are five columns of data containing counts of the number of hurricanes of each category type on the Saffir-Simpson scale (labeled `cat1`, `cat2`, `cat3`, `cat4`, or `cat5`). The data were obtained from <http://weather.unisys.com/hurricane/atlantic>.

- Model the total count of hurricanes per year as a function of time (the covariate) and report whether you find a significant time trend in the count. Of course, check graphically that the model you

propose seems reasonable, examine the residuals, and use Wald's test or a deviance test to make your decision.

- Repeat the process in Part A, using the count of intense hurricanes per year (category three or higher) as the response.

#### E.7. Assessing Skin Cancer Risk by Age

Data set: SkinCancer

Consider the data in Table 8.4, giving the incidence of non-melanoma skin cancer among women in Minneapolis–St. Paul by age group.<sup>16</sup>

**Table 8.4** Reported cases of non-melanoma skin cancer in Minneapolis–St. Paul.

Age Group	Number of Cases	Population Size
15–24	1	172,675
25–34	16	146,207
35–44	30	121,374
45–54	71	111,353
55–64	102	83,004
65–74	130	55,932
75–84	133	29,007
85+	40	7,538

The Data set SkinCancer contains the data in Table 8.4. The age groups have been recoded to have a value near the midpoint of each group (20, 30, 40, 50, 60, 70, 80, and 90).

- Consider age to be a continuous variable and model the non-melanoma skin cancer incidence as a function of age. Do not forget to include the population as an exposure offset in your model. Report whether you find a significant age trend in the count. Of course, check graphically that the model you propose seems reasonable, examine the residuals, and use Wald's test or a deviance test to make your decision.
- How well do the observed case counts match up with the counts predicted by the model?
- Try to improve on your model. Is the logarithm of the number of cases growing linearly with age or nonlinearly? Consider adding  $(\text{age})^2$  or even  $(\text{age})^3$  to your model. Try out some models and decide which one you think is best. Explain your choice.
- Report and interpret the coefficient for age in the model. How does the risk of cancer for a woman aged 50 years relate to the risk for a woman aged 40 years?

#### E.8. Skin Cancer by Age and Location

Data set: SkinCancer

- Use age as an explanatory variable to create a Poisson regression model for the incidence rate of skin cancer among women in Dallas–Fort Worth.
- Is age a significant predictor of the incidence rate of skin cancer among women in Dallas–Fort Worth?
- Using only the data from Dallas–Fort Worth, how does the risk of cancer for women aged 50 years relate to the risk for women aged 40 years?
- Create a model using both locations (Minneapolis–St. Paul and Dallas–Fort Worth), age, and an interaction term to estimate the incidence rate for skin cancer. Does one city have a higher risk of non-melanoma skin cancer for women than the other? Is there an interaction between age and location?

### E.9. Modeling the Number of Endemic Plant Species in the Galápagos

Data set: gala

You have already investigated Poisson regression models for the total number of plant species in the Galápagos islands. This exercise will examine Poisson models using the number of *endemic* plant species (species found only on that particular island) as the response.

- a. Fit a Poisson regression model to the endemic species count (`endemic`) with covariates `log(area)` and `log(elevation)`. Conduct the LRT to determine if at least one of the covariates is significant. Examine the deviance and Pearson statistics for this model, and comment on the model's goodness of fit.
- b. Determine if addition of the covariates `nearest`, `scruz`, and `adjacent` significantly improves the model's ability to predict the mean number of endemic species. Examine the deviance and Pearson statistics for this model, and comment on the model's goodness of fit.
- c. Based on your answers to Parts A and B, which final model would you select for the endemic species counts? For the model you chose, examine the Pearson residuals to check for overdispersion, and fit an overdispersed model if necessary.

## Endnotes

---

1. G. Polya, *How to Solve It*, 2nd ed. (Princeton: Princeton University Press, 1957), p. 114. Polya (1887–1985) was a Hungarian mathematician known for his work in problem solving.
2. Those who are interested can check out the Internet Movie Database <http://www.imdb.com>.
3. J. Harr, *A Civil Action* (New York: Vintage Books, 1996) is the classic book-length account of the cluster of childhood leukemia cases in Woburn and the court case that grew out of the medical investigation.
4. You can read more at <http://epi.grants.cancer.gov/LIBCSP>.
5. See R. Day, J. H. Ware, D. Wartenburg, and M. Zelen, “An Investigation of a Reported Cancer Cluster in Randolph, Massachusetts,” *Journal of Clinical Epidemiology*, 42 (1989):137–150.
6. D. Wartenburg, “Should We Boost or Bust Cluster Investigations?” *Epidemiology*, 6.6 (1995): 575–576.
7. A. P. Polednak, *Cancer Incidence in Connecticut 1980–1996*, October, 1999.
8. See A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed. (New York: Wiley, 2007).
9. Ibid.
10. Ibid.
11. Ibid.
12. M. P. Johnson and P. H. Raven, “Species Number and Endemism: The Galápagos Archipelago Revisited,” *Science*, 179.4076 (1973): 893–895.
13. P. McCullagh, and J. A. Nelder, *Generalized Linear Models*, 2nd ed. (Boca Raton, LA: Chapman and Hall, 1989), p. 200.
14. L. von Bortkewitsch, *Das Gesetz der kleinen Zahlen* (Leipzig: B. G. Teubner, 1898).
15. D. A. Preece, G. J. S. Ross, and P. J. Kirby, “Bortkewitsch’s Horse-Kicks and the Generalised Linear Model,” *Journal of the Royal Statistical Society: Series D (The Statistician)*, 37 (1988): 313–318.
16. D. Hand et al., *A Handbook of Small Data Sets* (London: Chapman and Hall, 1994).
17. M. Huber and A. Glen, “Modeling Rare Baseball Events: Are They Memoryless?” *Journal of Statistics Education*, 15.1 (2007), [http://www.amstat.org/publications/jse/v15n1/data\\_sets.huber.html](http://www.amstat.org/publications/jse/v15n1/data_sets.huber.html).

# Research Project: Hitting a Grand Slam in Baseball

## Reviewing the Literature

In this project, you will use Poisson regression to model rare events in Major League Baseball. Some rare events in baseball include pitching a no-hitter (a game without a hit) and hitting a grand slam. A grand slam is a home run with the bases loaded. In the five seasons between 2006 and 2010, the average number of grand slams per team per season was only about 4. Data on many commonly recorded characteristics of the 30 Major League Baseball teams were collected from the Website [http://mlb.mlb.com/stats/historical/team\\_stats.jsp](http://mlb.mlb.com/stats/historical/team_stats.jsp) and will be used to investigate possible models for the number of grand slams over the past five seasons. Huber and Glen discuss some applications of the Poisson distribution to model rare events in baseball, but they exclude grand slams and do not consider Poisson regression.<sup>17</sup>

## Exploring the Data

The primary response variable for this project will be the total number of grand slams each team has had over the past five seasons. The data set `Grandslam` contains measurements on several commonly recorded statistics (e.g., numbers of grand slams, home runs, and runs batted) for 30 Major League Baseball teams. The values for the variables `GrSlam` (number of grand slams), `Cycles` (number of cycles), and `Games` (number of games) are count totals over five seasons from 2006 to 2010. The values for the remaining variables are averages over the five seasons. Brief descriptions of the variables are provided below.

`Team.Name`: name of team

`GrSlam`: total number of grand slams (home runs with bases loaded)

`Cycles`: total number of cycles (when a player hits at least one single, double, triple, and home run in one game)

`League`: AL (American League Division) or NL (National League Division)

`Games`: total number of games played

`At.Bats`: number of players who came to the plate to hit in the given timeframe

`Runs`: runs scored

`Hits`: number of times the batter safely reached first base after hitting the ball into fair territory, without the benefit of an error or a fielder's choice

`Doubles`: doubles hit

`Triples`: triples hit

`Hm.Runs`: home runs

`RBI`: runs batted in (runs scored and credited to the batter)

`TB`: total number of bases, calculated by summing 1 for each single, 2 for each double, 3 for each triple, and 4 for each home run

`BB`: base on balls (walks due to 4 balls thrown by pitcher)

`SO`: strikeouts (3 strikes)

`SB`: stolen bases (runner advances without a hit or walk)

`CS`: caught stealing (runner tried to steal base and got thrown out)

`OBP`: on-base percentage (the total number of hits, bases on balls, and times hit by pitch, divided by the total number of at-bats, bases on balls, times hit by pitch, and sacrifice flies)

`SLG`: slugging percentage (the total bases recorded by the batter, divided by the total number of at-bats)

`AVG`: batting average (the number of base hits, divided by the total number of at-bats)

The following Website has the full glossary of abbreviations, and you can click on each acronym for a more complete definition: [http://mlb.mlb.com/mlb/official\\_info/baseball\\_basics/abbreviations.jsp](http://mlb.mlb.com/mlb/official_info/baseball_basics/abbreviations.jsp). To access additional MLB information and statistics, you can also visit the Website [http://mlb.mlb.com/stats/historical/team\\_stats.jsp](http://mlb.mlb.com/stats/historical/team_stats.jsp) and follow these instructions:

- In the left-hand column is a selectable search setting; select the “Historical Team Stats” tab.
- Select “Major League.”
- Select “All Teams.”
- Select “Hitting Stats.”
- Under the window labeled “Splits,” select “Bases Loaded,” which allows you to find the number of home runs with bases loaded (a.k.a. grand slams) that were hit by each team.
- Under “Season Stats,” highlight the season year desired, and then hit “Go.”

Once you have accessed the baseball data, address the following questions:

1. Construct a histogram of the grand slam counts. Describe the shape of the distribution. Does this graph look like a Poisson distribution?
2. Identify which explanatory variables in the data set you would expect to influence the number of grand slams. You may want to conduct some background research to determine which variables could be important.
3. Plot variables of interest against the logarithm of your response variable (number of grand slams), and examine the correlations between the response variable and the other covariates in the data set. Which covariate has the largest correlation with the logarithm of grand slams?
4. Examine the correlations between the explanatory variables. Are any of the explanatory variables highly correlated with each other?
5. Construct a preliminary Poisson regression model with the covariate that is most highly correlated with the logarithm of number of grand slams. Report and interpret Wald’s test for the coefficient, and assess the fit of the model with the deviance statistic. Check that most of the predicted Poisson means are greater than 5. What do you conclude based on the results of the tests?
6. Add additional explanatory variables to your Poisson regression model. Also consider interactions between covariates. Be sure to perform the likelihood ratio test (drop-in-deviance test) to determine whether the covariates you are considering for inclusion are significant. Interpret the exponentiated estimated coefficients.
7. When you are satisfied with the model you found in Question 6, create deviance residual plots to investigate potential lack of fit of the model and the form of the relationship between the covariate and the logarithm of grand slam count. Identify any variables that exhibit patterns in their residual plots. Based on the plots you have created, make any appropriate transformations of the response and/or explanatory variables.
8. Do your observations from Question 4 suggest that multicollinearity may be a problem for your model? If so, consider modifying your regression model. Explain the impact multicollinearity may have on the interpretation of your results.
9. Examine the Pearson residuals and determine if overdispersion may be a problem. If it is, then fit an overdispersed model to the data and re-examine the significance of the covariates.
10. The Poisson regression models you have investigated did not include an offset term. This assumes that the exposure levels for each team are the same. How might you describe exposure in the context of the data? How would you model the number of grand slams per unit of exposure? Re-fit the Poisson regression model for the *rate* of grand slams that incorporates an offset term. Summarize the results of your model.

## Presenting Your Own Model

11. Collect your data and meet with your professor to discuss your potential model assumptions, diagnostics, and analysis.
12. Analyze your data and use the discussions of prior work to write a 5- to 7-page research paper describing your analysis and discussing the results (see “How to Write a Scientific Paper or Poster” on the accompanying CD).

13. **Peer Review** Bring three copies of your research paper to class. Submit one to the professor. The other two will be randomly assigned to other students in your class to review.

## Final Revision

Make final revisions to the research paper. Then submit the first draft, other students' comments and checklists, the data set you used (in electronic format) along with descriptions of the variables in the data set, and your final paper.

## Other Project Ideas

Another rare phenomenon in baseball is a cycle. When a baseball player has a single, a double, a triple, and a home run in one game, he is said to have “hit for the cycle.” Use the *Grandslam* data to perform a Poisson regression analysis with the number of cycles in five years as the response variable.

Several of the homework activities can also be used to develop your own project ideas. There are many places where data are publicly available.

- Information on a variety of sports can be found at <http://cbs.sportsline.com>, <http://sportsillustrated.cnn.com>, <http://www.nfl.com/stats/team>, <http://www.ncaa.org>, and <http://www.baseball-reference.com>.
- Information from many federal agencies can be found at <http://www.fedstats.gov>, the Bureau of Labor Statistics (<http://www.bls.gov/cpi>), the Behavioral Risk Factor Surveillance System (<http://www.cdc.gov/brfss>), the National Center for Health Statistics (<http://www.cdc.gov/nchs>), and the U.S. Census Bureau (<http://www.census.gov>).
- College and university information can be found at <http://www.collegeboard.com>, <http://www.act.org>, and <http://www.clas.ufl.edu/au>.
- The National Center for Educational Statistics website is <http://nces.ed.gov>.
- Information about movies produced each year can be found at <http://www.the-numbers.com>, <http://www.boxofficemojo.com>, <http://www.imdb.com>, and <http://www.rottentomatoes.com>.

# Survival Analysis: Melting Chocolate Chips

*Far better an approximate answer to the right question,  
which is often vague, than the exact answer to the wrong  
question, which can always be made precise.*

—John Tukey<sup>1</sup>

**S**urvival analysis methods are used to investigate the time until a target event of interest (e.g., death, drug relapse, or college graduation) occurs, and they are used in a variety of disciplines such as medicine, sociology, and education. Although survival analysis techniques do not get the same amount of exposure in the literature as other statistical methods such as regression analysis or analysis of variance, it is important to consider their use whenever the response variable of interest is the *time* until an event occurs.

In this chapter, you will have the opportunity to perform a simple experiment to investigate the time required for different types of chocolate chips to melt. A set of activities related to this experiment will introduce you to a variety of methods for exploring the times until a target event occurs, also referred to as survival or time-to-event data. Upon completion of the activities, you should be able to do the following:

- Recognize special characteristics of survival data
- Summarize survival data and estimate survival probabilities using the Kaplan-Meier estimator
- Compute descriptive statistics, including the mean and percentiles, for a sample of event times
- Construct confidence intervals for survival probabilities
- Compare survival experiences for different groups of subjects
- Investigate periods of time when subjects are at low and high risk of experiencing an event of interest

The extended activities and research project provide opportunities to evaluate research articles discussing applications of survival analysis, hazard functions, and various types of incomplete data.

## 9.1 Investigation: How Long Does It Take for Chocolate Chips to Melt?

If you enjoy eating chocolate chips, then you can probably appreciate their sweet flavor and smooth texture as they melt in your mouth. Chocolatiers and food scientists are well aware that the material composition of chocolate affects the flavor, texture, and duration of the melting process, thereby resulting in a “good” or “bad” tasting experience. These individuals continually strive to improve the manufacturing process, as well as the properties of chocolate. By developing heat-resistant varieties that can withstand higher temperatures, they work to increase the time before the chocolate melts.<sup>2</sup> For food scientists and chocolate confectioners, melting chocolate can be serious business.

The purpose of the following investigation is to explore the time required for chocolate chips to melt. To perform the study, we will conduct an experiment that requires students to place a chip in their mouth and wait for it to melt. The experiment will be somewhat restrictive because the time allowed to run the study will be limited. Suppose we allow only 60 seconds for the study. It is possible that not all the chocolate chips will melt. For those chips that have not melted by the time 60 seconds has elapsed, we will have only partial information on melting times, and we will indicate that these times are incomplete. As we’ll see, incomplete times can create difficulties when we are trying to compute simple quantities like the average melting time or trying to estimate the proportion of chips that take longer than a particular time to melt.

To obtain a larger set of chip melting times that you can use for the activities and extended activities in this chapter, your class might participate in the following experiment to collect chocolate chip melting times. In addition, you will then be able to examine differences in the melting times by the type of chocolate chip (white or milk chocolate). In this experiment, you will place a chocolate chip in your mouth, hold it between your tongue and the roof of your mouth, and record the time required for it to completely dissolve (i.e., melt), without actually biting into the chip. Although this activity may appear rather trivial, it is meant to serve as a simple, yet illustrative, approach to generating real time-to-event data and exploring methods used to investigate survival data. Our hope is that when we examine additional examples of real survival data later in the chapter, you will be able to relate the features of those examples back to the features of the chip melting time data.

With the melting times for different types of chips and the methods you will learn in this chapter, you will be able to do the following:

- Estimate the proportion of chips that remain unmelted beyond a specific point in time
- Estimate the average time it takes for white or milk chocolate chips to melt
- Determine if the type of chip is related to the chip melting experience. For example, do milk and white chocolate chips melt at the same rates over time?
- For those chocolate chips that have not melted by a particular time, determine at what rate they are melting in the next “instant” of time

### NOTE

For this activity, you will need a time-keeping device that every student can clearly see, a bag of white chocolate chips, and a bag of milk chocolate chips.

### Activity Melting Chocolate Chips

1. Perform the chocolate chip melting activity outlined below. Be sure to record chip color, melting time, and whether the chip completely melted by 60 seconds. Combine all class data into one data file for future analysis purposes, and name this file `MeltingChips`.
  - a. Devise a system for randomly assigning each student to have a white or milk chocolate chip (this can be done by flipping a coin, for example).
  - b. When the instructor gives approval, place the white or milk chocolate chip in your mouth and record the time until it completely melts. Note that the group should come to a reasonable consensus on a clear definition of “completely melts” prior to the activity to ensure consistency in the recorded times.

- c. Create a data set in the following manner: Treat the study as if it could be done only for a specified period of time (you may need to experiment, but 60 seconds has worked well). If the actual time required for the chip to melt is less than 60 seconds, then the actual time will be complete and you submit (chip type, actual time, 1). If the chip has not melted by 60 seconds, then you regard the observation as incomplete and submit (chip type, 60, 0). Observations of any chips that are swallowed prior to 60 seconds should be regarded as incomplete as well, and you submit (chip type, actual swallowed time, 0).



#### NOTE

If your class does not perform the chip melting activity, then you can use the data set `MeltingChipsJS` supplied by the authors to complete the activities and extended activities. Be certain to verify which data set (`MeltingChips` or `MeltingChipsJS`) your instructor would like you to use in the following activities.\*

## 9.2 Overview of Survival Analysis Studies and Data

The collection of lengths of times required for the chips to melt is an example of **survival data**, also called **time-to-event data** or **failure-time data**. Survival data are the times until individuals experience an event of interest. The specific event of interest can be, for example, death, graduation, or test completion, while the individual experiencing the event may be living, such as a person or an animal, or inanimate, such as a light bulb, computer, or chocolate chip.

**Survival analysis** is a field of statistics covering methods and techniques for examining and investigating survival data, and it is used in diverse fields including medicine, education, and psychology. In studies that use survival analysis techniques, the response variable of interest, denoted  $T$ , is **the time until the event of interest occurs**, also called the **failure time, survival time, or time-to-event random variable**. For example, the time taken for a chocolate chip to melt is a time-to-event random variable, and the recorded melting times are the observed values of  $T$ . In the material that follows, we will discuss many ways to summarize and describe the observed values of  $T$  that you might get from an experiment or study. Additional examples of survival time random variables (with their related fields in parentheses) that will be discussed include

- time until drivers blocked by traffic honk their horn (sociology or psychology)
- time until students graduate from college (education)
- age at which first alcoholic drink is taken (public health)
- time until former inmates are rearrested (criminology)

When a study involves measuring the time until a target event occurs and when it possesses a clearly defined **beginning of time** as well as a **meaningful scale** for measuring time, then it is appropriate to use survival analysis methods and techniques. The beginning of time is a point at which no individual under study has yet to experience the event (e.g., the date on which a student enrolls in a post-secondary institution when the investigation concerns the time until college graduation), while a meaningful scale might be seconds, minutes, days, weeks, and so on.

#### NOTE

Time-to-event data are fundamentally different from time series data. **Time series data** are measurements on the same observational units collected at different time points. For example, time series data may be the number of chips that melt at 40 seconds, 50 seconds, and so on. Time-to-event data are **time durations until the observational unit experiences the target event**.

---

\*This data file contains the results of the chip melting activity administered by the second author of this textbook to his introductory survival analysis course. The maximum time allowed was 75 seconds.

**Key Concept**

The response variable in a survival analysis study is the time until the event of interest occurs. Survival analysis methods are appropriate for data from experiments or studies that possess a well-defined event of interest, a clearly defined beginning of time, and a meaningful scale for measuring time.

## Incomplete Event Times: Censoring

One feature common to many survival data sets that needs to be appropriately addressed is that some event times are **incomplete**; i.e., we have only partial information about the time until the event of interest occurs. We'll refer to event times that are known exactly as **complete**. There are several reasons why observations may be incomplete, but in this chapter we will focus on a mechanism called **right censoring**, which occurs when observation begins at a defined starting time and ends before the outcome of interest is observed. The chocolate chip data could have right-censored observations if a chip did not melt in the allotted time or if a student accidentally swallowed a chip. For example, a chip that did not melt by 60 seconds would have a right-censored (incomplete) time of 60 seconds. Other types of incomplete data will be discussed in the extended activities.

In written reports and journals, it is common to display a right-censored event time with a + placed to the right of the observed time. For example, a recorded melting time of 60+ seconds would indicate that the chip was observed for 60 seconds, but did not (completely) melt. Another way to record an event time is to assign a pair of numbers consisting of an observed value for the survival time variable,  $T$ , and a value for a censoring status variable,  $C$ . The censoring variable might, for example, take the value of 0 if the event of interest was not observed and the value of 1 if it was (this 0–1 coding choice is arbitrary, although common). Although more formal, this method of recording survival data is particularly relevant when the times need to be entered into data files. This coding scheme was adopted for the in-class chip melting activity performed earlier.

Incomplete observations can introduce systematic error, also called **bias**, into the estimated quantities (like the mean or median) if not handled appropriately—for example, if right-censored observations are treated as complete or are removed from the study. Descriptive statistics of survival time, such as the mean, may be grossly underestimated if right censoring is present but ignored. Survival analysis methods and techniques to accommodate data with incomplete observations have been developed, and some of these will be covered in the sections that follow.

**NOTE**

Be sure to read the description of any file containing survival data so that you know which value of the censoring variable corresponds to a right-censored time and which value corresponds to a complete time.

## Activity Melting Chocolate Chips

2. For the chip melting study, describe the event of interest, the time-to-event random variable  $T$ , the beginning of time, and the scale for measuring time.
3. Examine your class data. How many of your melting times are complete? How many are censored?

**Key Concept**

Survival data may contain right-censored observations; that is, an individual may not experience the event by the end of the study, or the individual may drop out of the study before the event time is observed. An event time that is right censored can be displayed with a + to the right of the recorded time or represented by using a pair of values that include the observed recorded time and a value to indicate that the time is censored.

## 9.3 The Survival Function

Now that we have discussed particular features of time-to-event data and looked at some notation and terminology, we turn to methods for examining and summarizing survival data. The primary function used to characterize the values of a time-to-event random variable  $T$  is the **survival function**  $S(t)$ , given by

$$S(t) = P(T > t)$$

$S(t)$  provides the probability that a randomly selected individual in the *population* will survive (not experience the event of interest) beyond time  $t$ . Another interpretation of  $S(t)$  is that it provides the proportion of subjects in a population who have yet to experience the event of interest by time  $t$ .

In the context of the chocolate chip melting times, the survival function  $S(t) = P(T > t)$  provides the probability that a randomly selected chip in the population of all chips will take longer than the specified time  $t$  to melt. So, for example,  $S(45) = P(T > 45)$  gives the probability that a randomly selected chip will take longer than 45 seconds to melt. Equivalently,  $S(45)$  provides the proportion of chips in the population that have not melted after 45 seconds.

At the beginning of time, no one has experienced the event, so the proportion of subjects in the population who have yet to experience the target event is 100% and  $S(0) = 1$ . Then as time progresses, individuals will experience the event (e.g., chocolate chips will melt, college students will graduate, former inmates will be arrested again), so the survival function will decline toward its lower bound of 0 (although it may not actually reach this value).

### Key Concept

The survival function provides the probability that an individual will survive beyond a given time  $t$ —that is, the probability that an individual does not experience the event of interest until after time  $t$ . *Important:*  $S(t)$  does not indicate the probability that an individual experiences the event at time  $t$ .

## The Empirical Survival Function

To determine the exact proportion of chips that take longer than  $t$  seconds to melt, we would need to know the melting times of the entire population of chips or know the exact probability distribution for  $T$ . In practice, we hardly ever know the exact probability distribution for  $T$ . In the real world, we will collect (or be given) a sample of survival times, and we will need to find an estimator for  $S(t)$ .

### NOTE

This is very similar to what was done in your first statistics course. An exact value, such as the population mean  $\mu$ , is estimated using a function of sample data,  $\bar{x}$ .

To illustrate the calculation of various quantities in the following sections, we will use a small sample of melting times (in seconds) of milk chocolate chips for 7 students, where the maximum time allowed for the experiment was 60 seconds. These times are displayed in Table 9.1. Statistical software will be used for the in-class chip melting activity data, as well as the additional data sets described in the extended activities.

**Table 9.1** Hypothetical chocolate chip melting times for a sample of 7 students.

Student	1	2	3	4	5	6	7
Time	35	30	60	45	25	55	30

To estimate the proportion of chocolate chips that have not melted after 45 seconds,  $\hat{S}(45)$ , we simply calculate the sample proportion:

$$\begin{aligned}\hat{S}(45) &= \frac{\text{number of chips that have not melted after 45 seconds}}{\text{total number of chips in the sample}} \\ &= \frac{2}{7}\end{aligned}$$

When all observations are complete, we can generalize this calculation to any time  $t$  using an estimator of  $S(t)$  called the **empirical survival function**, denoted by  $\hat{S}(t)_E$  and given by

$$\begin{aligned}\hat{S}(t)_E &= \frac{\text{number of individuals yet to experience the event at time } t}{\text{total number of individuals in the study}} \\ &= \frac{\text{number of event times greater than } t}{\text{total number of individuals in the study}}\end{aligned}\tag{9.1}$$

When all observations are complete, the empirical survival function works well. However, this estimator may not be as precise when the data contain censored (i.e., incomplete) observations.

Now suppose that some of the milk chocolate chip melting times displayed in Table 9.1 are incomplete. Let's assume that students 1 and 7 withdrew from the study (they accidentally swallowed the chips before they melted), while student 3 had not experienced a melted chip by the end of the experiment. Then students 1, 3, and 7 have censored times, and the melting times can be displayed as shown in Table 9.2, where the + denotes right-censored observations.

**Table 9.2** Hypothetical chocolate chip melting times for a sample of 7 students, with incomplete times for students 1, 3, and 7.

Student	1	2	3	4	5	6	7
Time	35+	30	60+	45	25	55	30+

## Activity Empirical Survival Function

- 4. Use Equation (9.1) and the 7 milk chocolate melting times in Table 9.1 to compute  $\hat{S}(25)_E$ ,  $\hat{S}(30)_E$ ,  $\hat{S}(40)_E$ , and  $\hat{S}(60)_E$ .
- 5. With the melting times provided in Table 9.2, use the following two approaches to calculate the estimated probability that it takes more than 45 seconds for a chocolate chip to melt, based on the empirical survival function  $\hat{S}(45)_E$ .
  - Treat all the censored times as complete (actual observed) times, and use Equation (9.1) to calculate  $\hat{S}(45)_E$ .
  - Eliminate all censored observations, and then use Equation (9.1) and the remaining complete observations to calculate  $\hat{S}(45)_E$ .

Note the different answers obtained in Parts A and B of Question 5. By treating the censored times as complete times, we assume that the event times are shorter than they actually are, thereby underestimating the true probability of survival (not melting). By disregarding the censored times, we lose information about melting times from the sample (consider the extreme case where all melting times are 60+). Treating the censored observations as complete or ignoring them will *bias* any estimates based on the remaining complete times.

## The Kaplan-Meier Estimator

When a data set contains incomplete observations, the best estimator of the survival function is the **Kaplan-Meier estimator**,  $\hat{S}(t)_{KM}$ . While this estimator is typically calculated with statistical software, this section will describe the logic behind how the Kaplan-Meier estimator is put together.

The first step in creating the Kaplan-Meier estimator is to establish a series of time intervals. Order the complete event times from smallest to largest and label the smallest complete time as  $t_1$ , the second smallest as  $t_2$ , and so on. We will denote the number of distinct *complete* event times by  $m$ , where  $m$  is less than or equal to  $n$ , the total number of *observed* event times (complete and incomplete).

The complete times,  $t_1$  through  $t_m$ , are used to define intervals beginning at one complete event time and ending just prior to the next complete event time, with some minor modifications for the first and last intervals as outlined below:

- By convention, the 0th interval begins at time  $t_0 = 0$  and ends just prior to the time when the first event occurs, time  $t_1$ . This interval is given by  $[0, t_1]$ .
- The next interval begins with the complete time  $t_1$  and ends just prior to next complete event time  $t_2$ . Time intervals of the form  $[t_i, t_{i+1})$  are created for  $i = 1, 2, \dots, m - 1$ .
- If the largest observed event time is censored, then this time is denoted by  $t_n$  and the interval extends to  $t_n$  and is open on the right; i.e., the interval is given by  $[t_m, t_n)$ . If the largest observed event time is complete, then the last interval is technically not an interval and just consists of a single point; that is, the interval is given by  $[t_m, t_m]$ .

### Activity Time Intervals for the Chip Melting Times

6. Consider the chocolate chip melting time data in Table 9.2. What is  $m$ ? List  $t_1$  through  $t_m$  for the chip melting times.
7. The first two intervals for the chocolate chip melting times are  $[0, 25)$  and  $[25, 30)$ . Write out the remaining intervals. Notice that any incomplete times, such as  $30+$  and  $35+$ , are ignored in creating intervals.
8. Determine  $d_i$ , the number of melted chips in each interval, and  $n_i$ , the number of chips at risk of melting in each interval (all chips with complete or censored times that have not yet occurred), for  $i = 0, 1, 2, 3, 4$ .

**Table 9.3** Counts and estimated probabilities of melting for melting times.

Interval $i$	Time Interval	Number at Risk ( $n_i$ )	Number Censored	Number of Events That Occurred ( $d_i$ )	$\hat{p}_i$	$1 - \hat{p}_i$	$\hat{S}(t)_{KM}$
0	$[0, 25)$	7	0	0	0/7	1	1
1	$[25, 30)$	7	0	1	1/7	6/7	6/7
2	$[30, 45)$	6	2	1	1/6	5/6	5/7
3	$[45, 55)$						
4	$[55, 60)$						

Table 9.3 displays some of the quantities required to compute the Kaplan-Meier estimates.

After the time intervals, the number of events of interest ( $d_i$ ) and the number at risk ( $n_i$ ) have been calculated, three more calculations are still needed for each interval:  $\hat{p}_i$ ,  $1 - \hat{p}_i$ , and  $\hat{S}(t)_{KM}$ .

After the time intervals have been appropriately defined, we estimate  $\hat{p}_i$ , the conditional probability of experiencing the event in the  $i$ th time interval, given that the event has not occurred by the start of the interval. That is, we compute

$$\hat{p}_i = \frac{d_i}{n_i}$$

where  $d_i$  is the number of subjects who experienced the target event in interval  $i$  and  $n_i$  is the total number of subjects (with complete and censored times) who are eligible (at risk) to experience the target event at the *beginning* of the  $i$ th time interval.

Now if  $\hat{p}_i$  is the probability of an individual experiencing the event in the  $i$ th time interval, given that the individual has not experienced the event in the previous time intervals, then  $1 - \hat{p}_i$  is the probability of *not* experiencing the event (i.e., *surviving*) through the  $i$ th time interval, given that the individual has not experienced the event prior to the  $i$ th time interval.

For example, the estimate of the conditional probability that a chip will *not* melt between the 25th second and the 30th second, given that it has remained unmelted through the 25th second, is given by

$$\begin{aligned}1 - \hat{p}_i &= 1 - d_1/n_1 \\&= 1 - 1/7 \\&= 6/7\end{aligned}$$

That is,  $6/7$ , or about 86%, of the chips that have not melted just prior to the 25th second will remain unmelted (survive) between the 25th and the 30th second.

## Activity ◀ Estimated Conditional Melting Probabilities

Use the chocolate chip data in Table 9.2 to answer the following questions:

9. What is the value of  $\hat{p}_0$ ? Interpret this value.
10.  $\hat{p}_1$  is the estimate of the conditional probability that a chip will melt between the 25th second and the 30th second, given that it has remained unmelted through the 25th second. Show that about 14% of the chips that have not melted just prior to the 25th second will melt between the 25th and the 30th second.
11. Calculate the remaining estimated conditional probabilities  $\hat{p}_3$  and  $\hat{p}_4$ . Place these values in the appropriate cells in Table 9.3 and interpret the values.
12. Calculate the remaining estimated conditional probabilities  $1 - \hat{p}_3$  and  $1 - \hat{p}_4$ . Place these values in the appropriate cells in Table 9.3 and interpret the values.

The final step in constructing the Kaplan-Meier estimated survival probabilities is to multiply together each conditional probability of surviving through the  $i$ th time interval to get the *unconditional* probability of surviving through the  $i$ th time interval.

For example, the Kaplan-Meier estimate of the probability that a chip will remain unmelted through the 25th second is given by

$$\begin{aligned}\hat{S}(25)_{\text{KM}} &= (1 - \hat{p}_0)(1 - \hat{p}_1) = (1)(1 - d_1/n_1) \\&= 1 - 1/7 \\&= 6/7\end{aligned}$$

Therefore,  $6/7$ , or about 86%, of the chips remain unmelted (survive) beyond the 25th second.

### Key Concept

The Kaplan-Meier estimator provides the proportions of subjects in the sample that survive beyond a given time. To compute the Kaplan-Meier estimator for a data set with  $n$  individuals, define the following quantities:

$m$ : the number of *distinct* uncensored event times, where  $m \leq n$ . By *distinct* we mean that two or more identical times contribute only once to determine  $m$ .

$t_1, t_2, \dots, t_m$ : the **ordered complete times** (i.e., those times when the event of interest actually occurred, ordered from smallest to largest. By convention,  $t_0 = 0$ ).

$n_i$ : the number at risk of experiencing the event at time  $t_i$  (i.e., just prior to the start of the time interval  $[t_i, t_{i+1})$ , for  $i = 0, 1, \dots, m - 1$ ).

$d_i$ : the number experiencing the event at time  $t_i$  (i.e., in time interval  $[t_i, t_{i+1})$ , for  $i = 0, 1, \dots, m - 1$ ).

Then the Kaplan-Meier estimator of the survival function is given by

$$\hat{S}(t)_{\text{KM}} = \prod_{t_i \leq t} (1 - d_i/n_i) \quad (9.2)$$

where  $\prod$  is the symbol for taking the products of terms  $(1 - d_i/n_i)$  for all  $i$  such that the complete event times  $t_1, \dots, t_m$  are less than or equal to the time of interest  $t$ . Also,  $\hat{S}(t)_{\text{KM}} = 1$  for all times  $t < t_1$ .

#### MATHEMATICAL NOTE

The Kaplan-Meier estimator is derived using the multiplication rule from introductory probability. Let  $A_i$  be defined as any event occurs after interval  $i$ . Then, for example,

$\hat{S}(25)_{\text{KM}} = (1 - \hat{p}_1)(1 - \hat{p}_0)$  is an estimate of  $P(A_1) = P(A_1 \text{ and } A_0) = P(A_1 | A_0)P(A_0)$   
and

$\hat{S}(30)_{\text{KM}} = (1 - \hat{p}_2)(1 - \hat{p}_1)(1 - \hat{p}_0)$  is an estimate of  $P(A_2) = P(A_2 \text{ and } A_1 \text{ and } A_0)$   
 $= P(A_2 | A_1)P(A_1 | A_0)P(A_0)$

### Activity

#### Kaplan-Meier Estimates

Refer to the entries in Table 9.3 to answer the following questions.

13. Use the remaining chocolate chip melting times to complete Table 9.3.
14. What is the estimate for  $S(45)$  in Table 9.3? That is, what proportion of chips in the sample has not melted after 45 seconds?
15. Use the entries in Table 9.3 to estimate the proportion of chips that have melted by 35 seconds.
16. Use the entries in Table 9.3 to estimate the proportion of chips that have not melted after 50 seconds.
17. Assume that no censoring is present in the melting times (see the entries in Table 9.1). Estimate  $S(25)$ ,  $S(30)$ ,  $S(45)$ , and  $S(55)$  using both the empirical survival function and the Kaplan-Meier estimator, and compare your answers. What do your answers suggest about the Kaplan-Meier estimator when no censoring is present?

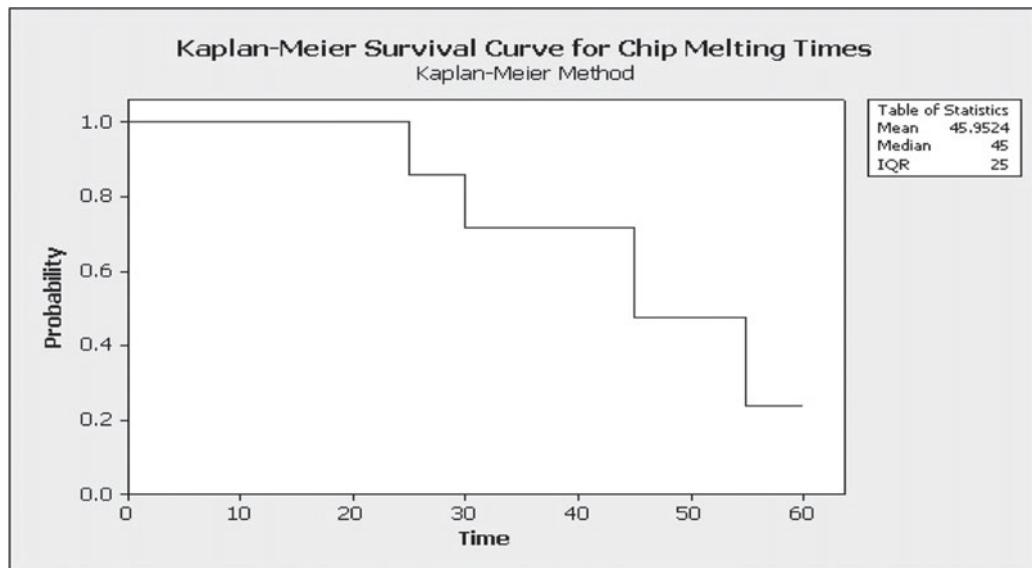
### Graphing the Kaplan-Meier Curve

Once the survival probabilities have been estimated, a graph of the **Kaplan-Meier curve** can be constructed to display the relationship between time and the estimated probability of surviving. The Kaplan-Meier curve is an approximation to the true **survival curve**, which is a graphical representation of  $S(t)$ . The values of  $\hat{S}(t)_{\text{KM}}$  are plotted against the complete event times  $t_1, t_2, \dots, t_m$ . Figure 9.1 shows the Kaplan-Meier curve for the chocolate chip melting times displayed in Table 9.2.

Notice that the value of  $\hat{S}(t)_{\text{KM}}$  remains the same across each time interval. From Table 9.3, we see that, for example,  $\hat{S}(30)_{\text{KM}} = 5/7$ ; i.e., the proportion of chips in the sample that have not melted after 30 seconds is  $5/7$ , or 71%. This value remains constant for time  $t$  in the interval  $[30, 45]$ . From Figure 9.1 we can observe that the Kaplan-Meier curve is a decreasing series of steps, with drops occurring at each complete event time  $t_i$ . This type of plot is called a **step function** because it looks like a series of steps. The height of each step corresponds to the value of  $\hat{S}(t)_{\text{KM}}$  for  $t$  inside  $[t_i, t_{i+1})$ , where  $i = 0, \dots, m - 1$  (with the convention that  $t_0 = 0$ ).

We can see that the estimated probability that a randomly selected chip remains unmelted decreases as time increases or, stated another (equivalent) way, the proportion of unmelted chips decreases over time. Another feature of Figure 9.1 is that the proportion of chips that remain unmelted after 60 seconds is nonzero, so the last step of the Kaplan-Meier curve extends to the right up to 60 seconds.

You'll notice a box to the right of the graph, labeled "Table of Statistics," that contains the mean and median survival times for the sample, as well as the interquartile range (IQR). We'll discuss these descriptive statistics in Section 9.4.



**Figure 9.1** Kaplan-Meier estimated proportions (estimated survival probabilities) of chocolate chips.

## Activity Kaplan-Meier Curves

- 18. Compare the values of  $\hat{S}(t)_{\text{KM}}$  in Table 9.3 to those in Figure 9.1. How would the Kaplan-Meier curve in Figure 9.1 change if the largest observed melting time were not censored?
- 19. Use the technology instructions provided on the CD to construct the Kaplan-Meier curve for the white and the milk chocolate chips from the chocolate chip melting activity. Compare these two curves. Do the melting proportions for the different types of chips appear similar over time? We'll discuss formal methods for comparing survival curves for the populations of white and milk chocolate chips in Section 9.6.

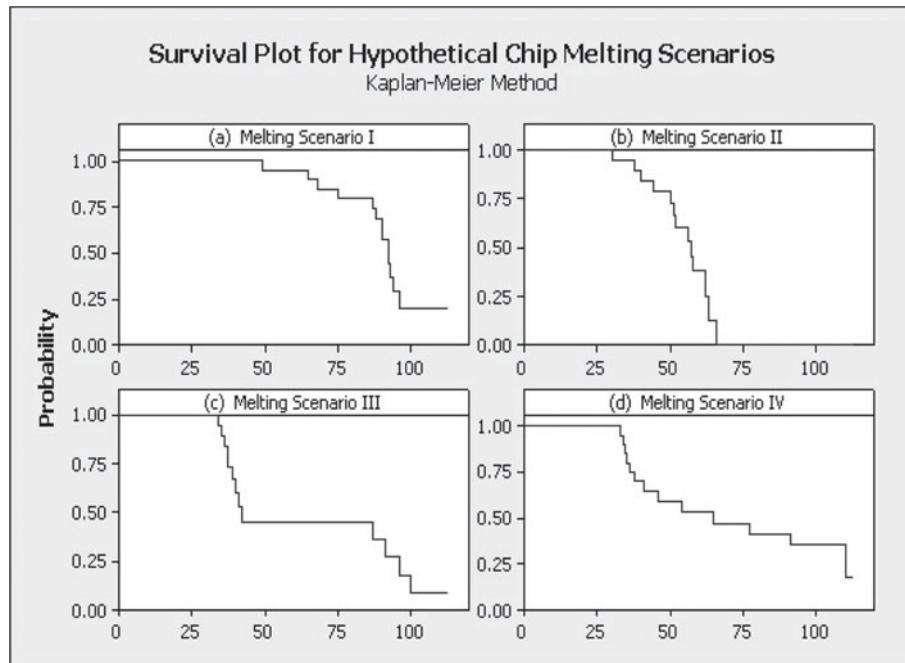
## Additional Chip Melting Experiences

In the class data, you observed two possible chip melting “experiences” over time, corresponding to the milk and white chocolate chips. Figure 9.2 presents four additional Kaplan-Meier curves corresponding to different chip melting experiences.

The Kaplan-Meier curve corresponding to the first scenario, displayed in Part (a), reveals that no chips melt until after the 49th second. Then for any time, say  $t$ , between the 49th and the 87th second, the estimated proportion of chocolate chips that remain unmelted decreases somewhat slowly, indicating a possible resistance by the chocolate chips to melt. For the period of time between the 87th and the 96th second, the estimated proportion of chips that remain unmelted decreases rapidly, suggesting that the chocolate chips are very susceptible to melting. For any time beyond the 96th second, the estimated proportion of chips that remain unmelted remains constant at about 20%. This means that there are some chips that have not melted by the end of the observation period, (i.e., their melting times are censored); however, note that 20% does not necessarily refer to the proportion of chips with censored melting times.

The Kaplan-Meier curve corresponding to the second melting scenario, displayed in Part (b), shows that chips begin completely melting after 30 seconds, and the proportion of chips that remain unmelted beyond any time between 30 and 65 seconds decreases rapidly, indicating a period of time when the chips are melting quite quickly. The estimated probability that a chip remains unmelted beyond the 65th second is 0, since all chips have melted by this point.

Part (c) displays the Kaplan-Meier curve for the third melting scenario, and we can observe that the chips do not begin completely melting until after 35 seconds. Then during the next 5 seconds, the estimated proportion of chips that remain unmelted declines rapidly (approximately 50% of the chips melt between



**Figure 9.2** Kaplan-Meier survival curves for different chip melting experiences.

the 35th and the 40th second). Then for an extended period of time between the 40th and the 85th second, the melting stabilizes, so the proportions of chips that survive are identical. Then melting resumes for the next 15 seconds. For any time beyond the 100th second, it is estimated that about 9% of the chips have not melted.

### Activity Chip Melting Experiences

- 20. Discuss the fourth chip melting scenario, illustrated by the Kaplan-Meier curve in Part (d) of Figure 9.2.

## 9.4 Descriptive Statistics for Survival Data

So far we have focused on using the Kaplan-Meier curve to estimate the proportion of chips that remain unmelted after 25 seconds, after 30 seconds, and so on. While the survival function is useful, we may also be interested in finding an estimate for the mean and median time-to-event for a population of individuals. For example, what is the typical time required for a chocolate chip to melt? Or how long does it take for half of the chips to melt?

It may seem odd that we have spent a great deal of time discussing survival probabilities and their estimates *before* discussing descriptive statistics. The primary reason for this ordering of topics in survival analysis is that descriptive measures of survival data require the Kaplan-Meier estimated survival probabilities.

It may be tempting to use the same calculations for means and medians that you learned in your previous statistics course. However, consider the mean time until a chocolate chip melts. If we use the sample mean  $\bar{x}$  to estimate this quantity, then once again we run into the same problem as when we estimated chip melting rates—**how are censored observations to be treated?** For example, if censored observations are treated as complete, then the resulting estimate of the mean melting time will *underestimate* the true average melting time. We will need to incorporate the Kaplan-Meier estimated survival probabilities to deal with censored observations.

## Estimating the Mean Survival Time

The estimated mean survival time is the total area under the Kaplan-Meier curve. This can be found by adding up the areas of the bars formed by the height of the curve between two adjacent complete survival times, with a slight adjustment if the largest observed event time is censored. Consider the rectangular bars displayed in Figure 9.3. The area of each bar is found by taking the width of each interval  $t_{i+1} - t_i$  (the duration of each time period) and multiplying by the estimated probability of surviving through the interval  $\hat{S}(t_i)_{\text{KM}}$  (the height of the bar). If the largest observed event time is censored (i.e.,  $t_n > t_m$ ), then the last interval extends from  $t_m$  to  $t_n$ . Hence, we have the following expressions for computing the estimated mean survival time:

If the largest observed event time is complete (i.e.,  $t_n = t_m$ ), then the estimator of the mean survival time, denoted by  $\hat{\mu}$ , is given by

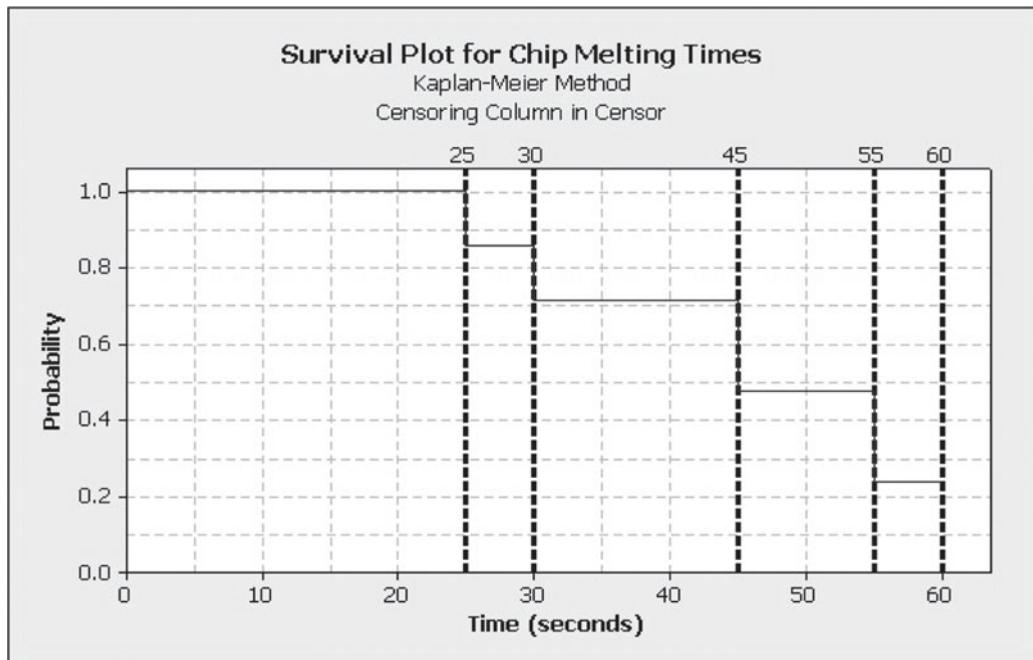
$$\begin{aligned}\hat{\mu} &= \sum_{i=0}^{m-1} \hat{S}(t_i)_{\text{KM}}(t_{i+1} - t_i) \\ &= \hat{S}(t_0)_{\text{KM}}(t_1 - t_0) + \hat{S}(t_1)_{\text{KM}}(t_2 - t_1) + \cdots + \hat{S}(t_{m-1})_{\text{KM}}(t_m - t_{m-1})\end{aligned}\quad (9.3)$$

where we make the provision that  $\hat{S}(t_0)_{\text{KM}} = 1$ .

If the largest event time  $t_n$  is censored (i.e.,  $t_n > t_m$ ), then the estimator of the mean survival time is given by

$$\hat{\mu} = \sum_{i=0}^{m-1} \hat{S}(t_i)_{\text{KM}}(t_{i+1} - t_i) + \hat{S}(t_m)_{\text{KM}}(t_n - t_m)\quad (9.4)$$

Equations (9.3) and (9.4) can be thought of as a weighted average of the time-to-event data, where the Kaplan-Meier curve  $\hat{S}(t_i)_{\text{KM}}$  acts as weights (probabilities) for the time-to-event intervals. Each term in the summation of Equations (9.3) and (9.4) represents the area of a rectangle formed by the width of the  $i$ th time interval  $t_{i+1} - t_i$  and the height of the Kaplan-Meier curve  $\hat{S}(t_i)_{\text{KM}}$ . For example, the width of the left-most rectangle in Figure 9.3 outlined by the vertical dotted line is 25 seconds, and the height of the rectangle is 1. So the first term in the sum of Equation (9.3) is  $\hat{S}(t_0)_{\text{KM}}(25 - 0) = (1)(25) = 25$ .



**Figure 9.3** Kaplan-Meier estimated proportions. Vertical lines appear at the complete melting times, and grid marks are overlaid on the graph. The dimension of each square in the grid is 5 seconds (width) by 0.1 (height).

**Key Concept**

The mean survival time is estimated using the time intervals and the Kaplan-Meier estimated probabilities.

### Activity ➔ Estimated Mean Chip Melting Time

21. Examine Figure 9.3. Visually estimate the mean survival time (i.e., the estimated average time taken for the chocolate chips in the sample to melt) by computing a rough approximation to the area under the Kaplan-Meier curve.
22. For the sample of chocolate chip melting times in Table 9.2, which equation, (9.3) or (9.4), is appropriate for estimating the mean survival time of the chips? Based on your answer, calculate the estimated mean using the quantities in Table 9.3.

### Estimating Percentiles of the Survival Time Distribution

Suppose we want the time at which 50% of the chocolate chips have melted. This quantity is called the **median survival time** (or, more precisely, the *estimated* median survival time, since we are working with a sample of survival times). Other times at which a certain percentage of subjects have experienced the event of interest are known as **percentiles** of the survival time distribution.

The  $p$ th percentile of the distribution for the survival time random variable  $T$  is defined to be the time by which  $p\%$  of the subjects in the population have experienced the target event. The estimate of the  $p$ th percentile, denoted by  $\hat{t}_{(p)}$ , is defined to be the smallest *complete* event time in the sample such that at least  $p\%$  of the subjects in the sample have experienced the event of interest before  $\hat{t}_{(p)}$  and no more than  $(100 - p)\%$  of the subjects in the sample experience the event after  $\hat{t}_{(p)}$ . Depending on the number of distinct complete event times, the value of  $\hat{t}_{(p)}$  can be found either by inspecting the Kaplan-Meier curve or by finding the solution to the following equation:

$$\hat{t}_{(p)} = \text{smallest complete event time } t_i \text{ in the sample such that } \hat{S}(t_i)_{\text{KM}} \leq 1 - \frac{p}{100} \quad (9.5)$$

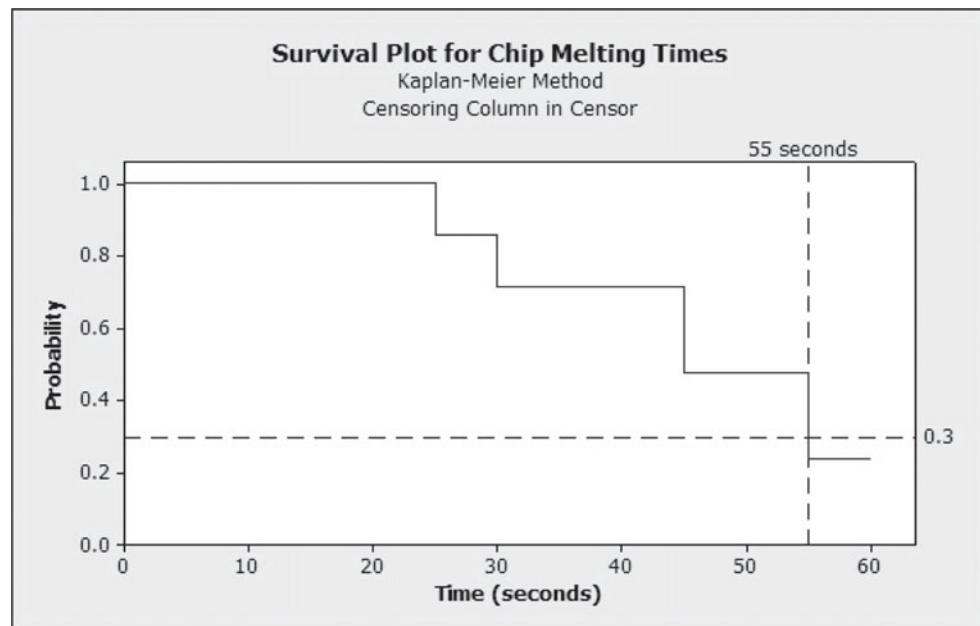
For example, to find the estimate for the 70th percentile of the chocolate chip melting times in Figure 9.1, we find the smallest complete event time  $t_i$  such that at least 70% of the chips have melted: Since  $\hat{S}(t_i)_{\text{KM}} \leq 1 - 0.7 = 0.3$ , we draw a horizontal line at  $\hat{S}(t)_{\text{KM}} = 0.3$ , and eventually we reach the vertical step that occurs at  $t_4 = 55$ . Since  $t_4 = 55$  is the smallest complete event time that satisfies  $\hat{S}(t)_{\text{KM}} \leq 0.3$ ,  $\hat{t}_{(70)} = 55$ . See Figure 9.4.

**Key Concept**

Percentile estimates of the survival time distribution are found using the Kaplan-Meier estimated probabilities.

### Activity ➔ Estimated Median Chip Melting Time

23. For the chocolate chip melting time data in Table 9.2, calculate the estimated median survival time  $\hat{t}_{(50)}$ . Since right-censored data are typically right skewed, the median survival time is usually preferred to the mean.
24. Refer back to the Kaplan-Meier curve displayed in Figure 9.1. The mean and median survival times for the sample of chip melting times are provided in the “Table of Statistics” box on the graph. The interquartile range (IQR) is also provided. Recall that the IQR is the third quartile (75th percentile) minus the first quartile (25th percentile). Verify by hand that the IQR for the sample of times is 25 seconds.
25. Use the technology instructions provided on the CD to determine the estimated mean and median survival times separately for the milk and white chocolate chip melting time data from your class activity. Discuss the differences you observe between the two descriptive measures and between the two types of chips.



**Figure 9.4** Kaplan-Meier estimated proportions. The horizontal dotted line represents  $\hat{S}(55)_{\text{KM}} = 0.3$ , and the vertical dotted line represents  $\hat{t}_{(70)} = t_4 = 55$ .

#### MATHEMATICAL NOTE

Based on the definition of the estimated percentile in Equation (9.5), it is possible that particular estimated percentiles do not exist. The  $p$ th percentile will not exist (and your software may return NA) if there does not exist a complete time  $t_p$  such that  $\hat{S}(t_p)_{\text{KM}} \leq 1 - p/100$ . For example, in the `MeltingChipsJS` data set, more than 25% of the milk chocolate chips had not melted by the end of the study (75 seconds). Thus, there is no estimate of the 75th percentile and the IQR cannot be calculated.

## 9.5 Confidence Intervals for Survival Probabilities

Just like any other sample statistic (e.g., the sample mean  $\bar{x}$ ), the estimates of the survival probabilities are subject to **sampling variability**. For example, different samples of chocolate chips (and/or students) would likely lead to different melting times, which would lead to different Kaplan-Meier estimates. Although a point estimate, such as  $\hat{S}(25)_{\text{KM}}$ , is useful for descriptive purposes, we may also want to construct a range of possible values that the estimated survival probability can take—i.e., a range that we can be reasonably sure contains the true survival probability  $S(25)$ . In other words, we want to construct a **confidence interval** for the true survival probability  $S(25)$  at time  $t = 25$  seconds.

In your first course in statistics, you saw that a confidence interval for a population parameter (e.g., a population mean  $\mu$ ) has the following form:

$$\text{point estimate} \pm \text{critical value} \times \text{standard error of the estimate} \quad (9.6)$$

where the point estimate is a single estimate of the parameter (such as  $\bar{x}$  for  $\mu$ ), the critical value is taken from a reference distribution like the standard normal distribution or  $t$ -distribution, and the standard error of the estimate is a measure of the variability in the point estimate. The expression for a confidence interval for  $S(t)$  at a fixed time  $t$  has a very similar form.

We know that the sampling distribution of  $\bar{x}$  is approximately normal for large sample sizes. A similar result can be stated concerning the sampling distribution of  $\hat{S}(t)_{\text{KM}}$  for fixed  $t$ . Some advanced theory tells

us that, for larger sample sizes, the sampling distribution of  $\hat{S}(t)_{\text{KM}}$  at a fixed time  $t$  is approximately normal. This fact allows us to use critical values from the standard normal distribution.

Using the critical value from a standard normal distribution and the standard error of  $\hat{S}(t)_{\text{KM}}$ , we can put together the confidence interval for  $S(t)$  at a fixed time  $t$ . The usual procedure is to construct a confidence interval for  $S(t)$  at all the complete event times  $t_1, \dots, t_m$ .

The  $100(1 - \alpha)\%$  confidence interval for  $S(t)$  at a fixed time  $t$  is given by

$$\hat{S}(t)_{\text{KM}} \pm Z_{\alpha/2} \text{se}(\hat{S}(t)_{\text{KM}}) \quad (9.7)$$

where  $Z_{\alpha/2}$  is the critical value from the standard normal distribution with  $\alpha/2$  area under the curve to its right [i.e., corresponding to confidence level  $100(1 - \alpha)\%$ ] and  $\text{se}(\hat{S}(t)_{\text{KM}})$  is the standard error of the Kaplan-Meier estimator at time  $t$ , discussed in the following section.

## The Standard Error of the Kaplan-Meier Estimator

The variability in the values of  $\hat{S}(t)_{\text{KM}}$  at a fixed time  $t$  in each interval  $[t_{i-1}, t_i]$  is measured by the estimated variance of the Kaplan-Meier estimator, denoted  $\hat{\text{Var}}(\hat{S}(t)_{\text{KM}})$  and calculated using the expression

$$\hat{\text{Var}}(\hat{S}(t)_{\text{KM}}) = (\hat{S}(t)_{\text{KM}})^2 \left( \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \right) \quad (9.8)$$

where  $d_i$  is the number of subjects who experienced the event of interest in time interval  $[t_{i-1}, t_i]$  and  $n_i$  is the number who are at risk at each complete event time  $t_i$ .<sup>3</sup>

The standard error of the Kaplan-Meier estimator at time  $t$ , denoted by  $\text{se}(\hat{S}(t)_{\text{KM}})$ , is the square root of  $\hat{\text{Var}}(\hat{S}(t)_{\text{KM}})$ .

$$\text{se}(\hat{S}(t)_{\text{KM}}) = \sqrt{\hat{\text{Var}}(\hat{S}(t)_{\text{KM}})} \quad (9.9)$$

As we will see, when we construct the confidence intervals for  $S(t)$ , the estimated variance  $\hat{\text{Var}}(\hat{S}(t)_{\text{KM}})$  [and standard error  $\text{se}(\hat{S}(t)_{\text{KM}})$ ] will need to be calculated for each complete event time  $t_1, \dots, t_m$ ; that is, we need to calculate

$$\hat{\text{Var}}(\hat{S}(t_1)_{\text{KM}}), \hat{\text{Var}}(\hat{S}(t_2)_{\text{KM}}), \dots, \hat{\text{Var}}(\hat{S}(t_m)_{\text{KM}})$$

For example, the Kaplan-Meier estimated survival probability at the complete time  $t = 25$  for the data in Table 9.2 is  $\hat{S}(25)_{\text{KM}} = 6/7$ . Using Equation (9.8) and the quantities in Table 9.3, we find that the variance estimate  $\hat{\text{Var}}(\hat{S}(t)_{\text{KM}})$  at time  $t = 25$  is given by

$$\begin{aligned} \hat{\text{Var}}(\hat{S}(25)_{\text{KM}}) &= (\hat{S}(25)_{\text{KM}})^2 \left( \sum_{t_i \leq 25} \frac{d_i}{n_i(n_i - d_i)} \right) \\ &= \left( \frac{6}{7} \right)^2 \left( \sum_{i=0}^1 \frac{d_i}{n_i(n_i - d_i)} \right) \\ &= \left( \frac{6}{7} \right)^2 \left( \frac{0}{7(7 - 0)} + \frac{1}{7(7 - 1)} \right) \\ &= 0.0175 \end{aligned}$$

Then the standard error of  $\hat{S}(t)_{\text{KM}}$  at time  $t = 25$  is  $\text{se}(\hat{S}(25)_{\text{KM}}) = \sqrt{0.0175} = 0.1323$ .

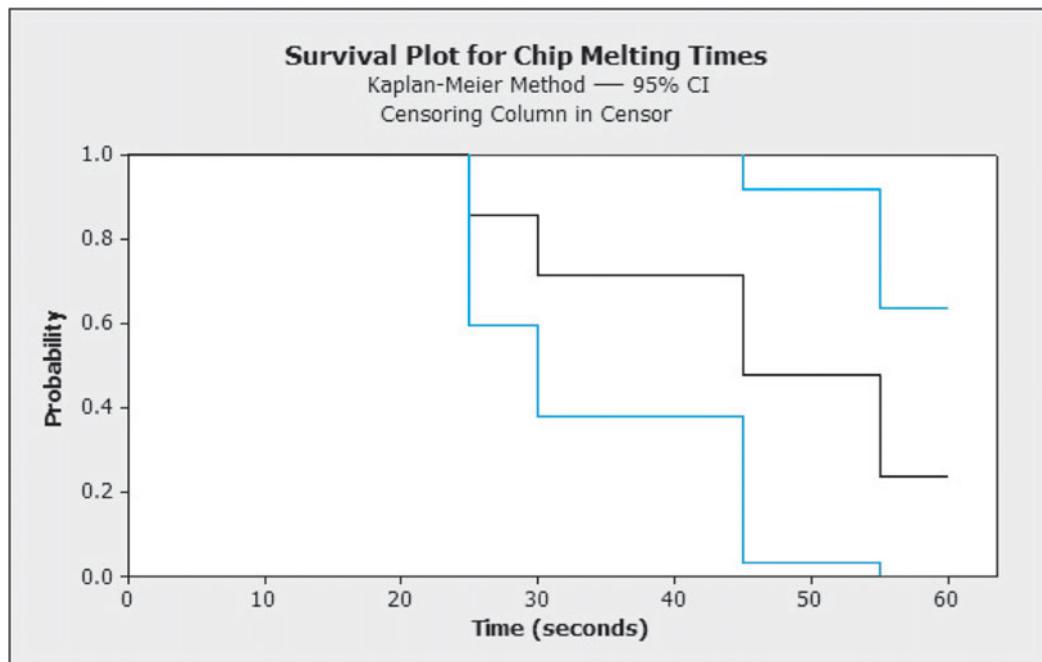
With the computed standard error, we can use Formula (9.7) to construct a confidence interval for a survival probability. For example, the 95% confidence interval for  $S(25)$ —the probability that a randomly selected milk chocolate chip takes longer than 25 seconds to melt—is calculated as

$$\begin{aligned} \hat{S}(25)_{\text{KM}} &\pm Z_{0.025} \text{se}(\hat{S}(25)_{\text{KM}}) \\ 6/7 &\pm 1.96(0.1323) \\ &= (0.60, 1.12) \end{aligned}$$

Observe that the upper limit for the confidence interval for  $S(25)$  exceeds 1. Since the confidence interval given by Formula (9.7) is for a survival *probability*, the interval limits should technically not fall outside the range  $[0, 1]$ . However, it is possible to obtain a lower limit less than 0 and/or an upper limit greater than 1, especially for smaller sample sizes. This is a drawback of using Formula (9.7) to construct confidence intervals. When confronted with a situation (as is the case here) where the interval limit(s) is (are) outside the range  $[0, 1]$ , it is common practice to truncate the limit(s) at the appropriate minimum value of 0 or maximum value of 1.

The interpretation of confidence interval limits for survival probabilities is similar to the standard interpretation of confidence intervals for other parameters of interest (think back to confidence intervals for a population mean or proportion). For the chip melting times provided in Table 9.2, we can state with 95% confidence that the probability that a chocolate chip will take longer than 25 seconds to melt is between 0.60 and 1.00. Or, equivalently, we are 95% confident that the true proportion of chips that have not melted after 25 seconds is between 60% and 100%.

The Kaplan-Meier curve and associated 95% confidence intervals for the chocolate chip melting time data are displayed in Figure 9.5. Observe the times where the upper limit has been constrained to equal 1 and the lower limit has been constrained to equal 0.



**Figure 9.5** Kaplan-Meier curve with 95% confidence intervals. Note that Minitab graphs use the same line type and color for both the Kaplan-Meier curve and the confidence limits; however, the middle line will always be the Kaplan-Meier curve.

#### NOTE

There are alternative formulas for computing confidence intervals for survival probabilities that constrain the limits to lie between 0 and 1, but we'll leave it to interested readers to investigate them on their own.<sup>4</sup>

#### Key Concept

Confidence intervals provide a range of values that we can be reasonably sure contain the true survival probabilities.

**Activity****Standard Errors and Confidence Intervals for Survival Probabilities**

Use the estimated survival probabilities and the entries in Table 9.3 to answer Questions 26 through 28, and use `MeltingChips` or `MeltingChipsJS` to answer Questions 29 through 31.

26. Provide a brief explanation of why the estimated variance of  $\hat{S}(0)_{\text{KM}}$ , and hence the standard error of  $\hat{S}(0)_{\text{KM}}$ , is equal to 0.
27. Find the remaining standard errors of  $\hat{S}(t)_{\text{KM}}$  at times  $t = 30$ ,  $t = 45$ , and  $t = 55$  for the chocolate chip melting time data.
28. Using your above answers and the appropriate entries in Table 9.3, construct 95% confidence intervals for the survival probabilities  $S(t)$  at each of the remaining complete melting times, and interpret their values.
29. Use software to construct the 95% confidence intervals for the survival probabilities for the milk chocolate chip melting time data from your class activity. Graph the Kaplan-Meier curve and the confidence intervals for  $S(t)$ .
30. Based on the width of the confidence intervals (upper limit minus lower limit), how useful do you think these intervals are for providing estimates of the true melting times? How could we improve the usefulness of these intervals? (*Hint:* Refer back to Equation (9.8). What could you change to make the intervals narrower?)
31. On the same graph, plot both Kaplan-Meier estimates and 95% confidence intervals for the white and milk chocolate chip melting time data from your class activity. Explain whether or not the two types of chocolate appear to have different survival functions.

**NOTE**

The confidence interval for  $S(t)$  is valid for a single fixed time at which the inference is to be made. For this reason, confidence intervals are sometimes referred to as pointwise intervals. Each interval constructed using Formula (9.7) is relevant only for  $t$  within the  $i$ th interval. Therefore, it is not correct to state that we are 95% confident that the entire true survival curve  $S(t)$  falls between the confidence bands in Figure 9.5. If we want a confidence band for an entire survival function  $S(t)$  within which we can guarantee with a specified level of confidence that the entire curve falls, alternative confidence bands need to be computed.<sup>5</sup>

## 9.6 Comparing Survival Functions

In Question 19 you were asked to comment on the differences in melting experiences between white and milk chocolate chips. Now we will look at how to formally compare the survival experiences of two or more groups. This is equivalent to investigating whether survival experience depends on another (categorical) variable of interest. To compare the survival experiences of milk and white chocolate chips, we will ask, “Does melting experience depend on the type of chip?”

To illustrate the comparison techniques, we’ll use a sample of *white* chocolate chip melting times for 9 other students:

45+ 35 48 64+ 72 42 55+ 43 62

First, we visually inspect the Kaplan-Meier curves for both groups. Figure 9.6 shows that the Kaplan-Meier curve for the white chocolate chip melting times tends to be above the milk chocolate chip curve. This means that, over time, the proportion of unmelted white chocolate chips is generally larger than the proportion of unmelted milk chocolate chips; that is, white chocolate chips generally take longer to melt than milk chocolate chips. While Figure 9.6 appears to show a difference in the curves based on our sample data, a formal hypothesis test is needed to determine if the difference in the *sample data* is large enough to conclude that the survival curves for the *populations* of white and milk chocolate chips are different.

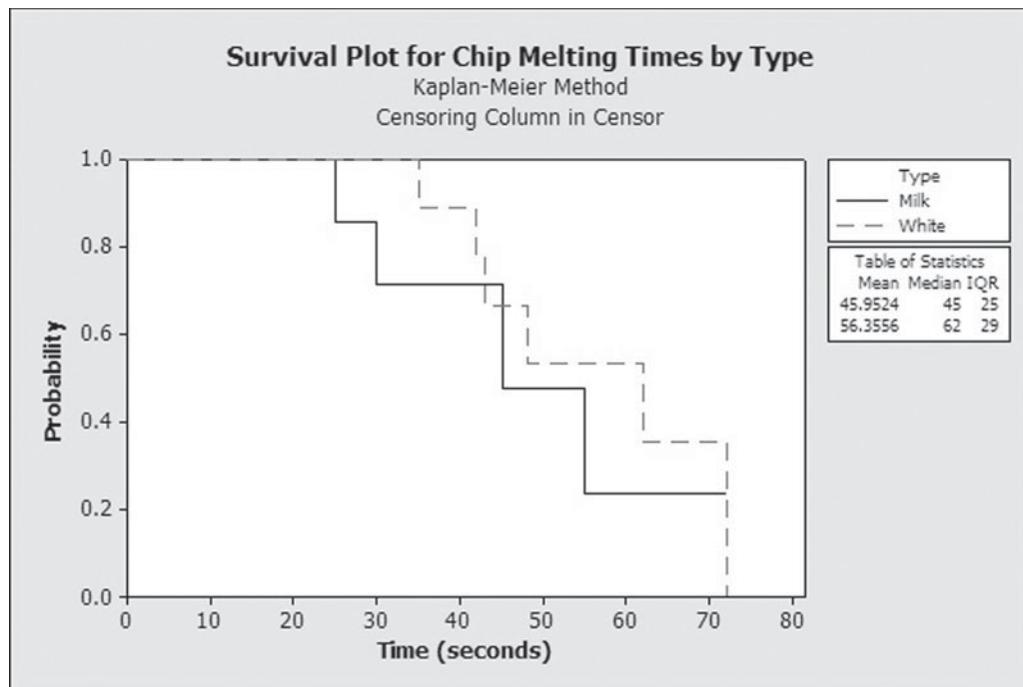


Figure 9.6 Kaplan-Meier curves for the hypothetical melting times of milk and white chocolate chips.

## The Log-Rank Test

The **log-rank test** is a formal statistical inference procedure used to determine if the population survival curves are significantly different over the range of time  $t$ —that is, to determine if the survival experiences significantly differ. The null hypothesis is that the survival experiences of both populations are equal—that is,  $H_0: S_1(t) = S_2(t)$  for all times  $t$  during which at least one of the groups has at least one subject at risk of experiencing the event. The alternative hypothesis is that the survival experiences are not identical for at least one value of  $t$  (i.e.,  $H_a: S_1(t) \neq S_2(t)$  for some value of time  $t$ ).

Assuming that we have two independent populations and that the survival experiences for group 1 and group 2 are identical, the log-rank test compares the total number of *observed* events to the total number of *expected* events in group 1. Observed counts that disagree significantly with the expected counts lead to rejection of the null hypothesis that the survival experiences are identical.

Table 9.4 contains the melting data for both the milk and the white chocolate chips. To calculate the log-rank test statistic, define the following terms:

$m$ : the number of distinct ordered complete event times  $t_1, t_2, \dots, t_m$  for both groups combined (we will not consider  $t_0 = 0$  in our calculations)

$n_{1i}$ : the number of subjects at risk of experiencing the event at time  $t_i$  in group 1 for  $i = 1, \dots, m$

$n_{2i}$ : the number of subjects at risk of experiencing the event at time  $t_i$  in group 2

$n_i$ :  $n_{1i} + n_{2i}$ , the total number of individuals in both groups at risk at time  $t_i$

$d_{1i}, d_{2i}$ , and  $d_i$ : the observed number of event occurrences at each of the complete event times for group 1, group 2, and the combined groups, respectively

Table 9.4 Hypothetical melting times of the milk and white chocolate chips.

<b>Group 1 (milk chocolate)</b>	35+	30	60+	45	25	55	30+		
<b>Group 2 (white chocolate)</b>	45+	35	48	64+	72	42	55+	43	62

The partially completed Table 9.5 contains the appropriate quantities to compute the log-rank test statistic. The total number of complete event times for both types of chips is  $m = 10$ . All 16 milk and white chocolate chips were at risk of melting prior to the beginning of the first time interval [25, 30), so  $n_{11} = 7$ ,  $n_{21} = 9$ , and  $n_1 = 16$ . Since 1 milk chocolate chip melted in the first time interval [25, 30) and no white chocolate chips melted in the first interval,  $d_{11} = 1$ ,  $d_{21} = 0$ , and  $d_1 = 1$ .

**Table 9.5** Selected quantities for the combined groups of chocolate chips.

Interval $i$	Time Interval	$n_i$	Number Censored	$d_i$	$d_{1i}$	$d_i/n_i$	$n_{1i}$	$n_{2i}$	$E_{1i}$	$V_{1i}$
1	[25, 30)	16	0	1	1	1/16	7	9	7/16	0.246
2	[30, 35)	15	1	1	1	1/15	6	9	6/15	0.240
3	[35, 42)	13	1	1	0	1/13	4	9	4/13	0.213
4	[42, 43)	11	0	1	0	1/11	3	8	3/11	0.198
5	[43, 45)	10	0	1	0	1/10	3	7	3/10	0.210
6	[45, 48)	9	1	1	1	1/9	3	6	3/9	0.222
7	[48, 55)									
8	[55, 62)									
9	[62, 72)								0	0
10	[72, 72]								0	0

The expected number of event occurrences in group 1 at time  $t_i$ , denoted  $E_{1i}$ , is given by

$$E_{1i} = \frac{n_{1i}d_i}{n_i} \quad (9.10)$$

The quantity  $d_i/n_i$  is the overall proportion of individuals at time  $t_i$  who experience the event, and  $n_{1i}$  is the number of individuals in group 1 who are at risk at time  $t_i$ .

#### MATHEMATICAL NOTE

Equation (9.10) is based on the assumption that the number of occurrences for group 1 at time  $t_i$  follows a hypergeometric distribution.<sup>6</sup>

Next, we need to compare the observed and expected counts at time  $t_i$ , which can be done by taking the difference between the counts and dividing by an appropriate scaling factor. This scaling quantity is the variance of the number of event occurrences at time  $t_i$ . Again, based on the assumption that the number of event occurrences at time  $t_i$  for group 1 follows a hypergeometric distribution, the variance for the number of event occurrences for group 1 at time  $t_i$ , denoted  $V_{1i}$ , is given by

$$V_{1i} = \frac{n_{1i}n_{2i}d_i(n_i - d_i)}{n_i^2(n_i - 1)}$$

### Activity Calculating the Expected Number and Variance of the Number of Melted Chips

Refer to the entries in Table 9.5.

32. Show that the expected number of melted milk chocolate chips in the first time interval is  $7/16 = 0.4375$ .
33. Show that the variance of the number of melted milk chocolate chips in the first time interval is 0.246.

Table 9.5 contains values of  $E_{1i}$  and  $V_{1i}$  for the first six time intervals for the milk chocolate chip and white chocolate chip melting time data. Remember that we need to compare the total observed and total expected counts over all the complete event times, so we compare the sum of the observed event occurrences,  $\sum_{i=1}^m d_{1i}$ , and the sum of the expected event occurrences,  $\sum_{i=1}^m E_{1i}$ , over all  $m$  complete event times. We put this together to obtain the statistic

$$\chi = \frac{\sum_{i=1}^m d_{1i} - \sum_{i=1}^m E_{1i}}{\sqrt{\sum_{i=1}^m V_{1i}}} \quad (9.11)$$

where  $\sum_{i=1}^m V_{1i}$  is the variance of the total number of event occurrences over the  $m$  complete event times. Basically, Equation (9.11) is nothing more than a  $z$ -score; i.e., it tells us how many standard deviations the total observed number of event occurrences is above or below its mean.

Although we can use Equation (9.11) to make our decision regarding the null hypothesis of equal survival functions, it is more common to see the square of the statistic in Equation (9.11) reported in statistical software output. The square of the statistic in Equation (9.11), called the **log-rank test statistic** (for two groups), is given by

$$\begin{aligned} \chi^2 &= \frac{(\text{total observed events in group 1} - \text{total expected events in group 1})^2}{\text{variance in the total number of all complete events}} \\ &= \frac{\left( \sum_{i=1}^m d_{1i} - \sum_{i=1}^m E_{1i} \right)^2}{\sum_{i=1}^m V_{1i}} \end{aligned} \quad (9.12)$$

If the sample size is reasonably large, then under the assumption that the two survival curves are identical,  $\chi^2$  approximately follows a chi-square distribution with 1 degree of freedom. If the observed and expected numbers of events are far apart, the test statistic will correspond to a small  $p$ -value; thus, we will reject  $H_0: S_1(t) = S_2(t)$  and conclude that the two population survival functions are different.

#### ► MATHEMATICAL NOTE ▼

The test statistic  $\chi^2$  is said to *asymptotically* follow a chi-square distribution, which means that as the total sample size approaches infinity, the distribution of the statistic will closely follow the chi-square distribution with 1 degree of freedom. Unfortunately, there is no generally accepted rule as to what constitutes a large sample size; however, simulation studies have shown that even for small sample sizes (10 per group), the statistic will be approximately chi-square if the amount of censoring is roughly the same in each group.<sup>7</sup>

### Activity (▶) Log-Rank Test Statistic

- 34. Calculate the missing quantities in Table 9.5. Note that  $E_{19}$ ,  $E_{1,10}$ ,  $V_{19}$ , and  $V_{1,10}$  are all equal to 0. This is because all the milk chocolate chips have melted by 62 seconds.
- 35. Use Equation (9.12) and the entries in Table 9.5 to verify that the value of the log-rank test statistic is 1.007. Note that your answer may vary slightly because of rounding.

The  $p$ -value corresponding to the value of  $\chi^2$  is 0.316. Hence, we do not have strong enough evidence to conclude that the melting experiences (survival curves) of the populations of white and milk chocolate chips significantly differ. Note that we should be cautious about interpreting the results of this test because

the sample size of 16 is fairly small. We should always supplement the results of the test with graphs of the Kaplan-Meier curves.

If the difference between the total observed and expected number of event occurrences is large, then the test statistic is much greater than 0 and the resulting  $p$ -value will be small. Then the null hypothesis will be rejected.

## Activity ◀ Log-Rank Test for Chip Melting Activity

36. Use software to conduct the log-rank test to determine if the melting experiences for the white and milk chocolate chips in your chip melting activity are different. What do you conclude about the survival experiences for the two types of chips?

## The Wilcoxon Test

Another test for comparing survival curves, commonly used in practice and reported in software output, is the **Wilcoxon test**. The Wilcoxon test is similar to the log-rank test in the sense that it also compares the observed and expected number of event occurrences; however, the Wilcoxon test statistic is a slight variant of the log-rank test statistic. It places more weight on differences in the survival curves at earlier times (when the number at risk will be larger), so it can be better at detecting differences in the survival curves at earlier time periods. Note that when survival curves cross, neither test may be able to detect a significant difference in survival experiences. This underscores the importance of graphing the survival curves and describing the survival experiences based on observation.

Although the results of the log-rank and Wilcoxon tests will typically agree, it is possible for the two tests to yield different results. In practice, when the tests yield different results, you should report the results of both tests. When the tests yield similar conclusions, results of either test can be provided.<sup>8</sup>

The log-rank test and the Wilcoxon test can be implemented with many statistical packages. For the chip melting data in Table 9.4, the results for the tests comparing survival curves are shown in Figure 9.7.

Test Statistics			
Method	Chi-Square	DF	P-Value
Log-Rank	1.00690	1	0.316
Wilcoxon	1.09402	1	0.296

**Figure 9.7** Software output for the log-rank and Wilcoxon tests.

Observe that the results of the two different tests are in agreement. Thus, even though the two curves based on the sample data look fairly different, we do not have enough evidence to conclude that the population survival functions for white and milk chocolate chips are different.

### Key Concept

The log-rank and Wilcoxon tests provide formal inferential methods for comparing the survival experiences of two or more groups.

### NOTE

The log-rank (and Wilcoxon) test can easily be applied to more than two groups. Although the details are not given here, the test statistic follows a chi-square distribution with  $k - 1$  degrees of freedom, where  $k = \text{number of groups}$ .<sup>9</sup>

## Beyond Different Survival Experiences: Regression Models for Survival Data

The log-rank and Wilcoxon tests are useful procedures for determining whether the survival experiences of two or more independent groups are different. This is basically equivalent to testing whether there is a significant effect of a categorical explanatory variable on survival. However, the tests do not allow us to determine the extent to which the survival experiences differ or, equivalently, to estimate the effects of the categorical explanatory variables on survival. Furthermore, the tests cannot be used to appropriately assess the effects of quantitative variables on survival. Consider the chocolate chip activity, and suppose we had additional information on the age of the chip. Are older chips likely to melt earlier than newer chips? Can we predict how long a chip that is a month old will take to melt? How do we incorporate censored melting times? These are questions that need to be addressed using regression-type models that exploit the relationship between the survival time random variable,  $T$ , and the explanatory variables of interest.

Regression models for survival data will not be discussed in this chapter; however, we will briefly mention two main approaches: regression models that explicitly model the log of the survival time variable,  $T$ , as a function of the explanatory variables, sometimes referred to as **parametric regression models** or **accelerated failure time models** and those that model the hazard function (to be discussed) as a function of the explanatory variables, known in the literature as the **proportional hazards model** or the **Cox regression model**.<sup>10</sup>

## 9.7 What Can We Conclude About Melting Chocolate Chips?

The goal in the previous sections and activities was to motivate introductory topics in survival analysis by conducting a simple experiment with chocolate chip melting times. As we have seen, the Kaplan-Meier estimator of the survival function forms the basis of several quantities used to summarize and investigate chip melting times, including descriptive measures like the mean and median.

We were also able to compare the melting experiences of white and milk chocolate chips using the log-rank test. Since the chip melting activity was an experimental study (students were randomly assigned to white or milk chocolate chips), if the result of the log-rank test is significant, then we can conclude that the type of chip affects the melting time (i.e., that type of chip causes differences in the melting experiences).

Note that the group sizes (i.e., the numbers of white and milk chocolate chips) could be small (depending on the number of students participating in the activity), and we need to be cautious about interpreting the results of the log-rank test if this is the case.

Although simple in scope, the chip melting time data you collected have the same general features as other survival data that will be introduced in the extended activities and that are of more practical importance and interest to researchers in various fields. Real studies that implement survival analysis methods are common in medicine, health, and the social sciences, and in the remainder of this chapter we will use data from real surveys and studies to see how survival analysis methods can be applied in a variety of disciplines.

### A Closer Look Survival Analysis

The remaining sections and extended activities introduce additional methods for analyzing survival data and additional types of incomplete data. Although the Kaplan-Meier estimator is useful for examining the proportion of individuals yet to experience the target event at a specific moment in time, it does not assess the **risk**, or **potential**, at that particular moment that an individual who has not previously done so will experience the target event. Risk can be addressed using the hazard function and the closely related cumulative hazard function, which will be discussed in the following sections. In the final section of this chapter, we'll introduce some other types of incomplete survival data, including left- and interval-censored data and truncated data.

## 9.8 The Hazard Function

Recall the chip melting experiment, and suppose that 45 seconds have passed. Among the chips that have not melted (have survived), at what rate will the chips melt (fail) in the next 5 seconds? In the next second? In the next instant of time? Would this rate be the same for chips that have not melted after 60 seconds? These

types of questions address when individuals (or chips, in this example) are likely to experience the target event, given that they have not previously experienced the event. To answer such questions, we will need to use another function of survival time called the hazard function, denoted  $h(t)$ .

In survival analysis, the (conditional) risk of experiencing the target event is associated with the rate at which individuals who have survived up to a particular time will experience failure at that moment. Earlier, when we discussed the construction of the Kaplan-Meier estimator, we introduced the risk set of individuals. The individuals in the risk set have survived (not experienced the event) up through time  $t_{i-1}$  and are available to experience the target event in the time interval  $[t_{i-1}, t_i]$  for  $i = 2, \dots, m$ . So anybody in the risk set at time  $t_{i-1}$  runs the risk of experiencing, or has the potential to experience, the target event in the next interval of time.

As the unit of time becomes finer and finer, until we are looking for a rate of failure in the next *instant* of time, we approach a quantity known as the **hazard rate** or **conditional failure rate**. The hazard rate at time  $t$ , also known as the (population) **hazard function**, denoted  $h(t)$ , is defined as the *instantaneous* rate of failure at time  $t$  for all subjects in the population who have survived until time  $t$ .

The formal definition of the population hazard function requires the use of calculus. Those of you who have the necessary background and wish to do so may read about  $h(t)$  in more detail in the next section. Otherwise, you can skim or skip the next section. At this point it is enough to understand that the hazard function (population or estimated) is used to describe when subjects are likely to experience the event of interest and can help us identify periods of high and low risk of event occurrence.

#### NOTE

If the notion of an “instantaneous” rate troubles you, imagine the situation when you are driving your car and you glance down at the speedometer to see how fast you are traveling. If the speedometer reads 65 miles per hour at the instant you glance at it, this is the instantaneous rate at which you are traveling in distance per hour.

## The Population Hazard Function\*

To reiterate, the population hazard function,  $h(t)$ , is the rate at which individuals in the population experience the target event in the next *instant* of time (time  $t$ ), conditional on having survived (not experienced the event) up to time  $t$ . Although you are probably comfortable with the idea of a small interval of time like one second or even a tenth of a second, it may be a bit harder to imagine an interval of time that lasts an instant. In calculus, an instant of time is regarded as an interval of time whose width approaches (but never reaches) 0. The hazard function,  $h(t)$ , is defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (9.13)$$

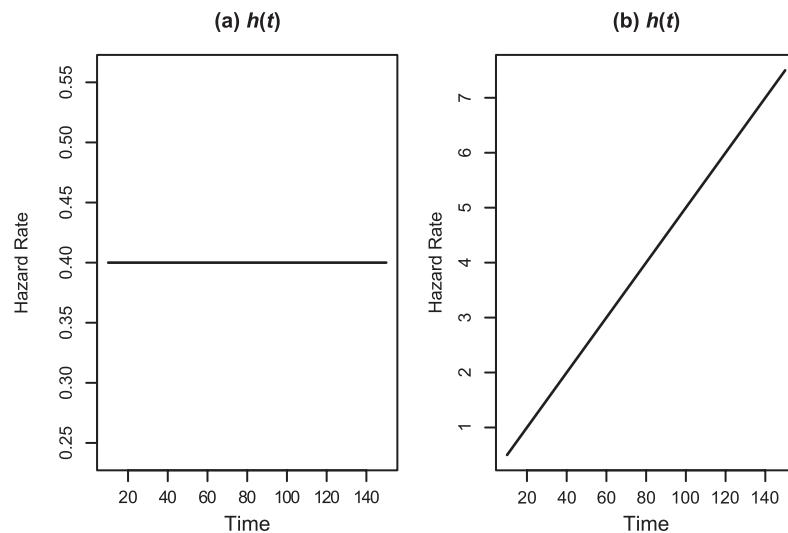
where  $\Delta t$  indicates a small change in time  $t$ .

From this expression, we see that the hazard is the conditional probability of experiencing the event per unit of time  $\Delta t$ . Note that since  $h(t)$  is a rate, it may take on values greater than 1. The only restriction on the hazard function is that  $h(t) \geq 0$ .

In some situations, a graph of  $h(t)$  can be constructed and examined for periods of high and low risk; however, an expression for  $h(t)$  requires an assumption about the probability distribution of the time-to-event random variable  $T$  (as well as some calculus and knowledge of intermediate concepts in mathematical statistics). In fields such as engineering, a distribution for  $T$  may be assumed, where, for example,  $T$  might be defined as the time until a computer processing chip fails, and then graphs of  $h(t)$  [as well as  $S(t)$ ] can be constructed.

Although we will not provide any computational examples of  $h(t)$ , we'll provide two hypothetical hazard curves and describe them in the context of melting chocolate chips.<sup>11</sup> Figure 9.8 displays two

\*Calculus is suggested for this section.



**Figure 9.8** Possible forms of the population hazard function.

hazard functions that are used in practice when simplifying assumptions about the data can be made. The hazard function in Part (a) illustrates constant risk, while the hazard function in Part (b) illustrates increasing risk. Part (a) displays a situation or study where the subjects have a constant risk of experiencing the target event. In the context of melting chips, the constant hazard rate of 0.40 (i.e.,  $h(t) = 0.4$ ) shown in Part (a) would correspond to a period throughout which events (melting chips) occur with equal frequency during any equal period of time. Now examine Part (b). The hazard function is increasing over time; furthermore, the hazard rate is linearly related to time. This implies that the risk of chips melting increases as a function of time, corresponding to a period of time during which the proportion of unmelted chips is decreasing.

Equation (9.13) may look something like the formal definition of the derivative of some function (although it probably isn't clear what that function is). It turns out that  $h(t)$  is related to the derivative of a function of the (population) survival function  $S(t)$  by the expression

$$h(t) = -\frac{d}{dt} \ln[S(t)] \quad (9.14)$$

That is, the hazard function is equal to the negative of the derivative of the natural logarithm of the survival function. This identity is not very obvious, and we skip the mathematical details here. The important thing to note is that since  $S(t)$  is a function based on the distribution of the time-to-event random variable  $T$ ,  $h(t)$  is also a function based on the distribution of  $T$ . Hence, if we know the exact distribution of  $T$ , we can calculate the hazard function using calculus.

### Key Concept

The population hazard function describes how the conditional risk of experiencing the event changes over time for individuals in the population.

## The Estimated Hazard Function

### NOTE

The *R* statistical package is suggested for calculating and graphing the estimated hazard and estimated cumulative hazard functions discussed in this section and Section 9.9.

We are not likely to know the true hazard rate at any particular moment in time unless we can make some assumption about the time-to-event variable  $T$ . Typically we have only observed event times that can be used to estimate the true hazard rates. Whereas the population hazard function,  $h(t)$ , is the rate at which individuals in the population experience the target event in the next *instant* of time (time  $t$ ), conditional on surviving to time  $t$ , the estimated hazard function assesses the conditional failure rate during the  $i$ th *interval* of time  $[t_i, t_{i+1})$  for those individuals in the sample who have survived to time  $t_i$ .

The estimates of hazard are based on quantities used to find the Kaplan-Meier estimated probabilities, so we will first construct time intervals  $[t_i, t_{i+1})$  for  $i = 1, \dots, m - 1$  as if we were constructing the Kaplan-Meier estimator, where  $t_0 = 0$  and  $m$  is the number of complete event times (see Section 9.3 for details).

Recall that the estimated conditional probability of experiencing the target event in the  $i$ th interval  $[t_i, t_{i+1})$  is given by

$$\hat{p}_i = d_i/n_i$$

for  $i = 0, 1, \dots, m - 1$ . Recall that  $\hat{p}_i$  represents the proportion of remaining subjects in the sample at the beginning of the  $i$ th interval who experience the target event in the time interval  $[t_i, t_{i+1})$ .

If we divide  $\hat{p}_i = d_i/n_i$  by the width of the time interval  $t_{i+1} - t_i$ , then the resulting quantity

$$\frac{\hat{p}_i}{t_{i+1} - t_i} \quad (9.15)$$

yields the *average* conditional probability of event occurrence per unit of time within  $[t_i, t_{i+1})$ . This quantity is the **estimated hazard rate** or **estimated hazard function** for  $t_i \leq t < t_{i+1}$ , denoted by  $\hat{h}(t)_{\text{KM}}$ , and measures the estimated conditional probability of experiencing the event *per* unit of time in  $[t_i, t_{i+1})$ . When no interval of time is specified,  $\hat{h}(t)_{\text{KM}}$  will be used to refer to the estimated hazard function for any  $t$  within the entire observed range of time.

Table 9.3 contained the observed number of melted chips, number of chips at risk, and so on, for the chip melting times. In Table 9.4, columns containing the width of each interval of time  $t_{i+1} - t_i$  and the estimated hazard rate  $\hat{h}(t)_{\text{KM}}$  for each time interval have been appended, and values have been computed for the first three rows. We interpret the estimated hazard rates  $\hat{h}(t)_{\text{KM}} = 0$  for  $t_0 \leq t < t_1$ ,  $\hat{h}(t)_{\text{KM}} = 0.0286$  for  $t_1 \leq t < t_2$ , and  $\hat{h}(t)_{\text{KM}} = 0.0111$  for  $t_2 \leq t < t_3$  in the following manner: Since no chips melt within the time interval  $[0, 25]$ , 0% of the unmelted chips melt per second during that interval. Within the time interval  $[25, 30)$ , we estimate that, of the chips that have not yet melted after 25 seconds, about 2.9% will melt per second. The estimated hazard rate is not defined for the last interval  $[55, 60)$ , since the width of the last interval cannot be determined because 60+ is a censored melting time.

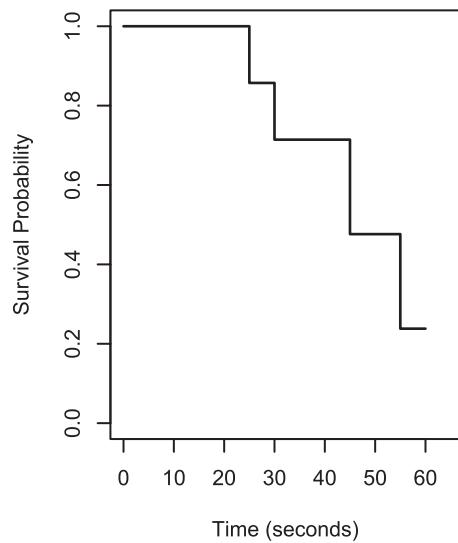
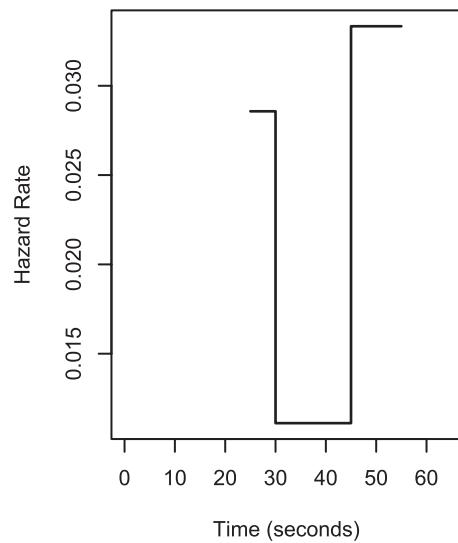
## Extended Activity ➔ Chip Melting Hazard Rates

- 37. Compute the missing entries in Table 9.6.
- 38. Among chips that have not yet melted after 30 seconds, estimate the rate at which the chips will melt in the next 15 seconds.
- 39. Excluding the first time interval, during which period (interval) of time are chips at their highest risk of melting? Lowest risk?
- 40. Do the values of  $\hat{h}(t)_{\text{KM}}$  suggest that the estimated hazard function is a strictly increasing or strictly decreasing function, or do they suggest that the function can increase and decrease over time?

We can plot the estimated hazard rates versus time to view the estimated **hazard curve** for the sample of survival times. The Kaplan-Meier curve and estimated hazard function for the chocolate chip melting times are shown in Figure 9.9. Note that, because of the small sample size, the plot is extremely rough and the pattern is somewhat erratic. This curve is an approximation to the corresponding true hazard function for the population of all chocolate chip melting times. From Figure 9.9 we observe that the lowest risk of melting occurs between 30 and 45 seconds (among chips that have remain unmelted up through 30 seconds). The highest risk of melting occurs between the 45th and the 55th second. Note

**Table 9.6** Selected quantities for calculating the estimated hazard rates.

Interval $i$	Time Interval	$n_i$	Number Censored	$d_i$	$\hat{p}_i$	$1 - \hat{p}_i$	$\hat{S}(t)_{KM}$	$t_{i+1} - t_i$	$\hat{h}(t)_{KM}$
0	[0, 25)	7	0	0	0	1	1	25	0
1	[25, 30)	7	0	1	1/7	6/7	6/7	5	0.0286
2	[30, 45)	6	2	1	1/6	5/6	5/7	15	0.0111
3	[45, 55)								
4	[55, 60)								NA

(a)  $\hat{S}(t)_{KM}$  for Time Until Chip Melts(b)  $\hat{h}(t)_{KM}$  for Time Until Chip Melts**Figure 9.9** Estimated survival probabilities and estimated hazard rates for the chocolate chip data.

that the estimated hazard rate is not defined before the first complete event time and after the last complete time.

## Extended Activity

### Estimated Hazard Rates

Examine Figure 9.9, the entries in Table 9.6, and Formula (9.15) to address the following questions:

41. In general, can  $\hat{h}(t)_{KM}$  take on a negative value in any time interval? Briefly explain. What does this suggest about the minimum value of  $\hat{h}(t)_{KM}$ ?
42. Is there a maximum value that  $\hat{h}(t)_{KM}$  can take of within a (finite) interval of time? Briefly explain.

Note some additional features of the estimated hazard function:

- The estimated hazard curve extends only to the last complete event time,  $t_m$ ; that is, if the last time interval is of the form  $[t_m, t_n]$ , where the largest observed event time,  $t_n$ , is censored, then no hazard rate is defined during that interval. The reason is that the width of this last interval cannot be determined. In this respect the estimated hazard curve is unlike the Kaplan-Meier curve, which would extend to the largest observed event time if it were censored.
- The estimated hazard curve  $\hat{h}(t)_{KM}$  will typically have an erratic pattern, especially when the number of event times is large. It can be difficult to provide a general summary of the hazard rate over time.

## Age at First Alcoholic Drink

When do individuals have their first drink of alcohol? The legal age for consuming alcohol is 21 years, but some individuals claim to have had their first alcoholic drink when they were as young as 1 year old! We can investigate the age at which individuals had their first drink of alcohol using data from the National Comorbidity Survey of 1990–1992. Participants were asked to recall the age at which they had their first drink of alcohol. Those who could recall the age at which they had their first drink had complete event times (i.e., the age at which they had their first drink). Individuals who had not had a drink by the time of the interview had a right-censored event time (their age at the time of the interview). Individuals who could not recall the age at which they had their first drink were not included in the sample.

### Extended Activity Age at First Alcoholic Drink

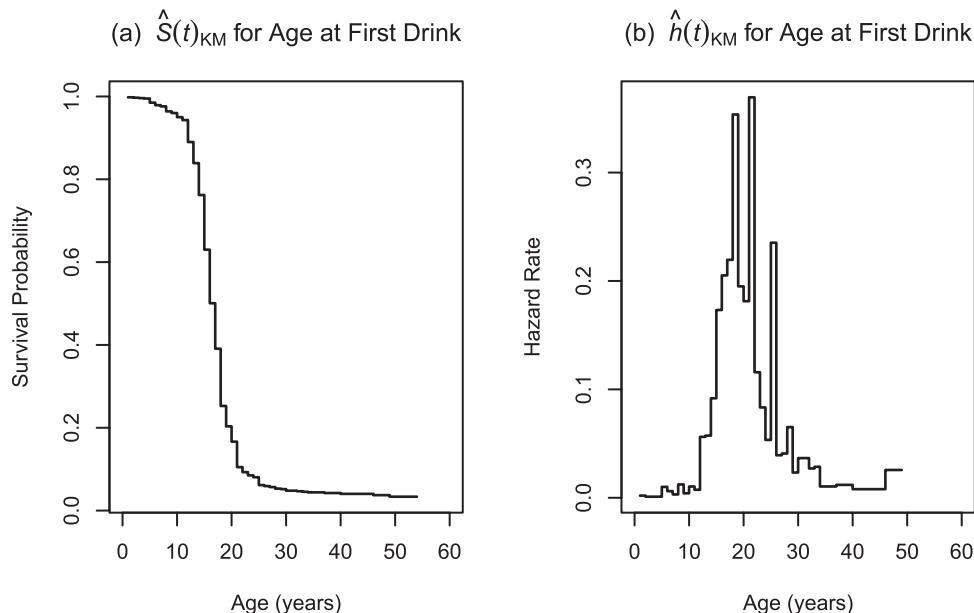
Data set: `Firstdrink`

43. Investigate the time until first drink by creating a Kaplan-Meier curve (with a confidence interval) for the data. What do you observe about the survival probabilities?
44. Use the software instructions provided to plot the estimated hazard rates for the age at first drink data. We will examine the plot in the following paragraphs. Note that Minitab computes the estimated hazard rates using an expression different from Formula (9.15) (it restricts the rate to fall between 0 and 1), so it is recommended that the R software be used to construct the plot of the estimated hazard function.

From the Kaplan-Meier curve in Figure 9.10, we can observe that the estimated proportion of individuals who had not taken their first drink decreases rapidly after age 13. What does this mean in terms of the estimated hazard rates? Examine Figure 9.10, which was produced using the R statistical software. Part (a) displays the Kaplan-Meier curve  $\hat{S}(t)_{\text{KM}}$ , while Part (b) displays the estimated hazard curve  $\hat{h}(t)_{\text{KM}}$ .

Although not practical, we could interpret the estimated hazard rate for each time interval as was done for the chip melting data. Instead, if possible, we should try to summarize some of the aspects of the estimated hazard curve in terms of periods of low and/or high risk of event occurrence.

As mentioned earlier, the estimated hazard curve exhibits a rough and erratic pattern. But we can observe a sharp increase in  $\hat{h}(t)_{\text{KM}}$  from about 13 years to about 18 years. This seems to suggest that individuals who



**Figure 9.10** Estimated survival probabilities and estimated hazard rates for the age at first drink data.

have not yet tried alcohol are at high risk of having their first drink during their adolescent and teenage years, possibly because of peer pressure from their friends in junior and senior high school. The estimated hazard curve drops right after age 18, but note the age at which it spikes back up again! For those individuals who have not yet had alcohol by age 21, the estimated hazard rate suddenly jumps to its highest point, indicating that drinking is likely to occur among 21-year-olds who have never had alcohol before. After age 22, the estimated hazard decreases until about 25 years, when there is another sudden increase. The curve generally decreases after age 26 and levels off at about 34 years of age.

### Key Concept

The estimated hazard rate assesses conditional risk at a specific moment in time for individuals in the sample. It is defined as the rate at which individuals in the sample who have not already experienced the target event will do so in the next small interval of time. The estimated hazard function shows how the conditional risk of event occurrence changes over time for a sample of subjects.

## College Graduation

How long does it take for students to graduate from college? At what time(s) during their college career are they most likely to graduate? Perhaps you have just started college and have a few years to go, or maybe you're planning to graduate soon. We can investigate the number of years needed to complete the requirements for a bachelor's degree using data from individuals who participated in the National Educational Longitudinal Survey (NELS) from 1988 to 2002. To illustrate how survival analysis techniques can be used to answer these questions, we will use data on a sample of 1000 participants from the NELS data set who began college prior to the year 2000. Note that this sample includes individuals who began study at any community college or bachelor's degree granting postsecondary institution, even though some students who began their postsecondary education at a two-year college may not have been intending to pursue a bachelor's degree.

The time-to-event random variable is the number of years taken to complete the requirements for a bachelor's degree. If an individual had obtained a bachelor's degree by the interview date, then her or his event time is complete. If the student had dropped out or had not graduated by the interview date, then her or his time is right censored.

### Extended Activity ▶ College Graduation

Data set: Graduate

45. Use the software instructions provided to plot the estimated hazard rates for the college graduation data.
46. Although the estimated hazard curve may not exhibit a distinguishable pattern, discuss some important features of the curve.
47. Indicate periods of time during their college career when students are at their lowest and highest risk of graduating college. Does your answer match your common understanding of when students typically graduate from college?

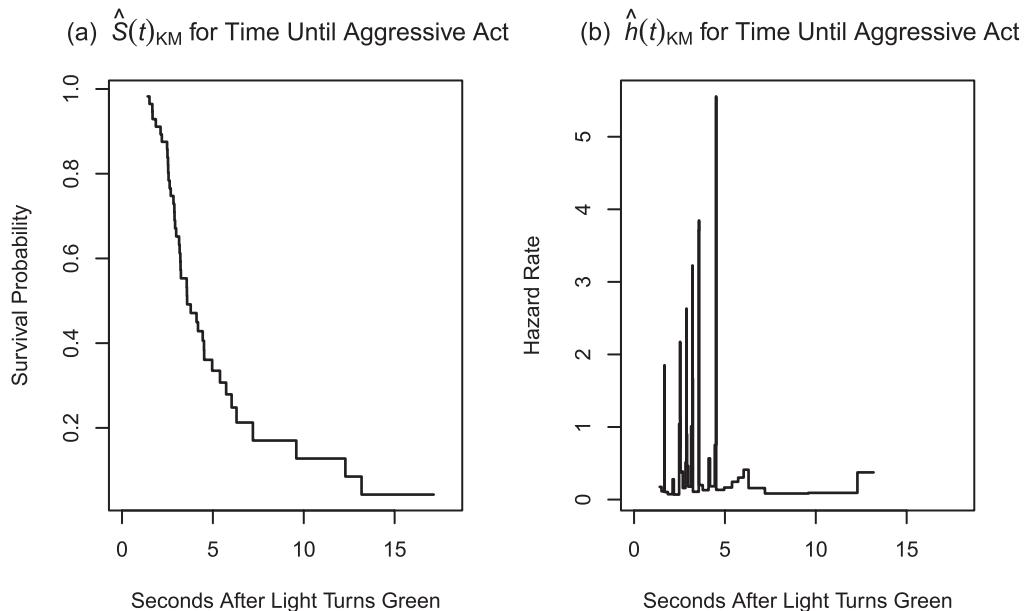
## 9.9 The Cumulative Hazard Function

The coarse nature of the estimated hazard function can make it difficult to describe or summarize how the conditional risk of event occurrence changes over time. An alternative method for assessing and describing how hazard rates change over time is to investigate the accumulation of the hazard rates over time and look for patterns in the **cumulative hazard**. The function that allows us to examine accumulated hazard over time is called the **cumulative hazard function**. By examining the cumulative hazard function, we can detect when the hazard is increasing, decreasing, or remaining relatively constant.

## Motorist Reaction Times

Have you ever been frustrated because you were stuck behind a car that was stopped in front of you for no apparent reason? Did you feel inclined to honk your horn (or make some other gesture)? In a study on aggressive behavior displayed by motorists, Diekmann and his colleagues investigated the time it took for 57 motorists intentionally blocked at a green light by a Volkswagen Jetta to show signs of aggression.<sup>12</sup> Signs of aggression included honking their horn or beaming their headlights at the Jetta. The time-to-event variable is the time (measured in seconds to the nearest hundredth of a second) until the motorist honked (or beamed the headlights) at the Jetta. If the motorist did not honk or flash the headlights by the time the Jetta moved, then the motorist's event time was right censored.

The Kaplan-Meier survival curve and the estimated hazard function for the motorist reaction time data are displayed in Figure 9.11. The estimated hazard function behaves erratically and displays several spikes, making it very difficult to assess how the hazard rate is changing over time.



**Figure 9.11** Estimated survival probabilities and estimated hazard rates for the time until the motorist displays aggressive behavior.

Similar to the survival function and hazard function, we can define a cumulative hazard function for a population and a sample. We will denote the population cumulative hazard function by  $H(t)$ , and the estimator of the cumulative hazard function by  $\hat{H}(t)_{KM}$ . The definition of  $H(t)$  is given by

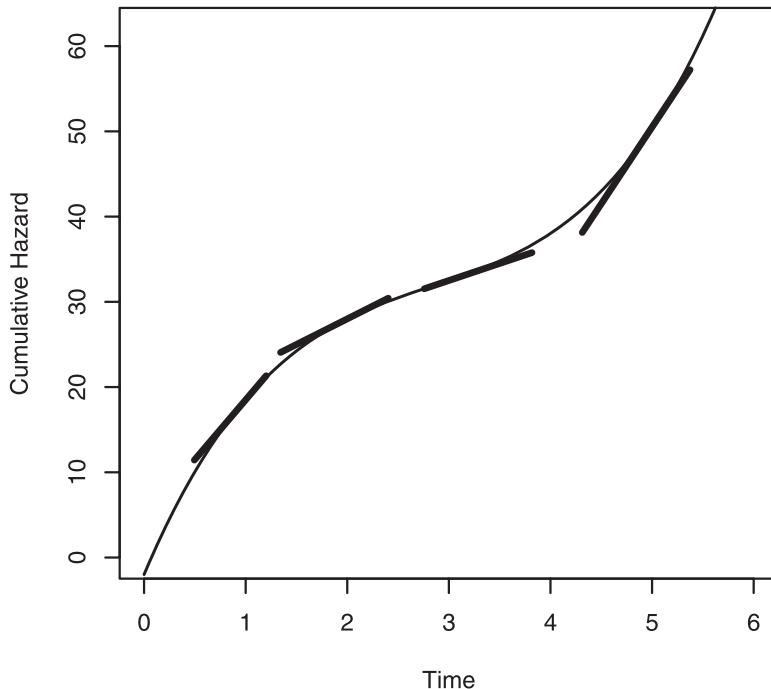
$$H(t) = \text{the accumulation of the hazard rate } h(t) \text{ for time } T \text{ between } 0 \text{ and } t \text{ for subjects in the population}$$

A background in calculus is necessary to fully understand and appreciate the definition of the cumulative hazard function  $H(t)$ . We present it at the end of this section as a mathematical note. In this section we will look at the reasons for examining the cumulative hazard function, and then in Section 9.8 we will discuss computation of the estimated cumulative hazard function.

An important point to make is that the cumulative hazard function  $H(t)$  is neither a probability nor a rate. It is an accumulation of (hazard) rates over time. Since  $H(t)$  is cumulative, it never decreases (and rarely remains constant). Furthermore, we can gain a better idea of how hazard changes over time by examining the nature of the change in  $H(t)$ —that is, whether the *rate* of change in  $H(t)$  is increasing or decreasing.

To better understand what is meant by an increase or decrease in the rate of change in  $H(t)$ , consider Figure 9.12. The dashed line represents the population cumulative hazard function  $H(t)$ , and the values of the

slopes of the four solid lines represent estimates of the rates of increase in  $H(t)$  over the corresponding four time periods. Observe that  $H(t)$  increases over the entire range of time, but the rates of increase (the values of the four slopes) vary over the course of time. The slopes of the first three lines are decreasing, implying that the rate of change in  $H(t)$  decreases until about time  $t = 4$ . The slope of the fourth line is greater than the previous slope, indicating that the rate of change *increases* after time  $t = 4$ .



**Figure 9.12** Cumulative hazard function displaying various rates of change over time.

In general, we can examine how the rate of change in the cumulative hazard function changes over time to understand how the hazard function,  $h(t)$ , changes over time.

- If the *rate of change* in  $H(t)$  is *increasing* (over an interval of time), then  $h(t)$  is increasing (over the same interval of time).
- If the *rate of change* in  $H(t)$  is *decreasing*, then  $h(t)$  is decreasing.
- If the *rate of change* in  $H(t)$  is *constant* (and greater than 0), then  $h(t)$  is constant (and greater than 0).
- If the *rate of change* in  $H(t)$  is 0, then  $h(t)$  is 0.

Figure 9.13 presents the hazard functions originally shown in Figure 9.8 and their corresponding cumulative hazard functions. We can see that if the hazard rate is constant over time [Part (a)], then the cumulative hazard function [Part (c)] will increase at a *constant* rate—that is, there is a perfect positive linear relationship

#### ► MATHEMATICAL NOTE ▼

Assuming a particular probability distribution for  $T$ , the mathematical definition of  $H(t)$  requires integral calculus and is given by

$$H(t) = \int_0^t h(x) dx \quad (9.16)$$

In general,  $H(t)$  increases as  $t$  increases [ $H(t)$  is constant only when  $h(t) = 0$ ]. Also note that  $H(t)$  is neither a rate nor a probability (it is an accumulation of rate over time).

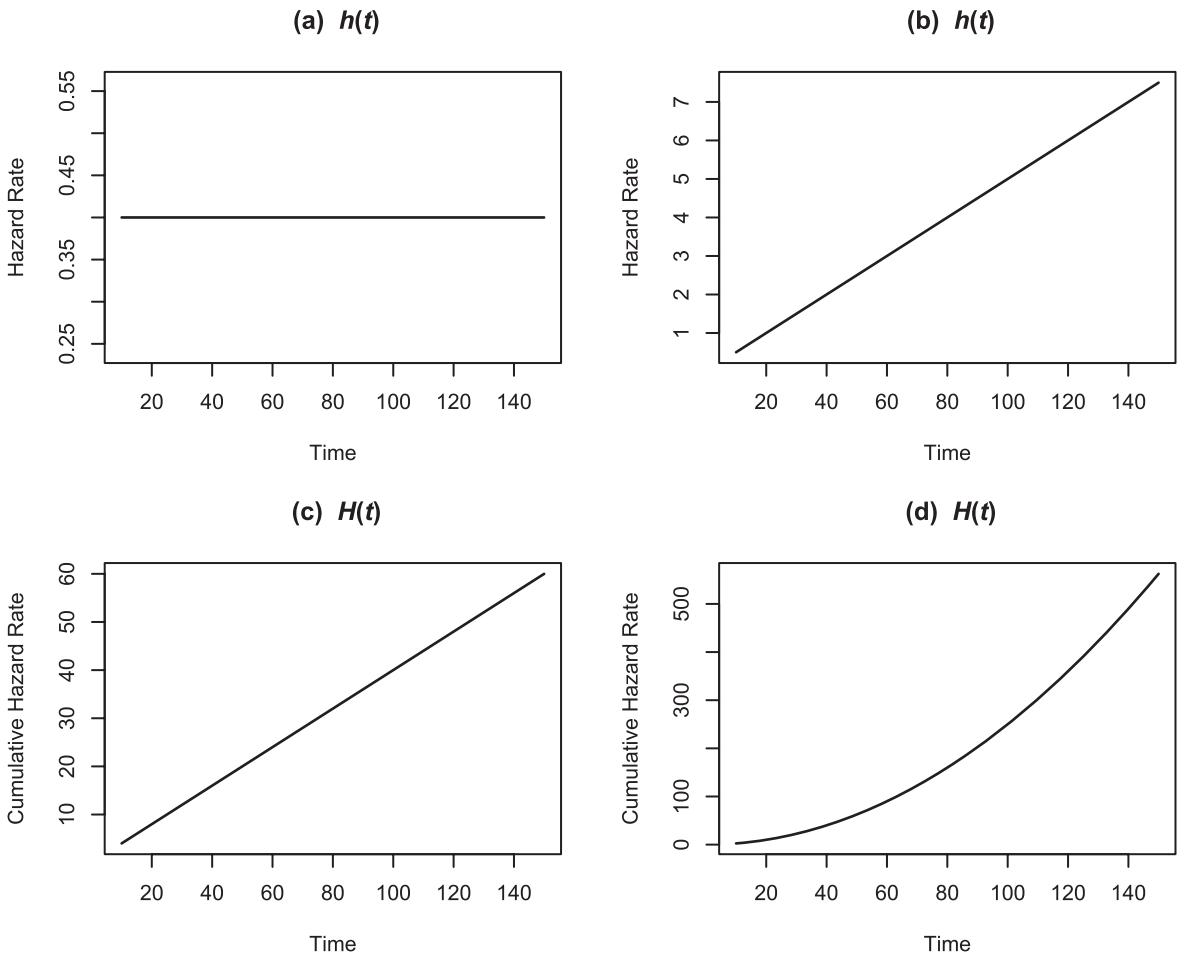


Figure 9.13 Population hazard functions and their corresponding cumulative hazard functions.

between values of  $t$  and values of  $H(t)$ . If the hazard rate is increasing linearly over time [Part (b)], then the rate of change in  $H(t)$  is increasing [Part (d)].

## Estimator of the Cumulative Hazard Function

Since  $H(t)$  is an accumulation of the population hazard  $h(t)$  between time 0 and time  $t$ , it makes intuitive sense that an estimator of  $H(t)$  should also accumulate (or sum up) the estimated hazard rates computed between time 0 and time  $t$ . This is the approach we will take.

Once again, we will start with intervals of time defined as if we were computing the Kaplan-Meier estimator. Then, given the estimated hazard rates  $\hat{h}(t)_{\text{KM}}$  for  $t_i \leq t_{i+1}$ , we can calculate the **total estimated hazard** during each time interval  $[t_i, t_{i+1})$  for  $i = 1, \dots, m - 1$ , given by

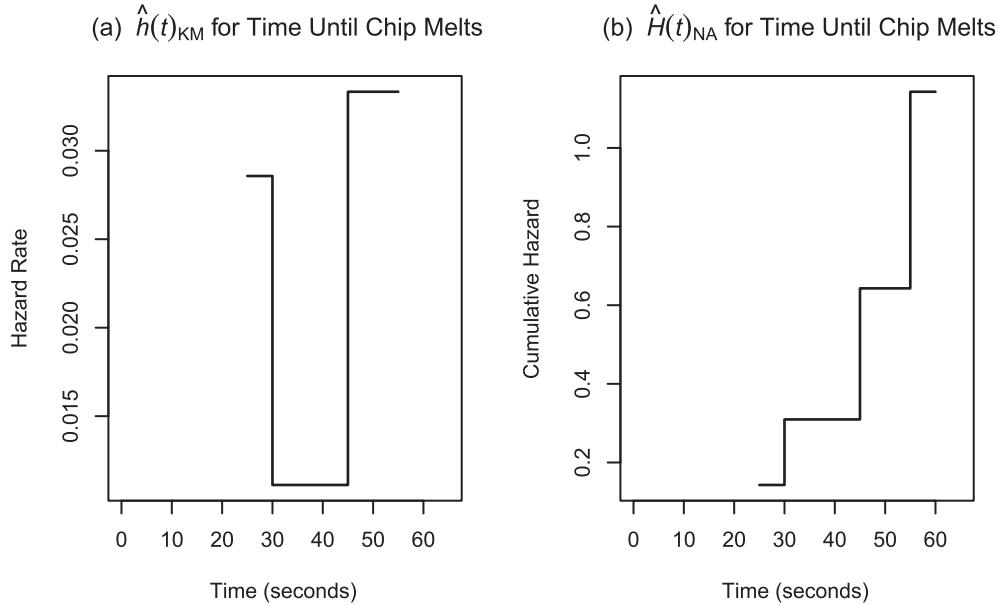
$$\text{Total estimated hazard during } [t_i, t_{i+1}) = \hat{h}(t)_{\text{KM}} \times (t_{i+1} - t_i)$$

where  $\hat{h}(t)_{\text{KM}}$  is the estimated hazard rate for  $t_i \leq t < t_{i+1}$  and  $t_{i+1} - t_i$  is the width of the  $i$ th interval. Then the **Nelson-Aalen estimator** of  $H(t)$ , denoted  $\hat{H}(t)_{\text{NA}}$ , is simply the sum of these total estimated hazard quantities up to a particular time  $t$ , given by

$$\hat{H}(t)_{\text{NA}} = \sum_{t_i \leq t} [\hat{h}(t)_{\text{KM}} \times (t_{i+1} - t_i)] \quad (9.17)$$

Note the following details about  $\hat{H}(t)_{NA}$ , depending on whether the last observed event time is censored or complete:

- If the largest observed event time, denoted  $t_n$ , is censored, then  $\hat{H}(t)_{NA}$  will peak (reach its highest point) at the last complete event time,  $t_m$ , and then extend to  $t_n$ —that is, it will be constant over the interval  $[t_m, t_n]$  (see Figure 9.14 for example).
- Otherwise, if the last observed event time is complete, then  $\hat{H}(t)_{NA}$  will simply reach its highest value at the complete time,  $t_m$ .



**Figure 9.14** Estimated hazard function and estimated cumulative hazard function for the chip melting data.

## Extended Activity

### ▶ Estimated Cumulative Hazard Function

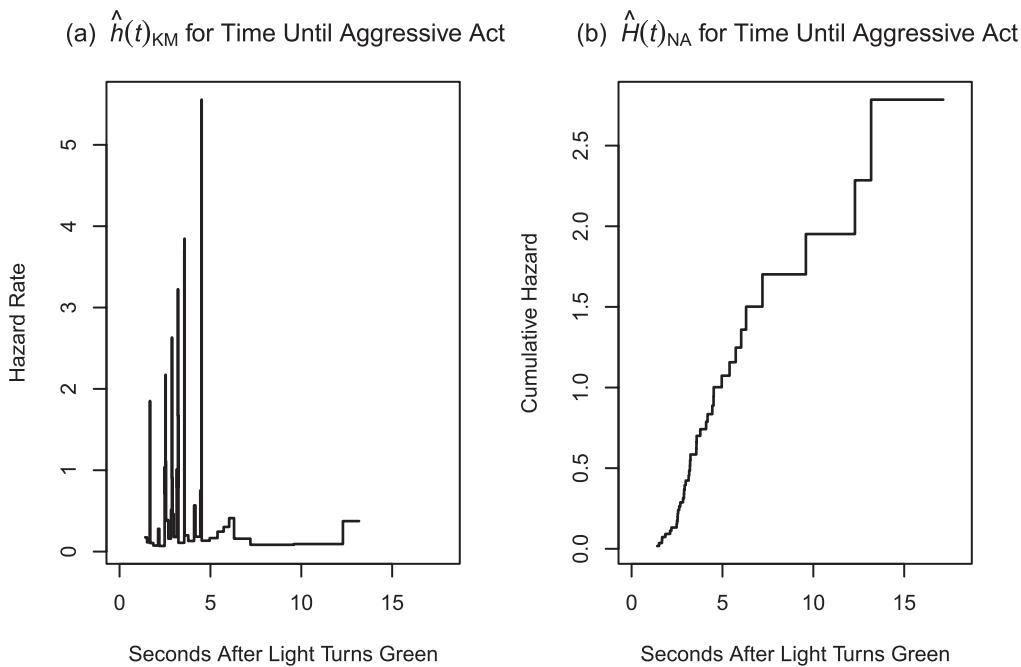
48. Use Equation (9.17) and the quantities in Table 9.6 to compute  $\hat{H}(t)_{NA}$  for the chip melting times from Table 9.2 by hand.

As discussed earlier in connection with the population cumulative hazard function  $H(t)$ , we can also examine how the rate of change in the estimated cumulative hazard,  $\hat{H}(t)_{NA}$ , changes over time, to investigate changes in the estimated hazard function over time.

- If the *rate of change* in  $\hat{H}(t)_{NA}$  is *increasing* (over an interval of time), then  $\hat{h}(t)_{KM}$  is increasing (over the same interval of time).
- If the *rate of change* in  $\hat{H}(t)_{NA}$  is *decreasing*, then  $\hat{h}(t)_{KM}$  is decreasing.
- If the *rate of change* in  $\hat{H}(t)_{NA}$  is *constant* (and greater than 0), then  $\hat{h}(t)_{KM}$  is constant (and greater than 0).
- If the *rate of change* in  $\hat{H}(t)_{NA}$  is 0, then  $\hat{h}(t)_{KM}$  is 0.

Figure 9.14 displays the estimated hazard curve for the chip melting times (shown earlier in Figure 9.9) in Part (a) and the estimated cumulative hazard function in Part (b). Because of the small number of intervals, it is difficult to determine if the rate of change in  $\hat{H}(t)_{NA}$  is increasing or decreasing.

Let's return to the motorist reaction time data. As we've already seen, the estimated hazard function displayed several spikes, but the overall pattern was difficult to summarize. Figure 9.15 displays the estimated hazard and cumulative hazard functions.



**Figure 9.15** Estimated hazard function and estimated cumulative hazard function for the aggressive motorist behavior.

We can summarize the rates of increase and decrease in the estimated cumulative hazard function shown in Part (b) of Figure 9.15 and describe the changes in the estimated hazard function. The rate of change in the estimated cumulative hazard is slowest between seconds 1.5 and 2.5 and is then followed by a higher rate of change for the next 2.5 seconds. Between the 5th and the 8th second the rate of change in the estimated cumulative hazard decreases, and then after the 8th second the rate of change remains fairly constant. From this general description of  $\hat{H}(t)_{NA}$  we can summarize the changes in the estimated hazard rates. Hazard increases between seconds 1.5 and 2.5, but is somewhat low, and then increases substantially between the 2.5th and the 5th second. We might consider this period of time to be the “boiling point” of frustration, when motorists who have not done so already are most likely to become impatient and decide to honk their horn or flash their high beams. Between the 5th and the 8th second, hazard decreases, after which it levels off.

#### Key Concept

The cumulative hazard function is a useful graphical display for describing the accumulation of hazard over time and for showing particular time periods when risk is high.

## Time to Rearrest for Former Inmates

For 36 months, Henning and Frueh followed criminal activities of 194 inmates released from a medium security prison.<sup>13</sup> We can use the data from their study to investigate the time until the former inmates were rearrested. If the former inmate had been rearrested for a criminal act before 36 months (after initial prison release) had passed, then that former inmate’s event time is complete. If the former inmate had not been rearrested for a criminal act after 36 months had passed or had completely dropped out of the study, then that former inmate’s event time is right censored. In addition to the time until rearrest, measurements are also available on the following variables:

person: a dichotomous variable identifying former inmates who had a history of person-related crimes—that is, those with one or more convictions for offenses such as aggravated assault or kidnapping

property: a dichotomous variable indicating whether former inmates had been convicted of a property-related crime

cenage: the “centered” age of the individual—that is, the difference between the age of the individual on release and the average age of all inmates in the study

## Extended Activity

### Estimated Cumulative Hazard Function

Data set: Rearrest

49. Use statistical software to construct the estimated hazard function and the cumulative hazard function for the time to rearrest data.
50. The estimated hazard function will be difficult to describe, but try to explain how the risk of rearrest changes over time based on the estimated cumulative hazard function. When does the risk of being rearrested appear to be the highest? The lowest?

#### NOTE

In this chapter, we have focused primarily on estimating the population survival, hazard, and cumulative hazard functions based on a sample of survival times. The models that we have fit to data are sometimes referred to as **nonparametric** since we haven’t assumed anything about the shape of the distribution of the survival time random variable  $T$  (e.g., whether  $T$  is normally distributed). You may wonder what we can do if the distribution of  $T$  is known. A class of models called **parametric** models can sometimes be used to represent the survival function, hazard function, and cumulative hazard function if we know (or have a pretty good idea about) the exact distribution of  $T$  (this was mentioned briefly in the discussion of the survival function). With the exception of a few simple graphical examples and brief descriptions, our discussion of parametric models has been very limited, and we have not provided any explicit formulas for  $S(t)$ ,  $h(t)$ , or  $H(t)$ .<sup>14</sup>

## 9.10 Additional Types of Incomplete Data

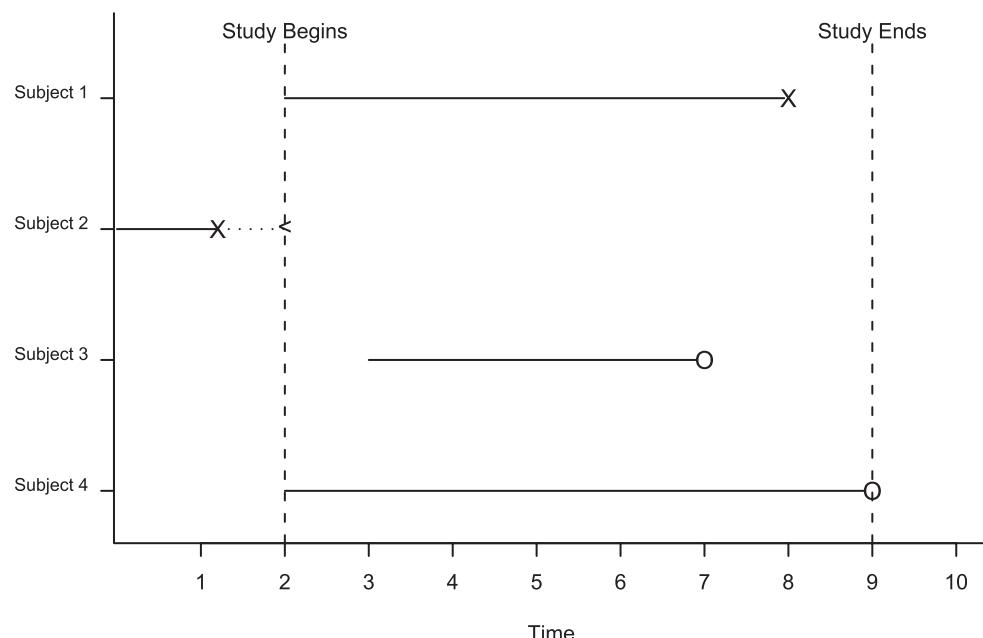
Throughout our discussions of survival analysis techniques, we’ve assumed that any incomplete data are due to right censoring. In Section 9.2, you learned that right censoring occurs when observation of an individual begins at a defined starting time and ends before the outcome of interest is observed. In this section, we’ll describe two other censoring mechanisms, left censoring and interval censoring, and two selection processes that determine whether individuals will be included in the study, left truncation and right truncation.

### Left Censoring and Interval Censoring

**Left censoring** occurs when the event of interest is known to have occurred before the study began.\* For an individual with a left-censored survival time, we know that the study start time is greater than the time at which the event of interest occurred.

To see the difference between right censoring and left censoring, examine Figure 9.16. Figure 9.16 displays the target event times of four subjects. For now, we’ll assume that time is measured in months. We’ll discuss subjects 1, 3, and 4 first. Subject 1 entered the study at time 2 and experienced the target event at time 8, so subject 1 has a complete target event time of 6 months. Subject 3 entered the study at time 3, but dropped out of the study at time 7. Hence, subject 3 has a right-censored event time of 4 months. This means that subject 3’s event time is *at least* 4 months. Finally, subject 4 entered the study at time 2 and had not experienced the event by the end of the study at time 9. Subject 4 has a right-censored event time of 7 months.

\*Technically, left censoring can occur after a study has started if the event of interest occurred prior to a particular recorded time. For example, a subject may experience an event of interest during the course of a study (such as contracting a disease) but not know the exact time of the event (knowing only that it occurred prior to testing positive for the disease).



**Figure 9.16** Observed survival times for four subjects. X indicates the target event occurred at displayed time; O indicates the target event was not observed (event time is censored); < indicates when Subject 2 was first observed in the study.

Subject 2 is rather interesting. Subject 2 experienced the event prior to the beginning of the study but was not observed until time 2. In other words, subject 2's observed event time is *greater* than her exact event time. Her event time is left censored.

**Interval censoring** occurs when the event of interest is known to have occurred between two time points, but the precise time is not known.

## Extended Activity

### Left Censoring and Interval Censoring

51. **Counting to Ten** A child development researcher is interested in the age at which children first learn to count to 10. A particular child in the study was 2 years old and had already learned to count to 10 when the researcher interviewed her parents; however, the parents couldn't remember exactly how old she was when she first counted to 10. Briefly explain why this particular child's event time is left censored.
52. **Type 2 Diabetes** In a study to determine whether exposure to oral contraceptives increases the risk of developing type 2 diabetes among Latina women with prior gestational diabetes milletus (GDM), women who had recently given birth and had been previously diagnosed with GDM were screened for type 2 diabetes at intervals ranging from every three months to every year during the period of the study (1987–1994).<sup>15</sup> Briefly explain why the times to develop type 2 diabetes were interval censored.

#### Key Concept

Left and interval censoring are two additional mechanisms that lead to incomplete observations. Left censoring occurs when the exact time an event occurred is unknown; it is known only that the event occurred prior to a particular time  $t$ . Interval censoring occurs when the event time is known only to have occurred between two time points.

## Truncation Mechanisms

In addition to censoring, truncation is another feature of a study that may cause survival data to be incomplete. **Truncation** is a mechanism inherent in particular studies that determines which individuals will be selected for observation. Individuals who have not experienced a specific truncation condition will not be included in the study. There are two types of truncation mechanisms to consider:

- **Left truncation**, also known as **delayed entry**, occurs when a subject is included in the study only *after* a specific condition has been met or a particular event has occurred (but not the target event of interest)—that is, the truncation condition has occurred. The individuals who are included are said to have left-truncated survival times. Those who do not experience the condition are not included in the study.
- **Right truncation** is a process by which only subjects who have experienced the target event are included in the study. In this situation, the truncation condition *is* the target event of interest. The individuals who are included are said to have right-truncated survival times. Any individual who has yet to experience the event is not included in the study.

### Extended Activity

#### Left and Right Truncation

53. **Drug Relapse** Among individuals who had undergone a treatment program for habitual drug use, researchers were interested in the time until subjects experienced their first relapse. Time until first relapse was measured from the beginning of the treatment program; however, individuals who did not successfully complete the treatment program were not included in the study. Briefly describe the truncation condition, and explain why those individuals included in the study have left-truncated event times.
54. **Handedness** In a study on the differences in mortality rates of left- and right-handed individuals, Panjer used data from *The Baseball Encyclopedia*, 8th Edition on the dates of birth and death for left- and right-handed professional baseball players.<sup>16</sup> Only those players who had died prior to the 1990 encyclopedia publication date were included in the study. What is the right truncation condition? Briefly explain why those individuals included in the study have right-truncated event times.

The fundamental difference between censoring and truncation mechanisms is that censoring pertains to when subjects *leave* a study (i.e., they drop out of the study, or the study ends before they experience the event) while truncation refers to when subjects can *enter* a study (e.g., patients who do not survive until the start of a study will not be included; furthermore, these patients will never be known to the investigator). Truncation is a selection process that applies to *all* individuals in the study; censoring applies to particular individuals in the study. Note that it is possible for a study to possess both truncated data and censored data (consider the activity below).

#### Key Concept

Truncation is another mechanism that leads to incomplete survival data. Right truncation is a screening procedure by which only individuals who have experienced the target event are included in the study. Left truncation occurs when only individuals who have experienced a selection condition (different from the target event) are included in the study.

### Extended Activity

#### Censoring and Truncation

55. Identify whether the time-to-event data in the following examples are subject to left, right, interval, or any combination of the three types of censoring schemes discussed, and briefly explain your answers:
- a. After a certain type of brain tumor is surgically removed, doctors are interested in the time it takes until the tumor recurs. A group of adult patients who had the brain tumor removed are examined six months after their operation to determine if the tumor recurred. Upon examination, it was found that some of the patients showed signs of tumor regrowth.
  - b. Consider investigating the lifetime (in hours) of light bulbs—that is, the time until the light bulb burns out. At the beginning of the study, the light bulbs are illuminated, and then they are inspected every 50 hours for a period of 2000 hours.

56. For the following studies, describe the truncation condition and determine whether the study involves left or right truncation:
- Gerontologists investigated the survival rates of elderly residents in a retirement community. Ages at death were recorded, as well as the ages at which individuals entered the retirement community.<sup>17</sup> Describe the truncation condition, and determine whether the study involves left or right truncation.
  - A study on the time to complete a PhD program in statistics at a university used data on students who had received a PhD prior to June 2009. Only data on students who had received a PhD by this date were available to the researchers. Describe the truncation condition, and determine whether the study involves left or right truncation.
57. Describe a possible scenario that includes both censoring *and* truncation mechanisms.
58. Describe a possible study in which particular types of censoring and/or truncation mechanisms *cannot* both occur in the same study. For example, would it be possible to have a study with individuals who are right truncated and have right-censored survival times? Why or why not?



#### **NOTE**

The survival analysis methods we have discussed in this chapter, like the Kaplan-Meier estimator, must be adjusted to accommodate left- and interval-censored data, as well as truncated data. Most survival analysis textbooks discuss methods for fitting various types of incomplete data.

## Chapter Summary

The goal of this chapter was to provide an overview of time-to-event data and expose you to some introductory techniques for exploring survival data. The activities in the first half of the chapter focused on characteristics of survival data, the survival function, and descriptive measures to summarize and describe survival data.

A sample of event times can be summarized using the Kaplan-Meier curve, in conjunction with descriptive measures such as the mean and selected percentiles. The **Kaplan-Meier estimator** of the probability of survival beyond time  $t$  is given by

$$\hat{S}(t)_{\text{KM}} = \prod_{t_i \leq t} \left( 1 - \frac{d_i}{n_i} \right)$$

Other descriptive statistics can be computed for the observed event times, including the mean survival time and percentiles of survival time. If the largest complete event time is identical to the largest observed event time, the estimator of the **mean survival time** is taken to be

$$\hat{\mu} = \sum_{i=0}^{m-1} \hat{S}(t_i)_{\text{KM}} (t_{i+1} - t_i)$$

If the largest event time is censored, the estimator of the mean survival time is

$$\hat{\mu} = \sum_{i=0}^{m-1} \hat{S}(t_i)_{\text{KM}} (t_{i+1} - t_i) + \hat{S}(t_m)_{\text{KM}} (t_n - t_m)$$

To estimate the time at which  $p\%$  of the subjects had yet to experience the target event, the  **$p$ th percentile of survival time** can be calculated as

$$\hat{t}_{(p)} = \text{smallest complete event time } t_i \text{ in the sample such that } \hat{S}(t_i)_{\text{KM}} \leq 1 - \frac{p}{100}$$

To provide a range of possible values for the true survival probabilities, confidence intervals based on the Kaplan-Meier estimates can be constructed for  $S(t)$  at fixed points in time. The  $100(1 - \alpha)\%$  **confidence interval for the survival probability**  $S(t)$  at fixed time  $t$  is given by

$$\hat{S}(t)_{\text{KM}} \pm Z_{\alpha/2} \text{se}(\hat{S}(t)_{\text{KM}})$$

where

$$\text{se}(\hat{S}(t)_{\text{KM}}) = \sqrt{(\hat{S}(t)_{\text{KM}})^2 \left( \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \right)}$$

The survival experiences of two or more groups of subjects can be compared with the **log-rank test** or **Wilcoxon test**. Given sample event times for two groups, the log-rank test statistic for comparing the survival curves for two independent populations is

$$\chi^2 = \frac{\left( \sum_{i=1}^m d_{1i} - \sum_{i=1}^m E_{1i} \right)^2}{\sum_{i=1}^m V_{1i}}$$

The **hazard function** assesses the risk that a subject will experience a target event in the next instant given that the subject has not previously experienced the event. The estimated hazard function, constructed from a sample of event times, is given by

$$\hat{h}(t)_{\text{KM}} = \frac{\hat{p}_i}{t_{i+1} - t_i}$$

where  $\hat{p}_i$  is the estimated conditional probability of experiencing the event of interest in the interval  $[t_i, t_{i+1})$ . The hazard function is useful for examining periods of time when an individual is at high or low risk of experiencing the event of interest.

One drawback of the estimated hazard function is that it can exhibit a very erratic pattern, which makes summarizing the data more difficult. An alternative function that can be used to identify periods of constant risk, as well as periods when risk is increasing slowly or rapidly, is the **cumulative hazard function**, given by

$$\hat{H}(t)_{\text{NA}} = \sum_{t_i \leq t} [\hat{h}(t)_{\text{KM}} \times (t_{i+1} - t_i)]$$

The estimated cumulative hazard function provides the accumulation of estimated hazard up to a particular time.

The equations provided for the techniques and methods discussed in this chapter apply when **right censoring** is present in the survival data (i.e., when the event of interest occurs after the observed event time). We should also be cautious of other mechanisms that give rise to incomplete data, including left and interval censoring and left and right truncation. **Left censoring** occurs when subjects have already experienced the event of interest before the study period begins, while **interval censoring** occurs when the event of interest is known only to have occurred within two given time points. **Left truncation** occurs when subjects enter the sample (or study) only after they have experienced a particular event or condition, while **right truncation** occurs when subjects come under observation only if they have experienced the specific target event. If left- or interval-censored event times or left- or right-truncated event times are present in your data, then it is not appropriate to use the formulas and equations discussed in this chapter.

## Exercises

### E.1. Journal Articles.

The *Journal of the American Medical Association* has published quite a few research articles that investigate survival data. The following are some suggested articles that you might examine:

- T. G. Liou, F. R. Adler, B. C. Cahill, S. C. FitzSimmons, D. Huang, J. R. Hibbs, and B. C. Marshall, “Survival Effect of Lung Transplantation Among Patients with Cystic Fibrosis,” *Journal of the American Medical Association*, 286 (2001): 2683–2689.

- K. Shear, E. Frank, P. R. Houck, and C. F. Reynolds III, “Treatment of Complicated Grief: A Randomized Controlled Trial,” *Journal of the American Medical Association*, 293 (2005): 2601–2608.
- M. S. Sulkowski, R. D. Moore, S. H. Mehta, R. E. Chaisson, and D. L. Thomas, “Hepatitis C and Progression of HIV Disease,” *Journal of the American Medical Association*, 288 (2002): 199–206.

Select an article that implements survival analysis methods, and answer the following:

- a. Provide a brief description of the objective of the survival analysis study.
  - b. Describe the time-to-event variable and define the beginning of time. Discuss whether right censoring is present in the data. If any other types of incomplete data were used, briefly explain. (Other types of incomplete data were discussed in Section 9.9.)
  - c. Describe any survival analysis techniques used in the paper that were covered in this chapter (e.g., the Kaplan-Meier estimator). Also list the names of other techniques used in the study that were not covered.
  - d. Briefly summarize the results and conclusions of the study.
- E.2. Six rats were exposed to carcinogens by injecting tumor cells into their feet. The times to develop a tumor of a given size were observed. The investigator decided to terminate the experiment after 30 weeks. Rats A, B, and D developed tumors after 10, 15, and 25 weeks, respectively. Rats C and E had not developed tumors by the end of the study. Rat F died accidentally without any tumors after 19 weeks of observation.
- a. Describe the time-to-event random variable.
  - b. Which rats had complete event times?
  - c. Which rats had censored event times? What type of censoring occurred? Be as specific as possible.
- E.3. For the studies described below:
- Describe the event of interest, beginning of time, time metric, and time-to-event random variable.
  - State which type(s) of censoring (i.e., left, right, or interval) may be present in each study, and briefly explain your answers.
- a. Survival/sacrifice experiments are designed to determine whether a suspected agent accelerates the time until tumor onset in experimental animals. For such studies, each animal is assigned to a prespecified dose of a suspected carcinogen and then examined at sacrifice or death for the presence or absence of a tumor. Since a lung tumor is occult (detectable only by microscopic examination or chemical analysis), the time until tumor onset is not directly observable. Instead, we observe only a time of sacrifice or death.
  - b. Beadle and colleagues report a study carried out to compare the cosmetic effects of radiotherapy alone versus radiotherapy and chemotherapy on women with early breast cancer.<sup>18</sup> To compare the two treatment regimes, a retrospective study of 46 radiation only and 48 radiation plus chemotherapy patients was made. After treatment, patients were observed initially every 4–6 months. As their recovery progressed, the interval between visits lengthened. At each visit, the clinician recorded a measure of breast retraction on a 3-point scale (none, moderate, severe). Researchers were particularly interested in moderate or severe breast retraction.
  - c. A study was conducted to determine the age at which marijuana was first used among high school boys in California.<sup>19</sup> Researchers asked 191 high school boys the question “When did you first use marijuana?” Possible answers were exact age, “I never used it,” and “I have used it but cannot recall just when the first time was.”
- E.4. Immediately after a heart transplant, patients are randomly assigned to two treatment therapies to improve recovery from the transplant, therapy 1 and therapy 2. The patients are then followed for up to 5 years after their surgery. Define the time-to-event random variable  $T$  as the time (in months) until recovery after a heart transplant. For each of the following study descriptions that involve  $T$ , sketch the graph of the survival curve (or curves) with as much detail as necessary. Please note that Parts A through D are independent of each other.
- a. Therapy 1 is not very effective shortly after surgery, but everybody recovers before the study period is over.
  - b. Therapy 2 is very effective shortly after surgery, but becomes less effective after 3 years. Not every patient fully recovers by the end of the study period.

- c. Two curves on the same plot: Therapy 1 is consistently more effective than therapy 2 over time.
- d. Two curves on the same plot: Therapy 1 is more effective than therapy 2 for the first  $2\frac{1}{2}$  years, and then therapy 2 is more effective than therapy 1 for the remaining duration of the study.

E.5. The times in minutes required for students at a West Coast university to get ready in the morning are

60 12 35 30+ 5 20

where the + denotes a right-censored time for a student who reported that he took at least 30 minutes.

- a. Construct the Kaplan-Meier estimator for  $S(t)$  using the observed times to get ready. Sketch a graph of the curve.
  - b. Using your answer to Part A, estimate the proportion of university students who take longer than 30 minutes to get ready.
  - c. Estimate the mean time to get ready in the morning using an appropriate expression.
- E.6. The Kaplan-Meier curve in Figure 9.17 displays hypothetical estimated survival probabilities of death due to brain cancer, where time (from diagnosis) until death is measured in months.
- a. Is the largest event time censored or complete? How do you know?
  - b. Use the curve to estimate the mean time until death due to brain cancer.

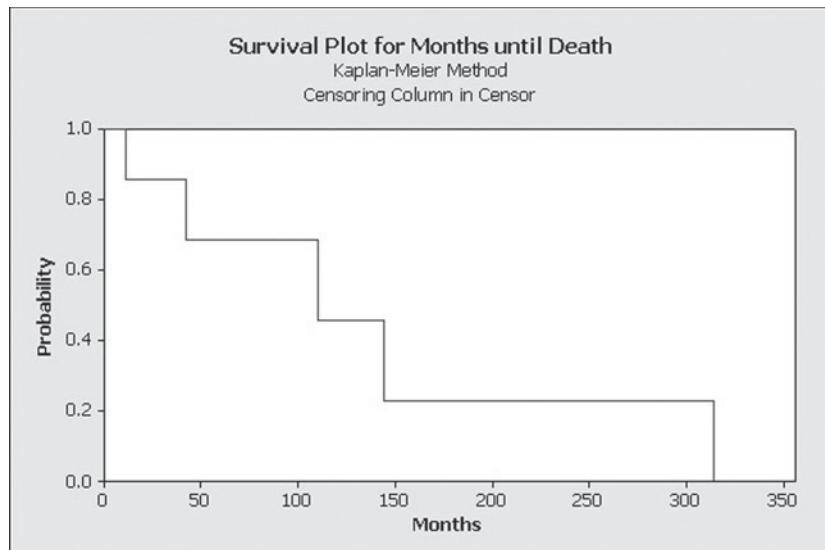


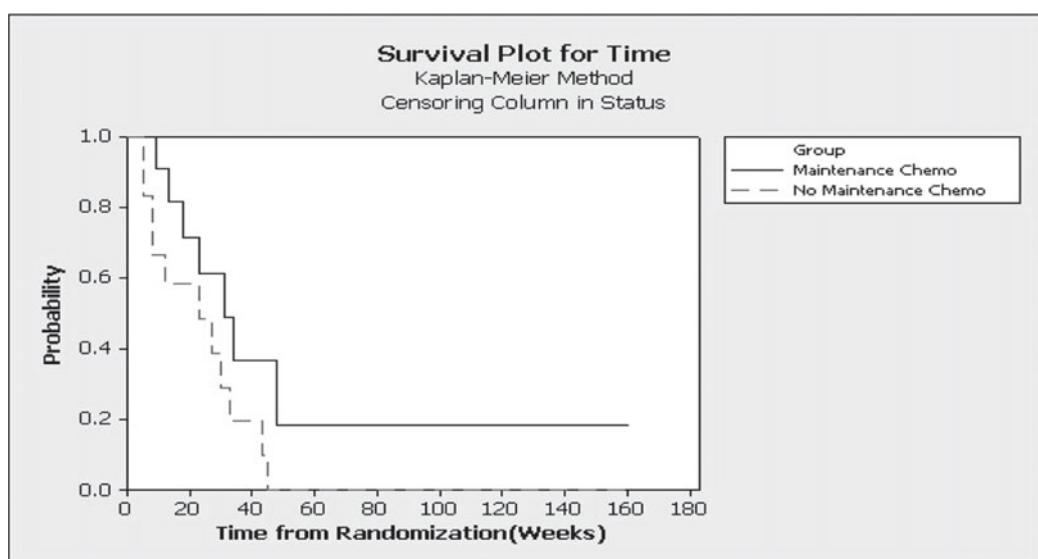
Figure 9.17 Kaplan-Meier curve for survival times after brain cancer diagnosis.

- E.7. Acute myelogenous leukemia (AML) is a cancer that starts inside bone marrow (the soft tissue inside bones that helps form blood cells). The cancer grows from cells that would normally turn into white blood cells. A clinical trial to evaluate the efficacy of maintenance chemotherapy for AML was conducted at Stanford University. After reaching a status of remission through treatment by chemotherapy, the patients who entered the study were randomly assigned to two groups; the first group received maintenance chemotherapy, and the second group did not. The objective of the study was to investigate whether maintenance chemotherapy prolonged time until relapse (measured in weeks from the time of randomization to the two groups). Table 9.7 contains quantities related to the construction of the Kaplan-Meier estimates for survival probabilities of patients with AML who received maintenance chemotherapy.
- a. Fill in the remaining quantities in Table 9.7.
  - b. Examine the table. How many patients in the study who received maintenance chemotherapy had censored event times?
  - c. Estimate the probability that a randomly selected patient who received maintenance chemotherapy remains in remission (does not relapse) for more than 40 weeks.

**Table 9.7** Quantities associated with patients with AML who received maintenance chemotherapy.

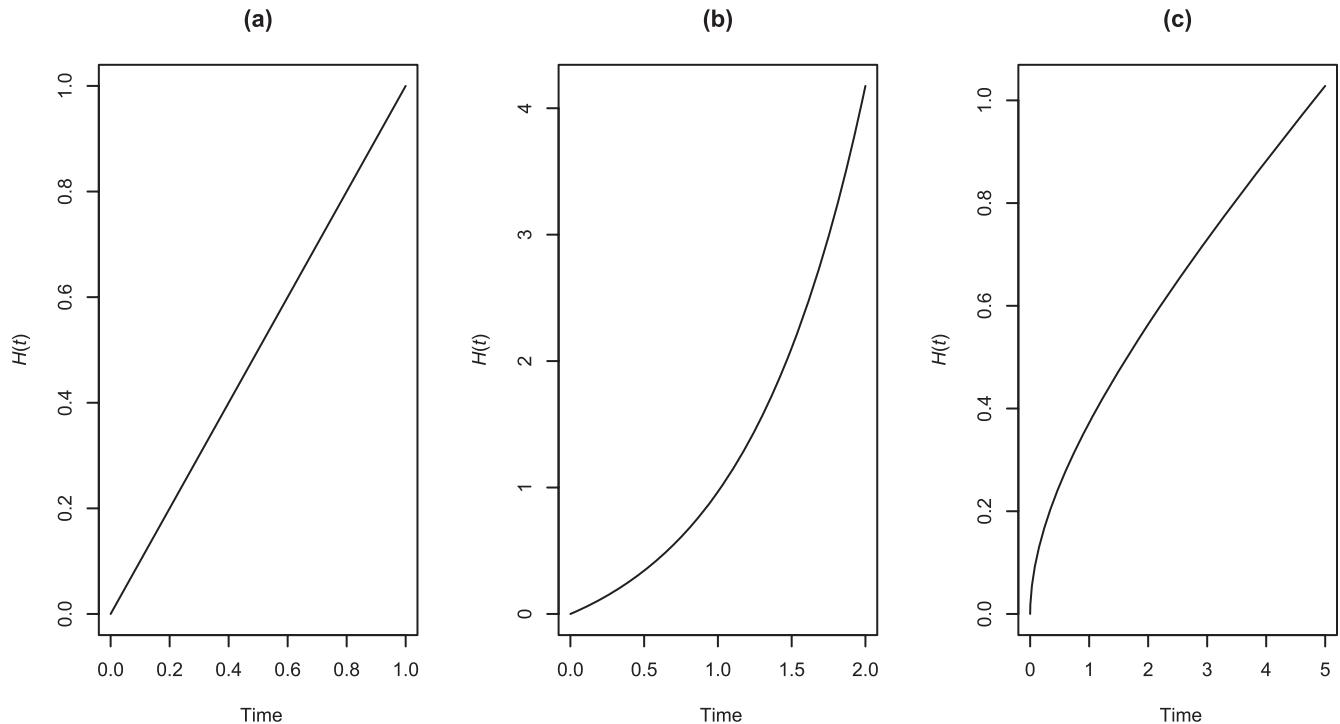
$i$	Interval	$t_i$	$n_i$	$d_i$	$n_i - d_i$	$\hat{p}_i$	$\hat{S}(t)_{KM}$
0	[0, 9)	0	11	0	11	1.00	1.00
1	[9, 13)	9	11	1	10	0.91	0.91
2	[13, 18)	13	10	1	9	0.82	0.82
3	[18, 23)	18	8	1	7	0.71	0.71
4	[23, 31)	23	7	1	6	0.60	0.60
5	[31, 34)	31	5	1	4	0.47	0.47
6	[34, 48)	34	4	1	3	0.38	0.38
7	[48, 161)	48	2	1	1	0.17	0.17

- d. Calculate the estimated mean time until relapse for those patients receiving maintenance chemotherapy. (Be sure to include the time metric in your answer.)
- e. Compute and interpret the estimated hazard rate  $\hat{h}(t)$  for the time interval  $13 \leq t < 18$ .
- f. Figure 9.18 displays the Kaplan-Meier curves for the maintained and nonmaintained groups. Does it appear that maintenance chemotherapy has an effect on time until AML relapse? Briefly explain.
- E.8. Using the information in Table 9.7, answer the following:
- Compute the standard error for the estimate of the probability that a patient receiving maintenance chemotherapy will relapse after 15 weeks.
  - Compute and interpret a 95% confidence interval [based on Equation (9.8)] for the probability that a patient receiving maintenance chemotherapy will relapse after 15 weeks. Note that the critical value is 1.96.
  - Suppose a doctor claims that fewer than 50% of patients receiving maintenance chemotherapy will relapse after 15 weeks. Based on your answer to Part B, how would you respond to this claim?



**Figure 9.18** Kaplan-Meier curves for AML relapse times for patients who did and did not receive maintenance chemotherapy.

- E.9. Sketch hazard functions that would correspond to the following time-to-event random variables. (You may want to do a little background research.)
- Lifetime of an individual measured from birth (don't assume anything about the health or demographics of this person)
  - Time until death after surgery to remove a cancerous tumor
- Be sure to label the time axis, and mark time points appropriately. Briefly explain your reasons for any changes in the shape of the hazard function over time.
- E.10. The graphs displayed in Figure 9.19 are population cumulative hazard functions for three distributions of the time-to-event random variable  $T$ . For each one, sketch a possible corresponding hazard function  $h(t)$ . Be sure to label the same time points on your sketches as are provided on the graphs of  $H(t)$ .



**Figure 9.19** Population cumulative hazard functions for three distributions of  $T$ .

### E.11. Male Fruit Fly Longevity

Data set: `Fruitfly`

One misconception that some students initially have about survival analysis methods is that they can be applied only to survival data that contain some censored observations. While survival analysis methods are appropriate for incomplete data, they are also perfectly acceptable for noncensored survival data. Earlier we noted that the empirical survival function and Kaplan-Meier estimator are identical when there are no censored event times.

The data set `Fruitfly`, introduced by Partridge and Farquhar and further analyzed by Hanley and Hanley and Shapiro,<sup>20</sup> was originally analyzed for the purpose of investigating the relationship between increased sexual activity of male fruitflies and longevity of life (in days) using regression and analysis of covariance techniques. However, survival analysis methods can also be used to study the life durations of male and female fruit flies. Brief descriptions of the variables are provided below:

`Partners`: number of companions (0, 1, or 8)

`Type`: type of companion (0 = newly pregnant female, 1 = virgin female, 9 = not applicable (when `Partners` = 0))

`Longevity`: lifespan, in days (This is the time-to-event variable.)

`Thorax`: length of thorax in mm

**Sleep:** percentage of each day spent sleeping

**Censor:** censoring status (Note that this variable takes only value 1, since the data are all complete. A censoring status variable is necessary for software implementation.)

- a. Construct the Kaplan-Meier curve with a confidence interval for the `Fruitfly` data and describe the survival pattern for the fruitflies over time. Use `Longevity` as the time-to-event variable.
- b. Construct the Kaplan-Meier curves for the lifetimes of the fruitflies by number of partners, using `Partners` as the grouping variable. Briefly comment on the observed relationship between survival and number of female partners.
- c. Perform the log-rank and Wilcoxon tests. Report the test statistics and *p*-values for both tests. State the conclusions for both tests. If the tests yield different conclusions, briefly explain why.

#### E.12. Veterans Administration (VA) Lung Cancer Study

Data set: `Veteran`

The U.S. Veterans Administration lung cancer trial studied 137 males with advanced inoperable lung cancer. The subjects were randomly assigned to either a standard chemotherapy treatment or a test chemotherapy treatment. Several additional variables were also measured on the subjects. The data in the file `Veteran` have been investigated in Kalbfleisch and Prentice.<sup>21</sup> A brief description of the variables is provided below:

`trt`: 1 = standard chemotherapy, 2 = test chemotherapy

`celltype`: 1 = squamous, 2 = smallcell, 3 = adeno, 4 = large

`time`: survival time in days

`status`: censoring status (1 = complete, 0 = censored)

`karno`: Karnofsky performance scale index score (100 = good) (This score is used to quantify cancer patients' functional impairment. The Karnofsky performance score is measured on a 0–100 scale in increments of 10, with 0 indicating that the subject is dead and 100 indicating that the subject shows no signs of the disease.)

`diagtime`: months from diagnosis to randomization

`age`: in years

`prior`: prior therapy (0 = no, 1 = yes)

- a. Create a graph with both Kaplan-Meier curves to compare the survival time (use the variable `time`) for subjects with the standard and the test chemotherapy treatment. What do you observe about the survival probabilities for the two groups of subjects?
- b. Conduct the log-rank test and the Wilcoxon test to compare the survival curves of both treatment groups. Interpret the results.
- c. It may be beneficial to incorporate health as a variable in the analysis. Patients with low Karnofsky scores are less healthy than patients with high Karnofsky scores. Create four groups with the `Veteran` data: `trt` = 1 and Karnofsky score low, `trt` = 1 and Karnofsky score high, `trt` = 2 and Karnofsky score low, and `trt` = 2 and Karnofsky score high. Recall that it is often best to keep sample sizes as equivalent as possible when you determine what is a low or high Karnofsky score. Create a Kaplan-Meier curve for each of the four groups. Conduct the log-rank test and the Wilcoxon test to compare the survival curves of the four groups. (While we have only discussed using these tests to compare two groups, they can easily be extended to more than two groups.) Did incorporating health into your analysis impact your conclusions?

#### E.13. Motorist Reaction Times Revisited

Data set: `Hornhonk`

Consider the motorist reaction time data introduced in Section 9.8.

- a. Construct the Kaplan-Meier curve with a confidence interval for the `Hornhonk` data and describe the survival pattern over time.
- b. Estimate the mean and median time until the motorist honked. At what time would you estimate that at least 40% of motorists will honk?

#### E.14. College Graduation Revisited

Data set: Graduate

Consider the college graduation data introduced in Section 9.7.

- Create a plot and compare the survival experiences to investigate whether there are differences in the time required to obtain a bachelor's degree by gender. What do you observe about the survival probabilities of the genders?
- Conduct the log-rank test and the Wilcoxon test to compare the survival curves of the two genders. Interpret the results.
- (R software is suggested for this problem.) Plot and examine the estimator of the cumulative hazard function. Discuss the changes in the rate of change of  $\hat{H}(t)_{NA}$  as each of the following time periods passes, and discuss corresponding changes in the hazard rate of graduation as college students move from one time interval to another.
  - 0–3.75 years
  - 3.75–4.75 years
  - >4.75 years

#### E.15. Time to Rearrest Revisited

Data set: Rearrest

Consider the time to rearrest data introduced in Section 9.8.

- Create a Kaplan-Meier curve with a confidence interval for the time until rearrest variable. Describe any patterns you see.
- Estimate the time at which half the released inmates have been rearrested.
- Do people with property crimes, person crimes, or both have a longer time before rearrest? You will need to create a new variable to answer this question. Use the new variable to create Kaplan-Meier curves as well as conduct the log-rank test and the Wilcoxon test. Interpret the results.
- (R software is suggested for this problem.) Plot the estimated hazard function (Kaplan-Meier type) for all the time-to-event data. (Include this graph in your assignment. You can copy and paste R graphs into Word documents.) What do you observe? (Don't worry about providing an interpretation. Just give your overall impression of the curve.)
- (R software is suggested for this problem.) Plot and examine the Nelson-Aalen estimator of the cumulative hazard function. (Include this graph in your assignment.) Discuss the changes in the rate of change of  $\hat{H}(t)$  as each of the following time periods passes:
  - 0–3.5 months
  - 3.5–8 months
  - 8–11 months
  - 11–20 months
  - >20 months

Also discuss corresponding changes in the estimated hazard as former inmates move from one time interval to another.

## Endnotes

1. "The Future of Data Analysis," *Annals of Mathematical Studies*, 33.1 (1962): 13.
2. See T. A. Stortz and A. G. Marangoni, "Heat Resistant Chocolate," *Trends in Food Science and Technology*, 2011.
3.  $\hat{\text{Var}}(\hat{S}(t)_{KM})$  is commonly referred to as *Greenwood's formula*. See M. Greenwood, "The Natural Duration of Cancer," *Reports on Public Health and Medical Subjects*, 33 (1926): 1–26.
4. For alternative confidence interval expressions with limits inside [0, 1], see D. W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed. (New York: Wiley, 2008).
5. For details on confidence bands for  $S(t)$ , see J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed. (New York: Springer, 2003).

6. Interested readers are referred to textbooks that discuss *contingency table* analysis—for example, W.J. Conover, *Practical Nonparametric Statistics*, 3rd ed. (New York: Wiley, 1999).
7. For details, see R. Latta, “A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data,” *Journal of the American Statistical Association*, 76 (1981): 713–719.
8. If you are interested in the mathematical details of the log-rank test, Wilcoxon test, and additional tests to compare survival experiences, see J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed. (New York: Springer, 2003).
9. For details on the log-rank and Wilcoxon tests for more than two groups, see, for example, D. W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed. (New York: J. Wiley, 2008).
10. Students currently interested in reading more about regression models for survival data can consult D. W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed. (New York: Wiley, 2008).
11. Readers interested in the calculation and derivation of  $h(t)$  for various distributions of  $T$  can consult a more mathematically oriented survival analysis textbook, such as J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed. (New York: Springer, 2003).
12. See A. Diekmann, M. Jungbauer-Gans, H. Krassnig, and S. Lorenz, “Social Status and Aggression: A Field Study Analyzed by Survival Analysis,” *Journal of Social Psychology*, 136 (1996): 761–768.
13. See K. Henning and C. Frueh, “Cognitive Behavioral Treatment of Incarcerated Offenders,” *Criminal Justice and Behavior*, 23 (1996): 523–541.
14. The level of mathematics required for adequate coverage of parametric methods is outside the scope of this book, but for an accessible discussion see M. Tableman and J. S. Kim, *Survival Analysis Using S: Analysis of Time-to-Event Data* (Boca Raton: Chapman & Hall, 2004).
15. See S. L. Kjos, R. K. Peters, A. Xiang, D. Thomas, U. Schaefer, and T. A. Buchanan, “Contraception and the Risk of Type 2 Diabetes Mellitus in Latina Women with Prior Gestational Diabetes Mellitus,” *Journal of the American Medical Association*, 280 (1998): 533–538.
16. See H. Panjer, “Mortality Differences by Handedness: Survival Analysis of a Right Truncated Sample of Baseball Players,” *Transactions of the Society of Actuaries*, 45 (1993): 257–274.
17. Adapted from J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed. (New York: Springer, 2003), p. 16.
18. See G. F. Beadle, S. Come, C. Henderson, B. Silver, and S. A. H. Hellman, “The Effect of Adjuvant Chemotherapy on the Cosmetic Results After Primary Radiation Treatment for Early Stage Breast Cancer,” *International Journal of Radiation Oncology, Biology, and Physics*, 10 (1984): 2131–2137; G. F. Beadle, J. R. Harris, B. Silver, L. Botnick, and S. A. H. Hellman, “Cosmetic Results Following Primary Radiation Therapy for Early Breast Cancer,” *Cancer*, 54 (1984): 2911–2918.
19. See B. W. Turnbull, and L. Weiss, “A Likelihood Ratio Statistic for Testing Goodness of Fit with Randomly Censored Data,” *Biometrics*, 34 (1978): 367–375.
20. L. Partridge and M. Farquhar, “Sexual Activity and the Lifespan of Male Fruitflies,” *Nature*, 294 (1981): 580–581; J. A. Hanley, “Appropriate Uses of Multivariate Analysis,” *Annual Review of Public Health*, 4 (1983): 155–180; and J. A. Hanley and S. H. Shapiro, “Sexual Activity and the Lifespan of Male Fruitflies: A Data set That Gets Attention,” *Journal of Statistics Education*, 2 (1994).
21. J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, 2nd ed. (New York: Wiley, 2002).
22. J. Theios, “Reaction Time Measurements in the Study of Memory Processes,” in H. Bower (ed.), *The Psychology of Learning and Motivation* (New York: Academic Press, 1973), pp. 44–85.
23. R. G. Pachella, “The Interpretation of Reaction Time in Information-Processing Research, in B. H. Kantowitz (ed.), *Human Information Processing—Tutorials in Performance and Cognition* (Hillsdale, NJ: Erlbaum, 1974), pp. 41–82.

# Research Project: Shapesplosion: A Study of Reaction Time

The chapter material and activities have introduced you to some techniques for examining time-to-event data. The following project will give you the opportunity to apply survival analysis methods to some real time-to-event data that you will collect on your own.

## Preparation

Computer lab space is needed to conduct this experiment, which can be done within or outside regularly scheduled class time. As in any experiment involving people, Institutional Review Board approval should be received from your institution. Assign each student subject (or group of student subjects) one student ID and decide what course ID to use. Any alphanumeric code without spaces is acceptable. However, human responses should be confidential and actual names should not be used as student IDs. You may want to search the database to determine if a course ID has already been used.

## Reviewing the Literature

In the paper assigned below, John Stroop tested the reaction time of college undergraduates identifying colors. He found that students took a longer time identifying ink colors when the ink was used to spell a different color. For example, if the word “green” was printed in blue ink, students took longer to identify the blue ink because they automatically read the word “green.” Even though students were told only to identify the ink color, the automatized behavior of reading interfered with the task and slowed their reaction time. Automatized behaviors are behaviors that can be done automatically, without carefully thinking through each step in the process.\* The Stroop effect demonstrates that automatized behaviors can interfere with other desired behaviors.

Another type of reaction time research question that cognitive psychologists are interested in is the speed-accuracy tradeoff. Theios conducted a study in which a digit was shown to a subject and the time it took the subject to name the digit was measured.<sup>22</sup> Theios varied the percentage of times a digit was shown (20% to 80%) and did not find any difference in the reaction time as the percentage changed. Pachella repeated the study by Theios; however, he measured both reaction time and accuracy. Pachella found that even though the reaction time stayed the same, the subject’s accuracy changed dramatically as the percentage varied.<sup>23</sup> In the following project, you will have the opportunity to develop your own reaction time experiment using the Shapesplosion game. Shapesplosion is a popular game in which a person is expected to place specifically shaped pegs into the appropriate holes within a short time period.

1. Read the paper by J. Stroop, “Studies of Interference in Serial Verbal Reactions,” *Journal of Experimental Psychology*, 12 (1935): 643–662. This paper can be found online at <http://psychclassics.yorku.ca/Stroop> or through other library resources such as <http://www.apa.org/psycinfo>. Focus primarily on the first two experiments. If there are any words that you do not understand, look them up and provide a short definition for each. Identify or answer each of the following for the second experiment, and be prepared to submit your answers as well as discuss this material in class.
  - a. Objective of the experiment
  - b. Any relevant background (from journals that were referenced)
  - c. Response variable and explanatory variables that were used

---

\*Note that many psychologists would call this procedural knowledge instead of automatized behavior. Both are processes that can be done without conscious thought, but automatized behaviors are processes that cannot be slowed down, do not decline with age, and show no gender differences.

- d. Variables that were held constant during the experiment
- e. Nuisance factors (i.e., factors that are not of interest but may influence the results)
- f. How many trials were run for each experiment?

## Designing the Study and Collecting the Data

Play the Shapesplosion game and develop a study involving an appropriate research design and using survival analysis methods. Play several games. Try a variety of conditions. Click the “Data Request Form” to view your results. Which factors do you believe will have the most significant effect on reaction time? For example, will games with distracting color take longer to solve than games with corresponding colors?

Develop your own experiment. Using any of the timed options will provide censored data. Submit a short proposal suggesting an experiment that can be done with these games. Note that these games allow you to develop three new explanatory variables of your own choice.

2. Clearly define a problem and state the objectives of your experiment:
  - a. Describe the event of interest.
  - b. Select at least two different conditions to compare. For example, “Shape Only,” “Color Only,” “Color and Shape,” and “Shape but No Color” are four possible conditions that could be compared.
  - c. Define the time-to-event random variable,  $T$ .
  - d. Define a clear beginning of time.
  - e. Provide a meaningful scale for measuring time for this study.
3. Identify what other conditions need to be controlled during the experiment to eliminate potential biases. Identify how measurements, material, and process may involve unwanted variability. What conditions would be considered normal for this type of experiment? Are these conditions controllable? If this condition changed during the experiment, how might it impact the results? Explain how these conditions will be controlled throughout the experiment, even if they are simply held constant. Will subjects be allowed a practice game before the actual experiment? How important is it to randomize the order in which the games are played?
4. Choose an experimental design.
  - a. Keep the design and analysis as simple as possible. A straightforward design is usually better than a complex design. If the design is too complicated and the data are not collected properly, even the most advanced statistical techniques may not be able to draw appropriate conclusions from your experiment.
  - b. How many trials will be run? Can you completely randomize all the trials, or do you need to account for timing, subject variability, and other nuisance factors?
5. Explain how your experimental design builds on previous research. Identify relevant background, such as theoretical relationships, expert knowledge/experience, or previous studies.
6. Discuss designs and decide on an experiment to be tested.
  - a. Prepare any questions you would like to ask cognitive psychologists or statisticians before you finalize your experimental design.
  - b. Write specific lab procedures that you will use while conducting the experiment. Determine who will collect the data at what time, how you will randomize the trials, how the data will be recorded, and exactly what will be measured.
  - c. Ensure that your group has received appropriate Institutional Review Board approval.
7. Conduct the study.
  - a. Meet with your professor to discuss your experimental design and potential analysis.
  - b. Collect data. While conducting the study, did you identify any additional sources of variability that could impact the results?

## Exploring the Data

Once you have collected your time-to-event data, address the following.

8. Create a Kaplan-Meier survival curve corresponding to each of the conditions tested. Comment on any similarities and differences you observe. Are there any groups whose survival curves are notably lower than those of other groups? What does this suggest about the reaction times under the different matching conditions?
9. To examine the risks of game completion under the different conditions, plot and comment on the estimated hazard functions for the different groups.
10. Conduct a formal log-rank and/or Wilcoxon test to determine if the survival experiences of the individuals in the various conditions are significantly different.
11. Based on the analyses you've conducted, is there any conclusive evidence of a Stroop effect (or other effect) in your experiment?

## Presenting Your Results

Write a research paper. (See “How to Write a Scientific Paper or Poster” on the accompanying CD.) Bring three copies of your research paper to class. Submit one to your instructor. The other two will be randomly assigned to other students in your class to review. Use the “How to Write a Scientific Paper or Poster” checklist to review each other’s papers and provide comments.

## Final Revision

Make final revisions to the research paper. Then submit the first draft, other students’ comments and checklists, the data set you used (in electronic format) along with descriptions of the variables in the data set, and your final paper.

## Other Project Ideas

The research articles, activities, and exercises in this chapter provide many project ideas. In addition, you might consider investigating the following questions with survival analysis techniques:

- Is there a difference between male and female students in their patience in a lunch line? Do males or females comment first? Do males or females budge first?
- Do hard or soft cheeses tend to get moldy more quickly?
- Does Roundup or another garden product keep out weeds longer?
- Do different brands of garden vegetable seeds germinate (or give rise to fruit) earlier than others?
- Are first-year or upper-class students more likely to finish an in-class assignment early?
- Does the type of class (e.g., major versus general education course) influence whether students enter the classroom late?

# Principal Component Analysis: Stock Market Values

*The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.*

—John Tukey<sup>1</sup>

Advancements in technology have made data collection and computationally intensive statistical techniques much more feasible. Researchers from many disciplines are now routinely using automated calculations to analyze data sets with dozens or even hundreds of variables. Principal component analysis (PCA) is an exploratory tool used to simplify a large and complex data set into a smaller, more easily understandable data set. It is different from ANOVA or multiple regression techniques that focus on statistical inference or model building to explain or predict relationships among variables. PCA summarizes complex data sets by creating new variables that are linear combinations (weighted sums) of the original data. In addition, each of these new variables, called a principal component, is constructed to be uncorrelated with all others. In this chapter, we will do the following:

- Create graphs to visualize principal components and compare principal components with an original data set
- Describe how each principal component is uncorrelated with the others
- Calculate and interpret eigenvalues and eigenvectors
- Show the process of creating principal components
- Determine how many principal components should be used
- Incorporate the use of principal component analysis into other statistical techniques
- Describe how more advanced mathematical calculations using matrix algebra could be used to calculate principal components

## 10.1 Investigation: Can a Single Variable Explain Patterns in the Stock Market?

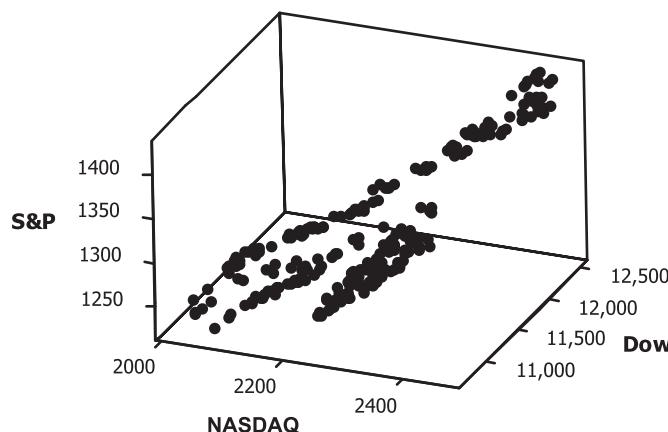
What do stock prices and climate change have in common? Superficially, the two seem unrelated, but if you take a more abstract, statistician's-eye view, you'll find that two important common features stand out. First, both the stock market and climate change involve time series—stock prices and environmental temperatures are numerical measures that change as time passes. This feature might be important in a different context, but here it is largely coincidental. Second, and more important for our purposes, both stock prices and temperatures lend themselves to multiple indirect measurements. Stock prices are measured by the sales records of individual stocks (Google, GM, etc.) and by indices like the Dow (Dow Jones Industrial Average), the S&P (Standard and Poor's 500), and the NASDAQ (National Association of Securities Dealers Automated Quotations). For the study of climate change, temperatures can be measured directly at any of hundreds of weather stations around the world. Historical temperatures can also be measured indirectly through methods involving ice cores, tree rings, and coral growth patterns.

For both situations—market and climate—we have a set of many measurements that we can regard as trying to tell us about a single underlying true, albeit intangible, number representing either asset value or global temperature. Large data sets (i.e., data sets with a large number of variables) can be difficult to interpret, and it is likely that several of the variables are correlated. In this chapter, we will discuss using principal component analysis to draw important information from large data sets. **Principal component analysis (PCA)** uses the variability (i.e., spread) of each variable and the correlation between variables to create a simplified set of new variables (components), which consist of uncorrelated linear combinations of the original variables. The goal of PCA is to reduce the number of variables in a data set and still account for as much of the total variation in the original data as possible.

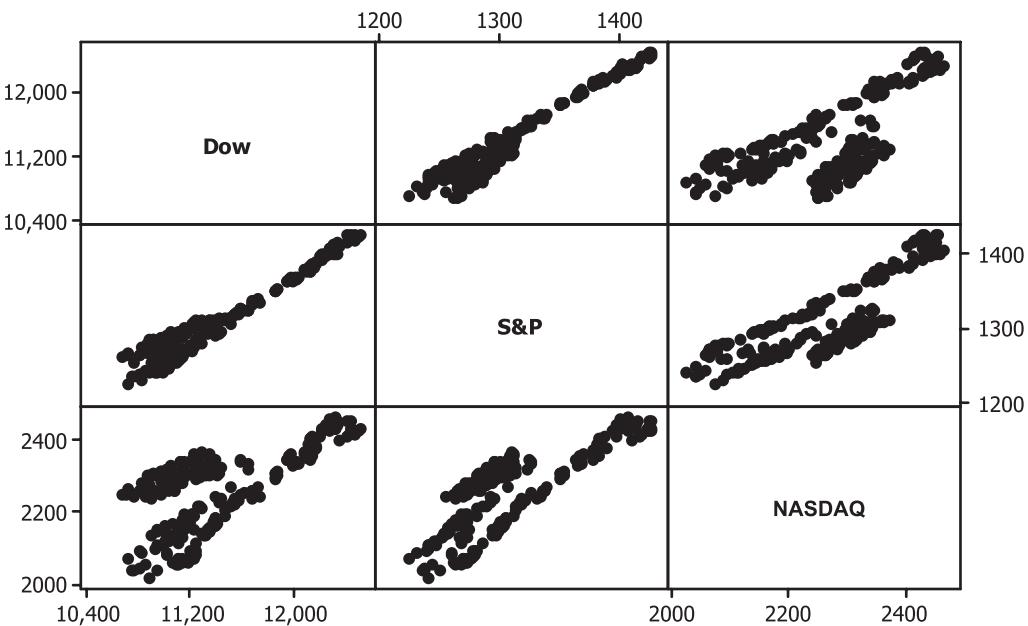
In this chapter, we will introduce PCA by combining stock market indices into a single overall stock market variable. In the corresponding project, we will transform several indirect temperature variables into a single principal component that can be used to measure change in global surface temperatures. The stock market example in the following sections contains only a few variables. Low-dimensional data (i.e., data with only a few variables) typically would not need to be simplified. However, this small data set was specifically chosen to emphasize conceptual understanding of PCA.

## 10.2 A Visual Interpretation of PCA Scatterplots

Three of the most closely watched indices tracking targeted stock market activity are the Dow, S&P, and NASDAQ. Each of the 2006 daily stock market values of these indices is shown in the three-dimensional scatterplot in Figure 10.1. When there are multiple variables of interest, analysts often create matrix plots, as shown in Figure 10.2, to assess the pairwise relationships between variables.



**Figure 10.1** Three-dimensional scatterplot of the 2006 Dow, S&P, and NASDAQ closing values.



**Figure 10.2** Matrix plot of the 2006 Dow, S&P, and NASDAQ closing values.

→ **NOTE** →

Given a set of  $k$  variables, a **matrix plot** is the set of two-dimensional plots for each pair of variables. This set of plots is shown in a matrix format where each column contains the same  $x$ -axis and each row the same  $y$ -axis. For example, in Figure 10.2 each scatterplot in the first row has the Dow variable as the  $y$ -axis. Similarly, each scatterplot in the first column has the Dow variable as the  $x$ -axis. While matrices of scatterplots, often called **scatterplot matrices**, are the most common, other plots can also be made. Matrix plots are useful for visualizing outliers, clusters, or other pairwise relationships between variables.

Each of the two-dimensional scatterplots shown in Figure 10.2 can be thought of as a projection (or a shadow) of the multivariate data onto a two-dimensional space. In other words, the graph in Figure 10.1 can be rotated to look like each of the two-dimensional scatterplots. While the graphs shown in Figures 10.1 and 10.2 are useful, there may be another rotation of Figure 10.1 that would more clearly show patterns within the data. This rotation will correspond to the first principal component.

Figures 10.1 and 10.2 show that the stock market indices are highly correlated. Since the variables share a similar pattern, a single variable may be able to encapsulate most of the information contained in these three variables. Principal component analysis creates a linear combination that represents a rotation of the original data onto the new axis that best emphasizes patterns in the data.

Only one principal component is needed in the stock market example to summarize most of the information contained in the original data. However, in many studies more than one principal component is useful. Later sections will show that if there are  $k$  variables in a data set, it is possible to create up to  $k$  principal components. As the number of variables (i.e., the number of dimensions) increases, it becomes more and more challenging to ensure that the graphs are rotated in a way that best allows the researcher to visualize meaningful features within the data. Fortunately, with computer software, the mathematical calculations described in this investigation can easily be extended to multiple studies with multiple variables.

**Key Concept**

Principal component analysis is used to create linear combinations of the original data that summarize much of the information contained in the original data set. Each linear combination represents a rotation onto a new plane that best reveals patterns or structures within the data.

## Time Series Plots

The stock market data set also contains dates for each business day in 2006. The following activity uses time series plots to visually compare two stock market variables, the Dow and the S&P, to the first principal component. Later sections will demonstrate how to calculate principal components.

### Activity ▶ Visualizing the Data

1. Open the stock market data file called `2006Stocks`. Create time series plots of the Dow and S&P. Both time series should be on the same graph. This graph helps to demonstrate that when variables are on very different scales (such as the Dow and the S&P 500), the first step in simplifying the data is often to standardize each variable.
2. Standardize the Dow column (i.e., for each element in the Dow column, subtract the Dow mean and divide by the Dow standard deviation). Save the standardized Dow values in a new column labeled  $\mathbf{z}_1$ . Repeat this process for the S&P 500 column, and store the standardized data in  $\mathbf{z}_2$ . You may choose to use different labels, but for the remainder of this chapter,  $\mathbf{z}_1$  and  $\mathbf{z}_2$  will refer to the standardized Dow and S&P, respectively.
  - a. Create time series plots of the standardized Dow and S&P,  $\mathbf{z}_1$  and  $\mathbf{z}_2$ . Both time series should be on the same graph.
  - b. Do you see similar patterns in  $\mathbf{z}_1$  and  $\mathbf{z}_2$  (i.e., are the variables correlated)? Describe why you would or would not expect stock market indices to be correlated.
  - c. Explain why the time series plot using  $\mathbf{z}_1$  and  $\mathbf{z}_2$  is more useful for comparing patterns in the stock market than the time series plot created in Question 1.

#### NOTE

The bold notation for  $\mathbf{z}_1$  and  $\mathbf{z}_2$  is used to show that these are not just single values—they are entire columns (i.e.,  $\mathbf{z}_1$  is a vector with 251 elements).

3. In later sections, we will show that the first principal component based only on the Dow and S&P columns is calculated as  $\mathbf{PC1} = \mathbf{y}_1 = 0.707\mathbf{z}_1 + 0.707\mathbf{z}_2$ .
  - a. Use software to calculate  $\mathbf{PC1}$  and submit a time series plot of  $\mathbf{z}_1$ ,  $\mathbf{z}_2$ , and  $\mathbf{PC1}$ . All three series should be on one plot.
  - b. Describe the relationship between  $\mathbf{PC1}$  and the standardized stock market data.

Figure 10.3 provides a time series plot similar to the one calculated in Question 3, except Figure 10.3 also includes a third standardized stock market index corresponding to the 2006 daily NASDAQ ( $\mathbf{z}_3$ ) values. Notice that the first principal component based on these three variables provides similar information (shows a similar pattern) to each of the original variables. Instead of all three variables, one term  $\mathbf{PC1} = \mathbf{y}_1 = 0.582\mathbf{z}_1 + 0.608\mathbf{z}_2 + 0.540\mathbf{z}_3$ , can be used as a simplified overall measure of patterns in the 2006 stock market values.

Notice in Figure 10.3 that  $\mathbf{PC1}$  is not simply an average of the three stock market variables—the average value would be a measure of center for all three indices, while  $\mathbf{PC1}$  emphasizes the key patterns in the data. For example, when all three stock values are increasing,  $\mathbf{PC1}$  increases at a faster rate than the other values. Similarly, when all three values are simultaneously decreasing,  $\mathbf{PC1}$  decreases more quickly than the other terms.

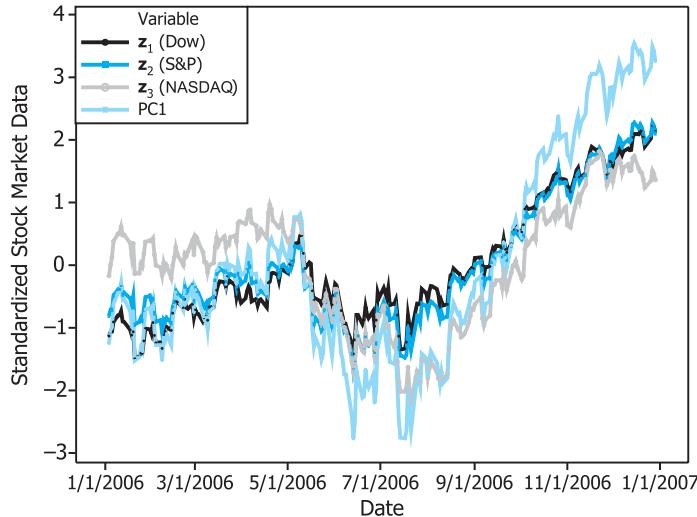
The pattern in the stock market example is quite easy to see without  $\mathbf{PC1}$ , since there are only three variables and they are highly correlated. When there are more variables, principal components are much more useful. The following activity shows that various linear combinations can be created to emphasize different characteristics of the data.

### Activity ▶ Interpreting Linear Combinations

4. Each of the following linear combinations creates a new variable  $\mathbf{c}_i$ . For each equation listed below, create one time series plot similar to Figure 10.3 that includes the standardized Dow, S&P, and

NASDAQ closing values ( $\mathbf{z}_1$ ,  $\mathbf{z}_2$ ,  $\mathbf{z}_3$ ) and  $\mathbf{c}_i$ . Then explain what characteristic of the data is emphasized by that linear combination.

- a.  $\mathbf{c}_1 = 1\mathbf{z}_1 + 0\mathbf{z}_2 + 0\mathbf{z}_3$
- b.  $\mathbf{c}_2 = 1\mathbf{z}_1 - 1\mathbf{z}_2 + 0\mathbf{z}_3$
- c.  $\mathbf{c}_3 = \frac{1}{3}\mathbf{z}_1 + \frac{1}{3}\mathbf{z}_2 + \frac{1}{3}\mathbf{z}_3$
- d.  $\mathbf{c}_4 = 0\mathbf{z}_1 + 0\mathbf{z}_2 - 1\mathbf{z}_3$



**Figure 10.3** Time series plot of the standardized Dow, S&P, and NASDAQ closing values ( $\mathbf{z}_1$ ,  $\mathbf{z}_2$ , and  $\mathbf{z}_3$ ) and the first principal component PC1.

## 10.3 Calculating Principal Components for Two Variables Vector Notation

When researchers are working with data sets that have many variables, they often use techniques based on matrix algebra. While this chapter does not require a prior knowledge of matrix algebra, we will introduce some terms often used in this type of analysis.

The first five rows of the data file labeled `2006Stocks` are shown in Table 10.1. In this file, there are four variables: Dow, S&P, NASDAQ, and Date. The first column, Dow (also called vector  $\mathbf{x}_1$ ), gives the closing Dow stock market value for every business day in 2006. The second column, vector  $\mathbf{x}_2$ , gives the closing S&P 500 value for every business day in 2006.

To keep the calculations simple, this section will use only the first two columns of data. Thus, throughout this section the first two columns of this data set (not including labels) will be called matrix  $\mathbf{X}$ . Each column will be treated as a vector  $\mathbf{x}_i$  that represents one variable of interest and consists of the  $n$  observations for

**Table 10.1** First five rows of stock market data.

Dow	S&P	NASDAQ	Date
10847.4	1268.80	2243.74	1/3/2006
10880.2	1273.46	2263.46	1/4/2006
10882.2	1273.48	2276.87	1/5/2006
10959.3	1285.45	2305.62	1/6/2006
11011.9	1290.15	2318.69	1/9/2006

that variable. In this example,  $n = 251$  is the number of business days in 2006 and  $k = 2$  is the number of variables in the data set. Matrix  $\mathbf{X}$  consists of 251 rows and 2 columns and thus is called a  $251 \times 2$  matrix.

**NOTE**

A matrix is a rectangular array of numbers; rows of a matrix are row vectors and columns of a matrix are column vectors. In this chapter, each row represents a business day and each column represents a variable of interest. The extended activities provide more details on notation and mathematical calculations involving matrices.

## Creating a Correlation Matrix

The strong linear relationship between the Dow and the S&P in Figure 10.4 shows that the two variables are highly correlated. Often the letter  $r$  is used to represent the sample correlation between two variables. Subscripts are often added to represent the correlation between specific pairs of variables for data sets with more than two variables. For example,  $r_{12}$  is the sample correlation between data vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$ .

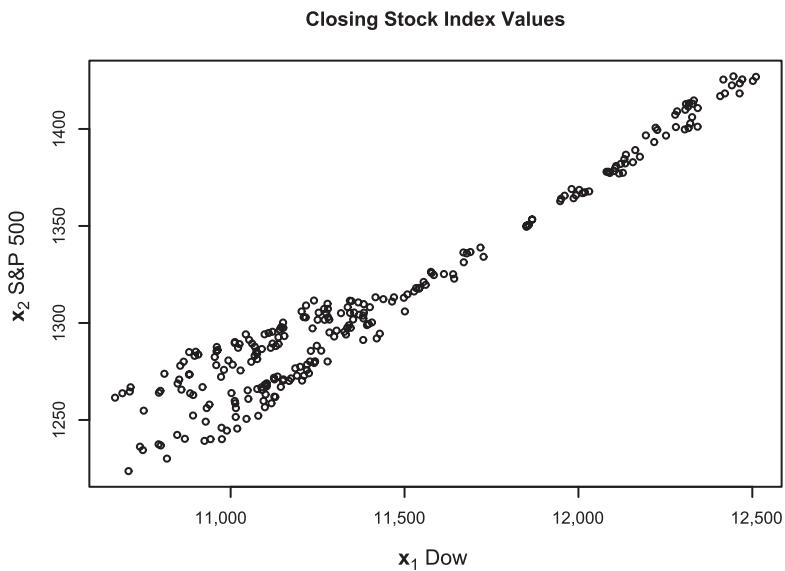


Figure 10.4 2006 daily closing values for the Dow and S&P 500 indices.

**NOTE**

The correlation between two variables, written  $r_{ij}$ , is a measure of the linear relationship between the  $i$ th and the  $j$ th variable. The correlation can be considered a measure of how much variability in one variable can be attributed to the variation of another variable. While studying linear regression, you may have seen that correlation doesn't make a distinction between the explanatory and the response variable. In other words, the correlation between  $\mathbf{x}_1$  and  $\mathbf{x}_2$  equals the correlation between  $\mathbf{x}_2$  and  $\mathbf{x}_1$ :  $r_{12} = r_{21}$ . In addition, the correlation between  $\mathbf{x}_1$  and  $\mathbf{x}_2$  is equal to the correlation between  $\mathbf{z}_1$  and  $\mathbf{z}_2$ , since a linear scale change does not affect the value of the sample correlation. Finally, the scatterplot of  $\mathbf{x}_1$  versus  $\mathbf{x}_1$  will give a perfectly straight line, so  $r_{11} = 1$ . Similarly,  $r_{22} = 1$ .

When you have several variables, it is beneficial to summarize their relationships with a correlation matrix  $\mathbf{R}$ . This matrix is always square, with the number of rows and the number of columns equal to the number of variables in the data set. In addition, this matrix will always have ones on the diagonal and will be symmetric above and below the diagonal.

$$\text{corr}(\mathbf{X}) = \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \quad (10.1)$$

## Activity Calculating the Correlation Matrix

-  5. Calculate the  $2 \times 2$  correlation matrix for the Dow and S&P variables.
-  6. Calculate the  $2 \times 2$  correlation matrix for the Dow and NASDAQ variables.

### Key Concept

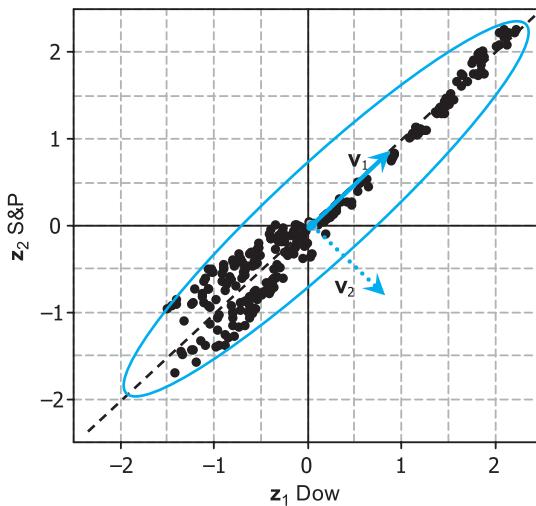
A common first step in exploring a data set with many variables is to check for dependencies among the variables. One way to do this is to calculate a correlation matrix of all variables to identify any strong linear relationships.

The following paragraphs will discuss how a correlation matrix,  $\mathbf{R}$ , can be used to create the single vector that best explains the direction of the data (i.e., the pattern best representing the relationships between all the variables).

## Finding the Direction of Largest Variability

The direction of largest variability (most spread of the data) is easy to visualize in two dimensions. Figure 10.5 is a plot of the standardized Dow and S&P data,  $\mathbf{z}_1$  and  $\mathbf{z}_2$ . An oval has been drawn around the data. The direction of the most variability (where the oval is the most spread out) is shown as a vector,  $\mathbf{v}_1$ .

The vector  $\mathbf{v}_1$  is called the first eigenvector of the correlation matrix,  $\mathbf{R}$ . While it can be difficult to visualize multidimensional data, the first eigenvector of the correlation matrix can always be considered the direction of the most spread in a multidimensional cluster of data points.



**Figure 10.5** Standardized daily closing values for the Dow and S&P 500 indices. The first eigenvector,  $\mathbf{v}_1$ , is drawn from the origin  $(0, 0)$  to the point  $(0.707, 0.707)$  and represents the direction of the most variation in the data.

### MATHEMATICAL NOTE

The definition of an **eigenvector** (sometimes called the characteristic vector) of the correlation matrix,  $\mathbf{R}$ , is any vector  $\mathbf{v}$  that is parallel to  $\mathbf{R}\mathbf{v}$ . Mathematically,  $\mathbf{R}\mathbf{v}$  and  $\mathbf{v}$  are considered parallel if one is a constant multiple of the other. In other words, the process of finding an eigenvector consists of finding any  $\mathbf{v}$  that satisfies the equation  $\mathbf{R}\mathbf{v} = \lambda\mathbf{v}$ , where  $\lambda$  is a constant called an **eigenvalue**. In essence, this means that multiplying the vector  $\mathbf{v}$  by the matrix  $\mathbf{R}$  may stretch or shrink the vector, but the direction of vector  $\mathbf{v}$  is the same as  $\mathbf{R}\mathbf{v}$ . The extended activities describe eigenvalues and eigenvectors in more detail.

Figure 10.5 also displays a second eigenvector,  $\mathbf{v}_2$ , that is drawn from the origin  $(0, 0)$  to the point  $(0.707, -0.707)$ . In any data set, the second eigenvector,  $\mathbf{v}_2$ , is always perpendicular to  $\mathbf{v}_1$  and expresses the direction of the second largest amount of variability.

A technique involving matrix algebra can be used to find the eigenvectors of the correlation matrix,  $\mathbf{R}$ . Computer software is typically used to calculate eigenvectors. However, since this example consists of a very small data set, the next activity allows you to find eigenvectors by hand and by using software (the extended activities provide more advanced mathematical details).

## Activity ▶ Calculating Eigenvectors

Calculating the first eigenvector of the correlation matrix,  $\mathbf{R}$ , involves solving the equation  $\mathbf{R}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ . Using matrix algebra, it is possible to show that

$$\mathbf{R}\mathbf{v}_1 = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} + rv_{12} \\ rv_{11} + v_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 v_{11} \\ \lambda_1 v_{12} \end{bmatrix} = \lambda_1 \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \lambda_1 \mathbf{v}_1 \quad (10.2)$$

7. Equation (10.2) can be written as a system of two equations:

$$\begin{aligned} v_{11} + rv_{12} &= \lambda_1 v_{11} \\ rv_{11} + v_{12} &= \lambda_1 v_{12} \end{aligned}$$

where  $r = r_{12} = 0.97135$  is the correlation between  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . Find the first eigenvector,  $\mathbf{v}_1$ , by adding the two equations and simplifying to show that  $\lambda_1 = 1 + r = 1.97135$ . Then substitute  $\lambda_1 = 1.97135$  into either equation to show that  $v_{11} = v_{12}$ .

8. Find the second eigenvector by solving the equation  $\mathbf{R}\mathbf{v}_2 = \lambda_2\mathbf{v}_2$ . This can also be written as a system of two equations:

$$\begin{aligned} v_{21} + 0.97135v_{22} &= \lambda_2 v_{21} \\ 0.97135v_{21} + v_{22} &= \lambda_2 v_{22} \end{aligned}$$

The second eigenvector is found by subtracting  $0.97135v_{21} + v_{22} = \lambda_2 v_{22}$  from  $v_{21} + 0.97135v_{22} = \lambda_2 v_{21}$  to solve for  $\lambda_2$  and then showing that  $v_{21} = -v_{22}$ .

### ► MATHEMATICAL NOTE ▲

It is possible to show that the same solutions can be found whenever there are only two variables (i.e., the correlation matrix  $\mathbf{R}$  has only two rows and two columns). In other words, when there are only two variables and  $r$  is positive, the equation  $\mathbf{R}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$  will always result in  $v_{11} = v_{12}$ . Similarly,  $\mathbf{R}\mathbf{v}_2 = \lambda_2\mathbf{v}_2$  can be solved to show that  $v_{21} = -v_{22}$ . When there are only two variables and  $r$  is negative, the two solutions will be  $v_{11} = -v_{12}$  and  $v_{21} = v_{22}$ .

9. Use software to calculate the eigenvectors of  $\mathbf{R}$ .

### ► MATHEMATICAL NOTE ▲

As long as none of the  $k$  variables in a data set are perfectly correlated (i.e., no variable is an exact linear combination of the other variables), the correlation matrix can be used to calculate  $k$  uncorrelated eigenvectors and corresponding eigenvalues.

In Questions 7 and 8, the solutions to  $\mathbf{v}_1$  and  $\mathbf{v}_2$  were not unique. The first eigenvector,  $\mathbf{v}_1$ , can be any vector with two elements  $v_{11}$  and  $v_{12}$ , where  $v_{11} = v_{12}$ . Figure 10.5 provides an intuitive explanation. Any vector starting at the point  $(0, 0)$  and going to any point  $(a, a)$  would represent the direction of most variability.

The specific values of each eigenvector are typically chosen by normalizing each vector. A vector is **normalized** when the sum of the squared elements equals one. In this example,

$$\mathbf{v}_1 = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} \sqrt{1/2} \\ \sqrt{1/2} \end{bmatrix}$$

Thus, the sum of the squared elements is  $v_{11}^2 + v_{12}^2 = 0.707^2 + 0.707^2 = 1$ . Similarly,

$$\mathbf{v}_2 = \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix} = \begin{bmatrix} \sqrt{1/2} \\ \sqrt{1/2} \end{bmatrix}$$

and  $v_{21}^2 + v_{22}^2 = 0.707^2 + (-0.707)^2 = 1$ .

The elements of each eigenvector are normalized so that their squared values sum to one. This normalization is done so that the total variability of the principal components is equal to the total variability of the original (or standardized) data.

Notice that the sign of the eigenvectors is arbitrary. In Figure 10.5, the vector  $\mathbf{v}_1$  could start at the origin  $(0, 0)$  and go to  $(0.707, 0.707)$  or  $\mathbf{v}_1$  could start at  $(0, 0)$  and go to  $(-0.707, -0.707)$ . In either case,  $\mathbf{v}_1$  is in the direction of the most variability in the data.

$$\begin{aligned}\mathbf{v}_1 &= \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \quad \text{or} \quad \mathbf{v}_1 = \begin{bmatrix} -0.707 \\ -0.707 \end{bmatrix} \\ \mathbf{v}_2 &= \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix} \quad \text{or} \quad \mathbf{v}_2 = \begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix}\end{aligned}$$

### Key Concept

Eigenvectors  $\mathbf{v}$  of the correlation matrix  $\mathbf{R}$  are found by solving the matrix equation  $\mathbf{R}\mathbf{v} = \lambda\mathbf{v}$ . The first eigenvector points in the direction of most variability within a multivariate cluster of data. The second eigenvector represents the direction of the largest amount of variability in the data set that is perpendicular to the first eigenvector.

## Creating Principal Components

To calculate principal components, multiply the standardized data by each eigenvector using the following formula:

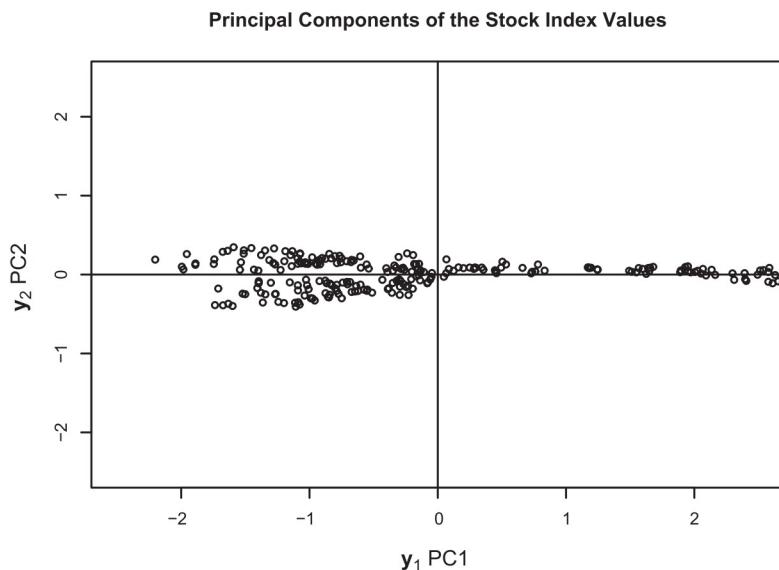
$$\text{PC}_i = \mathbf{y}_i = \mathbf{Z}\mathbf{v}_i = v_{i1}\mathbf{z}_1 + v_{i2}\mathbf{z}_2 + \dots + v_{ik}\mathbf{z}_k \quad \text{for } i = 1, 2, \dots, k \quad (10.3)$$

In our example, there are only two columns of data; thus,  $k = 2$ . This will create two new variables: the first principal component,  $\text{PC1} = \mathbf{y}_1 = v_{11}\mathbf{z}_1 + v_{12}\mathbf{z}_2 = 0.707\mathbf{z}_1 + 0.707\mathbf{z}_2$ , and the second principal component,  $\text{PC2} = \mathbf{y}_2 = v_{21}\mathbf{z}_1 + v_{22}\mathbf{z}_2 = 0.707\mathbf{z}_1 + (-0.707)\mathbf{z}_2$ .

Figure 10.6 is a scatterplot of  $\text{PC1}$  and  $\text{PC2}$ . In essence, principal components provide weights that rotate the data in Figure 10.5 onto a new axis. In other words, Figures 10.5 and 10.6 contain exactly the same data; however, the data have been shifted (rotated) so that the first eigenvector in Figure 10.5 is along the horizontal axis ( $\text{PC1}$ ) in Figure 10.6. Thus, the spread of the data is now represented along  $\text{PC1}$ , the new horizontal axis. The second component,  $\text{PC2}$ , is along the vertical axis of Figure 10.6 (perpendicular to  $\text{PC1}$ ) and explains the rest of the variability in the data. There is a key reason why each of the principal components is perpendicular to each other principal component. Any two variables that are standardized and perpendicular are uncorrelated. Thus, instead of producing a large data set of highly correlated variables, PCA reduces the number of variables that are needed to explain patterns within the data and each principal component is uncorrelated with every other principal component.<sup>2</sup>

### Key Concept

Eigenvectors of the correlation matrix are used to weight the original variables in order to rotate the data onto new axes.

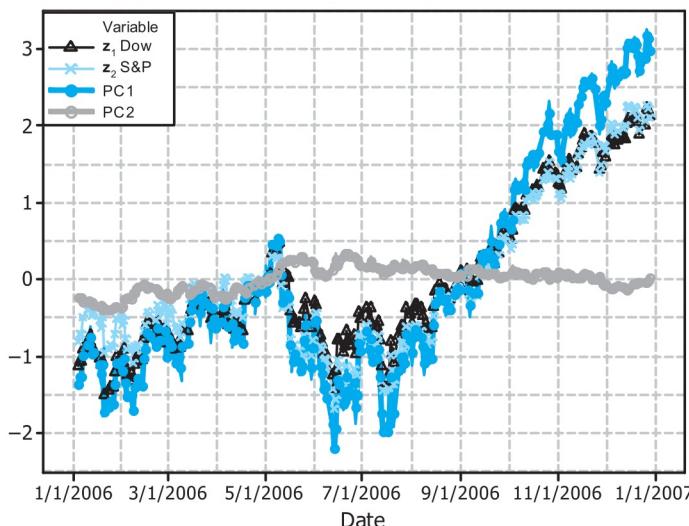


**Figure 10.6** Principal components of the daily closing values for the Dow and S&P 500 indices.

### Activity Calculating Principal Components

10. Use software to calculate the principal components  $\text{PC1} = \mathbf{y}_1 = v_{11}\mathbf{z}_1 + v_{12}\mathbf{z}_2$  and  $\text{PC2} = \mathbf{y}_2 = v_{21}\mathbf{z}_1 + v_{22}\mathbf{z}_2$ .
  - a. Create a scatterplot of  $\text{PC1}$  versus  $\text{PC2}$ . What is the correlation between  $\text{PC1}$  and  $\text{PC2}$ ?
  - b. Calculate the variance of  $\text{PC1}$  and  $\text{PC2}$  and explain how the variances of the principal components are related to the eigenvalues,  $\lambda_1$  and  $\lambda_2$ .
11. Create a time series plot of  $\mathbf{z}_1$ ,  $\mathbf{z}_2$ ,  $\text{PC1}$ , and  $\text{PC2}$ . All four series should be on one plot. Explain how  $\text{PC1}$  and  $\text{PC2}$  are related to  $\mathbf{z}_1$  and  $\mathbf{z}_2$ .

Your time series plot in Question 11 should look like Figure 10.7. Notice in Figure 10.7 that  $\text{PC1}$  follows a similar pattern to  $\mathbf{z}_1$  and  $\mathbf{z}_2$ .  $\text{PC1}$  also increases at a faster rate when both  $\mathbf{z}_1$  and  $\mathbf{z}_2$  are increasing. In



**Figure 10.7** Principal components of the daily closing values for the Dow and S&P 500 indices.

addition, when  $\text{PC1}$  increases at a faster rate than  $\mathbf{z}_1$  and  $\mathbf{z}_2$ ,  $\text{PC2}$  somewhat compensates by decreasing at those times.

## 10.4 Understanding Eigenvalues

In addition to calculating eigenvectors for the correlation matrix, Questions 7 and 8 also involved calculating constant values  $\lambda_1$  and  $\lambda_2$ , called eigenvalues. Corresponding to each eigenvector  $\mathbf{v}_i$  for  $i = 1, 2, \dots, k$  is an eigenvalue,  $\lambda_i$ . As discussed in Question 10, the eigenvalue  $\lambda_i$  is equal to the variance of the corresponding principal component. In other words,  $\lambda_1 = \text{Var}(\text{PC1})$  and  $\lambda_2 = \text{Var}(\text{PC2})$ .

Principal components are ordered according to their variances (the size of their eigenvalues). The eigenvector corresponding to the largest eigenvalue is considered the first eigenvector,  $\mathbf{v}_1$ . The second largest eigenvalue determines the second eigenvector,  $\mathbf{v}_2$ . If there are more variables in the data set, this process continues for all  $k$  variables.

Thus,  $\text{PC1} = \mathbf{y}_1$  is always the linear combination of the variables that encapsulates most of the variability. In other words, the first principal component represents a rotation of the data along the axis representing the largest spread in the multidimensional cluster of data points. The second principal component is the linear combination that explains the most of the remaining variability while being uncorrelated with (i.e., perpendicular to) the first principal component. If there were a third principal component, it would explain most of the remaining variability while being uncorrelated with the first two principal components. This pattern continues for all consecutive principal components.

### Key Concept

Principal components are ordered according to their variances. Thus, the principal component with the largest corresponding eigenvalue is called the first principal component,  $\text{PC1}$ . The second principal component is uncorrelated with  $\text{PC1}$  and has the second largest variance. This process continues for all principal components calculated from the data set of interest.

### NOTE

Recall that the eigenvalues were normalized so that the sum of the variances of all the principal components would be equal to the sum of the variances of the original standardized data. In other words, the sum of the variances of the principal components,  $\text{Var}(\text{PC1}) + \text{Var}(\text{PC2}) = \lambda_1 + \lambda_2$ , is equal to the total variance of the original standarized variables,  $\text{Var}(\mathbf{z}_1) + \text{Var}(\mathbf{z}_2) = 1 + 1$ .

The proportion of the total variation that is explained by the first principal component can be found by dividing the first eigenvalue by the sum of all eigenvalues (only two in this example). In this example, the percentage of variation explained by the first principal component is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.9714}{1.9714 + 0.0286} = 0.986 \quad (10.4)$$

Similarly, we can show that only 1.4% of the variation is explained by the second principal component. Since  $\text{PC1}$  explains over 98% of the variation in this example, there is no need to use  $\text{PC2}$ . Thus,  $\text{PC1}$  can be used to reduce the two-dimensional data set to just one dimension.

## 10.5 A Three-Dimensional Example

The file `2006Stocks` contains the daily closing stock prices of the Dow, S&P 500, and NASDAQ financial indices for all business days in 2006. In this data set,  $n = 251$  represents the number of business days in 2006 and  $k = 3$  represents the three variables of interest. In this section, we briefly conduct a principal component analysis on three variables in order to show how this technique can be extended to more than just

two variables. The same steps will be used as in the previous section; however, the sizes of the matrices and number of vectors will increase.

Step 1: Calculate the correlation matrix,  $\mathbf{R}$ , by finding the correlation between each pair of variables. For example, the correlation between the Dow and the NASDAQ is  $r_{13} = 0.683$ .

$$\text{corr}(\mathbf{X}) = \mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.971 & 0.683 \\ 0.971 & 1 & 0.809 \\ 0.683 & 0.809 & 1 \end{bmatrix} \quad (10.5)$$

The correlation matrix shows that all three variables are positively correlated. The strongest linear relationship is between the Dow and the S&P; the Dow and the NASDAQ have the weakest linear relationship. Figures 10.1 and 10.2 agree with the information found with the correlation matrix.

Step 2: Calculate the  $k = 3$  eigenvectors and eigenvalues of  $\mathbf{R}$ . The first eigenvector provides a vector in the direction of the largest variability in the cluster of three-dimensional data points. Computers are used to find each eigenvector ( $\mathbf{v}_i$ ) and each eigenvalue ( $\lambda_i$ ) by solving the following equation:

$$\mathbf{R}\mathbf{v}_i = \begin{bmatrix} 1 & 0.971 & 0.683 \\ 0.971 & 1 & 0.809 \\ 0.683 & 0.809 & 1 \end{bmatrix} \begin{bmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \end{bmatrix} = \lambda_i \begin{bmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \end{bmatrix} = \lambda_i \mathbf{v}_i \quad (10.6)$$

The order of the eigenvectors is identified by their corresponding eigenvalues. The first eigenvector has the largest eigenvalue, the second eigenvector has the second largest eigenvalue, and so on.

Step 3: Use the eigenvectors to transform the standardized data into principal components. These principal components represent a rotation onto the new axis that best summarizes the information in the original data. The first principal component in this example is

$$\begin{aligned} \text{PC1} &= \mathbf{Z}\mathbf{v}_1 \\ &= v_{11}\mathbf{z}_1 + v_{12}\mathbf{z}_2 + v_{13}\mathbf{z}_3 \\ &= 0.582\mathbf{z}_1 + 0.608\mathbf{z}_2 + 0.540\mathbf{z}_3 \end{aligned} \quad (10.7)$$

where  $\mathbf{Z}$  represents a  $251 \times 3$  matrix containing vectors  $\mathbf{z}_1$ ,  $\mathbf{z}_2$ , and  $\mathbf{z}_3$ , the standardized Dow, S&P, and NASDAQ values, respectively.

## Activity Calculating Principal Components for Three Dimensions

12. Standardize the NASDAQ column into a new column,  $\mathbf{z}_3$ , and conduct a principal component analysis on all three variables ( $\mathbf{z}_1$ ,  $\mathbf{z}_2$ , and  $\mathbf{z}_3$ ).
  - a. Calculate the eigenvectors and eigenvalues.
  - b. Calculate (but do not submit) the three principal components.
13. Create a three-dimensional plot or matrix plot of the standardized data. Draw the appropriate eigenvectors onto the graph by hand or with computer software. Convince yourself that the eigenvectors are always perpendicular (and that the principal components are uncorrelated).
14. Create a time series plot with  $\mathbf{z}_1$ ,  $\mathbf{z}_2$ ,  $\mathbf{z}_3$ ,  $\text{PC1}$ ,  $\text{PC2}$ , and  $\text{PC3}$ . Remember that it might be more appropriate to plot a negative principal component (e.g.,  $-\text{PC1}$  instead of  $\text{PC1}$ ), since the sign of each eigenvector is arbitrary. Submit this graph and explain how the time series patterns of  $\text{PC1}$ ,  $\text{PC2}$ , and  $\text{PC3}$  relate to  $\mathbf{z}_1$ ,  $\mathbf{z}_2$ , and  $\mathbf{z}_3$ .
15. What percentage of the variability is explained by the first principal component? What percentage of the variability is explained by the first two principal components combined?

### CAUTION

It is important to recognize that PCA is scale sensitive. Notice that the Dow has a much larger magnitude and range than the S&P and NASDAQ. The extended activities show that if the data were not standardized before PCA was conducted (i.e., if the covariance matrix was used instead of the correlation matrix), the Dow would have much more influence than the other variables.

## 10.6 What Can We Conclude from the Stock Market Investigation?

The statistical analysis described in this chapter is a little different from those in other chapters. We did not conduct an experiment or take a simple random sample from a larger population. In this study, we had several variables that contained all observations for the entire year (the entire population). PCA tends to be more reliable when the number of observations is large (greater than 100). A general guideline is that the number of observations should be at least five times the number of variables.

The goal of PCA is to re-express a large and complex data set so that only the first few variables (dimensions) account for as much of the variability as possible. In addition to reducing the number of variables, principal component analysis creates uncorrelated variables.

In this investigation, three variables were combined into one overall stock market value. While working with three variables is fairly manageable, the eigenvalues tell us that using just one principal component instead of all three variables still explains 88.3% of the variability. Instead of looking at three variables, we are now able to use just one variable to understand key patterns in the stock market.

In this example, the first principal component assigns nearly equal weight to each stock market index. Thus, the first principal component can be thought of roughly as an overall average standardized stock market index value. Caution should be used in interpreting the principal components. If PCA was repeated on a similar new data set, the coefficients (i.e., the weights in each linear combination) would change. The extended activities discuss how to interpret principal components and how to determine the number of principal components to retain. PCA is most effective if just a few linear combinations of a large data set explain most of the variability.

### A Closer Look Statistical Models

## 10.7 The Impact of Standardizing Each Variable

In the previous investigation, the correlation matrix was used. However, there are times when it is appropriate to use the covariance matrix instead of the correlation matrix (i.e., the unstandardized data instead of the standardized data). Principal components are very sensitive to differences in scale.

#### MATHEMATICAL NOTE

The variance-covariance matrix of  $\mathbf{X}$ , often called the **covariance matrix**, is a  $k \times k$  symmetric matrix and is typically shown as

$$\text{Var}(\mathbf{X}) = \begin{bmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{bmatrix} \quad (10.8)$$

where the diagonal contains the variance of each of the data vectors (e.g.,  $\text{Var}(\mathbf{x}_1) = s_1^2$ ,  $\text{Var}(\mathbf{x}_2) = s_2^2$ ) and  $s_{ij}$  represents the covariance of vectors for any two vectors  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . Later activities use matrix algebra to show that the correlation matrix is equal to the covariance matrix of the standardized data. In the diagonal of the correlation matrix,  $1 = s_1^2/s_1s_1$  and  $r_{ij} = s_{ij}/s_i s_j$ .

### Extended Activity

#### Using the Covariance Matrix in PCA

Data set: 2006Stock

16. Calculate the three eigenvectors and eigenvalues of the unstandardized 2006Stock data (i.e., use the covariance matrix instead of the correlation matrix).
  - a. How do the eigenvector weights change when the data are not standardized?
  - b. Provide an interpretation of the first eigenvalue. In other words, does the first eigenvalue using the unstandardized data indicate that PC1 will explain more or less of the variability than when the data were standardized?

- c. Use the covariance matrix of  $\mathbf{X}$  to compare the variances of the original data [ $\text{Var}(\mathbf{x}_1)$ ,  $\text{Var}(\mathbf{x}_2)$ , and  $\text{Var}(\mathbf{x}_3)$ ] to the variances of the principal components [ $\text{Var}(\text{PC1})$ ,  $\text{Var}(\text{PC2})$ , and  $\text{Var}(\text{PC3})$ ].
17. Create a time series plot of the original data ( $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ ) and the first principal component (PC1) from Question 16. All four series should be on one plot. Describe the relationship between PC1 and the stock market data.
18. Assume that the NASDAQ values are measured on a different scale, as the NASDAQ is actually a composite measure of over 5000 different stocks. Multiply the NASDAQ values by 5000 and label the new variable NewNASDAQ.
- Conduct PCA using the covariance matrix on the Dow, S&P, and NewNASDAQ. How does changing the scale influence the eigenvalues and eigenvectors?
  - Conduct PCA using the correlation matrix on the Dow, S&P, and NewNASDAQ. How does changing the scale influence the eigenvalues and eigenvectors?

The variables with the largest variances will have the most influence on the principal components. Using the correlation matrix ensures that each variable is centered at zero and has a variance equal to one. The eigenvalues and eigenvectors clearly depend on whether the correlation or the covariance matrix is used.

#### ► MATHEMATICAL NOTE ▼

The dependence of the principal components on the size of the variance can be explained geometrically. Figure 10.11 in Section 10.10 shows that PCA minimizes the distance between each point and the first principal component. The minimum distance is measured from a point to PC1 by a line perpendicular to PC1. If the original points are plotted instead of the standardized points, the perpendicular lines (and thus the sum of squared distances) also change.

#### Key Concept

Principal components are heavily influenced by the magnitude of the variances. If the covariance matrix is used, the principal components will tend to follow the direction of the variables that have the largest variability. If the correlation matrix is used, the principal components represent the strongest patterns of intercorrelation within the data.

There are situations where it may be appropriate to use the covariance matrix. For example, assume several students are given a 20-question survey that attempts to measure their attitudes toward statistics. For each question, students select one of eleven responses (0 = “Strongly Disagree” to 10 = “Strongly Agree”). All questions (all 20 variables) are originally measured on the same scale. However, there may be a few questions that are confusing or unclear and so most students give them a value of 5. The researcher may want these unclear questions to have less influence on the principal components than questions to which students provided stronger responses (i.e., students tended to respond with a 0 or 10). Using the covariance matrix would put less emphasis on the questions that had small variances (i.e., all students answered them similarly) and more emphasis on the questions with more distinct responses (i.e., larger variances). Caution should be used, however, since even small differences in variances can dramatically influence the principal components and the corresponding weightings, and the interpretations are typically very different.

#### Key Concept

The covariance matrix can be used in PCA if all the original variables are measured on a similar scale and there is reason to believe that the magnitude of the variance should influence the interpretation of the data.

## 10.8 Determining the Number of Components to Retain

In data sets with many variables, it is often appropriate to retain more than just one principal component. There is no exact test to determine how many principal components to retain, but there are several general guidelines based on an understanding of the variance in the data.

The variance of each standardized variable is equal to one (these ones are represented in the diagonal of the correlation matrix). Thus, the sum of the variances in the standardized data set is equal to  $k$ , the number of variables in the data. Similarly, in principal component analysis, each eigenvector is normalized so that the sum of all the variances of the principal components equals the sum of all the variances represented in the correlation matrix. In other words, for any data set with  $k$  variables,

$$\begin{aligned} k &= \text{Var}(\mathbf{z}_1) + \text{Var}(\mathbf{z}_2) + \cdots + \text{Var}(\mathbf{z}_k) \\ &= \text{Var}(\text{PC1}) + \text{Var}(\text{PC2}) + \cdots + \text{Var}(\text{PC}_k) \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_k \end{aligned} \quad (10.9)$$

where  $\mathbf{z}_i$  for  $i = 1, 2, \dots, k$  represents a standardized variable and  $\text{PC}_i$  represents the  $i$ th principal component. When each variable is standardized, the percentage of the variation explained by the  $i$ th principal component can be found by dividing the  $i$ th eigenvalue by  $k$ :

$$\frac{\lambda_i}{k} \quad \text{for } i = 1, 2, \dots, k \quad (10.10)$$

### MATHEMATICAL NOTE

While it is typically advised to standardize the data, even without standardization it can be shown that  $\sum_{i=1}^k \text{Var}(\mathbf{x}_i) = \sum_{i=1}^k \text{Var}(\text{PC}_i)$ . In other words, when the covariance matrix is used in PCA instead of the correlation matrix, the sum of the variances of the variables is equal to the sum of the variances of the principal components calculated with eigenvectors.

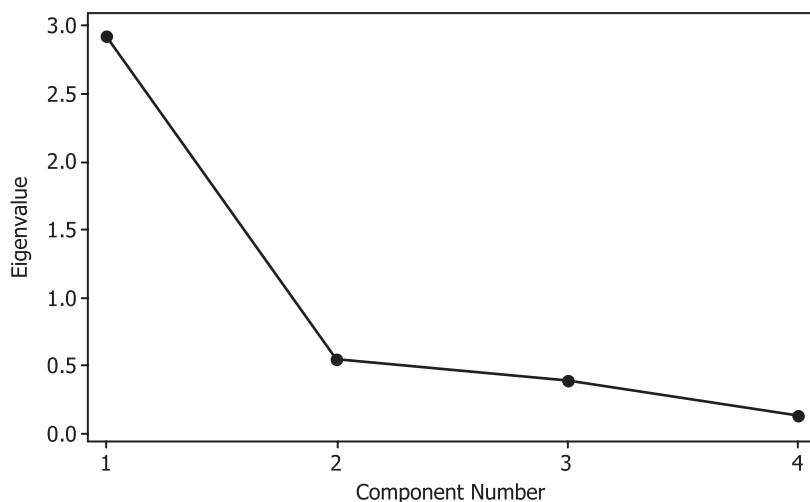
### Extended Activity

#### Fisher's Iris Data

Data set: `Versicolor`

19. Ronald Fisher published an analysis on the sizes of iris sepals and petals.<sup>3</sup> The data, collected over several years by Edgar Anderson, were used to show that these measurements could be employed to differentiate between species of irises. The data set `Versicolor` contains four measurements (sepal length, sepal width, petal length, and petal width) on 50 versicolor irises. The petals are colorful and flashy, while the sepal (under the petals) tends to be less colorful. Conduct a principal component analysis on the `Versicolor` data. Be sure to use the correlation matrix instead of the covariance matrix.
  - a. List the four eigenvalues. How many eigenvalues are greater than one? Kaiser recommended retaining the principal components with corresponding eigenvalues greater than one.<sup>4</sup> However, this should be a general guideline, not an absolute rule.
  - b. What is the percentage of the variation explained by the first principal component?
  - c. What is the percentage of the variation explained by the first two principal components combined?
  - d. A **scree plot** is a simple graphic used to display the eigenvalues for each successive principal component.<sup>5</sup> In this graph, the number of each principal component is plotted on the horizontal axis and the corresponding cumulative eigenvalue is plotted on the vertical axis. Use the software instructions to create a scree plot for the data.

Since the variance of each standardized variable is equal to one, Kaiser's greater than one rule retains all components that account for more variability than any individual variable and ignores (removes) any component accounting for less variance than an individual variable. The general rule for evaluating scree plots is to look for a change in slope. The principal components corresponding to a steep curve at the beginning of the plot should be kept. Components where the slope appears flat are not retained. In Figure 10.8, it is clear that



**Figure 10.8** Scree plot for the *Versicolor* iris data. There is a clear change in slope after the second component, suggesting that only one component is needed.

the first eigenvalue corresponds to a very large proportion of the overall variance:  $2.926/4 = 0.732$ . Thus, only the first principal component should be used. Not all scree plots are this clear. Often the scree plot looks more like a smooth curve, and the determination of how many components to include is more challenging. If there is more than one clear bend in the scree plot, keep all components to the left of the last big bend before the eigenvalues begin to level off.

Other researchers suggest that the number of components should be determined by the cumulative percent variation. However, the acceptable percentage of variation that is explained depends on the context of the problem. Only 80% may be needed for descriptive purposes, but a larger percentage may be needed if further analysis is to be done.

#### CAUTION

Eigenvalues very close to zero should not automatically be ignored. When eigenvalues are equal to zero, one or more variables are redundant (perfect linear combinations of other variables). Rounding error may make the eigenvalues very small instead of zero. Ignoring redundant variables may cause problems in interpreting future analysis.

## 10.9 Interpreting Principal Components

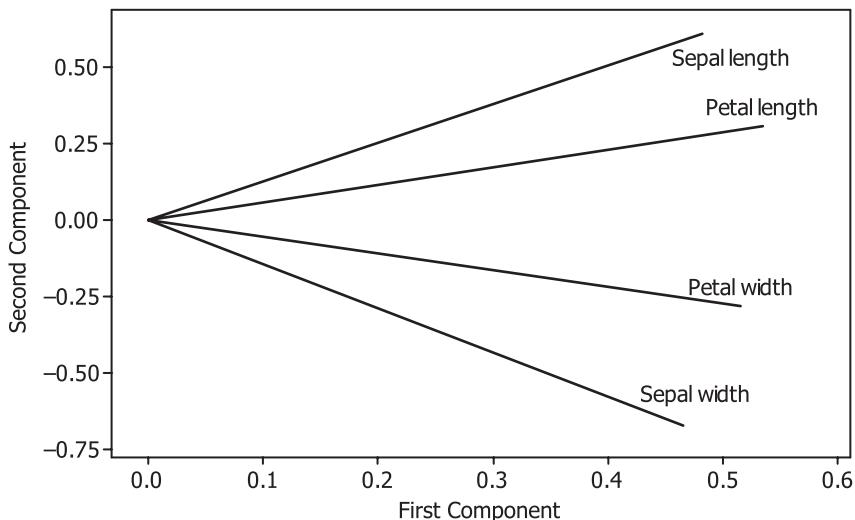
The four eigenvectors for the correlation matrix for the *Versicolor* iris data set are given in Table 10.2.

**Table 10.2** Eigenvectors for iris data.

Variable	v <sub>1</sub>	v <sub>2</sub>	v <sub>3</sub>	v <sub>4</sub>
Sepal length	0.482	0.611	-0.491	0.392
Sepal width	0.465	-0.673	-0.540	-0.199
Petal length	0.535	0.307	-0.340	-0.710
Petal width	0.515	-0.283	0.593	0.550

Principal component loading plots are often calculated to better understand how principal components are characterized by the original variables. Figure 10.9 provides a loading plot for the first two components. This plot is used to visualize the first and second eigenvectors. It is simply a scatterplot of the first two eigenvectors,

with lines drawn from the origin to each point on the scatterplot. The  $x$ -axis represents the weights for PC1, and the  $y$ -axis represents the weights for PC2. For example, the line representing sepal width starts at the point  $(0, 0)$  and ends at  $(0.465, -0.673)$ , corresponding to the weights provided in the first two eigenvalues. Figure 10.9 shows that all weights (all elements of the eigenvector) on the first principal component are positive, while PC2 has positive weights for lengths but negative weights for widths.



**Figure 10.9** Loading plot for the first two principal components for the *Versicolor* iris data.

The first eigenvector shows roughly equivalent coefficients. Thus, the first principal component can be considered an overall size measurement. When the iris has larger than average sepal and petal values, the first principal component will also be larger than average. When all the measurements are smaller than average, the first principal component will be smaller than average. If an iris has some original measurements larger than average and some smaller (or if all the iris's measurements are near average), the first principal component will be close to average. The first eigenvector shows that this one principal component accounts for over 73% of the variability in the data.

Some researchers may choose to also retain PC2. The second eigenvector also has a nice intuitive interpretation. The second principal component will be large when the iris is longer than average but has shorter widths. In addition, the sepal has more influence than the petal on the value of PC2.

## Extended Activity

### Fisher's Iris Data Again

Data set: *Versicolor*

20. Interpret the third principal component in Fisher's iris data. What percentage of the variation in the original data is explained by this component?
21. With software or by hand, create a loading plot of the second and third principal components (eigenvectors). Write a brief interpretation of the plot.

The loading plot shown in Figure 10.9 is based on plotting values of the eigenvectors. **Factor loadings** (also called unrotated factor loadings or component loadings in PCA) are the correlations between the original data and the principal components. The factor loadings for the first two principal components are shown in Table 10.3.

Factor loadings show similar patterns to the eigenvectors, except they have the added benefit that they can be easily interpreted as correlations. The squared correlation (similar to  $R^2$  in regression) represents the percent of variance in that variable explained by the principal component.

**Table 10.3** Factor loadings for iris data.

Variable	PC1	PC2
Sepal length	$0.826 = \text{corr}(\text{PC1}, \mathbf{z}_1)$	$0.451 = \text{corr}(\text{PC2}, \mathbf{z}_1)$
Sepal width	$0.795 = \text{corr}(\text{PC1}, \mathbf{z}_2)$	$-0.497 = \text{corr}(\text{PC2}, \mathbf{z}_2)$
Petal length	$0.914 = \text{corr}(\text{PC1}, \mathbf{z}_3)$	$0.227 = \text{corr}(\text{PC2}, \mathbf{z}_3)$
Petal width	$0.882 = \text{corr}(\text{PC1}, \mathbf{z}_4)$	$-0.209 = \text{corr}(\text{PC2}, \mathbf{z}_4)$

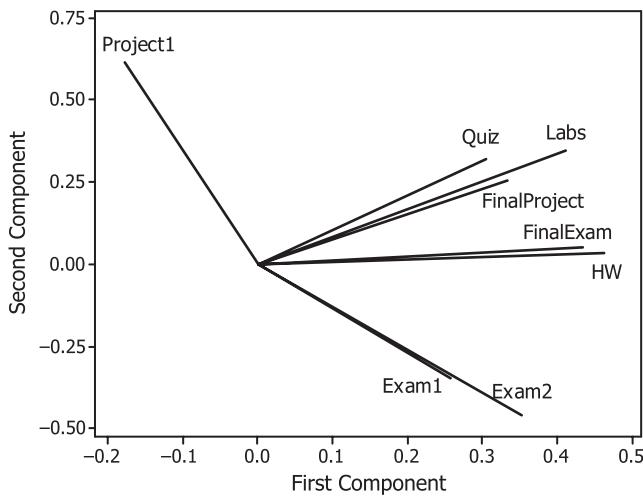
PCA and **factor analysis** are very similar. PCA attempts to summarize variability within a data set with a smaller set of uncorrelated variables. Factor analysis assumes that there is an underlying structure to the data—in other words, there are variables that can't be directly measured but are influencing the variables in the data set. While some researchers treat these methods as interchangeable, PCA is useful for data reduction, while factor analysis is useful in detecting structure within a data set.<sup>6</sup>

## Extended Activity

### Understanding Course Grades

Data set: Grades

There are several ways to measure a student's understanding of statistics in an introductory course. The data set *Grades* contains total scores for students' homework, labs, quizzes, projects, and exams. This course had two 1-hour exams and a 3-hour final exam. In addition to a final project worth 20% of the final grade, there was an initial small project worth only 5% of the final grade. The exams were fairly standard statistics exams, but both projects required students to read peer-reviewed research projects, collect their own data, conduct an appropriate statistical analysis, and present their results. While several students in this introductory class had seen a little statistics in previous courses, this was typically the first time students were asked to read journal articles or conduct a realistic research project. Figure 10.10 provides a loading plot created from PCA on the *Grades* data. In both the initial stock market investigation and Fisher's iris data, the first principal component could be considered some type of overall average. However, the first principal component of the *Grades* data has a slightly different interpretation. Conduct a principal component analysis on the *Grades* data.



**Figure 10.10** Loading plot for the first two principal components for the *Grades* data.

22. Which variables have large positive loadings on PC1? Write a general interpretation of the first principal component using the first eigenvalue and the loading plot.
23. Which variables have large positive or large negative loadings on PC2? Write a general interpretation of the second principal component.

Figure 10.10 shows that PC1 has positive weights (loadings) for all variables except Project1. PC2 has a very high loading for Project1 and strong negative loadings for Exam1 and Exam2.

While interpretation of principal components is subjective, we see that HW (homework), FinalExam, and Labs seem to have very high loadings. Thus, it seems reasonable to consider the first principal component as representing an overall effort level in the course. In this class, students were expected to complete daily homework and weekly labs, even though these two variables were not weighted highly for the overall course grade. In addition, students who did well on the final exam typically worked through suggested, but ungraded, sample questions.

The second principal component could be thought of as a measure of initial comfort with research. The exams tended to focus on standard textbook-type problems, which clearly contrast with the open-ended nature of research problems.

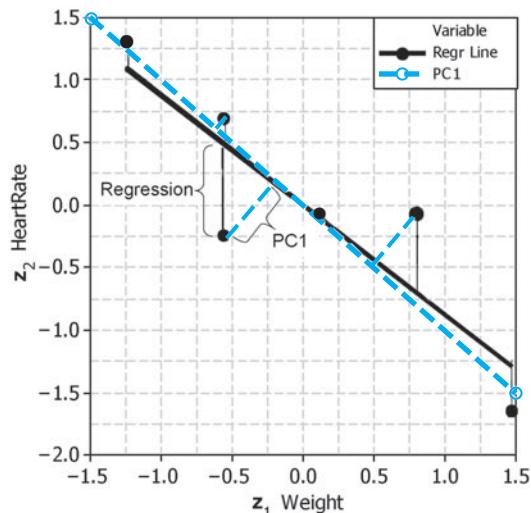
Together, the first two principal components represent 66.3% of the total variability. One or two additional principal components should be included to capture most of the variability within the data. In addition, this data set has a small number of observations, so the results are likely to be somewhat unreliable.

## 10.10 Comparing Regression and Principal Components

A small data set based on tarantula weights and heart rates is shown in Figure 10.11 to emphasize the difference between regression and PCA. Carrel and Heathcoat collected data on the weight and resting heart rate of several species of spiders.<sup>7</sup> They evaluated the influence of the body size of spiders on their heart rate.

Figure 10.11 is a scatterplot of the data with a linear regression line and a line in the direction of the first eigenvector. Both lines minimize the sum of squared distances.

In regression, residuals measure the vertical distance between each observed value and a corresponding expected value, while in PCA distance is measured as the shortest distance between a point and a line. Algebra can be used to show that the shortest distance between a point and a line is always perpendicular to the line.



**Figure 10.11** Comparing regression and PC1 for standardized tarantula weights and heart rates. The least squares regression line is  $\text{HeartRate} = 0.000 - 0.876(\text{Weight})$ . The first principal component corresponds to a vector that goes from the point  $(0, 0)$  to  $(0.707, -0.707)$ .

### Key Concept

PCA focuses on finding the line closest to all the points in the data set—even if there are multiple dimensions. If the two-dimensional data set in Figure 10.11 could only be seen in one dimension (a single line), the most informative line (the line closest to all points) would be along the direction of the first eigenvector,  $\mathbf{v}_1$ .

## 10.11 Incorporating Principal Components into Other Statistical Methods\*

You may have worked with a large number of variables before, such as in conducting multiple regression analysis. Unlike multiple regression, PCA has no response variable. Principal component analysis is often a first step in simplifying a more complex data set in order to make further analysis easier. PCA is beneficial in multiple regression analysis because each component is uncorrelated with the others, so principal components can be used as explanatory variables (eliminating any problems with multicollinearity).

### NOTE

Chapter 3 describes that **multicollinearity** exists when two or more explanatory variables in a multiple regression model are highly correlated with each other. If two explanatory variables are highly correlated, it can be very difficult to identify which variables are actually responsible for influencing the response variable.

## Extended Activity



### Using Principal Components in Regression

Data set: Cars

24. The data set `Cars` contains the make, model, equipment, mileage, and Kelley Blue Book suggested retail price of several used 2005 General Motor cars. Kelley Blue Book (<http://www.kbb.com>) has been an accurate resource for pricing cars for over 80 years and was used as a source to collect these data. In this activity, you will create a regression model that will describe the association of several explanatory variables (car characteristics) with the retail value of a car.
- In the `Cars` data set, liters and cylinders are highly correlated, as they both are a measure of engine size. Instead of using either the liter or the cylinder variable, create the first principal component obtained from a PCA of just these two variables. Use this first principal component, plus mileage, Buick, Cadillac, Chevrolet, Pontiac, and SAAB, in a regression analysis to predict the natural log of retail price, `LnPrice`.
  - Run a regression analysis with liter, cylinder, mileage, Buick, Cadillac, Chevrolet, Pontiac, and SAAB to predict the natural log of retail price, `LnPrice`.
  - In Part A, using the first principal component from the PCA of liter and cylinder eliminated multicollinearity, but how did it impact the  $R^2$  value? Did using `PC1` instead of both liter and cylinder cause you to miss a key explanatory variable? Which model would you suggest is better?

### CAUTION

While most of the variability is usually explained by the first few principal components, principal components with smaller eigenvalues may also be important. For example, a few principal components are often calculated and then used in a regression analysis. However, researchers should be aware that a meaningful linear combination or single explanatory variable may be omitted that would be highly correlated with the response variable.

### Key Concept

PCA is often used to avoid multicollinearity among several explanatory variables in multiple regression.

## 10.12 Calculating Eigenvectors and Eigenvalues Using Matrix Algebra†

This section provides detailed mathematical calculations of principal component analysis. The supplemental material provides a brief review of matrix calculations, including the calculation of variance-covariance and correlation matrices.

\*Requires multiple regression.

†Matrix algebra required.

A scalar quantity  $\lambda$  is an eigenvalue (also called a characteristic root) of a matrix  $\mathbf{A}$  if there is a nonzero vector  $\mathbf{v}$  that satisfies the equation  $\mathbf{Av} = \lambda\mathbf{v}$ . Such a vector  $\mathbf{v}$  is called an eigenvector of  $\mathbf{A}$  corresponding to  $\lambda$ .

For invertible  $k \times k$  matrices, the  $k$  eigenvalues are distinct and the  $k$  eigenvectors are orthogonal, which will make them more applicable (and interpretable) for statistical methods. Covariance and correlation matrices for data collected on  $k$  variables are invertible square matrices as long as each random variable is not a linear combination of the other variables.

## Calculating the Eigenvalues and Eigenvectors for a $2 \times 2$ Matrix

Let  $\mathbf{A}$  be any  $2 \times 2$  matrix  $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ . The two eigenvalues of  $\mathbf{A}$  and their corresponding eigenvectors must satisfy

$$\mathbf{Av} = \lambda\mathbf{v} \quad (10.11)$$

or, equivalently,

$$\mathbf{Av} - \lambda\mathbf{v} = [\mathbf{A} - \mathbf{I}\lambda]\mathbf{v} = \mathbf{0} \quad (10.12)$$

where  $\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$  is the  $2 \times 2$  identity matrix,  $\mathbf{v}$  is any  $2 \times 1$  column vector, and  $\mathbf{0}$  is the  $2 \times 1$  vector of zeros.

To solve this equation for  $\lambda$ , we write out the details:

$$\begin{aligned} &[\mathbf{A} - \mathbf{I}\lambda]\mathbf{v} = \mathbf{0} \\ &\left[ \begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \lambda \right] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ &\begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned} \quad (10.13)$$

This matrix equation can be written as a system of two equations:

$$\begin{aligned} (a - \lambda)v_1 + bv_2 &= 0 \\ cv_1 + (d - \lambda)v_2 &= 0 \end{aligned} \quad (10.14)$$

This system of equations can be solved using techniques similar to those in Questions 7 and 8. However, instead of the equation  $\mathbf{Av} = \lambda\mathbf{v}$ , an equivalent equation using the determinant  $|\mathbf{A} - \mathbf{I}\lambda| = 0$  is often used to solve for the eigenvalues.

The **determinant** of any  $2 \times 2$  matrix  $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$  is  $ad - bc$ , written  $|\mathbf{A}| = ad - bc$ . The solution to  $|\mathbf{A} - \mathbf{I}\lambda| = (a - \lambda)(d - \lambda) - bc = 0$  is given as

$$\lambda = \frac{1}{2}[(a + d) \pm \sqrt{(a - d)^2 + 4bc}] \quad (10.15)$$

Thus, the two eigenvalues for any  $2 \times 2$  matrix  $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$  are

$$\lambda_1 = \frac{1}{2}[(a + d) + \sqrt{(a - d)^2 + 4bc}] \quad \text{and} \quad \lambda_2 = \frac{1}{2}[(a + d) - \sqrt{(a - d)^2 + 4bc}] \quad (10.16)$$

Once the eigenvalues have been found, the corresponding eigenvectors can be calculated. The eigenvectors of  $\mathbf{A}$ ,  $\mathbf{v}_1 = [v_{11} \quad v_{12}]^T$  and  $\mathbf{v}_2 = [v_{21} \quad v_{22}]^T$ , satisfy the equations

$$[\mathbf{A} - \mathbf{I}\lambda_1]\mathbf{v}_1 = \mathbf{0} \quad \text{and} \quad [\mathbf{A} - \mathbf{I}\lambda_2]\mathbf{v}_2 = \mathbf{0} \quad (10.17)$$

Typically, each eigenvector is normalized so that it has length one. That is, we set the eigenvector  $\mathbf{v}_i$  corresponding to  $\lambda_i$  such that  $\mathbf{v}_i^T \mathbf{v}_i = v_{i1}^2 + v_{i2}^2 = 1$ .

## Eigenvalues and Eigenvectors for a $2 \times 2$ Correlation Matrix

For any two variables from a random sample of  $n$  observational units  $\mathbf{x}_1 = [x_{11} \ x_{12} \ \cdots \ x_{1n}]^T$  and  $\mathbf{x}_2 = [x_{21} \ x_{22} \ \cdots \ x_{2n}]^T$ , two standardized vectors can be calculated as  $\mathbf{z}_1 = [z_{11} \ z_{12} \ \cdots \ z_{1n}]^T$  and  $\mathbf{z}_2 = [z_{21} \ z_{22} \ \cdots \ z_{2n}]^T$  where:

$$z_{1j} = \frac{x_{1j} - \bar{x}_1}{s_1} \quad \text{and} \quad z_{2j} = \frac{x_{2j} - \bar{x}_2}{s_2} \quad \text{for } j = 1, 2, \dots, n \quad (10.18)$$

Then it can be shown that

$$\mathbf{R} = \text{cov}(\mathbf{z}_1, \mathbf{z}_2) = \text{corr}(\mathbf{z}_1, \mathbf{z}_2) = \text{corr}(\mathbf{x}_1, \mathbf{x}_2) = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

where cov represents the variance-covariance matrix, corr represents the correlation matrix, and  $r$  is the sample correlation between  $\mathbf{x}_1$  and  $\mathbf{x}_2$ .

### Extended Activity ▶ Calculating Eigenvectors and Eigenvalues

- 25. Show that the eigenvalues for any  $2 \times 2$  correlation matrix are  $\lambda_1 = 1 + r$  and  $\lambda_2 = 1 - r$ .
- 26. Show that the eigenvectors for any  $2 \times 2$  correlation matrix are

$$\mathbf{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T \text{ and, similarly, } \mathbf{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}^T$$

### Calculating Principal Components

Principal components are linear combinations of the standardized data vectors  $\mathbf{z}_1$  and  $\mathbf{z}_2$ :

$$\mathbf{y}_1 = \text{PC1} = \mathbf{v}_1^T \mathbf{z} = v_{11}\mathbf{z}_1 + v_{12}\mathbf{z}_2 \quad \text{and} \quad \mathbf{y}_2 = \text{PC2} = \mathbf{v}_2^T \mathbf{z} = v_{21}\mathbf{z}_1 + v_{22}\mathbf{z}_2 \quad (10.19)$$

These new variables are just linear combinations of the original data, using the eigenvectors as the weights. If the correlation is positive ( $r > 0$ ), the larger eigenvalue is  $\lambda_1 = 1 + r$  and the first principal component is  $\mathbf{y}_1 = (\mathbf{z}_1 + \mathbf{z}_2)/\sqrt{2}$ .

If the correlation is negative, the larger eigenvalue is  $\lambda_2 = 1 - r$  and the first principal component is  $\mathbf{y}_2 = (\mathbf{z}_1 - \mathbf{z}_2)/\sqrt{2}$ .

### Extended Activity ▶ Understanding the Variability of Principal Components

- 27. Show that  $\text{Var}(\mathbf{y}_1) = \lambda_1$ .
- 28. Show that  $\lambda_1 + \lambda_2 + \cdots + \lambda_k = k$ , where  $k$  is the number of variables in the original data set.
- 29. Show that  $\mathbf{y}_1$  is the linear combination of the original standardized data that has the most variance. That is, show that  $\text{Var}(w_1\mathbf{z}_1 + w_2\mathbf{z}_2)$  is maximized (subject to the constraint  $w_1^2 + w_2^2 = 1$ ) by the weights  $w_1 = 1/\sqrt{2} = v_{11}$  and  $w_2 = 1/\sqrt{2} = v_{12}$ .

## Chapter Summary

Principal component analysis is one of many exploratory tools that are useful for providing a general understanding of a large and complex data set. PCA transforms a large number of original variables into a smaller set of uncorrelated components that capture most of the variability in the data. The following steps are carried out in order to calculate principal components for a data set consisting of  $k$  variables.

**Step 1: Calculate the correlation matrix,  $\mathbf{R}$ .** It is important to recognize that PCA is scale sensitive. Unless each variable in a data set was originally collected using the same scale, the data should be

standardized (i.e., the correlation matrix is used). If the data are not standardized (i.e., the covariance matrix is used), the principal components will follow variables with larger variances instead of representing the correlation structure of the entire data set.

**Step 2: Calculate the  $k$  eigenvectors and eigenvalues of  $\mathbf{R}$ .** Computer software is typically used to find eigenvectors ( $\mathbf{v}_i$ ) and their corresponding eigenvalues ( $\lambda_i$ ) by solving the equation  $\mathbf{R}\mathbf{v}_i = \lambda_i\mathbf{v}_i$ , where  $i = 1, \dots, k$ . The first eigenvector points in the direction of most variability within a multivariate cluster of data. The second eigenvector represents the direction of the second largest amount of variability perpendicular to the first eigenvector. This process continues, with each eigenvector perpendicular to all prior eigenvectors, for all  $k$  dimensions.

**Step 3: Use the eigenvectors to create principal components, uncorrelated linear combinations of the standardized data.** The eigenvectors of the correlation matrix are used to transform the original variables through the equation  $\mathbf{PC}_i = \mathbf{Z}\mathbf{v}_i = v_{i1}\mathbf{z}_1 + v_{i2}\mathbf{z}_2 + \dots + v_{ik}\mathbf{z}_k$ . These eigenvectors serve as weights for the first principal component,  $\mathbf{PC}_1$ , to rotate the data onto a new axis representing the direction of the largest amount of variation in the data. Each principal component is ordered by the size of its corresponding eigenvalue,  $\lambda_i = \text{Var}(\mathbf{v}_i)$ . In addition, each principal component is uncorrelated with all others. Thus,  $\mathbf{PC}_1$  explains most of the variability, and  $\mathbf{PC}_2$  represents a perpendicular axis with the second largest amount of variability. This continues for all  $k$  principal components.

Once the principal components have been calculated, we determine how many principal components to retain by considering the following guidelines:

- Retain all components with eigenvalues greater than one.
- Retain enough principal components to explain 70%–90% of the variability within the original data.
- Create a **scree plot** (a graph showing each principal component ordered by the eigenvalue) and retain all variables to the left of sharp bends in the plot.
- Retain components that provide a conceptual understanding of true underlying factors.

Principal component analysis is a nonparametric technique, since no assumptions are made about the distribution of the data. Caution should be exercised when interpreting and using principal components. Each variable in the original data set is typically expected to contain natural variability. A new sample of data from the same population would result in different coefficients (i.e., the linear combinations would change). Principal components are typically created to be used as variables in future data analysis methods, and the technique is most effective if just a few linear combinations of a large data set explain most of the variability.

PCA is typically based on the correlation matrix, and correlations measure only linear relationships. Thus, PCA may miss nonlinear patterns within a data set. PCA is also influenced by outliers. Transformations on the original data can be attempted to address nonlinear patterns, outliers, or skewness; however, caution should be used in interpreting the results. More advanced techniques can also be used to address these issues, but they are beyond the scope of this text.

## Exercises

---

### E.1. 2006 Stocks Again

Data set: 2006Stock2

In addition to the daily closing values, the file 2006Stock2 contains the daily opening, high, and low values.

- a. Conduct principal component analysis on the data set. Can a majority of the variability in the data be explained with just one or two principal components? Explain.
- b. Which, if any, principal component measures the difference between opening and closing values?
- c. Which, if any, principal component measures the overall stock value?
- d. Which, if any, principal component measures the difference between high and closing values?
- e. Which, if any, principal component measures the difference between high and low values?
- f. Create a scatterplot of  $\mathbf{PC}_1$  versus  $\mathbf{PC}_2$  where each point is grouped by stock type. This type of scatterplot is often called a **score plot**. When the first two components explain most of the variance in the data, a score plot can detect clusters, outliers, and trends. What do the clear groupings of points indicate in this plot?

## E.2. Why Not Just Average the Data?

Data set: `Test`

- To summarize the overall pattern within a data set, some may consider it appropriate to simply calculate the average values. The file `Test` contains several standardized variables. Create (but do not submit) a new variable, `Avg`, which is the average of each row in the `Test` data set.
- Create scatterplots of `Avg` versus each of the original variables (or create an appropriate matrix plot).
- Conduct principal component analysis on the data set. Can a majority of the variability in the data be explained with just one or two principal components? Explain.
- Create scatterplots of `PC1` versus each of the original variables (or create an appropriate matrix plot). Does `PC1` or `Avg` better explain patterns in the data? Explain.

## E.3. 2010 Stock Market

Data set: `2010Stock`

The data set `2010Stock` contains 2010 daily closing stock market values for several well-known companies.

- Conduct PCA on the `2010Stock` data set. What percentage of the variation is explained by the first principal component?
- Create a scree plot. How many components should you include? Provide an explanation for your choice.
- Create a loading plot. Which variables have large positive or large negative loadings on `PC1`?
- Which variables have large positive or large negative loadings on `PC2`?
- In 2010, Toyota received a great deal of bad publicity and recalled over 1.5 million vehicles because of a brake problem.<sup>8</sup> How does this information influence your interpretation of the first two principal components?

## E.4. Course Grades

Data set: `Grades`

In an introductory statistics course, the following weights are applied to determine each student's overall grade for the course.

Homework: 10%

Labs: 10%

Project 1: 5%

Final Project: 20%

Exam 1: 15%

Exam 2: 15%

Final Exam: 20%

Quiz: 5%

- Use the assigned weightings to give an overall course score to each of the 32 students in the `Grades` data set. The top 15% of the students should get an A, the next 25% a B, the next 35% a C, the next 15% a D, and the final 10% an F.
- Use the first principal component from a PCA of the `Grades` data instead of the assigned weights to create an overall score for each student in the class. What percentage of the total variability in grade components is explained by this score? Compare the weightings from the PCA to the instructor's weightings. If the first principal component were used instead of the predefined weights, how many students would get a different grade in the course? Explain how changing the weightings influenced students' final grades.
- Repeat Part C, but do not standardize the data (use the covariance matrix instead of the correlation matrix). Explain why homework now has a much higher weight and is the primary variable in a student's grade calculation. (Hint: Which variable has the highest variability in the unstandardized data set?)
- What percentage of the variation in the data is explained by `PC1` in Parts B and C? Do higher eigenvalues indicate a better analysis? Explain.

### E.5. Turtle Shells

Data set: *Turtles*

Jolicoeur and Mosimann measured the length, width, and height of 48 painted turtle shells and conducted PCA on the data. Their work has been influential in the field of allometry, the study of the relative growth of a part of an organism in relation to an entire organism.<sup>9</sup>

- a. Conduct principal component analysis on the entire data set (ignore gender). Can a majority of the variability in the data be explained with just one or two principal components?
- b. Calculate and interpret the principal component loadings.
- c. Create a scatterplot of PC1 versus PC2 where each point is grouped by gender. Was PCA useful in differentiating the genders? Other multivariate techniques are designed to classify or cluster data into clear groups.
- d. Conduct principal component analysis on all three variables using only the male data. Repeat the process using only the female data. Compare the loadings for the two analyses. Would you expect the results to be similar? Why or why not?

### E.6. Intelligence

Data set: *Intelligence*

Measuring intelligence is another area where a principal component analysis is often used. There is not just one test that can be used to accurately measure a person's intelligence—there are several tests that are known to be related to intelligence. Often intelligence tests can be classified into three groups: spatial reasoning, vocabulary, and memory. PCA is used to combine multiple measurements into an overall measure of intelligence. Psychologists call the first component the general intelligence factor, or *g-score*.

As Stephen Jay Gould pointed out, intelligence testing and principal component analysis have been used to offer a “scientific” basis for racist, classist, and sexist theories of “intelligence.”<sup>10</sup> There is little doubt that the U.S. Congressional Immigration Act of 1924 was racially motivated, to limit the entrance of several unwanted groups of people into the United States. Kamin states that intelligence tests were used to verify that “83% of the Jews, 80% of the Hungarians, 79% of the Italians, and 87% of the Russians were ‘feeble-minded’.”<sup>11</sup> Snyderman and Herrnstein question whether or not psychologists had any influence on the act.<sup>12</sup> They claim that intelligence tests were misused as “evidence” of why some people should not be allowed to emigrate to the United States.

- a. The data set *Intelligence* contains test results for 214 people who have each taken several intelligence tests. Use PCA on the *Intelligence* data set to find one overall measure of intelligence. How much of the total variability in the several measures of intelligence is explained by this *g-score*?
- b. Intelligence test results are often standardized into an IQ score so that the mean score is 100 with a standard deviation of 15. Transform your data and find the IQ of the 10th person in the *Intelligence* data set.
- c. Explain why intelligence tests (such as vocabulary tests) may have biased results when measuring the intelligence of newly arriving immigrants on Ellis Island.

#### CAUTION

While vocabulary may be a very appropriate measure of intelligence in some cultures, the same test can be completely invalid if applied to a new population. The ability of PCA to construct one easy-to-understand variable from a large complex data set is powerful and useful, but like almost any powerful tool, PCA can be put to morally repugnant uses. Even the most carefully constructed and complex statistical analysis will reach incorrect conclusions if the data are not collected properly.

- d. Write a 2- to 3-page report to Congress. Describe to the congresspeople, who have not had a statistics class, how a *g-score* is derived. Then explain why, even though the concept of *g-score* is accepted by most psychologists, the test may not accurately describe an overall intelligence measure for all populations.

### E.7. Crime

Data set: Crime

The FBI's Uniform Crime Reporting (UCR) program gathers crime data from more than 17,000 law enforcement agencies throughout the United States. The Crime data set contains crime counts for several years in Iowa. Crimes are listed by type of crime (murder, burglary, larceny, etc.).

- a. Conduct PCA on the crime variables in this data set. What percentage of the variation can be explained by the top two principal components?
- b. How would you interpret these two principal components in terms of crime characteristics?
- c. Some may suggest simply totaling the crime rates, in essence giving a weight of 1 to every unstandardized variable. Give an example where calculating a "Total" value may be more appropriate than using PCA. Give an example where using PCA may be more appropriate than calculating a "Total" value.
- d. Repeat Parts A and B using a rate instead of using the raw numbers of crimes in each category. Explain why either counts or rates would be more appropriate.

### E.8. Corn Suitability Rating

Data set: Corn

Iowa is one of the leading states in corn production. Each year, approximately 12 to 13 million acres of corn are planted in Iowa. To evaluate the overall quality of the soil, an Iowa Corn Suitability Rating (CSR) is developed for each location based on the kind of soil, slope of the land, erosion potential, water infiltration, and ease of machine operation in the area. The CSR can affect the value of an acre of Iowa farmland by thousands of dollars, since it is a measure of the inherent productivity of the land. CSRs are listed by county in the Iowa Soil Properties and Interpretation Database at <http://www.agronext.iastate.edu>.

Use multiple regression and the explanatory variables in the Corn data set to estimate CSR. Note that some of the variables are highly correlated, so PCA may be useful for some subsets of explanatory variables. Also some quadratic terms may be appropriate. For example, the best cropland may have average drainage, not excessive or very poor drainage. Submit a model that "best" estimates CSR while having a relatively small number of terms. Provide a justification for your model.

## Endnotes

---

1. J. Tukey, "Sunset Salvo," *American Statistician*, 40.1 (February 1986): 72–76.
2. For a brief summary of these relationships, see J. L. Rodgers, W. A. Nicewander, and L. Toothaker, "Linearly Independent, Orthogonal, and Uncorrelated Variables," *The American Statistician*, 38.2 (May 1984): 133–134.
3. R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7 (1936): 179–188.
4. H. F. Kaiser, "The Application of Electronic Computers to Factor Analysis," Symposium on the Application of Computers to Psychological Problems, American Psychology Association, 1959.
5. R. B. Cattell, "The Screen Test for the Number of Factors," *Multivariate Behavioral Research*, 1 (1966): 245–276.
6. The factor loadings are often rotated in order to better understand key underlying factor structures. For more details, see R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. (Englewood Cliffs, NJ: Prentice Hall, 2001).
7. J. E. Carrel and R. D. Heathcoat, "Heart Rate in Spiders: Influence of Body Size and Foraging Energetics," *Science*, July 9, 1976, 148–150. They used the resting heart rate of spiders to calculate their metabolic rates and showed that some species of spiders have lower metabolic rates than others. Greenstone and Bennett found contradictory results; see M. H. Greenstone and A. F. Bennett, "Foraging Strategy and Metabolic Rate in Spiders," *Ecology*, 61.5: (1980): 1255–1259.
8. <http://www.cnn.com>, accessed 4/4/11.
9. P. Jolicoeur and J. E. Mosimann, "Size and Shape Variation in the Painted Turtle: A Principal Component Analysis," *Growth*, 24 (1960): 339–354.
10. S. J. Gould, *The Mismeasure of Man* (New York: Norton, 1981).
11. L. J. Kamin, *The Science and Politics of IQ* (Potomac, MD: Erlbavrn, 1974), p. 16.
12. M. Snyderman and R. J. Hermstein, "Intelligence Tests and the Immigration Act of 1924," *American Psychologist*, 38 (1983).

13. Intergovernmental Panel on Climate Change, *Climate Change 2001: Third Assessment Report: The Scientific Basis*, [http://www.grida.no/climate/ipcc\\_tar](http://www.grida.no/climate/ipcc_tar), accessed 1/23/07.
14. M. E. Mann, R. S. Bradley, and M. K. Hughes, "Global-Scale Temperature Patterns and Climate Forcing over the Past Six Centuries," *Nature*, 392 (1998): 779–787.
15. Wegman's testimony is available at <http://republicans.energycommerce.house.gov/108/Hearings/07192006hearing1987/Wegman.pdf>. The Wegman Ad Hoc Committee full report is available at [http://energycommerce.house.gov/108/home/07142006\\_Wegman\\_Report.pdf](http://energycommerce.house.gov/108/home/07142006_Wegman_Report.pdf). This extract is from pages 2–3 of the full report, accessed 3/2/07.
16. M. E. Mann, R. S. Bradley, and M. K. Hughes, "Northern Hemisphere Temperatures During the Past Millennium: Inferences, Uncertainties, and Limitations," *Geophysical Research Letters*, 26.6 (1999): 759–762; M. E. Mann and P. D. Jones, "Global Surface Temperature over the Past Two Millennia," *Geophysical Research Letters*, 30.15 (2003): 1820, doi:10.1029/2003GL017814; Michael E. Mann, S. Rutherford, E. Wahl et al., "Testing the Fidelity of Methods Used in Proxy-Based Reconstructions of Past Climate," *Journal of Climate* 18.20 (2005): 4097–4105; H. Goosse, H. Renssen, A. Timmermann, R. S. Bradley, and M. E. Mann, "Using Paleoclimate Proxy-Data to Select Optimal Realisations in an Ensemble of Simulations of the Climate of the Past Millennium," *Climate Dynamics*, 27 (2006): 165–184.
17. M. E. Mann, R. S. Bradley, and M. K. Hughes, "Global-Scale Temperature Patterns and Climate Forcing over the Past Six Centuries," *Nature*, 392 (1998): 779.
18. Intergovernmental Panel on Climate Change, *Climate Change 2001: Third Assessment Report*, [http://www.grida.no/climate/ipcc\\_tar](http://www.grida.no/climate/ipcc_tar).
19. S. McIntyre and R. McKittrick, "Corrections to the Mann et al. (1998) Proxy Data Base and Northern Hemisphere Average Temperature Series," *Energy and Environment*, 14.6 (2003): 751–772.
20. S. McIntyre and R. McKittrick, "Hockey Sticks, Principal Components, and Spurious Significance," *Geophysical Research Letter*, 32 (2005), <http://www.climateaudit.org>.
21. Wegman Ad hoc Committee full report.
22. Wegman Ad Hoc Committee full report, pp. 61–63.
23. S. McIntyre and R. McKittrick, "Hockey Sticks, Principal Components, and Spurious Significance," *Geophysical Research Letters*, 32 (2005): p. 2, <http://www.climateaudit.org>.
24. Wegman Ad Hoc Committee full report, pp. 48–49. MM03 is Stephen McIntyre and Ross McKittrick, "Corrections to the Mann et al. (1998) Proxy Data Base and Northern Hemisphere Average Temperature Series," *Energy and Environment*, 14.6 (2003): 751–771. MM05a is S. McIntyre and R. McKittrick, "The M&M Critique of MBH98 Northern Hemisphere Climate Index: Update and Implications," *Energy and Environment*, 16.1 (2005): 69–100. MM05b is Stephen McIntyre and Ross McKittrick, "Hockey Sticks, Principal Components, and Spurious Significance," *Geophysical Research Letters*, 32 (2005), doi:10.1029/2004GL021750. MBH98 is Michael E. Mann, Raymond S. Bradley, and Malcolm K. Hughes, "Global-Scale Temperature Patterns and Climate Forcing over the Past Six Centuries," *Nature*, 392 (1998): 779–787. MBH99 is Michael E. Mann, Raymond S. Bradley, and Malcolm K. Hughes, "Northern Hemisphere Temperatures During the Past Millennium: Inferences, Uncertainties, and Limitations," *Geophysical Research Letters*, 26.6 (1999): 759–762.
25. Union of Concerned Scientists, Citizens and Scientists for Environmental Solutions, "Global Warming," Dec. 2, 2006, [http://www.ucsusa.org/global\\_warming/science](http://www.ucsusa.org/global_warming/science), accessed 1/4/06.
26. "Wegman Ad Hoc Committee full report, p. 50. C. Wunsch, "Ocean Observations and the Climate Forecast Problem," in R. P. Pearce (ed.), *Meteorology at the Millennium* (London: Academic Press, 2002), p. 233; C. Wunsch, "Abrupt Climate Change: An Alternative View," *Quaternary Research*, 65 (2006): 191–203.
27. Wegman Ad Hoc Committefull report, p. 51.

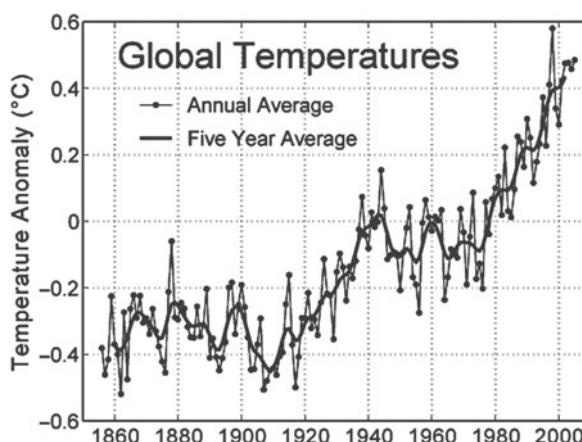
# Research Project: The Global Warming Hockey Stick Controversy

Now that you can create and interpret principal components, it is time to conduct your own research project. The following pages provide guided steps to help you evaluate a heated debate about the use of principal component analysis to model climate change.

## The Debate over Statistical Techniques Used in the Study of Climate Change

Since the 1850s, scientists have been able to measure the temperature of Earth's atmosphere and oceans. Observed temperature readings from multiple locations around the world have been used to estimate the global average near-surface temperatures. The Intergovernmental Panel on Climate Change (IPCC) used the data, called the Instrumented Temperature Record, to conclude, "Average global surface temperature has increased by approximately  $0.6^{\circ}\text{C}$  since the late 19th century, with 95% confidence limits of close to 0.4 and  $0.8^{\circ}\text{C}$ " ( $1.1 \pm 0.4^{\circ}$  Fahrenheit).<sup>13</sup>

Figure 10.12 shows the average global surface temperature for the last 150 years. Clearly the data from 1850 to 2005 show a strong increasing trend. However, it is much more difficult to reconstruct temperatures before 1850. Researchers are interested in determining if Earth's temperature naturally fluctuates every few centuries or if Earth's temperature has been relatively stable for thousands of years and only recently started to increase.



**Figure 10.12** The instrumental record of global average temperatures. Compiled by the Climatic Research Unit of the University of East Anglia and the Hadley Centre of the UK Meteorological Office.

In 1998, Mann, Bradley, and Hughes used principal component analysis on a set of proxy measurements to reconstruct surface temperatures over the past six centuries.<sup>14</sup> The practice of using proxies to determine past temperatures has become widely accepted by those researching global warming. A **proxy measurement** is an indirect measurement. Instrumentation was not available to accurately measure temperatures before 1850; however, there are many natural phenomena that are known to fluctuate with

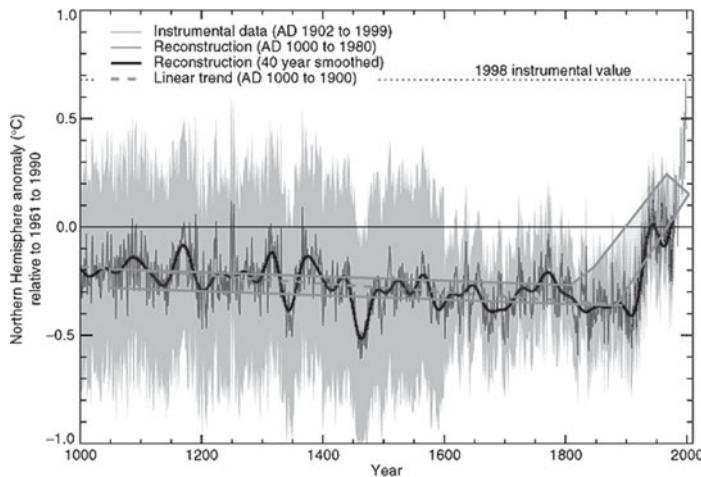
temperature. Three of the main proxies used to predict surface temperatures are found in tree rings, glacial ice cores, and coral skeletons:

The width and density of tree rings vary with climatic conditions (sunlight, precipitation, temperature, humidity, and carbon dioxide and nitrogen oxides availability), soil conditions, tree species, tree age, and stored carbohydrates in the trees. However, tree ring density is useful in paleoclimatic temperature reconstructions because in mature trees, tree rings vary approximately linearly with age. The width and density of tree rings are dependent on many confounding factors, making it difficult to isolate the climatic temperature signal. It is usually the case that width and density of tree rings are monitored in conjunction in order to more accurately use them as climate proxies.

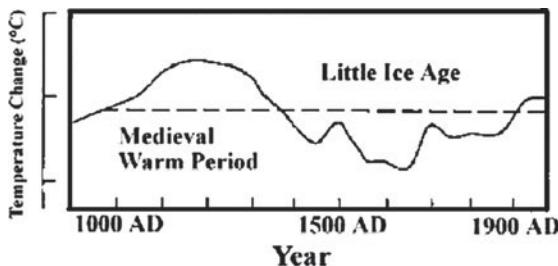
Ice cores are the accumulation of snow and ice over many years that have recrystallized and have trapped air bubbles from previous time periods. The composition of these ice cores, especially the presence of hydrogen and oxygen isotopes, provides a picture of the climate at the time. Because isotopes of water vapor exhibit a lower vapor pressure, when the temperature falls, the heavier water molecules will condense faster than the normal water molecules. The relative concentrations of the heavier isotopes in the condensate indicate the temperature of condensation at the time, allowing for ice cores to be used in global temperature reconstruction. In addition to the isotope concentration, the air bubbles trapped in the ice cores allow for measurement of the atmospheric concentrations of trace gases, including greenhouse gases carbon dioxide, methane, and nitrous oxide. The air bubbles may also contain traces of aerosols, which are produced in great concentrations during volcanic eruptions.

Coral is similar to trees in that the growth and density of the coral is dependent upon temperature. X-rays of coral cross sections show the relative density and growth over time. High density layers of coral are produced during years of high ocean surface temperatures. Hence, corals can be calibrated to estimate sea surface temperatures.<sup>15</sup>

Mann, Bradley, and Hughes's principal component analysis on temperature proxies in the Northern Hemisphere produced the "hockey stick graph" shown in Figure 10.13. The graph shows a sharp increase in global temperature at the end of the 20th century, while reconstructed temperatures before 1850 are fairly constant. We have superimposed a hockey stick shape on Figure 10.13 to illustrate. In more recent work, Mann and others have extended their graph to reconstruct global temperatures for the last 2000 years.<sup>16</sup> All of Mann's reconstructions show very stable temperatures until the mid 20th century and then a steep rise resembling a hockey stick. In their study, Mann, Bradley, and Hughes state that they



**Figure 10.13** Mann's 1998 hockey stick graph. From M. E. Mann, R. S. Bradley, and M. K. Hughes, "Global-Scale Temperature Patterns and Climate Forcing over the Past Six Centuries," *Nature*, 392 (1998): 783.



**Figure 10.14** World climate history according to the IPCC in 1990. From Climate Change: The IPCC Scientific Assessment 1990, [http://www.ipcc.ch/ipccreports/far/wg\\_1/ipcc\\_far\\_wg\\_1\\_full\\_report.pdf](http://www.ipcc.ch/ipccreports/far/wg_1/ipcc_far_wg_1_full_report.pdf), p. 202.

take a new statistical approach to reconstructing global patterns of annual temperature back to the beginning of the fifteenth century, based on the calibration of multiproxy data networks by the dominant patterns of temperature variability in the instrumental record.<sup>17</sup>

Figure 10.13 has been widely distributed as evidence of global warming. The 2001 Intergovernmental Panel on Climate Change (IPCC) and advocates of the Kyoto Protocol prominently featured Mann's 1998 graph.<sup>18</sup> However, Figure 10.13 is very different from generally accepted global reconstructions created just a few years before. Figure 10.14 is a graph published in 1990 where the IPCC showed a temperature reconstruction clearly without any hockey stick shape. Figure 10.14 still exhibits a sharp increase at the end of the 20th century; however, this graph seems to indicate that temperatures rise and fall every few centuries. It is important to note that research in this area has improved dramatically since 1990, and Figure 10.14 is not very specific and does not even provide a scale for the temperature axis.

Greenhouse gases are becoming an increasingly important issue in global warming. If Figure 10.13 is accurate, it may be argued that the increases in greenhouse gases and global warming have been induced by humans, since the rise in both corresponds to the industrial revolution. Clearly statisticians know that correlation does not show causation. However, the similarities between the trends in anthropogenically generated greenhouse gas emissions (specifically carbon dioxide) and global surface temperatures appear suspicious.

If the IPCC's 1990 graph (Figure 10.14) is accurate, it appears that Earth's temperature naturally varies every few centuries. While human technology may have some influence, global warming and cooling trends occur naturally and are not induced primarily by human technology.

After Mann, Bradley, and Hughes's 1998 hockey stick graph was repeatedly cited as evidence for global warming, Stephen McIntyre and Ross McKittrick attempted to use Mann, Bradley, and Hughes's algorithm to reproduce the hockey stick graph.<sup>19</sup> In a 2003 paper, McIntyre and McKittrick claim that there are clear statistical errors in the hockey stick graph construction. McIntyre and McKittrick state that they are not looking to deny the general patterns of recent global warming, but based on the statistical methods implemented, Mann, Bradley, and Hughes had no valid reason to conclude that the 20th century was uniquely warm.

In a 2005 paper, McIntyre and McKittrick critique several aspects of the Mann, Bradley, and Hughes approach which produced the hockey stick graph, such as statistical techniques and proxies used.<sup>20</sup> Websites created by supporters of the Mann, Bradley, and Hughes paper (<http://www.realclimate.org>) and websites created by supporters of the McIntyre and McKittrick paper (<http://climateaudit.org>) clearly have many points of disagreement. In 2006, Edward Wegman, Chair of the National Academy of Sciences' Committee on Applied and Theoretical Statistics, testified before Congress about the heated debate over the accuracy of one widely used graph reconstructing 600 years of global temperatures.<sup>21</sup>

Often, it is difficult to identify patterns in multivariate data if each variable is recorded in different units. To account for the difference in scale, each series should be standardized to have a mean of zero and a standard deviation of one. Mann, Bradley, and Hughes studied a time period from 1400 to 1980 (581 years of data for each of 70 proxy variables). Each of the 70 variables should have been standardized by subtracting the series mean (based on 581 entries in this example) and dividing by the series standard deviation (based on the 581 entries). However, Mann, Bradley, and Hughes improperly standardized each variable. Instead of using the means based on the entire time period (all 581 entries), they used truncated means, means from just the 1902–1980 time period (just the last 79 entries). They considered the use of truncated means a "new

statistical approach.” However, this project will show that using truncated means will create very biased results.\*

Suppose, just by chance, that for one series out of a large set of data, the mean from the last 79 years is larger than the overall mean (based on 581 years). Improperly centering the data with truncated means will cause this series to have a much larger variance than the other series. Remember that variance is calculated by squaring the difference between the mean and the observed value. In every series, the farther the overall mean is from the truncated mean (i.e., the mean from the last 79 years), the larger the variance will tend to be. PCA is designed to select series with the largest variance in the first few principal components. Therefore, when truncated means are used, PCA is biased to select series (proxy variables) that show the hockey stick shape, since these series have falsely inflated variances. If just one of the 70 proxy variables has a higher mean in the last 79 years, the corresponding “standardized” variable using a truncated mean is likely to have a larger standard deviation than other variables, and thus is likely be weighted very heavily in the first principal component.

Mann, Bradley, and Hughes have denied errors in their reconstruction techniques and have continued with similar methodologies in later papers to create similar graphs that display a hockey stick shape. Instead of providing a mathematical proof, as shown in Wegman’s report,<sup>22</sup> McIntyre and McKittrick conducted a simulation study to show that Mann, Bradley, and Hughes’s technique is invalid. In the simulation study, McIntyre and McKittrick repeated the following steps 10,000 times:

- Instead of using the 70 series of actual proxy data, McIntyre and McKittrick generated 70 series of random red noise data. Red noise data are data with no systematic pattern, but some correlation between adjacent observations (years) in the series. Each of the 70 series had 581 simulated random red noise “observations” representing the years 1400–1980.
- The Mann, Bradley, and Hughes truncated standardization was applied to each of the 70 random series. The 1902–1980 mean (instead of the overall mean) was subtracted, and then the series was divided by the 1902–1980 standard deviation (instead of the overall standard deviation).
- McIntyre and McKittrick conducted PCA on the 70 series to find the first principal component (i.e., the one variable that best captured the overall pattern of the data).

If random data are generated and data are properly standardized, a clear pattern in the first principal component should not emerge very often. However, using only random data, McIntyre and McKittrick’s simulation showed that Mann, Bradley, and Hughes’s technique identified a principal component with a hockey stick shape 99% of the time. With a properly centered standardization (i.e., using the standard and accepted PCA technique), a hockey stick shape occurred only about 15% of the time.<sup>23</sup> This simulation shows that Mann, Bradley, and Hughes created a technique that selectively chooses proxies that show a strong increasing (or decreasing) trend in more recent years.

Note that since the data were randomly generated in this simulation, about half the time the truncated mean (mean based on the last 79 entries in the series) was larger than the overall mean and about half the time the truncated mean was smaller than the overall mean. Thus, in PCA about half the hockey sticks turned upward and about half the hockey sticks turned downward.

Wegman’s report found results very similar to those of McIntyre and McKittrick; however, many people continue to defend the original hockey stick graph. Wegman states,

In general, we find the criticisms by MM03, MM05a and MM05b to be valid and their arguments to be compelling. We were able to reproduce their results and offer both theoretical explanations and simulations to verify that their observations were correct. We comment that they were attempting to draw attention to the deficiencies of the MBH98-type methodologies and were not trying to do paleoclimatic temperature reconstructions. . . . Generally speaking, the paleoclimatology community has not recognized the validity of the MM05 papers and has tended to dismiss their results as being developed by biased amateurs. The paleoclimatology community seems to be tightly coupled as indicated by our social network analysis, has rallied around the MBH98/99 position, and has issued an extensive series of alternative assessments most of which appear to support the conclusions of MBH98/99.<sup>24</sup>

---

\*After the data were standardized with a truncated mean and standard deviation, Mann, Bradley, and Hughes divided the improperly standardized variables by their entire standard deviation (581 entries). However, this additional calculation has no impact on the analysis we are conducting, and for simplicity we will ignore it. You may choose to explore this issue in a later section of this project.

While there are valid reasons to question the temperatures before 1850 reconstructed by Mann, Bradley, and Hughes, there is clear evidence that the average surface temperature has increased dramatically in the last 150 years. Many other studies show evidence of recent global warming.<sup>25</sup> Ocean temperatures have been increasing over the past 45 years, and plants and animals have been changing their habitation patterns, living in places previously foreign to them. Siberian peat lands are thawing, releasing greenhouse gases (including carbon dioxide and methane) into the atmosphere. These greenhouse gases are warming the atmosphere, causing more thawing and a feared cyclical pattern. As of now, it is thought that the most significant determinant of global warming is greenhouse gases. The levels of carbon dioxide in the atmosphere are, according to researchers, the most intense they have been in over 650,000 years. They have risen by 30% in the past 150 years and a full 15% in the previous 30 years alone. The ice shelf in Antarctica is losing volume, and the surface area of the Arctic Sea ice sheet has recently declined by 8%. If this trend in global warming continues, there could be a rise in sea levels and increased frequency of severe weather events, such as tornados, floods, and hurricanes.

According to experts at NASA's JPL [Jet Propulsion Lab], the average ocean height is increasing by approximately 1 millimeter per year, half of which is due to melting of polar ice and the other half due to thermal expansion. The latter fact implies that the oceans are absorbing tremendous amounts of heat, which is much more alarming because of the coupling of ocean circulation to the atmosphere.<sup>26</sup>

Wegman goes on to say, "It is clear that average global temperature increases are not the real focus. It is the temperature increases at the poles that matter and average global or Northern Hemisphere increases do not address the issue."

## Discussion Questions

1. Why do you think the IPCC and supporters of the Kyoto Accord prominently featured Mann, Bradley, and Hughes's graph?
2. Assuming we accept the above-cited reasons to believe that Mann, Bradley, and Hughes's graph was developed inappropriately, does this mean that there is no global warming?
3. State specifically how you might expect proponents and opponents to respond to McIntyre and McKittrick's and Mann, Bradley, and Hughes's work for their own political/personal benefit.
4. When there are flaws in a statistical technique, do you believe the inaccuracies in mathematical details should remain hidden if the resulting analysis could result in creating a better environment? Climate scientists have used other data sets with properly developed statistical techniques (such as expectation-maximization, or EM algorithms) and have found evidence of a hockey stick shape.
5. Wegman's report<sup>27</sup> and McIntyre and McKittrick describe the difficulty of obtaining the original data (and algorithm) from Mann, Bradley, and Hughes. Do you believe that researchers and journals should be required to share data after an article has been published? Would your opinion change if the data collection were paid for by the U.S. government?
6. Do you believe that research involving new/advanced statistical techniques should be reviewed by statisticians before it is published?
7. What can be done to ensure that proper information is appropriately communicated to the public? What are the consequences of inaccurate data being highly publicized?

## Simulating Hockey Stick Graphs

In the following exercises, you will have the opportunity to conduct simulation studies using PCA on properly standardized data and improperly standardized data. This project will empirically show that if a randomly generated series is improperly standardized by using a subset of the data, PCA will consistently result in a hockey-stick-shaped graph. The following R code and instructions for simulating data will be similar to what was done in McIntyre and McKittrick's global warming simulation study.

As in McIntyre and McKittrick's work, you will first generate 70 sets of random red noise data in a  $581 \times 70$  matrix (581 rows or observations in each series and 70 columns or variables/series). Statisticians typically call red noise data an "autoregressive time series model," or AR(1):

$$x_{i,j} = \beta x_{i,j-1} + \varepsilon_{i,j}$$

where  $\mathbf{x}_i$ ,  $i = 1, 2, \dots, 70$ , represents each of the 70 variables. Each vector  $\mathbf{x}_i$  consists of 581 elements, and each  $x_{i,j}$  represents the temperature in one year (a scalar). Within each series of data, each observation is dependent on the prior observation. The instructions below use this AR(1) model because Mann, Bradley, and Hughes's actual data consist primarily of tree ring data. According to paleoclimatologists, the growth of tree rings depends on conditions of the current year ( $x_{i,j}$ ) as well as conditions of the prior year ( $x_{i,j-1}$ ). In other words, if conditions were good in 2008, we could expect a tree to store energy that would have some impact of its growth in 2009.

The file `betassigmas.txt` gives the actual slope ( $\beta$ ) and the standard deviation for the random errors ( $\varepsilon_{i,j}$ ) used by McIntyre and McKittrick for each of the 70 series based on Mann, Bradley, and Hughes's original data. To simplify this project, we will use the same slope and standard deviation for all 70 series. The slope and standard deviation used in this project are the average values in `betassigmas.txt`:  $\beta = 0.415$  and the standard deviation  $sd(\varepsilon_{i,j}) = 0.276$ . Note that since our project is simplified, the results may be a little different than McIntyre and McKittrick's results.

Once the random series have been generated, you will be able to conduct principal component analysis on improperly centered data.

Start with a smaller example (5 randomly generated red noise series with just 20 observations each).

1. Generate a  $20 \times 5$  random matrix  $\mathbf{X}$  of simulated AR(1) data.
  - a. In **R**, create a  $20 \times 5$  matrix filled with zeros to get started.

```
> X = matrix(0, nrow=20, ncol=5)
```

Look at this matrix:

```
> X
```

- b. Create a random number in the first element of each column. Draw five values from a normal probability density with mean 0, standard deviation 0.276, and place them into the first row (year) of the “data” matrix.

```
> X[1, ] = rnorm(n=5, mean=0, sd=0.276)
```

Look at the data:

```
> X
```

All rows are still filled with zeros except the first.

- c. Create five sets of random AR(1) time series data of length 20. Loop over all of the remaining rows (years) and assign the next data value in each series to be  $\beta = 0.415$  times the previous data value in that series plus a random error (normally distributed with mean = 0,  $sd = 0.276$ ).

```
> for (i in 2:20) { X[i, ] = 0.415*X[i-1, ] + rnorm(n=5, mean=0, sd=0.276) }
```

We filled in the data one entire row (five values) at a time. Look at the data now:

```
> X
```

- d. Create a time series plot of these five random variables.

```
> plot.ts(X)
> plot.ts(X, plot.type="single")
```

- e. Save the time series plot. Click on the plot to make the graphics window the “active window” in R. On the File menu, choose “Save as” and save the file in pdf format or other format. Submit the plot with a short description. Do you see any trends?

2. Standardize each series to form the matrix  $\mathbf{Z}$ .

```
> Z = apply (X, 2, function(x) { (x-mean(x))/sd(x) } )
```

The 2 in the above function standardizes the data by column (the 2nd dimension of the matrix). Each column will be standardized using the mean and standard deviation of that column instead of using the mean and standard deviation of the entire matrix of data.

Look at the data now:

```
> Z
> plot.ts(Z, plot.type="single")
```

Submit the plot and describe any differences from the time series plot of the unstandardized data.

3. Create a new  $581 \times 70$  random matrix  $\mathbf{X}$  of simulated AR(1) data and find the first principal component ( $\mathbf{PC1}$ ) on properly standardized data. Does the plot of the first principal component show any trend?

```
> X = matrix(0, nrow=581, ncol=70)
> X[1, ] = rnorm(n=70, mean=0, sd=0.276)
> for (i in 2:581) { X[i, ] = 0.415*X[i-1, ] + rnorm(n=70, mean=0,
  sd=0.276) }
> X = data.frame(X)
> PC1 = prcomp(X, center=mean(X), scale=sd(X))$x[,1]
> plot.ts (PC1)
```

4. Create a function called `MBHsim` that repeats this process ten times. Print the time series plot of the ten  $\mathbf{PC1}$ s that you calculated. Identify how many times a hockey stick shape appears.

In R, first assign `MBHsim` to be a function and then edit the function using the text editor within R:

```
> MBHsim = function() {}
> MBHsim = edit(MBHsim)
```

When the editor opens, you will see just one line of R code:

```
function() {}
```

Place the cursor between the curly braces and type the following code to perform the needed simulation (10 times). It is always a good idea to add comments to your program, as we have here (start a comment line with the `#` character).

```
# Create a matrix to hold the 10 first principal components
PC1 = matrix(0, nrow=581, ncol=10)
# Repeat the process of data simulation and PCA 10 times
for (simnum in 1:10) {
  # Set up a matrix to hold the simulated data, fill it with zeros
  X = matrix(0, nrow=581, ncol=70)
  # Draw the first entry for each of the 70 series at random
  X[1, ] = rnorm(n=70, mean=0, sd=0.276)
  # Fill in the rest of each series using an AR(1) time series model
  # Determine the first principal component and save it for later output
  for (i in 2:581) { X[i, ] = 0.415*X[i-1, ] + rnorm(n=70, mean=0,
    sd=0.276) }
  X = data.frame(X)
  PC1[, simnum] = prcomp(X, center=mean(X), scale=sd(X))$x[,1]
}
# Output the matrix of 10 first principal components
return (data.frame(PC1))
```

Close the editor window and click “Yes” or “Save” to save the changes. Take a look at the program you have written:

```
> MBHsim
```

If you want to make changes, just edit the function again:

```
> MBHsim = edit (MBHsim)
```

To run the function, save the results in  $\mathbf{PC1}$  and make the time series plot:

```
> PC1 = MBHsim()
> plot.ts(PC1)
```

5. Create a function called `MBHsim2` that repeats ten times, as in step 4. However, this time use improperly centered data (“decentered” data), where the mean and standard deviation used to “standardize” each series are calculated only from the last 79 “years” of data. Print the time series plot of the ten first principal components that you calculate. Identify how many times the hockey stick appears.

Since just a few modifications are needed to create the new function, start with a copy of the original function:

```
> MBHsim2 = edit(MBHsim)
```

You only need to change one line of the R code. Change

```
PC1[, simnum] = prcomp(X, center=mean(X), scale=sd(X))$x[,1]
```

to read

```
PC1[, simnum] = prcomp(X, center=mean(X[(581-79+1):581,]), scale=sd(X[(581-79+1):581,]))$x[,1]
```

Close the editor and save the changes. To run the function, save the results in `PC1` and make the time series plot:

```
> PC1 = MBHsim2()
> plot.ts(PC1)
```

Print the time series plot of the ten `PC1`s that you calculated. Identify how many times a hockey stick shape appears. Remember that a hockey stick may be facing up or down.

6. Repeat step 5 several times, trying out new means and standard deviations for the random term ( $\epsilon_{i,j}$ ). Create a table with a column of your suggested means, a column of your suggested standard deviations, and a column of “number of hockey sticks observed out of 10 tries” for each scenario you try. Try several different mean and standard deviation combinations. Does changing the mean and standard deviation appear to affect the number of hockey sticks observed?

To complete this exercise, you could create a new program, such as `MBHsim3`, that allows you to input a different mean and standard deviation each time you run the program. Start with a copy of the `MBHsim2` function:

```
> MBHsim3 = edit(MBHsim2)
```

Make three changes. Add variables `m` and `s` at the top of the function (with default values `m = 0` and `s = 0.276` supplied), by changing the first line of the function from

```
function() {
  to
  function(m=0, s=0.276) {
```

Then replace the mean or standard deviation with the variable names `m` and `s` in two separate places:

```
X[1, ] = rnorm(n=70, mean=m, sd=s)
for (i in 2:581) { X[i, ] = 0.415*X[i-1, ] + rnorm(n=70, mean=m, sd=s) }
```

Now you can try out any mean and standard deviation you like for the 70 series. For example,

```
> PC1 = MBHsim3(m=1, sd=2)
> plot.ts(PC1)
```

7. Repeat step 5 with a new AR(1) parameter,  $\beta$ . Create a table with a column for  $\beta$  and a column for “number of hockey sticks observed out of 10 tries.” Does changing  $\beta$  appear to affect the number of hockey sticks observed?

Again, you can complete this exercise by editing an existing program in R:

```
> MBHsim4 = edit (MBHsim3)
```

Make two changes. Add `beta = 0.415` to the first line, so that you have the flexibility to change `beta` but also have a default value, and change each occurrence of `0.415` in the R code to read `beta` instead. Now you can try out different beta-values like this:

```
> PC1 = MBHsim4 (beta=0.7)
> plot.ts (PC1)
```

Note that  $\beta$  represents the strength of the relationship between an observation ( $x_{i,j}$ ) and the preceding observation ( $x_{i,j-1}$ ). In the context of temperature proxy measurements, reasonable values for  $\beta$  would probably be positive and less than one.

8. Read Section 2 of McIntyre and McKittrick’s 2005 *Geophysical Research Letters* paper, posted at <http://www.climateaudit.org>. How many simulations did they conduct? Describe their criterion for determining if a hockey stick exists in the first principal component.

## Creating Your Own Simulation Study

9. Modify the above code to conduct 1000 simulations using MBHsim4. Instead of printing graphs, develop your own criteria to determine if the “decentered” technique has a tendency to create hockey-stick-shaped graphs. Turn in your code for this exercise as well as the actual count of hockey stick graphs.

For example, the following code (1) generates random data, (2) conducts a centered and decentered PCA analysis, and then (3) counts the number of times the mean of the last 79 values in PC1 is significantly different from the overall mean of PC1 in both the centered and the decentered technique.

```
> MBHcounts = function (beta=0.415, m=0, s=0.276) {}  
> MBHcounts = edit (MBHcounts)
```

Within the [&{ }&] in the editor,

```
centered = 0 # INITIALIZE COUNT TO ZERO  
decentered = 0 # INITIALIZE COUNT TO ZERO  
for (simnum in 1:1000) {  
  X = matrix (0, nrow=581, ncol=70)  
  X[1, ] = rnorm(n=70, mean=m, sd=s)  
  for (i in 2:581) { X[i, ] = beta*X[i-1, ] + rnorm(n=70, mean=m, sd=s) }  
  X = data.frame(X)  
  PC1 = prcomp (X, center=mean (X), scale=sd(X))$x[,1]  
  if ( abs(mean(PC1[(581-79+1):581]) - mean(PC1)) > 0.5* (mean(PC1)+sd(PC1)) )  
    centered = centered + 1  
  PC1= prcomp (X, center=mean (X[(581-79+1):581,]),  
  scale=sd(X[(581-79+1):581,]))$x[,1]  
  if ( abs(mean(PC1[(581-79+1):581]) - mean(PC1)) > 0.5* (mean(PC1)+sd(PC1)) )  
    decentered = decentered + 1  
}  
return(c(centered, decentered))  
> myresults = MBHcounts()
```

Running 1000 simulations will take a few minutes:

```
> myresults
```

10. While it should be clear that Mann, Bradley, and Hughes’s 1998 technique needs modifications, it may be useful to conduct PCA on an actual data set. In fact, many researchers (including Mann) are now using different modeling algorithms. There are also several other issues that need to be addressed in creating reliable climate change data sets (such as calibrating data), so it is difficult to clearly define each of the original variables. However, PCA can be conducted using a data set provided at <http://www.cgd.ucar.edu/ccr/ammann/millennium>. To better understand the nuances involved in this controversy, we suggest you read the following article, available on the above website:

E. R. Wahl and C. M. Ammann, “Robustness of the Mann, Bradley, Hughes Reconstruction of Northern Hemisphere Surface Temperatures: Examination of Criticisms Based on the Nature and Processing of Proxy Climate Evidence,” *Climatic Change*, 85 (2007): 33–69.

## Presenting Your Results

Use a simulation to develop your own statistical technique to determine if Mann, Bradley, and Hughes’s decentered PCA analysis has a tendency to create hockey-stick-shaped graphs. Meet with your professor to discuss your model assumptions, diagnostics, and analysis.

Bring a draft of your poster to class. This draft should have all the elements of your final poster, but need not be in final, presentation-quality form.

Be prepared to pair up in teams and spend class time critiquing posters. Use the “How to Write a Scientific Paper or Poster” checklist, on the accompanying CD, to review other students’ posters and provide comments.

## Other Project Ideas

Several of the homework activities can also be used to develop your own project ideas. In addition, since principal component analysis is often integrated into other types of studies, many projects from previous chapters may spur project ideas. The following sites are useful in collecting data:

- Information on a variety of sports can be found at <http://cbs.sportsline.com>, <http://SportsIllustrated.cnn.com>, <http://www.nfl.com/stats/team>, <http://www.ncaa.org>, and <http://www.baseball-reference.com>.
- Information from many federal agencies can be found at <http://www.fedstats.gov>, the Bureau of Labor Statistics (<http://www.bls.gov/cpi>), the Behavioral Risk Factor Surveillance System (<http://www.cdc.gov/brfss>), the National Center for Health Statistics (<http://www.cdc.gov/nchs>), and the U.S. Census Bureau (<http://www.census.gov>).
- College and university information can be found at <http://www.collegeboard.com>, <http://www.act.org>, and <http://www.clas.ufl.edu/au>.
- The National Center for Educational Statistics website is <http://nces.ed.gov>.
- Information about movies produced each year can be found at <http://www.the-numbers.com>, <http://www.boxofficemojo.com>, <http://www.imdb.com>, and <http://www.rottentomatoes.com>.

# Bayesian Data Analysis: What Colors Come in Your M&M's® Candy Bag?

*There are two equally dangerous extremes: to shut reason out and to let nothing else in.*

—Blaise Pascal<sup>1</sup>

In previous chapters, sample data were collected to estimate the values of parameters. The parameters were unknown quantities, but they were treated as fixed constants. Bayesian statistical methods treat parameters as random variables with probability distributions, and estimates of the parameters are obtained by combining observed data with prior knowledge. The Bayesian strategy begins with a prior (often subjective) belief about how the parameter might be distributed and then updates the distribution based on observed sample data. This updated distribution allows us to compute various estimates for the parameter and construct intervals for the unknown parameter.

In this chapter, we will use a simple example involving chocolate M&M's to describe how to use the Bayesian statistical approach. By working through the activities in this chapter, you will do the following:

- Compare the Bayesian approach to the frequentist approach when assigning probability to events
- Apply Bayes' rule to combine observed data with prior (subjective) beliefs to obtain updated estimates of parameters or probability distributions
- Develop posterior estimates from a single prior estimate
- Develop posterior distributions and posterior estimates using discrete prior distributions
- Explore the impact of using different discrete prior distributions

The extended activities, end-of-chapter exercises, and project provide additional examples that describe details of Bayes' rule, working with continuous distributions, and constructing Bayesian credible intervals for parameters.

## 11.1 Investigation: Do Prior Beliefs Improve Your Estimate of the Proportion of Brown or Orange M&M's?

When you were learning to ride a bike, hit a baseball or perform basically any other task, you probably didn't succeed on the first try. Your learning process was most likely a series of trials and errors, and perhaps successes and failures. Essentially you started with what you knew or what you believed to be true (which may not have been much), and you performed the task to the best of your ability. You observed the outcome (maybe swung the bat and missed), and then you tried again based on what you had learned from your mistakes. Ideally, you performed the task better the second time (or third time, etc.)—maybe you didn't fall off your bike or maybe you finally got a base hit.

The **Bayesian approach** to data analysis can be viewed in a similar manner. You start with your prior knowledge or a belief about the value(s) of an unknown parameter (for example, a population proportion  $\pi$ ), observe some data, and then update your original belief about the parameter. In contrast, the **frequentist approach** uses the observed sample data to make decisions about  $\pi$ , ignoring information outside the study or experiment. The primary difference between the two strategies is that the frequentist approach treats the parameter  $\pi$  as a fixed, but unknown, quantity, while Bayesian analysis assigns a single probability or a probability distribution to  $\pi$  based on prior beliefs, knowledge, or information.

Bayesian methods have existed since Thomas Bayes introduced his famous theorem almost 250 years ago.<sup>2</sup> These methods have been considered controversial by some statisticians because of the (seemingly) subjective manner in which a parameter is assigned a probability distribution; however, there are reasonable assignment strategies, and these will be discussed in this chapter. In addition, Bayesian statistical methods are quite appealing because they mirror the scientific approach of acquiring new knowledge by correcting and integrating previous knowledge.<sup>3</sup>

In this chapter, we will start with a prior belief about M&M's candies (even if that knowledge is limited) and update it with observed data. When the Mars company first sold M&M's candies in the 1940s, the only color was dark brown, but after many color experiments and the addition of special seasonal colors (like red and green during the December holiday time), the Mars company now even holds M&M's surveys asking fans to vote for their favorite colors.

Even if you have not eaten M&M's candies, it is likely that several of your friends and classmates have (and you will get a chance to eat them in the activity described in Question 2). We often hear fairly strong opinions from students about the proportion of brown or orange candies, such as “The proportion is certainly less than 1/2” or “Surely, more than 1/4 of the candies are either brown or orange.” The question we will consider in the first half of the chapter is this: What is the probability that a randomly selected M&M is brown or orange? Denoting  $\pi$  as the probability that a randomly selected candy from the population of all chocolate M&M's is brown or orange, we will estimate  $\pi$  using several approaches.

### Activity A Subjective Estimate of $\pi$

1. Currently, plain milk chocolate M&M's come in six colors: red, yellow, green, blue, orange, and brown. Consider all milk chocolate M&M's in production. What is your initial guess at  $\pi$ , the probability of selecting a brown or orange candy from all chocolate M&M's in production? Compare your guess with the guesses of some of your classmates.

Your estimate of the true probability of selecting a brown or orange candy is based on personal beliefs and/or prior information. Maybe you recall a sizeable proportion of orange candies the last time you had some M&M's, which resulted in a high guess for the probability, or maybe since a probability must take a value between 0 and 1, you simply used the midpoint, 0.5, as your initial guess. Although this might not be as objective as directly collecting data, it is still a valid, defendable way to estimate the probability of selecting a brown or orange M&M. This is an example of a **subjective** definition of probability. The subjective probability of an event is based on personal beliefs, experiences, expertise, and other information available prior to observing the event.

A familiar nonsubjective approach to estimating the probability of selecting a brown or orange candy from the population of all chocolate M&M's in production might be to take a sample of candies from a bag of chocolate M&M's and then calculate the proportion (also called a relative frequency) of brown or orange candies. Using the relative frequency as our estimate for the probability is the **frequentist approach**. As the

sample size increases, the frequentist estimate will get closer to the true probability. Let  $\hat{p}$  denote the frequentist estimate of  $\pi$ , which is given by

$$\hat{p} = \frac{x}{n} \quad (11.1)$$

where

$x$  = the number of brown or orange M&M's candies in the sample

$n$  = the total number of M&M's candies in the sample

## Activity A Relative Frequency Estimate of $\pi$

You will use the frequentist approach to estimate the probabilities in this activity.

- Get a small (1.5 oz. to 3 oz.) bag of M&M's Milk Chocolate Candies, and randomly select one candy from your bag. With your sample size of one, is your estimated probability of getting a brown or orange candy 0 or 1—i.e., 0% or 100%? Select another candy. Now, with your two candies, is your estimated probability of getting a brown or orange candy 0%, 50%, or 100%? Repeat this process until you have selected all the candies from your bag. Create a plot (either by hand or using software) with your relative frequency estimates of  $\pi$  on the vertical axis and sample sizes 1 through  $n$  on the horizontal axis, where  $n$  is the total number of candies in the bag. The estimates may tend to vary wildly at first and then “settle down” to the true probability of selecting a brown or orange candy. Be sure to save your results in a data file called `MyMMs` for future activities and extended activities. When you're done with this problem, feel free to eat your candies!

### NOTE

If your class does not perform the M&M's activity, then you can use the data set `MMs` supplied by the authors to complete the activities and extended activities. Be certain to verify which data set (`MMs` or `MyMMs`) your instructor would like you to use in the following activities.\*

It is important to collect sample data, but can you also include your own prior experience and knowledge in the final data analysis? This is exactly the process by which many natural and social scientists acquire new knowledge—by starting with all that is known up to this point and then proposing a hypothesis about a phenomenon that seems to defy current theories or goes beyond what current understanding can explain. After experimenting or collecting other information about the phenomenon of interest, they update or revise the previous knowledge base in light of the new data.

### Key Concept

The (subjective) probability of an outcome is determined from personal beliefs, experiences, expertise, and other information available prior to observing the outcome. People with differing beliefs or other information may reasonably assign different subjective probabilities to the same outcome. The frequentist approach to assigning probabilities to outcomes relies on counting outcomes from observed data.

## 11.2 Combining Prior Information About $\pi$ with Data

In many situations, relative frequency estimates are unreliable or impossible to collect. For example, what is the probability that the Chicago Cubs will win the World Series in baseball next year? This is not an event that is repeatable, so we cannot obtain a relative frequency estimate.

---

\*The `MMs` data set contains the number of brown or orange M&M's in samples of size 1, 2, 3, etc., selected from a 1.69-ounce bag, as well as the proportion of brown or orange M&M's in each sample.

Subjective estimates and beliefs about a population or probability distribution (model) may not be entirely reliable, but as you obtain more information (data), it is possible to keep updating the probability distribution to describe your current belief about the population. This is the approach of the Bayesian statistician (rather than the frequentist). The Bayesian will first bring all prior knowledge to bear on the probability model and then update that (subjective) **prior model** in light of new information (data). Finally, she or he will utilize the updated probability model, called the **posterior model**, to describe the current state of knowledge about the phenomenon of interest. In the following activity you will combine a prior belief about the probability of selecting a brown or orange M&M with actual data to create a new estimate of the probability.

## Activity Prior and Posterior Probabilities

3. Let  $\hat{\pi} = 0.30$  represent the subjective probability that a randomly selected M&M is brown or orange. This subjective probability is often called the **prior probability**. Let  $\hat{p}$  represent the frequentist estimate of the probability that a selected M&M is brown or orange, which you obtained in Question 2 (based on all the candies that were selected from the bag). Then combine your prior estimate and your frequentist estimate to calculate a **posterior estimate** of  $\pi$ , denoted  $p^*$ , using the expression  $p^* = 0.5\hat{\pi} + 0.5\hat{p}$ . This expression gives your prior beliefs and the data equal weight.
4. Repeat Question 3, but give your sample data a weight of 75% and your prior estimate a weight of 25%, using  $p^* = 0.25\hat{\pi} + 0.75\hat{p}$ .

### Key Concept

Bayesian statistics begins with a prior probability of an outcome, based on all current information available. Sample data are collected to update a prior probability into a new posterior probability.

### NOTE

To introduce you to the notion of assigning subjective probabilities to outcomes, we treated  $\pi$  as the probability of selecting a brown or orange M&M. However, it also makes sense to view  $\pi$  as the proportion of all chocolate M&M's in production that are brown or orange. Especially in the context of the M&M's example, it may be more intuitive to inquire about the proportion of particular colors of M&M's rather than the probability of selecting particular colors. In the remainder of the activities, we will be referring to  $\pi$  as the proportion of all chocolate M&M's in production that are brown or orange. Regardless of whether we view  $\pi$  as a population proportion or a probability, it is an unknown parameter.

Now let's consider a second, more formal approach to combining prior beliefs and observed data to obtain a posterior estimate for  $\pi$ . To get started, let's agree to use  $\hat{\pi} = 1/3 = 0.33$  as our prior estimate for  $\pi$ . One approach to representing our prior belief about  $\pi$  as "prior data" is to use the following function, where  $a$  represents successes (brown or orange candies) and  $b$  represents failures (any other colors). Then

$$\hat{\pi} = \frac{a}{a + b}$$

is our prior estimate for  $\pi$  (prior to the data).

Note that we still have many options for  $a$  and  $b$ . Since our prior belief about  $\pi$  for the M&M's is that  $\hat{\pi} = 0.33$ , we might choose to set the prior information as

$$\hat{\pi} = \frac{1}{1 + 2} = 0.33 \text{ or } \hat{\pi} = \frac{10}{10 + 20} = 0.33 \text{ or } \hat{\pi} = \frac{100}{100 + 200} = 0.33$$

or any other choices for  $a$  and  $b$  that result in  $\hat{\pi} = a/(a + b) = 0.33$ .

The new posterior estimate of  $\pi$  is calculated as the total number of successes (from both prior belief and sample data) divided by the total number of successes and failures:

$$\begin{aligned} p^* &= \frac{x + a}{n + a + b} \\ &= \left( \frac{n}{n + a + b} \right) \left( \frac{x}{n} \right) + \left( \frac{a + b}{n + a + b} \right) \left( \frac{a}{a + b} \right) \\ &= \frac{n}{n + a + b} \hat{p} + \frac{a + b}{n + a + b} \hat{\pi} \end{aligned} \quad (11.2)$$

The posterior estimate  $p^*$  is an update for our prior belief about  $\pi$ , given that we now also have sample data. You can see from Equation (11.2) that  $p^*$  is a weighted average of the frequentist and prior estimates of  $\pi$ .

#### ► MATHEMATICAL NOTE ▾

In Section 11.7 in the extended activities, we will show that this posterior estimate for  $\pi$  is derived from the **beta distribution**.

## Activity (▶) Calculating the Posterior Estimate

5. Using the M&M's data that you collected and the prior estimate  $\hat{\pi} = 0.33$  with  $a = 1$  and  $b = 2$ , find the posterior estimate for  $\pi$  using Equation (11.2).
6. Using the M&M's data that you collected and the prior estimate  $\hat{\pi} = 0.33$  with  $a = 10$  and  $b = 20$ , find the posterior estimate for  $\pi$  using Equation (11.2).
7. Using the M&M's data that you collected and the prior estimate  $\hat{\pi} = 0.33$  with  $a = 100$  and  $b = 200$ , find the posterior estimate for  $\pi$  using Equation (11.2).
8. As  $a$  and  $b$  increased, what did you observe about the posterior estimate for  $\pi$ ? What connection can you draw between our prior belief and the values of  $a$  and  $b$ ?

The stronger your belief in the prior information, the more weight you will want this information to have in your final estimate for  $\pi$ . In the author's bag of M&M's,  $x = 23$  brown or orange candies were observed out of a total of  $n = 55$  candies in the bag. We will use this data in the worked-out examples of this chapter, but you should use your own data in the exercises. Later, you will combine your data with ours. So our completely data-based estimate for  $\pi$  is  $\hat{p} = x/n = 23/55 = 0.418$ . Our three posterior estimates in Questions 5 through 7 are  $p_1^* = 0.414$  ( $a = 1, b = 2$ ),  $p_2^* = 0.388$  ( $a = 10, b = 20$ ), and  $p_3^* = 0.346$  ( $a = 100, b = 200$ ).

You can see that  $p_1^* = 0.414$  is closest to the data estimate  $\hat{p} = 0.418$  and that  $p_3^* = 0.346$  is closest to the prior belief  $\hat{\pi} = 0.33$ . When we set  $a = 100$  and  $b = 200$  as the prior data, we are stating that we are more certain about our prior belief, and the posterior estimate is pulled toward our prior estimate  $\hat{\pi} = 0.33$ .

### Key Concept

A simple posterior estimate for a proportion  $\pi$  is

$$p^* = \frac{x + a}{n + a + b}$$

where our prior estimate is  $\hat{\pi} = a/(a + b)$  (as if we had  $a$  successes and  $b$  failures prior to collecting data) and the data are  $x$  successes out of  $n$  observations. Larger values for  $a$  and  $b$  indicate more certainty in our prior belief about the value of  $\pi$ .

## 11.3 Prior Distributions for $\pi$

In our earlier analysis, we simply provided  $\hat{\pi} = 0.33$  as our prior estimate for the proportion of all chocolate M&M's in production that were brown or orange. A more flexible approach is to specify our belief in the form of an entire probability distribution for several possible values for  $\pi$ . As mentioned in Section 11.1, assignment of a probability distribution to a parameter is a major distinction between the Bayesian and frequentist approaches to data analysis. A probability distribution for  $\pi$  will still make our belief about the parameter clear, but it will allow for more flexibility and expression of our uncertainty.\* For example, Table 11.1 illustrates a prior probability distribution that equally weights three reasonable possibilities for  $\pi$  (we will denote them as  $\pi_1, \pi_2, \pi_3$ ) with probabilities  $p(\pi_k) = 1/3$  for  $k = 1, 2, 3$ .

**Table 11.1** One possible prior distribution of  $\pi$  = the probability of selecting a brown or orange M&M.

$\pi$	0.28	0.33	0.38
$p(\pi)$	1/3	1/3	1/3

### MATHEMATICAL NOTE

The proposed prior distribution for  $\pi$  in Table 11.1 is a valid probability distribution for a discrete random variable. Recall that a valid probability distribution will satisfy the following two conditions:

1. Each probability  $p(\pi) \geq 0$  for any value of  $\pi$ .
2. The sum of the probabilities must be equal to 1.

Table 11.1 is just one example of a possible prior distribution that we can use for  $\pi$ . The prior distribution should incorporate our beliefs about how the values of  $\pi$  might vary and how much probability should be assigned to the values that  $\pi$  might take. Other prior distributions might be preferable if they display less variability in the values of  $\pi$ . Other criteria can be used to select prior distributions, and those will be discussed in the extended activities.

Some simple tools that we can use to compare prior distributions are the mean and standard deviation. From your introductory statistics course, recall that if we can list the possible values that the random variable  $X$  can take [i.e.,  $x_1, x_2, \dots, x_n$ ] and their corresponding probabilities [ $p(x_1), p(x_2), \dots, p(x_n)$ ], then

- The mean or expected value of  $X$ , denoted  $E(X)$ , is given by

$$E(X) = \sum_{k=1}^n x_k p(x_k) \quad (11.3)$$

- The variance of  $X$ , denoted  $\text{Var}(X)$ , is given by

$$\text{Var}(X) = \sum_{k=1}^n [x_k - E(X)]^2 p(x_k) \quad (11.4)$$

Computing the mean and variance of the prior distributions allows us to compare the typical values of  $\pi$  under different beliefs about  $\pi$ , as well as compare the amount of certainty in our beliefs. For example, the mean of the prior probability distribution for  $\pi$  displayed in Table 11.1 is given by

$$E(\pi) = \sum_{k=1}^3 \pi_k p(\pi_k) = 0.28(1/3) + 0.33(1/3) + 0.38(1/3) = 0.33 \quad (11.5)$$

\*Some might question if it is entirely appropriate to assign a probability distribution to the true proportion of brown or orange M&M's, since this should be a value that is fixed by the manufacturer, Mars, Inc. While there may be a particular value of  $\pi$  that Mars, Inc. has selected, it is likely that there are still fluctuations in the value of  $\pi$  during different production runs.

The mean of the prior distribution will be referred to as the **prior estimate** for  $\pi$ ,<sup>\*</sup> and as we can observe, the prior estimate for  $\pi$  is the same as our prior belief in earlier examples. The variance of the prior distribution for  $\pi$  given in Table 11.1 is

$$\begin{aligned}\text{Var}(\pi) &= \sum_{k=1}^3 [\pi_k - E(\pi)]^2 p(\pi_k) \\ &= (0.28 - 0.33)^2(0.33) + (0.33 - 0.33)^2(0.33) + (0.38 - 0.33)^2(0.33) \\ &= 0.00165\end{aligned}$$

## Activity Comparing Prior Distributions

9. Table 11.2 provides another possible prior probability distribution for the proportion  $\pi$  of brown or orange M&M's in the population. Show that the proposed prior distribution in Table 11.2 also has mean  $E(\pi) = 0.33$ .

**Table 11.2** A second possible prior distribution for  $\pi$ .

$\pi$	0.28	0.33	0.38
$p(\pi)$	1/4	1/2	1/4

10. Calculate the variance of the proposed prior distribution displayed in Table 11.2. Compare the variances of the two proposed prior distributions shown in Tables 11.1 and 11.2. Which variance implies more precision about the prior estimate of  $\pi$ ?
11. Assume that no other information regarding the prior distributions for  $\pi$  is available beyond the mean and variance. Based on your answer to Question 10, would you recommend the prior distribution in Table 11.1 or in Table 11.2 if you wanted to show more certainty in the prior estimate of 0.33? Briefly explain.

## 11.4 Calculating the Posterior Distribution for $\pi$

Earlier we updated our prior belief about the proportion of brown or orange M&M's by including information about the observed proportion of brown or orange M&M's to obtain a posterior estimate of  $\pi$ . Now we'll do something slightly more challenging and update the entire prior distribution for the proportion of brown or orange M&M's by incorporating observed data. Just like the prior distributions in Tables 11.1 and 11.2, the posterior distribution will give the probability that the proportion of brown or orange M&M's is equal to  $\pi_k$  for  $k = 1, 2, 3$ . However, the probabilities in the posterior distribution have been updated based on sample data (the observed proportion of brown or orange M&M's). These updated probabilities are known as **posterior probabilities**, and the set of all the posterior probabilities is referred to as the **posterior distribution** for  $\pi$ . Since these posterior probabilities depend on the observed data, they are described as the **conditional probability**  $p(\pi_k | x)$  for  $k = 1, 2, 3$ .

The conditional probability formula discussed in most introductory statistics courses states that for any two events  $A$  and  $B$ , the probability that  $A$  occurs, given that  $B$  has already occurred, denoted  $P(A | B)$ , is given by

$$\begin{aligned}P(A | B) &= \frac{P(A \text{ and } B)}{P(B)} \\ &= \frac{P(B | A)P(A)}{P(B)}\end{aligned}\tag{11.6}$$

<sup>\*</sup>Other characteristics of the prior distribution, such as the median or mode, can also be used as prior estimates, but will not be discussed in this chapter.

where  $P(B) > 0$ . Equation (11.6) is referred to as Bayes' rule and is the primary expression used in Bayesian statistics. This rule is discussed in much more detail in the extended activities.

When we are working with probability distributions, Bayes' rule is written using slightly different notation. If  $p(\pi)$  is the prior distribution assigned to  $\pi$  [i.e.,  $p(\pi_k)$  is the probability assigned to the outcome  $\pi_k$  for  $k = 1, 2, 3$ ], then Bayes' rule can be used to find the posterior probabilities for  $\pi$  by solving

$$\begin{aligned} p(\pi_k | x) &= \frac{p(\pi_k \text{ and } x)}{p(x)} \\ &= \frac{p(x | \pi_k)p(\pi_k)}{p(x)} \quad \text{for } k = 1, 2, 3 \end{aligned} \quad (11.7)$$

Equation (11.7) is key to all Bayesian statistics in this chapter. Thus, it is important to have an intuitive understanding of each term in this equation:

- $p(x | \pi_k)$  is the probability of observing  $x$  successes (e.g.,  $x$  brown or orange candies) in  $n$  independent and identical trials (e.g., a random sample of  $n$  chocolate M&M's) when the probability of success is  $\pi_k$ ;  $p(x | \pi_k)$  is also called the likelihood of observing the data given success probability  $\pi_k$ . You may recall that with a fixed probability of success ( $\pi_k$ ), the probability of observing  $x$  successes in  $n$  trials is just a binomial probability:

$$p(x | \pi_k) = \binom{n}{x} \pi_k^x (1 - \pi_k)^{n-x} \quad (11.8)$$

- $p(\pi_k)$  is the prior probability assigned to a potential value of  $\pi_k$  for  $k = 1, 2, 3$ . In Table 11.1,  $p(\pi_k) = 1/3$  for  $k = 1, 2, 3$ .
- $p(x)$  is the average likelihood weighted by the prior probabilities  $p(\pi_k)$  for  $k = 1, 2, 3$ . In other words,  $p(x)$  is a weighted average probability of observing  $x$  brown or orange candies, given values of  $\pi_k$ , and is given by

$$p(x) = \sum_{k=1}^3 p(x | \pi_k)p(\pi_k) \quad (11.9)$$

$p(x)$  is sometimes referred to as the **marginal probability of the data** or **marginal likelihood of the data**, since we are summing probabilities of the data  $x$  based on different values of  $\pi_k$ .

## Likelihoods for the Observed Data

To calculate the posterior distribution using Equation (11.7), we will first find  $p(x | \pi_k)$ , the likelihood of observing the data under the various prior values for  $\pi$ , for  $k = 1, 2, 3$ . Using Table 11.1, for  $k = 1$  the number of brown or orange M&M's in a sample of  $n$  candies follows a binomial distribution with success probability  $\pi_1 = 0.28$ . Then the likelihood of observing 23 brown or orange M&M's in a sample of 55 candies when  $\pi_1 = 0.28$  can be found using Equation (11.8):

$$p(x | \pi_k) = \binom{n}{x} \pi_k^x (1 - \pi_k)^{n-x} = \binom{55}{23} (0.28)^{23} (0.72)^{55-23} = 0.00978$$

### Activity Computing Likelihoods

12. Find  $p(x | \pi_k)$  for  $k = 2, 3$ ; that is, find the likelihood of observing 23 brown or orange M&M's in a sample of 55 when  $\pi_2 = 0.33$  and when  $\pi_3 = 0.38$ .

## Marginal Likelihood of the Data

The marginal likelihood is the probability of observing a particular outcome, taking into consideration all possible probabilities in the prior distribution ( $\pi_1 = 0.28$ ,  $\pi_2 = 0.33$ , and  $\pi_3 = 0.38$  in our example). This likelihood is calculated as the sum of the products between the priors and the likelihoods. We can find  $p(x)$  by solving

$$\begin{aligned}
 p(x) &= \sum_{j=1}^3 p(x | \pi_j)p(\pi_j) \\
 &= \binom{n}{x}(0.28)^x(0.72)^{n-x}(1/3) \\
 &\quad + \binom{n}{x}(0.33)^x(0.67)^{n-x}(1/3) \\
 &\quad + \binom{n}{x}(0.38)^x(0.62)^{n-x}(1/3)
 \end{aligned} \tag{11.10}$$

when  $x$  brown or orange M&M's are observed in a sample of  $n$  candies.

#### MATHEMATICAL NOTE

We observe  $x$  brown or orange M&M's in a sample of  $n$  M&M's only under one of the following values of  $\pi$ :  $\pi_1 = 0.28$ ,  $\pi_2 = 0.33$ , or  $\pi_3 = 0.38$ . Hence,

$$\begin{aligned}
 p(x) &= p(x \text{ and } \pi_1 = 0.28) + p(x \text{ and } \pi_2 = 0.33) + p(x \text{ and } \pi_3 = 0.38) \\
 &= \sum_{j=1}^3 p(x \text{ and } \pi_j) \\
 &= \sum_{j=1}^3 p(x | \pi_j)p(\pi_j)
 \end{aligned} \tag{11.11}$$

This result is an example of an application of the **law of total probability**. Suppose that the event  $A$  is defined on a sample space  $S$ . Then the law of total probability states that for  $n$  events  $E_1, E_2, \dots, E_n$  that completely partition the sample space  $S$  (i.e.,  $E_i \cap E_j = \emptyset$  for all  $i \neq j$ ),

$$P(A) = \sum_{j=1}^n P(A \cap E_j) = \sum_{j=1}^n P(A | E_j)P(E_j) \tag{11.12}$$

#### Activity (▶) Computing the Marginal Likelihood

13. Use Equation (11.10) to find  $p(x)$ , the marginal likelihood of observing 23 brown or orange M&M's in a sample of 55 candies.

#### Applying Bayes' Rule: Posterior Probabilities for $\pi$

Once we have the likelihoods, priors, and probability of observing the data, we can use Bayes' rule in Equation (11.7) to update our prior probabilities for  $\pi$  and find the posterior probabilities. For example, the posterior probability that  $\pi_1 = 0.28$  when we observe  $x$  brown or orange M&M's in a sample of  $n$  candies is

$$\begin{aligned}
 p(\pi_1 | x) &= \frac{p(x | \pi_1)p(\pi_1)}{p(x)} \\
 &= \frac{p(x | \pi_1)p(\pi_1)}{\sum_{j=1}^3 p(x | \pi_j)p(\pi_j)} \quad (\text{by the law of total probability}) \\
 &= \frac{\binom{n}{x}(0.28)^x(0.72)^{n-x}(1/3)}{\binom{n}{x}(0.28)^x(0.72)^{n-x}(1/3) + \binom{n}{x}(0.33)^x(0.67)^{n-x}(1/3) + \binom{n}{x}(0.38)^x(0.62)^{n-x}(1/3)} \\
 &= 0.068
 \end{aligned}$$

## Activity Computing the Posterior Probabilities

14. Table 11.3 displays some of the work required to find the posterior probabilities for  $\pi$ . Find the missing entries in the table and find the remaining posterior probabilities. For this question, use the M&Ms data with 23 brown or orange M&M's out of a sample of 55.

**Table 11.3** Partial posterior probability calculations for the proportion of brown or orange M&M's.

$\pi$	Prior $p(\pi)$	Likelihood = $p(x   \pi)$	Prior $\times$ Likelihood	Posterior $p(\pi   x)$
0.28	1/3	$\binom{55}{23}(0.28)^{23}(0.72)^{55-23} = 0.0098$	0.0033	0.068
0.33	1/3			
0.38	1/3			
Sum	1			1.000

### Key Concept

The posterior probability distribution for  $\pi$  is found by calculating posterior probabilities for the possible values of  $\pi_k$ , given the data:

$$p(\pi_k | x) = \frac{p(x | \pi_k)p(\pi_k)}{\sum_j p(x | \pi_j)p(\pi_j)} \quad (11.13)$$

Once all the posterior probabilities for  $\pi$  have been computed, you can examine the posterior probability distribution for  $\pi$  shown in Table 11.4. Observe that the possibilities for  $\pi$  are no longer equally weighted, as they were in the prior distribution. Since the data ( $\hat{p} = x/n = 23/55 = 0.418$ ) imply that  $\pi > 0.33$ , much of the probability for  $\pi$  has now moved away from the lowest prior possibility that we had considered ( $\pi_1 = 0.28$ ) and onto the larger value ( $\pi_3 = 0.38$ ).

**Table 11.4** Posterior distribution of  $\pi$ , given the prior distribution in Table 11.1 and our observed M&M's data with  $x = 23$  brown or orange candies out of  $n = 55$ .

$\pi$	0.28	0.33	0.38
$p(\pi   x)$	0.068	0.296	0.636

## Activity Constructing the Posterior Distribution

Answer the following question only if you have your own data. Otherwise this problem is identical to Question 14.

15. Find the Bayesian posterior probability distribution for  $\pi$  using the prior distribution from Table 11.1 and your own observed M&M's data  $x$  and  $n$ .

## 11.5 The Posterior Mean

The posterior distribution can be used to suggest a “good” estimate for  $\pi$ . What value of  $\pi$  shall we use as our estimate? Just as with the prior probability distribution, there are many quantities based on the posterior distribution that can be used as the posterior estimate of  $\pi$ .<sup>\*</sup> We will use the **posterior mean** in this chapter as our Bayesian posterior estimate for  $\pi$ . The posterior mean, denoted  $p^*$ , is calculated by

$$p^* = E(\pi | x) = \sum_{j=1}^n \pi_j p(\pi_j | x) \quad (11.14)$$

We can interpret  $p^*$  as the expected value of  $\pi$  after observing the sample data. For the posterior distribution in Table 11.4, the Bayesian posterior estimate is

$$p^* = E(\pi | x) = 0.28(0.068) + 0.33(0.296) + 0.38(0.636) = 0.358$$

One way to think about the level of certainty that we have placed in our prior distribution is to note that the posterior mean ( $p^* = 0.358$ ) is somewhere between our earlier posterior estimates  $p_3^* = 0.346$  (using prior “data” of  $a = 100$  successes and  $b = 200$  failures) and  $p_2^* = 0.388$  (using prior “data” of  $a = 10$  successes and  $b = 20$  failures). If we had simply stated that we believed (prior to the data) that  $\pi = 0.33$  and had been willing to attach prior “data” at the level, for example, of  $a = 45$  successes and  $b = 90$  failures to this belief, then, from Equation (11.2), the simple Bayesian posterior estimate for  $\pi$  would be

$$p^* = \frac{x + a}{n + a + b} = \frac{23 + 45}{55 + 45 + 90} = 0.358$$

That is, using the prior probability distribution in Table 11.1 to attach a level of certainty to our prior belief is like having a “prior sample” with  $a = 45$  successes and  $b = 90$  failures even before collecting any data. These values of  $a$  and  $b$  are very subjective, since they are based on the strength of our belief in the prior estimate. If we had felt less certain about our prior estimate of 0.33, then we might have used smaller values of  $a$  and  $b$ . If we had been confident in our prior estimate, then we might have chosen larger values.

### Key Concept

A common Bayesian posterior estimate for a proportion  $\pi$  is the mean of the posterior distribution (or the posterior mean), given by

$$p^* = E(\pi | x) = \sum_{j=1}^n \pi_j P(\pi_j | x) \quad (11.15)$$

### Activity Computing the Posterior Mean

Data set: MyMMS (or MMs) and full class data set

16. Find the posterior mean for the posterior distribution of  $\pi$  that you found in Question 15 using your own data and the prior probability distribution in Table 11.1.
17. For the second proposed prior distribution displayed in Table 11.2, find the Bayesian posterior estimate for  $\pi$  (the mean of the posterior probability distribution) using the M&M's data ( $x = 23, n = 55$ ). Compare this to our estimate  $p^* = 0.358$  using the quantities in Table 11.1. Which estimate is closer to  $E(\pi) = 0.33$ ? Does this make sense? Explain.
18. Using the combined data from the entire class, re-calculate your Bayesian posterior estimate. Compare your new posterior distribution to the posterior distribution shown in Table 11.4. How do larger sample sizes impact the posterior probability distribution and the estimate for the posterior mean?
19. At one time, the M&M's website <http://www.mms.com> claimed that the proportion of brown or orange candies in production was 33% (i.e.,  $\pi = 0.33$ ). Based on your own M&M's data, compare your original frequentist estimate  $\hat{p}$  to your Bayesian posterior estimate  $p^*$ . Which is closer to  $\pi$ ? As long as a reasonable prior distribution has been assigned, even if it is a little off, the Bayesian estimate generally tends to improve the frequentist estimate.

<sup>\*</sup>For example, the mode, median, and the mean of the posterior distribution are commonly used posterior estimates.

## 11.6 What Can We Conclude About Colors of M&M's?

Using a relatively simple study, we explored possible approaches to estimating the true proportion of brown or orange M&M's. The frequentist estimate of the population proportion,  $\pi$ , of brown or orange M&M's is simply the proportion of brown or orange candies in a sample of M&M's. From our sample, we found that the frequentist estimate of  $\pi$  was  $23/55 = 0.418$ . The completely data-based estimate did not incorporate information beyond our existing sample.

In both the frequentist and the Bayesian approaches, it is important to consider how the sample data are collected. If the sample is not reflective of the entire population (an appropriate random sample is not collected), the result may contain biases which may invalidate the results. It is likely that a single bag of M&M's (which is made at a particular time and place) may not truly reflect the population of all M&M's in the world. Thus, we must be cautious about extending the results from our sample data to a larger population.

The Bayesian approach to estimating  $\pi$  was slightly more involved, but more flexible. We began with an initial estimate for  $\pi$ , based either on a subjective guess or on the mean of a prior probability distribution. After observing sample data (23 brown or orange M&M's in a sample of 55), we updated our initial estimate or prior distribution to find the posterior estimate of  $\pi$ . The posterior estimate was taken to be the mean of the posterior distribution for  $\pi$ . Depending on the amount of belief we put into our prior estimate and the amount of data collected, we found that the posterior estimate of  $\pi$  may deviate substantially from the prior estimate. When we have more certainty in the prior estimate (i.e., there is less variability in the prior distribution), the posterior estimate will be closer to the prior estimate. In the end, we examined various choices based on different prior distributions, but did not settle on one "correct" (posterior) estimate for  $\pi$ . Our final Bayesian posterior estimate should be a compromise between the frequentist estimate and an estimate based on prior knowledge or beliefs. When prior distributions are reasonably well chosen, the Bayesian estimate generally tends to improve the frequentist estimate.

### A Closer Look

### Bayesian Data Analysis

The previous activities with M&M's were meant to expose you to the Bayesian strategy of updating your prior beliefs about a parameter by incorporating observed data. In the extended activities, you will cover Bayes' rule in more depth and see how Bayesian methods can be applied in a variety of fields. We'll work through some interesting examples in the health sciences and parapsychology, and we'll also examine more advanced topics like continuous distributions for priors, as well as selecting prior distributions for a binomial proportion using relatively simple criteria. Finally, we'll discuss interval estimates for a binomial proportion called Bayesian credible intervals, which are constructed from the posterior distribution after observing data.

#### NOTE

In most of the extended activities, our discussion is limited to Bayesian methods for estimating a population proportion  $\pi$ ; however, Bayesian methods can also be used to estimate other parameters, such as a population mean  $\mu$  or a population standard deviation  $\sigma$ .<sup>4</sup>

## 11.7 Screening for the HIV Virus in the U.S. Blood Bank Supply: Applications of Bayes' Rule

The American Red Cross uses an initial screening questionnaire and blood tests to ensure the safety of the blood bank in the United States. The initial test for the HIV virus performed on donated blood at the blood bank is the Enzyme-Linked ImmunoSorbent Assay (ELISA) test. This test has 99.7% **sensitivity**, which is the rate of correct positive test results for infected individuals, and 98.5% **specificity**, which is the rate of correct negative test results for non-infected individuals.<sup>5</sup>

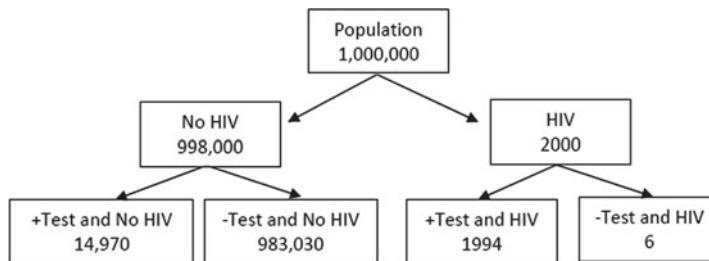
We do not have data on the rate of HIV infection among prescreened donors, but we do have related information: As of 2003, the prevalence of HIV in the United States was estimated to be 0.6% of the general

population. It is believed that the prevalence rate is lower for blood donors passing the prescreening survey, but incidence rates (new infections) must also be considered. In all the following examples, we will set the rate of HIV in the prescreening blood donor pool to 0.2%. Note that this is a subjective probability based on the synthesis of several expert opinions and different rates may also reasonably be selected.

Using Bayesian statistics, we can directly combine the three available sources of information:

- The prior model for the prevalence of HIV says that the probability that a randomly selected prescreened donor is infected with HIV is  $P(\text{HIV}) = 0.002$  and the probability that a randomly selected prescreened donor is *not* infected with HIV is  $P(\text{No HIV}) = 0.998$ .
- The ELISA test sensitivity information provides the chance of observing a positive test result given the prior model (i.e., given that the prescreened donor is infected with HIV):  $P(+\text{Test} | \text{HIV}) = 0.997$ .
- The ELISA test specificity information  $P(-\text{Test} | \text{No HIV}) = 0.985$  can be used to determine the chance of a positive test result, given that the donor does not have HIV:  $P(+\text{Test} | \text{No HIV}) = 1 - 0.985 = 0.015$ .

It is often beneficial to view the counts through a diagram or a table. Let's assume we have a population of 1,000,000 people. Based on the probabilities provided above,  $1,000,000 \times 0.002 = 2000$  would be infected with HIV and 998,000 would not. Of the 998,000 who are not infected, 1.5% (14,970) will test positive, while 98.5% (983,030) will test negative. Also, of the 2000 who are infected, 99.7% (1994) will test positive, while 0.3% (6) will test negative. The counts are organized in Figure 11.1.



**Figure 11.1** Diagram displaying counts of individuals in the population, categorized by HIV status and ELISA test results.

While these tests are fairly accurate, they are not perfect. In this section, we will address two key questions of great interest at the blood bank:

1. What is the probability that a randomly selected prescreened blood donor who tests positive actually has HIV:  $P(\text{HIV} | + \text{ Test})$ ?
2. What is the probability that a randomly selected prescreened blood donor who tests negative actually has HIV:  $P(\text{HIV} | - \text{ Test})$ ?

The question of the meaning of a positive ELISA test result is very serious, given the gravity of the diagnosis of HIV infection and the stigma surrounding this virus. The first question is of more interest to the individual blood donor, and the second has consequences for anyone receiving a blood transfusion from the blood bank.

The second question concerns testing errors resulting in the presence of HIV-infected blood in the blood supply and has been extensively studied. Many safeguards are in place in addition to the prescreening questionnaire and the ELISA test, so the current estimate of the chance of HIV infection by blood transfusion is just 1 in every 2.3 million.<sup>6</sup>

As discussed in the earlier activities, the Bayesian statistician is interested in the posterior distribution. In this example, we are interested in the posterior model  $P(\text{HIV} | + \text{ Test})$ . Bayes' rule in Equation (11.6) shows that the posterior model can be expressed in terms of the prior model. Equation (11.16) displays Bayes' rule written in terms of our example:

$$P(\text{HIV} | + \text{ Test}) = \frac{P(+\text{Test} | \text{HIV})P(\text{HIV})}{P(+\text{Test})} \quad (11.16)$$

where the law of total probability [see Equation (11.12)] tells us that

$$P(+\text{Test}) = P(+\text{Test} | \text{HIV})P(\text{HIV}) + P(+\text{Test} | \text{No HIV})P(\text{No HIV}) \quad (11.17)$$

## Extended Activity

### Finding the Probability of HIV Infection After a Positive ELISA Test

20. Use the counts in Figure 11.1 to complete Table 11.5 and answer the following questions.

**Table 11.5** Counts of individuals in the population (1,000,000 individuals), categorized by HIV status and ELISA test results.

	HIV	No HIV	Total
+Test			
-Test			
<b>Total</b>			1,000,000

21. The **false-positive** rate is the proportion of times that the ELISA test will indicate a positive result for an individual who is *not* infected by HIV [i.e.,  $P(+\text{Test} \mid \text{No HIV})$ ]. Use the counts in Table 11.5 to compute the false-positive rate.
22. Use the counts in Table 11.5 to find  $P(+\text{Test} \mid \text{HIV})$ . Note that this value is the sensitivity of the ELISA test.
23. Use the law of total probability [specifically, see Equation (11.17)] directly to find the probability that a donor tests positive; that is, find  $P(+\text{Test})$ .
24. Use Bayes' rule [Equation (11.16)] directly to find the probability that a randomly selected prescreened blood donor who tests positive actually has HIV; that is, find  $P(\text{HIV} \mid +\text{Test})$ .

Prior to collecting the data (the result of the ELISA test), we believed that the chance that a prescreened donor had HIV was about 0.2%. After running the ELISA test and determining that the result was positive (i.e., after collecting the data), we updated that chance to 11.75% (**conditional** on testing positive) from an “unconditional” (overall) probability of 0.2%.

### Bayesian Model Updating: What If We Re-Test the Blood Samples?

You might have found it unsettling to discover, from our analysis of the U.S. blood bank supply thus far, that about 7 in every 8 ( $1 - 0.1175 = 0.8825 = 88.25\%$ ) prescreened blood donors who test positive for HIV on the ELISA test do not actually have the disease. Given the stigma surrounding HIV and the serious health consequences of actually acquiring AIDS, the U.S. blood bank certainly would not want to alarm donors on the basis of one ELISA test result. Indeed, the blood bank does a second test (called the Western blot test) as part of its regular testing protocol. This test is the “gold standard” (most accurate) among HIV tests, but is much more expensive than the ELISA test.

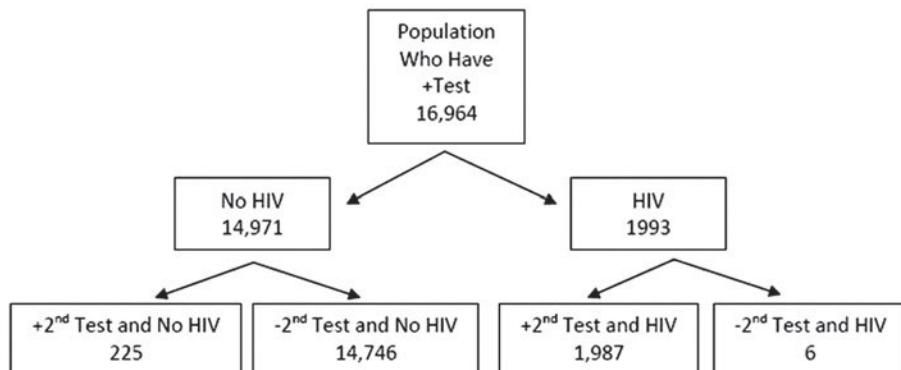
Here, however, let's investigate another way to be more careful: double check the blood work (that is, run the ELISA test twice for each blood donor). We are interested in finding the probability that a randomly selected prescreened blood donor whose blood sample tests positive *twice* actually has HIV; we'll denote this probability as  $P(\text{HIV} \mid +\text{2nd Test})$ . The Bayesian strategy will allow us to simply update our current posterior model for prevalence (among donors with a single positive test result), given new data (a second positive test result). For this population of donors who have already tested positive, we will directly combine the two available sources of information.

- The prior model is the posterior model for prevalence that we calculated for prescreened donors testing positive *once* for HIV. The prior probability of having HIV is 0.1175 for this subpopulation,  $P(\text{HIV}) = 0.1175$ , and the prior probability that a prescreened donor is not infected with HIV is  $P(\text{No HIV}) = 0.8825$ .

- The original ELISA test sensitivity will be used for the probability of testing positive a second time, given that the donor has HIV:  $P(+\text{Test} | \text{HIV}) = 0.997$ . Furthermore, the original ELISA test specificity information will be used to find the probability of a positive test result, given that the donor does not have HIV:  $P(+\text{Test} | \text{No HIV}) = 1 - 0.985 = 0.015$ . \*

We can create a figure of counts similar to Figure 11.1 and a table similar to Table 11.5, which you completed in Question 20. Starting with the population of 16,964 individuals who tested positive once already (from your completed Table 11.5) and using the same sensitivity and specificity rates as before, individuals can be classified according to their HIV status and their results on the second ELISA test. The updated counts are displayed in Figure 11.2 and Table 11.6.

Since the counts in Table 11.6 reflect all donors who tested positive once, they can be used to find the posterior probability that a donor has HIV after observing a second positive test result.



**Figure 11.2** Diagram displaying counts of individuals in the population of donors who tested positive once (16,964 individuals), categorized by HIV status and second ELISA test results.

**Table 11.6** Counts of individuals in the population of donors who tested positive once (16,964 individuals), categorized by HIV status and second ELISA test results.

	HIV	No HIV	Total
+2nd Test	1987	225	2212
-2nd Test	6	14,746	14,752
<b>Total</b>	<b>1993</b>	<b>14,971</b>	<b>16,964</b>

### Key Concept

In the Bayesian statistics paradigm, when additional data become available, we can use **Bayesian model updating**. The posterior distribution based on the first round of data becomes the prior distribution (prior to observing the additional data). This prior distribution is updated using the new data via Bayes' rule as usual, and we obtain a new posterior distribution that is based on all of the data observed thus far.

\*By using the same sensitivity and specificity rates, we are assuming that the result of the second test does not depend on the result of the first test (i.e., the tests are independent events). It is possible that this simplifying assumption may not be entirely appropriate, since some individuals may be more prone to false-positive results or false-negative results (negative test result when the individual actually has HIV).

## Extended Activity

### Bayesian Model Updating

Use the counts in Table 11.6 to answer the following questions.

25. Find the posterior probability of a donor having HIV given that the individual has tested positive twice on the ELISA test; that is, find  $P(\text{HIV} \mid + \text{ 2nd Test})$ .
26. The ELISA test is inexpensive, so it could be run a third time to reduce the chance of unnecessarily alarming donors even further. Starting from the results above after two ELISA tests, use the Bayesian model-updating approach to determine the posterior probability that a randomly chosen prescreened donor actually has HIV after three positive ELISA results. Assume that the same sensitivity and specificity rates apply.

Based on your answer to Question 25, you can check that there is still about a 10% chance that a donor does not have HIV after testing positive twice ( $1 - 0.898 = 0.102$ ), but this is substantially lower than the 88.25% chance of needlessly alarming a healthy donor based on just one test. Furthermore, from your answer to Question 26, you should observe that after three positive ELISA results, the chance that a donor does not have HIV is much smaller.

The ELISA example illustrated the concept of refining a probability calculation based on additional information. In the next section, we will apply Bayesian methods to parapsychology studies. In addition, we will investigate continuous prior distributions and how to select a prior distribution for a parameter.

## 11.8 Ganzfeld Experiments: Continuous Prior Distributions for $\pi^*$

Are you a skeptic, or are you a believer in psychic abilities such as telepathy (the ability to transfer information on thoughts or feelings between individuals without using the five known senses) or clairvoyance (the ability to obtain information about an object, person, location, or physical event through means other than the five senses)? These abilities, and others like precognition and retrocognition, are forms of extrasensory perception (ESP). A 2007 poll conducted by the Associated Press and Ipsos (a global market research company) found that approximately 48% of Americans believe in some form of ESP.<sup>7</sup> Since many individuals have some prior beliefs about the possible existence of psychic abilities, parapsychology studies lend themselves well to Bayesian methods. Parapsychologists who study paranormal psychic phenomena such as ESP claim that controlled experiments, such as **ganzfeld experiments**, provide compelling evidence for the existence of ESP. In the following extended activities, you will incorporate your subjective beliefs into an estimate of the probability that an individual can correctly identify a target image that has been “sent” by telepathic means.

In a typical ganzfeld experiment, a receiver is seated in a room and put into a state of mild sensory deprivation. For example, the receiver typically places halved ping-pong balls over the eyes, is seated under a red light, and wears a set of headphones through which white or pink noise (static) is played. Meanwhile a sender observes a randomly selected target image (a static photo or a short movie segment) for 15–30 minutes and then attempts to mentally send this information to the receiver. The receiver describes what can be seen, and this information is recorded by a third individual who is blind to the target. The recorded information is used to help the receiver during the judging procedure. In the judging procedure, the receiver is removed from the ganzfeld state and is asked to view a set of possible targets (e.g., a set of four pictures) and decide which one most resembles the images he or she witnessed. Choosing the correct target picture or video segment is treated as a hit. We can characterize the ganzfeld experiment in the following way:

- Since there are typically three decoys included with the true target, the expected overall hit rate is 25%, assuming there is no ESP occurring.
- Let  $X$  = the number of hits in a ganzfeld experiment involving  $n$  sessions ( $n$  targets to hit) and  $\pi$  = the probability of a hit (i.e., the probability that the receiver correctly identifies the target for each session). If the expected hit rate is 25% per session, then  $X$  can be modeled with a binomial distribution with  $\pi = 0.25$ .

From what we have seen so far about the Bayesian approach, we know to start with a prior distribution for  $\pi$  that reflects our beliefs about the possible values  $\pi$  can take, and then update it based on observed data to obtain the posterior distribution for  $\pi$  given the data. But what prior distribution should we use for  $\pi$ ? In our earlier discussion of the proportion of brown or orange M&M's, the prior distribution for  $\pi$  (the true

<sup>\*</sup>Calculus is suggested for this section, but not required.

proportion of brown or orange M&M's) was discrete (i.e., the potential values of  $\pi$  could be listed), resulting in a discrete posterior distribution as well. We certainly want to be able to assign probabilities for more than just a few prior choices for  $\pi$ . For an even more flexible Bayesian analysis, we would like to allow the prior distribution for  $\pi$  to include many more possibilities for  $\pi$  in the interval from 0 to 1. In the next section, we will consider continuous prior distributions for  $\pi$  that allow for these possibilities.

Before we discuss continuous priors for  $\pi$ , let's briefly review some characteristics of continuous density functions. You probably had some experience working with continuous probability distributions such as the normal and uniform distributions in your first statistics course, but you may not have studied properties of the **probability density function**,  $p(x)$ , that is used to find the probability that the continuous random variable is between two values.

If  $p(x)$  is a valid continuous density function, then the following conditions must be satisfied:

- $p(x) \geq 0$  for all values of  $x$
- $\int_{\text{all } x} p(x)dx = 1$

When working with continuous prior distributions, we will use a version of Bayes' rule given in Equation (11.7) slightly modified for continuous prior distributions. Let the continuous posterior distribution of  $\pi$  given the data  $x$  be denoted  $p(\pi | x)$ . Then the posterior distribution of  $\pi$  given the sample data  $x$ ,  $p(\pi | x)$ , is given by

$$p(\pi | x) = \frac{p(x | \pi)p(\pi)}{p(x)} \quad (11.18)$$

where

- $p(\pi)$  is the prior distribution assigned to  $\pi$ .
- $p(x | \pi)$  is the likelihood of observing the data given success probability  $\pi$ .
- $p(x) = \int_{\text{all } \pi} p(x | \pi)p(\pi)d\pi$  is the average likelihood weighted by the prior distribution, sometimes referred to as the marginal likelihood.

In more general terms, Bayes' rule given by Equation (11.18) can be expressed as

$$p(\text{parameter} | \text{data}) = \frac{p(\text{data} | \text{parameter})p(\text{parameter})}{p(\text{data})} \quad (11.19)$$

The marginal likelihood can be difficult to compute by hand; computer software may be required to perform the integration. However, under some simplifying assumptions, it is not always necessary to compute this quantity.

## The Uniform Prior

Let's assume that we want to be as objective as possible about potential values of  $\pi$ , so we will impose very few restrictions on the prior distribution for  $\pi$ . We know only that  $\pi$  must be between 0 and 1, and we'll assume that the chance that  $\pi$  takes a value within a particular interval in  $[0, 1]$  will be identical to the chance that  $\pi$  is within another interval of equal length. This means, for example, that the probability that  $\pi$  is between 0.1 and 0.2 is the same as the probability that  $\pi$  takes a value between 0.25 and 0.35.

These prior beliefs suggest that the uniform probability distribution over the interval from 0 to 1 could be a potential prior distribution for  $\pi$ :

$$p(\pi) = \begin{cases} 1 & \text{for } 0 \leq \pi \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

This is a flat **objective prior distribution**, also called a **non-informative prior distribution** for  $\pi$ , and it allows  $\pi$  to take on *any* value between 0 and 1 prior to collecting data. This is an appropriate prior distribution to use if we have little or no prior knowledge about  $\pi$ .

If you have studied the uniform distribution on the interval from  $a$  to  $b$ , then you may recall that the density function for a uniform random variable  $Y$  on  $[a, b]$  is given by

$$p(y) = \begin{cases} \frac{1}{b - a} & \text{for } a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

and that the mean and variance of  $Y$  are given by

$$E(Y) = \frac{a + b}{2} \text{ and } \text{Var}(Y) = \frac{(b - a)^2}{12} \quad (11.20)$$

## Extended Activity

Suppose the uniform prior density function on the interval  $[0, 1]$  is assigned to the proportion of hits  $\pi$ .

27. Sketch the density curve for this prior distribution. Provide the values that  $\pi$  can take along the horizontal axis, and the height of the curve on the vertical axis.
28. Compute the prior estimate for  $\pi$ ,  $E(\pi)$ , the mean of the prior distribution for  $\pi$ . Also compute the variance of the prior distribution,  $\text{Var}(\pi)$ .

With our prior distribution for  $\pi$  in place, we can determine the posterior distribution for  $\pi$  once we've observed the results of a ganzfeld experiment—i.e., a particular number of hits (correctly identified targets) in a fixed number of sessions. The likelihood of observing  $x$  hits in  $n$  sessions, given the true proportion of hits  $\pi$ , is binomial. Then, from Equation (11.18), the posterior probability density function for  $\pi$  ( $0 \leq \pi \leq 1$ ), given the observed data, is

$$\begin{aligned} p(\pi | x) &= \frac{p(x | \pi)p(\pi)}{\int_0^1 p(x | \pi)p(\pi)d\pi} = \frac{\binom{n}{x}\pi^x(1 - \pi)^{n-x}}{\int_0^1 \binom{n}{x}\pi^x(1 - \pi)^{n-x}d\pi} \\ &= \frac{\pi^x(1 - \pi)^{n-x}}{\int_0^1 \pi^x(1 - \pi)^{n-x}d\pi} \end{aligned} \quad (11.21)$$

Note that  $\pi$  in the numerator and denominator of Equation (11.21) have slightly different meanings. In the numerator,  $\pi$  is a fixed value for the proportion of all brown or orange M&M candies, while in the denominator we are integrating over all possible values of  $\pi$ . Let's suppose that, in a carefully controlled ganzfeld experiment, 18 hits were observed in 50 sessions. That's an observed hit rate of  $\hat{p} = 0.36$ . In light of the observed data, we'll investigate whether prior beliefs about psychic abilities are altered by computing the posterior estimate of  $\pi$ .

## Extended Activity

### Constructing the Posterior Distribution

29. Suppose that the prior distribution for  $\pi$  is uniform on the interval  $[0, 1]$ , and consider a particular ganzfeld experiment where 18 hits were observed in 50 sessions. Set up, *but do not evaluate*, the expression for the posterior distribution for  $\pi$  given in Equation (11.21) if 18 hits were observed in 50 sessions.

Before we take a closer look at the posterior distribution that you set up in Question 29, let's look at the numerator in the expression—that is, the likelihood function, which was found to be

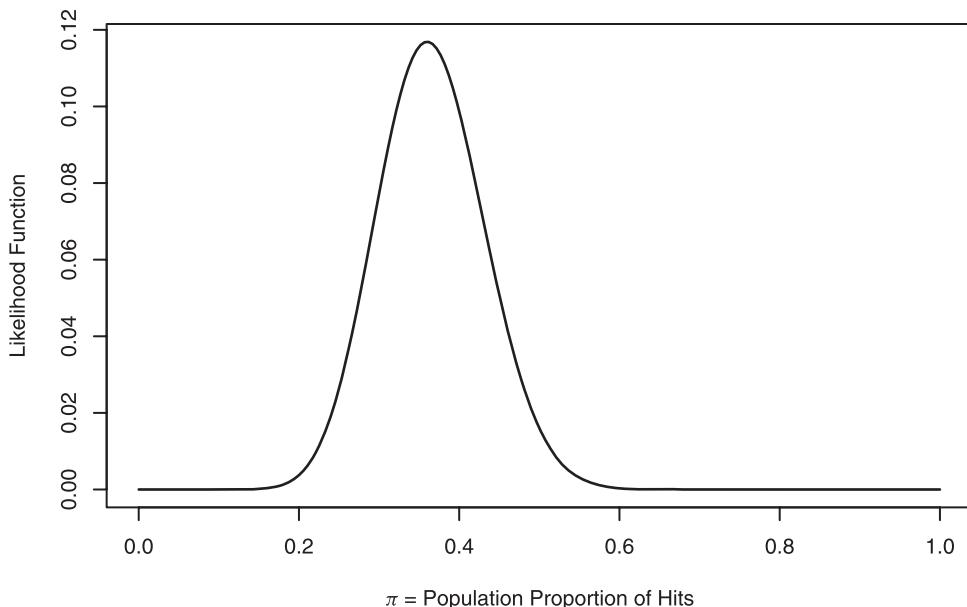
$$p(x | \pi) = \binom{50}{18} \pi^{18}(1 - \pi)^{32} \quad (11.22)$$

We can examine the graph of  $p(x | \pi)$  in Figure 11.3. The height of the curve in Figure 11.3 is found by plugging values of  $\pi$  between 0 and 1 into Equation (11.22). Note that the curve is *not* a density curve—that is, the total area under the curve does not equal one. The function tells us that if we observed 18 hits in 50 sessions, then likely values of  $\pi$  are between roughly 0.2 and 0.6, and the most likely value of  $\pi$  occurs where the peak of the curve is highest, at 0.36. Hence, the likelihood function is maximized at the observed hit rate.

Although this is certainly an estimate of  $\pi$ , it is not the posterior estimate,  $p^*$ . To find  $p^*$ , we first need the posterior distribution for  $\pi$ . The full posterior distribution you set up in Question 29 is given by

$$p(\pi | x) = \frac{\pi^{18}(1 - \pi)^{32}}{\int_0^1 \pi^{18}(1 - \pi)^{32}d\pi} \quad (11.23)$$

<sup>\*</sup>The estimate 0.36 is known as a maximum likelihood estimate of  $\pi$ . Maximum likelihood estimates, introduced in Chapters 7 and 8, are the values of the parameter that maximize the likelihood function.



**Figure 11.3** The likelihood function,  $p(x | \pi)$ , over possible values of  $\pi$ , given that 18 hits were observed in 50 sessions.

and we can show (details will follow) that the mean of this posterior distribution (i.e., the posterior estimate of  $\pi$ ) is given by

$$p^* = E(\pi | x) = 19/52 = 0.365 \quad (11.24)$$

Recall that the prior estimate of  $\pi$  (the mean of the prior distribution for  $\pi$ ) found in Question 28 was 0.5. This meant that, without any prior beliefs or data, the midpoint between 0 and 1 was taken as the estimate. It is not surprising that, after observing much fewer than half of the targets hit, the prior estimate of  $\pi$  has been down-weighted to 0.365.

It may seem like a daunting task to move from the posterior distribution in Equation (11.23) to the posterior mean in Equation (11.24). However, we will see that there are several simplifying steps that can be taken to avoid complex mathematics.

## The Beta Distribution

The integral in the denominator of the posterior distribution in Equation (11.21) can be difficult to compute, but it turns out that its value will be a constant that does not depend on  $\pi$ . Therefore, another way to express the posterior density function is

$$p(\pi | x) = C\pi^x(1 - \pi)^{n-x} \quad 0 \leq \pi \leq 1 \quad (11.25)$$

where  $C$  is the **normalizing constant** that will ensure that the posterior density  $p(\pi | x)$  integrates to 1.

Furthermore, as we'll see shortly, we typically don't need to compute  $C$ , so we can express the posterior distribution in an even more general form:

$$p(\pi | x) \propto \pi^x(1 - \pi)^{n-x} \quad 0 \leq \pi \leq 1 \quad (11.26)$$

where the symbol  $\propto$  means “is proportional to” or “are equal up to a constant multiplier.” At this point we can see that the posterior distribution  $p(\pi | x)$  is proportional to the likelihood function  $p(x | \pi)$ . This proportional relationship is important because it means that we will be able to find the general form of the posterior distribution without doing any complicated integration to find  $C$ .

Equation (11.26) is an example of the general approach taken in Bayesian statistics to find the basic form of the posterior distribution,  $p(\pi | x)$ :

$$p(\pi | x) \propto p(x | \pi) \times p(\pi) \quad (11.27)$$

or, more basically,

$$p(\text{parameter} \mid \text{data}) \propto p(\text{data} \mid \text{parameter}) \times p(\text{parameter})$$

Then much of the work to find the posterior distribution is concentrated on computing  $p(\text{data} \mid \text{parameter}) \times p(\text{parameter})$ , which is the numerator of Equation (11.19).

Now, the form of the posterior probability density in Equation (11.26) may be new to you even if you have had both a first course in statistics and a calculus class. However, we will show that Equation (11.26) has the basic form of a common probability distribution called the **beta distribution**. We say that a random variable  $Y$  has a beta( $\alpha, \beta$ ) probability distribution [the notation beta( $\alpha, \beta$ ) is read “beta distribution with parameters  $\alpha$  and  $\beta$ ”] if its probability density function is

$$p(y) = \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} y^{\alpha-1}(1 - y)^{\beta-1} \quad 0 \leq y \leq 1 \quad (11.28)$$

for integers  $\alpha > 0$  and  $\beta > 0$ . There is a more general form for the beta probability density when  $\alpha$  and  $\beta$  are not integers, but we will use only the simpler form for integers, shown in Equation (11.28).<sup>8</sup>

Although the formula for the beta density function in Equation (11.28) may at first appear a bit intimidating, the quantity

$$\frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} \quad (11.29)$$

in the equation is only a constant that ensures that the density function  $p(y)$  integrates to 1. Hence, the density function of a beta( $\alpha, \beta$ ) random variable  $Y$  is proportional to  $y^{\alpha-1}(1 - y)^{\beta-1}$ ; that is,

$$p(y) \propto y^{\alpha-1}(1 - y)^{\beta-1} \quad 0 \leq y \leq 1$$

Now let's re-examine the posterior distribution for  $\pi$  in Equation (11.26). You can see that (ignoring the constant out front) the posterior distribution for  $\pi$  also has the same form as a beta density, since

$$\begin{aligned} p(\pi \mid x) &\propto \pi^x(1 - \pi)^{n-x} \\ &= \pi^{(x+1)-1}(1 - \pi)^{(n-x+1)-1} \quad 0 \leq \pi \leq 1 \end{aligned}$$

So, starting with a uniform prior distribution on  $[0, 1]$ , the posterior probability distribution for  $\pi$  has a beta distribution with parameters  $\alpha = x + 1$  and  $\beta = n - x + 1$ . This is an important property of our posterior distribution for  $\pi$  because it will facilitate calculation of the posterior estimate of  $\pi$ .

## Extended Activity ▶ Finding the Beta Parameter Values

30. For the beta posterior distribution that you found in Question 29, determine the values for the parameters  $\alpha$  and  $\beta$ .

Earlier it was suggested that the mean of the posterior distribution,  $E(\pi \mid x)$ , be used as the Bayesian posterior estimate for  $\pi$ . If the posterior distribution for  $\pi$  is beta, then the posterior mean (the Bayesian posterior estimate for  $\pi$ ) is

$$p^* = E(\text{model} \mid \text{data}) = E(\pi \mid x) = \frac{\alpha}{\alpha + \beta} = \frac{x + 1}{(x + 1) + (n - x + 1)} = \frac{x + 1}{n + 2} \quad (11.30)$$

### NOTE

It can be shown that if a random variable  $Y$  has a beta( $\alpha, \beta$ ) distribution, then the mean and variance of  $Y$  are

$$E(Y) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var}(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (11.31)$$

## Extended Activity

### Computing the Posterior Estimate of $\pi$

31. Suppose that the prior distribution for  $\pi$  is uniform on the interval  $[0, 1]$  and that 18 hits were observed in 50 sessions in a ganzfeld experiment. Compute the posterior estimate of the proportion of hits,  $\pi$ , when the uniform prior density is used. How did your answer change from the prior estimate of  $\pi$  that you found in Question 28?

You may have noticed that your Bayesian posterior estimate of  $\pi$  is not much different from the frequentist estimate  $\hat{p} = x/n = 18/50$ . This makes sense, since the uniform prior incorporates very little information about  $\pi$  and does not favor any particular values of  $\pi$  more than others. If we want to include more beliefs in the prior distribution for  $\pi$  (i.e., how “much” we believe in ESP abilities), then we will need to use a more flexible “collection” of prior distributions.

#### Key Concept

If we assign a uniform prior distribution (on  $[0, 1]$ ) for the true proportion of hits  $\pi$  and assume that the likelihood of observing  $x$  hits in  $n$  sessions given  $\pi$  is a binomial distribution, then the posterior distribution for  $\pi$  is beta with parameters  $\alpha = x + 1$  and  $\beta = n - x + 1$ . In addition, the posterior estimate of  $\pi$  is  $E(\pi | x) = \frac{x + 1}{n + 2}$ .

## Extended Activity

### Relationship Between the Uniform and Beta Distributions

32. Plug values of  $\alpha = 1$  and  $\beta = 1$  into the formula for the beta density given in Equation (11.28). What expression do you get for the density function? Recall that  $0! = 1$ .
33. What does this imply about a random variable that is uniform on  $[0, 1]$ ? In other words, what is another name for the uniform prior distribution on the interval  $[0, 1]$ ?

$$p(\pi) = \begin{cases} 1 & \text{for } 0 \leq \pi \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

### Conjugate Priors

As you saw in Question 33, the prior distribution  $p(\pi) = 1$  on  $0 \leq \pi \leq 1$  is also a beta density with parameters  $\alpha = 1$  and  $\beta = 1$ . So does this mean if we use *any* beta density for the prior distribution (and a binomial distribution for the likelihood function), the posterior distribution will also be beta? The answer is yes! Let’s see how this works. Suppose that the prior distribution for  $\pi$  is beta with parameters  $\alpha$  and  $\beta$ , and the likelihood function  $p(x | \pi)$  is the binomial distribution. Then, using Equation (11.27), we have

$$\begin{aligned} p(\pi | x) &\propto p(x | \pi) \times p(\pi) \\ &= \pi^x (1 - \pi)^{n-x} \times \pi^{\alpha-1} (1 - \pi)^{\beta-1} \\ &= \pi^{(x+\alpha)-1} (1 - \pi)^{(n-x+\beta)-1} \end{aligned} \tag{11.32}$$

Hence, if the prior distribution is  $\text{beta}(\alpha, \beta)$  and the likelihood function is the binomial distribution, then the posterior distribution is  $\text{beta}(x + \alpha, n - x + \beta)$ .

Although the parameters of the posterior distribution are not the same as those of the prior distribution, the form of the density function is the same. A prior distribution that results in a posterior distribution with the same form is called a **conjugate prior**. In the past, conjugate priors were extremely convenient to use because they did not require difficult computations to find the posterior distribution and characteristics of the distribution like the mean or standard deviation. Presently, this is less of a concern because of software and computing power, but it can still be helpful when updated posterior quantities are needed quickly.

Then if a beta( $\alpha, \beta$ ) prior is used for  $\pi$ , by Equation (11.31) we have the following results:

- The prior estimate of  $\pi$  (i.e., the mean of the prior distribution for  $\pi$ ) and the variance of the prior distribution for  $\pi$  are given by

$$E(\pi) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var}(\pi) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

- The posterior estimate of  $\pi$  and the variance of the posterior distribution are given by

$$p^* = E(\pi | x) = \frac{\alpha + x}{\alpha + \beta + n} \quad \text{and} \quad \text{Var}(\pi | x) = \frac{(x + \alpha)(n - x + \beta)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

From Section 11.2 you may recall a very similar expression for the posterior estimate of  $\pi$ , given in Equation (11.2). This simple posterior estimate was basically a weighted average of the frequentist estimate  $\hat{p} = x/n$  and the prior estimate  $\hat{\pi} = a/(a + b)$ , where  $a$  represented the number of successes and  $b$  was the number of failures. Hence, we can view the beta posterior estimate of  $\pi$  as a weighted average of the frequentist estimate  $\hat{p} = x/n$  and the beta prior estimate given by  $E(\pi) = \alpha/(\alpha + \beta)$ .

### Key Concept

If the prior distribution for  $\pi$  is beta ( $\alpha, \beta$ ) and the likelihood function is binomial with  $x$  successes in  $n$  trials, then the posterior distribution of  $\pi$  is also beta with parameters  $x + \alpha$  and  $n - x + \beta$ . When the prior and posterior distributions have the same form, the prior is called a conjugate prior.

### ► MATHEMATICAL NOTE ▶

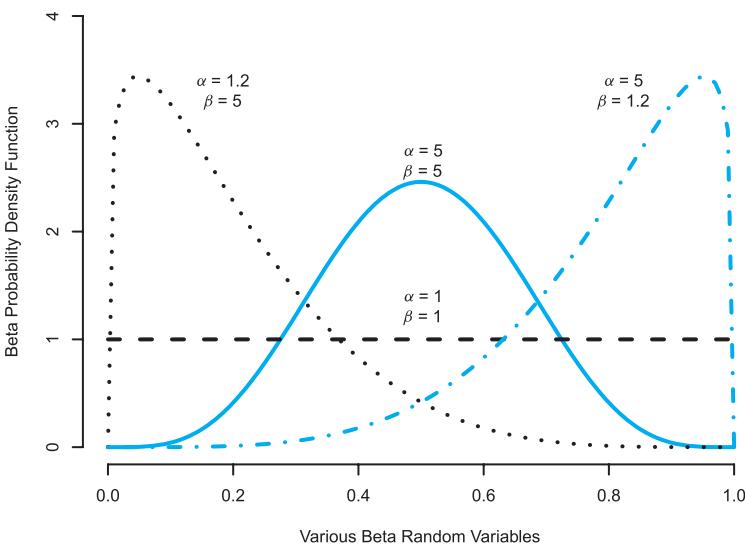
The existence of a conjugate prior distribution depends on the form of the likelihood function, which, in turn, depends on the original distribution of the data  $x$ . For example, if the data are normal with unknown mean  $\mu$  and known variance  $\sigma^2$ , then it can be shown that the normal distribution is also a conjugate prior.<sup>9</sup>

## Choosing a Prior Probability Distribution for $\pi$

Figure 11.4 displays the variety of shapes attainable by beta probability density curves. When  $\alpha$  and  $\beta$  are the same, the distribution is symmetric. When  $\alpha > \beta$ , the distribution is skewed left, and the opposite holds when  $\alpha < \beta$ . The larger  $\alpha$  and  $\beta$  are, the smaller the spread in the distribution. As we can observe, there is quite a bit of flexibility in the shape of the beta distribution, which allows for a variety of prior models to be explored.

Since we know that the beta density is convenient to work with, we will assign a beta density to the prior distribution for  $\pi$ . But this still leaves us with the task of deciding what parameter values to use for  $\alpha$  and  $\beta$ . This is the tricky part and will depend on our degree of belief in psychic abilities—that is, what we think reasonable values of the hit rate  $\pi$  should be. So first ask yourself: Are you a skeptic, an open-minded individual, or a believer in ESP?

We have already stated that if no ESP is occurring, then the proportion of hits in  $n$  sessions of a ganzfeld experiment is expected to be  $\pi = 0.25$ . Now if ESP occurs, then the expected hit rate should be higher than chance alone, and  $\pi > 0.25$ . Based on your prior beliefs, what are some likely or unlikely values for  $\pi$ ? One reasonable strategy is to consider two values of the hit rate  $\pi$ : a value that we believe  $\pi$  is not too likely to fall below and another value that we believe  $\pi$  is not too likely to fall above. In other words, we determine minimum and maximum values of  $\pi$  such that there is small probability that  $\pi$  is beyond these two values. Then based on these restrictions, we can determine the values of the parameters of the beta prior distribution that give rise to these two values of  $\pi$ .



**Figure 11.4** Several beta( $\alpha, \beta$ ) probability density functions.

To keep things simple, we will consider minimum and maximum values of  $\pi$ , denoted by  $\pi_{\min}$  and  $\pi_{\max}$ , for which values outside are not very likely to occur. So our beliefs about the beta prior distribution are summarized in the following two statements:

- We believe there is only a 10% chance that  $\pi$  is less than  $\pi_{\min}$ .
- We believe there is only a 10% chance that  $\pi$  is greater than  $\pi_{\max}$ .

The 10% is somewhat arbitrary, and it certainly is not necessary to make the lower tail and upper tail probabilities identical; however, for computational ease, we will set them equal to each other. Based on this approach for selecting low and high values of  $\pi$ , the following  $\pi_{\min}$  and  $\pi_{\max}$  values might be assigned to the skeptic, open-minded individual, and believer:

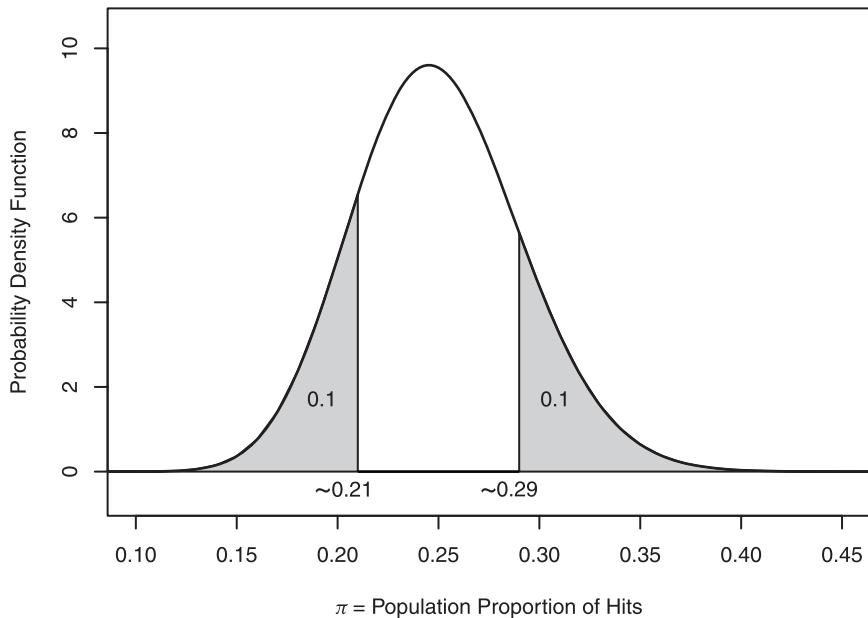
- The skeptic doesn't believe that the hit rate  $\pi$  should be too much lower than 0.25 or too much higher than 0.25, so  $\pi_{\min} = 0.21$  and  $\pi_{\max} = 0.29$ .
- The open-minded individual doesn't believe that  $\pi$  should be too much lower than 0.25 but is definitely open to the possibility that  $\pi$  could be higher than 0.25, so  $\pi_{\min} = 0.21$  and  $\pi_{\max} = 0.5$ .
- The believer is fairly convinced that  $\pi$  is higher than 0.25 and could be *much* higher than 0.25, so  $\pi_{\min} = 0.3$  and  $\pi_{\max} = 0.7$ .

Once the values of  $\pi_{\min}$  and  $\pi_{\max}$  have been determined, the accompanying software instructions can be followed to find the *approximate* values of the parameters of the beta distribution corresponding to your selected minimum and maximum values of  $\pi$ . They are approximate because we are considering only integer values of  $\alpha$  and  $\beta$ . For the skeptic, for example, the (approximate) parameter values for the beta prior distribution are  $\alpha = 27$  and  $\beta = 81$ .

## Extended Activity Using Software to Determine the Parameters of the Prior Distribution

34. Use the accompanying software instructions to verify that the parameters for the beta prior distribution for the skeptic are  $\alpha = 27$  and  $\beta = 81$ . Note: You do not need to submit any output for this problem.  
The exercise is intended to introduce you to the software code.

This distribution corresponds to the skeptic's prior beliefs that there is approximately 0.10 area under the beta(27, 81) density curve to the left of  $\pi = 0.21$  and 0.10 area under the curve to the right of  $\pi = 0.29$ . Figure 11.5 displays the beta distribution with these parameters.



**Figure 11.5** The skeptic's prior distribution for the proportion of hits in a ganzfeld experiment: a beta prior distribution with parameters  $\alpha = 27$  and  $\beta = 81$ .

Since the prior distribution is beta, the prior estimate of  $\pi$  is

$$\begin{aligned} E(\pi) &= \frac{\alpha}{\alpha + \beta} \\ &= \frac{27}{27 + 81} \\ &= 0.25 \end{aligned}$$

This makes sense to the skeptic, since (on average) the hit rate shouldn't be any different from what would be expected by random chance. Furthermore, the variance of the prior distribution is

$$\begin{aligned} \text{Var}(\pi) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\ &= \frac{(27)(81)}{(27 + 81)^2(27 + 81 + 1)} \\ &= 0.0017 \end{aligned}$$

If we assign a beta(27, 81) prior distribution to  $\pi$ , then since the beta distribution is a conjugate prior, we know by Equation (11.32) that the posterior distribution will also be a beta distribution. The values of the parameters for the beta posterior distribution will be  $x + 27$  and  $n - x + 81$ , where  $x$  will be the observed number of hits in  $n$  sessions.

Again consider a ganzfeld experiment in which we observe 18 hits in 50 sessions for a hit rate of  $18/50 = 0.36$ . Then for the skeptic, the parameter values of the beta posterior distribution will be

$18 + 27 = 45$  and  $50 - 18 + 81 = 113$ . Hence, the posterior estimate of  $\pi$ , after observing the results of a ganzfeld experiment, is

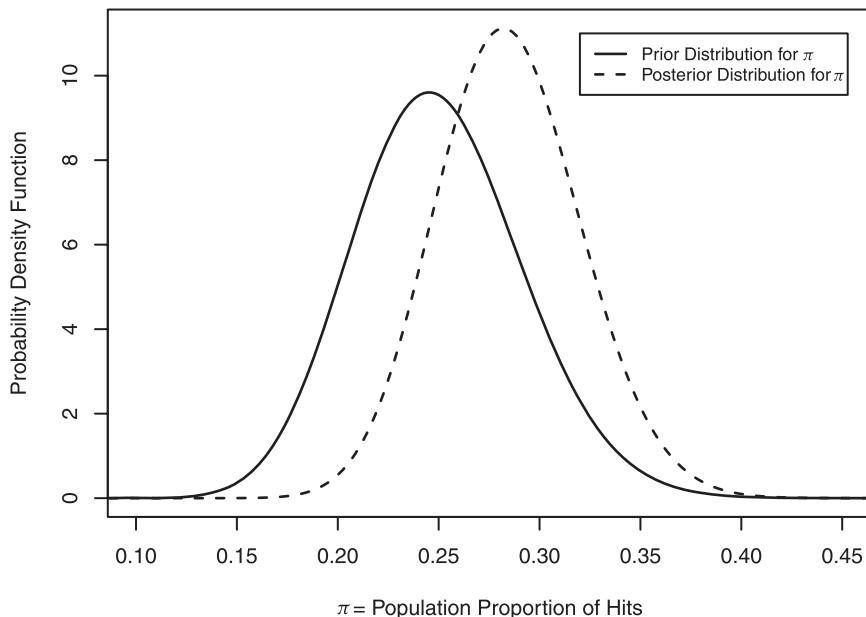
$$\begin{aligned} p^* &= \frac{\alpha + x}{\alpha + \beta + n} \\ &= \frac{27 + 18}{27 + 81 + 50} \\ &= 0.285 \end{aligned}$$

and the variance of the posterior distribution is

$$\begin{aligned} \text{Var}(\pi | x) &= \frac{(x + \alpha)(n - x + \beta)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \\ &= \frac{(18 + 27)(50 - 18 + 81)}{(27 + 81 + 50)^2(27 + 81 + 50 + 1)} \\ &= 0.0013 \end{aligned}$$

Observe that the posterior variance is less than the prior variance.

The prior and posterior distributions for the skeptic are displayed in Figure 11.6. Notice that the posterior distribution shifted only slightly higher than the prior distribution.



**Figure 11.6** The skeptic's prior and posterior distribution for the proportion of hits in a ganzfeld experiment after observing 18 hits in 50 sessions.

### Extended Activity

35. For the skeptic, how does the posterior estimate change if we observed 9 hits in 25 sessions instead of 18 hits in 50 sessions? Find the posterior estimate if we observed 36 hits in 100 sessions. What pattern do you see emerging in the posterior estimates as the numbers of hits and sessions increase, even though the observed hit rate remains constant at 0.36?

From Question 35, you should have observed that with a relatively small amount of observed data, our posterior estimate did not change much from the prior estimate. As the amount of observed data increases, the posterior estimate will get further from the prior estimate.

## Extended Activity

### ► Computing the Prior and Posterior Estimates for the Open-Minded Individual and Believer

36. Use the accompanying software instructions to find the parameters of the beta prior distribution that correspond to the open-minded individual and the believer.
37. Assuming 18 hits are observed in 50 sessions, use the parameters that you found in Question 36 to find the posterior estimates of  $\pi$  for the open-minded individual and the believer.
38. Plot the prior and posterior distributions for the open-minded individual on the same graph and then do the same for the believer. Comment on the differences between the prior and posterior distributions—that is, how did the center and variability change?
39. For which individual (skeptic, open-minded individual, or believer) did the posterior estimate of  $\pi$  change the most from the prior estimate?

Based on your responses to Questions 38 and 39, you were able to examine how the perceptions of skeptics, open-minded individuals, and believers in psychic abilities changed after observing the data from a ganzfeld experiment. For the skeptic and the open-minded individual, the posterior estimates of the true proportion of hits were only slightly higher than prior estimates; however, the posterior estimate of  $\pi$  decreased for the believer (as well as changed the most). We conclude that, after observing the results of the ganzfeld study, the perceptions of the skeptic and the open-minded individual were not swayed very much; however, the believer may have less faith in psychic abilities.

Although the mean of the posterior distribution provides an updated estimate of  $\pi$  after the data are observed, it would be useful to have a range of plausible values for  $\pi$ . This can be achieved by constructing an interval of values for  $\pi$  based on the posterior distribution; the process is discussed in the next section.

## 11.9 Return to M&M's: Bayesian Credible Intervals

When you think back to the first time you worked with confidence intervals in your introductory statistics course, you may have been a bit annoyed by their slightly awkward interpretation. You learned that the confidence level—for example, 95%—is *not* the probability that the interval contains the unknown parameter value. It seems as if it would make more intuitive sense to say something about the likelihood of the value of the parameter based on data that we have already observed. With the tools you learned in your first statistics course, you couldn't do this, but with the techniques covered in this chapter, you will be able to construct intervals that contain a parameter with a specified probability.

Now that we have covered some additional topics in Bayesian statistics, we can continue our discussion of the proportion of brown or orange M&M's. In this section, we will develop an interval for  $\pi$ , called a **Bayesian credible interval**, to which we can assign a probability that  $\pi$  will reside in the interval.

As we've seen, the beta distribution is a good choice for a prior for  $\pi$ , since the posterior estimates can be easily computed. Let's start with the beta prior distribution with parameters  $\alpha = 1$  and  $\beta = 1$  (i.e., the uniform distribution on the interval  $[0, 1]$ ) for the prior distribution for  $\pi$ . This prior incorporates the least amount of prior information about  $\pi$ . Later we'll see how our intervals change depending on the choice of prior.

## Extended Activity

### ► A Uniform Prior for the Proportion of Brown or Orange M&M's

Data set: MyMMs or MMs

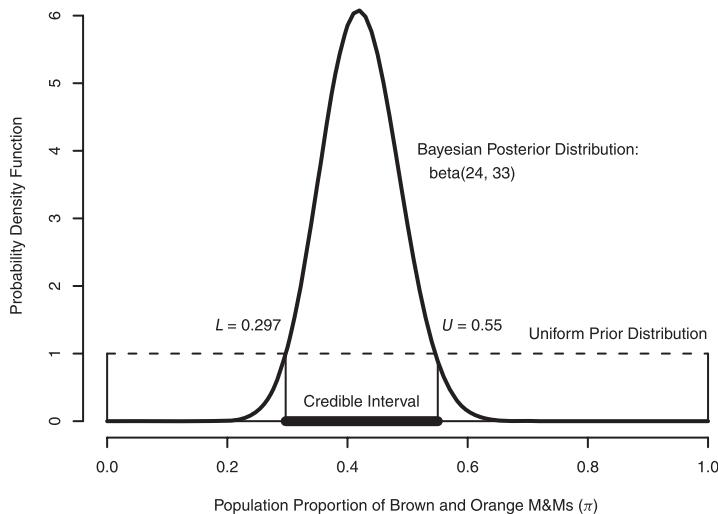
40. Use the data you collected on the number of brown or orange M&M's in your bag and assume a uniform prior distribution for  $\pi$  to determine the posterior distribution for  $\pi$  and calculate the Bayesian posterior estimate for  $\pi$ .

Once we have the posterior distribution for  $\pi$ , we can use it to construct an interval estimate for  $\pi$ . In particular, we can find an interval  $(L, U)$  such that  $P(L < \pi < U | \text{data}) = 0.95$ . This is called a 95% Bayesian credible interval or a Bayesian posterior interval, since the limits are found from the posterior distribution for  $\pi$  given that we have observed the data.

There are infinitely many ways to determine the values for the lower and upper limits,  $L$  and  $U$ , of a Bayesian posterior interval for  $\pi$  that will all satisfy  $P(L < \pi < U | \text{data}) = 0.95$ . However, we will find an interval that sets  $P(\pi < L | \text{data}) = 0.025$  and  $P(\pi > U | \text{data}) = 0.025$ .

Let's assume that we used a uniform prior for  $\pi$ . Then for the sample data,  $x = 23$  and  $n = 55$ , the Bayesian posterior probability density is beta with parameters  $x + 1 = 24$  and  $n - x + 1 = 33$ . We can use software such as R or Minitab to find  $L$  and  $U$  such that there is 0.025 probability in the lower and upper tails of the posterior probability distribution for  $\pi$ . For example, if the posterior distribution for  $\pi$  is beta(24, 33), then  $L = 0.297$  and  $U = 0.550$ .

Hence, a 95% Bayesian credible interval for  $\pi$  is  $(0.297, 0.550)$ . Figure 11.7 shows the uniform prior probability density, the beta posterior density, and the 95% Bayesian credible interval after observing data  $x = 23$  and  $n = 55$ . Notice that with this interval, we can state that there is a 95% chance (or 0.95 probability) that  $\pi$  is between 0.297 and 0.550, given that we have observed 23 brown or orange M&M's in a sample of 55 candies.



**Figure 11.7** The uniform prior probability density, Bayesian posterior density, and 95% Bayesian credible interval for population proportion  $\pi$  after observing data  $x = 23$  and  $n = 55$ .

### NOTE

You might recall that the confidence intervals for  $\pi$  that you studied in your first statistics course did not have such a clear interpretation. You stated with “95% confidence” that the true proportion  $\pi$  was captured by (or contained within) the interval limits, but it would have been incorrect to state that  $\pi$  fell inside the interval with 0.95 probability.

### Key Concept

A 95% Bayesian credible interval for  $\pi$  is a probability statement about  $\pi$  calculated from the posterior probability distribution

$$P(L < \pi < U | \text{data}) = 0.95$$

One way to choose the endpoints  $L$  and  $U$  is to satisfy  $P(\pi < L | \text{data}) = 0.025$  and  $P(\pi > U | \text{data}) = 0.025$ .

## Extended Activity

### Bayesian Credible Intervals for Your M&M's Data Using a Uniform Prior

Data set: MyMMs or MMs

41. From the data that you collected on the number of brown or orange M&M's in your bag, use computer software to calculate a Bayesian credible interval for  $\pi$  using the objective (uniform) prior. Use the method that sets the lower and upper limits of your interval such that  $P(\pi < L | \text{data}) = 0.025$  and  $P(\pi > U | \text{data}) = 0.025$ .
42. Return to your classical frequentist training in statistics for just a moment. Using only your M&M's data, calculate an approximate 95% confidence interval for  $\pi$  of the form

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

43. You should find that the confidence interval is very similar to your Bayesian credible interval for  $\pi$ . Explain why this is not surprising when the uniform prior distribution for  $\pi$  is used.
44. Provide interpretations of the Bayesian credible interval you constructed in Question 41 and the confidence interval you constructed in Question 42. Which interpretation seems more natural to you?
45. At the start of this discussion about the proportion of brown and orange M&M's (before we collected any data), we mentioned that some of you or your friends might have felt that the proportion is certainly less than 1/2. From the data that you collected, use computer software to calculate the Bayesian posterior probability that  $\pi < 1/2$  using the objective (uniform) prior.

### Selecting a Beta Prior Distribution for $\pi$

If you are one of those students who believes it is fairly unlikely that the proportion of brown or orange M&M's is less than 0.20 and fairly unlikely that it is more than 0.50 prior to collecting your data, then you may be unsatisfied with the choice of the uniform prior distribution used in the previous section.

Fortunately, to find the parameters of the beta prior, we can use a strategy similar to the one that was used to determine the parameters for the prior distribution for the true proportion of hits in a ganzfeld experiment. For example, we may believe that there is only a small chance (i.e., 10%) that  $\pi$ , the true proportion of brown or orange M&M's, is below 0.2 or above 0.5 (i.e.,  $\pi_{\min} = 0.2$  and  $\pi_{\max} = 0.5$ .) Then using statistical software, the (approximate) parameter values of the beta prior distribution are  $\alpha = 5$  and  $\beta = 10$ . The graph of the beta prior distribution with parameters  $\alpha = 5$  and  $\beta = 10$  is displayed in Figure 11.8. Furthermore, the mean and variance for  $\pi$  prior to collecting any data are

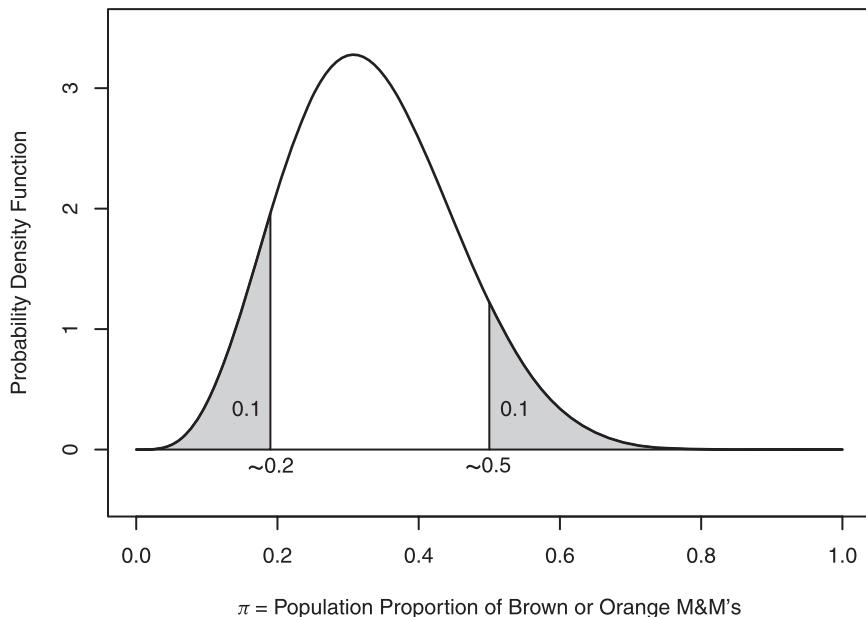
$$E(\pi) = \frac{5}{5 + 10} = \frac{1}{3} = 0.33 \quad \text{and} \quad \text{Var}(\pi) = \frac{5(10)}{(5 + 10)^2(5 + 10 + 1)} = 0.0139$$

Based on this prior distribution for  $\pi$ , we can update our estimate after observing  $x = 23$  brown or orange M&M's in a sample of  $n = 55$  candies. The Bayesian posterior estimate is

$$\begin{aligned} p^* &= \frac{\alpha + x}{\alpha + \beta + n} \\ &= \frac{5 + 23}{5 + 10 + 55} \\ &= 0.4 \end{aligned}$$

and the variance of the posterior distribution is

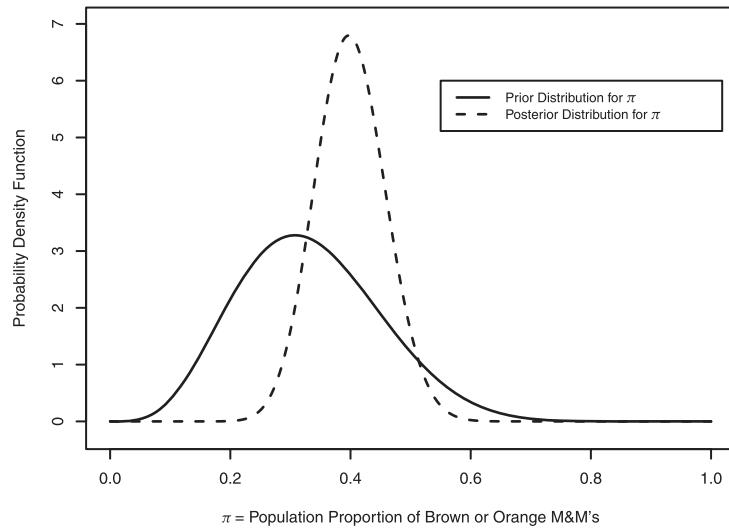
$$\begin{aligned} \text{Var}(\pi | x) &= \frac{(x + \alpha)(n - x + \beta)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \\ &= \frac{(23 + 5)(55 - 23 + 10)}{(5 + 10 + 55)^2(5 + 10 + 55 + 1)} \\ &= 0.0034 \end{aligned}$$



**Figure 11.8** Beta(5, 10) prior distribution for  $\pi$  based on our prior beliefs that  $P(\pi < 0.2) = 0.1$  and  $P(\pi > 0.5) = 0.1$ .

The posterior distribution for  $\pi$  after observing  $x = 23$  brown or orange M&M's in a sample of  $n = 55$  candies is shown in Figure 11.9. Notice how the posterior mean has shifted up toward the observed proportion  $\hat{p} = 23/55$  and the variance of the posterior distribution has decreased.

Let's see how using the beta prior distribution with parameters  $\alpha = 5$  and  $\beta = 10$  changes the posterior estimate and the 95% credible interval for  $\pi$  compared to when we used a uniform prior. For the uniform prior,  $p^* = 0.421$ . For the beta prior,  $p^* = (23 + 5)/(5 + 10 + 55) = 0.4$ .



**Figure 11.9** Beta(5, 10) prior distribution (solid line) and beta(28, 42) posterior distribution for  $\pi$  (dashed line) after observing that  $x = 23$  and  $n = 55$ .

This makes sense. The beta(5, 10) incorporates more information about  $\pi$  and is less variable, so the estimate should be pulled toward the prior mean [ $E(\pi) = 0.33$ ]. Using statistical software, we find that the 95% credible interval was (0.289, 0.516). With the uniform prior, the credible interval was (0.297, 0.550), which is (not surprisingly) wider and higher (closer to the frequentist estimate  $\hat{p} = 0.418$ ).

### Key Concept

To determine an estimate for  $\pi$ , the population proportion of brown or orange M&M's candies, you should (1) choose your prior distribution for  $\pi$  using a suitable strategy like the one we have discussed and (2) calculate a Bayesian posterior estimate and 95% credible interval for  $\pi$  based on all data available to you.

## Sensitivity Analysis

Although we have discussed how to use subjective information (prior beliefs) about data to select a prior distribution, it is also a good idea to consider the sensitivity of your analysis to your choice of a prior distribution. A **sensitivity analysis** entails examining the results from the posterior distribution (e.g., the posterior mean, posterior variance, or credible interval) when the prior distribution is changed. If the posterior distribution is insensitive to a variety of priors that are regarded as *reasonably* believable and comprehensive, we should be fairly confident of our results. This typically occurs when the amount of data is large or the data are very precise. If the posterior is sensitive to the choice of prior, this suggests that we have not learned enough from the data to be confident of our results.

Table 11.7 is a summary of the Bayesian (prior and posterior) estimates resulting from several potential beta prior distributions that all have prior mean  $E(\pi) = 0.33$  but different minimum and maximum values of  $\pi$ . The results for the beta(20, 40) and beta(25, 50) prior distributions have intentionally been left blank (see Question 46).

**Table 11.7** Summary table of changes in the posterior distribution of the true proportion of brown or orange M&M's, with  $\pi$  corresponding to changes in the prior distribution of  $\pi$ .

$\alpha$	$\beta$	Properties of the Prior Distribution			$p^* = E(\pi   x)$	Posterior Variance	95% Credible Interval	
		$\text{Var}(\pi)$	$\pi_{\min}$	$\pi_{\max}$				
5	10	0.0139	0.20	0.50	0.4000	0.0034	0.2891	0.5163
10	20	0.0072	0.23	0.45	0.3882	0.0028	0.2880	0.4934
15	30	0.0048	0.25	0.42	0.3800	0.0023	0.2878	0.4767
20	40		0.26	0.41				
25	50		0.26	0.40				

Each prior distribution for  $\pi$  has differing levels of variability in  $\pi$  and strength of belief in terms of the minimum ( $\pi_{\min}$ ) and maximum ( $\pi_{\max}$ ) values that  $\pi$  is unlikely to fall beyond. The posterior estimate and a 95% credible interval are reported for each prior when the observed data are  $x = 23$  successes in  $n = 55$  observations. Note that the primary change to the prior distributions involves decreasing the variance. We might have also considered varying the prior estimate,  $E(\pi)$ . We can see from the last column in Table 11.7 that change in the beta prior distribution from our initial beta(5, 10) prior does not substantially change the width of the 95% credible interval. Therefore, we should feel reasonably confident about our 95% credible interval (0.289, 0.516) for  $\pi$ .

## Extended Activity

### Bayesian Credible Intervals for Your M&M's Data Using Other Beta Priors

46. Complete Table 11.7. Are there any substantial changes to the original 95% credible interval (0.289, 0.516) when the beta(20, 40) and beta(25, 50) prior distributions are used? What does this suggest about the amount of data collected?
47. **For Students Who Performed the M&M's Activity** Combine the class data and count the number of brown or orange M&M's in the entire collection. Complete a sensitivity analysis similar to Table 11.7 using your larger combined data set. Are the posterior estimates more sensitive or less sensitive to the choice of prior distribution than what is calculated in Table 11.7 using only our data ( $x = 23, n = 55$ )? Briefly explain.

## Chapter Summary

Unlike the frequentist approach, which entails estimating parameters based only on the observed data, the Bayesian approach incorporates prior knowledge and beliefs into the estimate. By starting with what we currently know to determine a prior estimate and then updating our estimate based on observed data, we follow the traditional scientific approach and the natural process by which we acquire knowledge. With the Bayesian approach, a probability distribution can be assigned to a parameter (e.g., a proportion  $\pi$ ) based on prior beliefs.

The primary rule in Bayesian statistics is Bayes' rule, which states for events  $A$  and  $B$ ,

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

This rule allows us to update  $P(A)$  after observing event  $B$ . The Bayesian approach to estimating a parameter  $\pi$  begins with assigning a **prior distribution**,  $p(\pi)$ , to  $\pi$  and observing the sample data  $x$ . Then the **posterior distribution**,  $p(\pi | x)$ , is computed using the following version of Baye's rule:

$$p(\pi | x) = \frac{p(x | \pi)p(\pi)}{p(x)}$$

where  $p(x | \pi)$  is the **likelihood** of the data and  $p(x)$  is the **marginal likelihood**. The updated estimate of  $\pi$  is taken to be the mean of the posterior distribution,  $E(\pi | x)$ . Different prior distributions will result in different posterior estimates.

If  $\pi$  is treated as discrete (we can list its potential values), then we will need to compute a posterior probability for each prior value of  $\pi$ . If  $\pi$  is treated as continuous, then integration may be required to find the posterior density function; however, calculus may be avoided if we can use a **conjugate prior** for  $\pi$  (i.e., the prior and posterior distributions have the same form). An example of a conjugate prior for  $\pi$  is the beta( $\alpha, \beta$ ) distribution when the likelihood function,  $p(x | \pi)$ , is binomial. If the prior distribution for  $\pi$  is beta( $\alpha, \beta$ ) and  $x$  successes are observed in  $n$  trials, then the posterior distribution for  $\pi$  is beta( $x + \alpha, n - x + \beta$ ) and the posterior estimate of  $\pi$  is  $p^* = E(\pi | x) = (\alpha + x)/(\alpha + \beta + n)$ . One strategy for selecting an appropriate beta prior distribution is to first determine a small value and a large value of  $\pi$  beyond which there is only a small chance of observing  $\pi$ . Then the beta parameters  $\alpha$  and  $\beta$  corresponding to the small and large values of  $\pi$  can be determined using software.

We can form a range of values for the parameter  $\pi$  with a Bayesian credible interval for  $\pi$  that is constructed from the posterior distribution. The lower and upper limits of the 95% credible interval,  $L$  and  $U$ , respectively, are found to satisfy  $P(\pi < L | \text{data}) = 0.025$  and  $P(\pi > U | \text{data}) = 0.025$ . The Bayesian credible intervals are straightforward to interpret, unlike the confidence intervals that you covered in your first statistics course. A sensitivity analysis can be performed to examine the effects of different prior distributions on the width of the credible interval.

## Exercises

---

### E.1. M&M's

Data set: MyMMs or MMs

In Question 9 of the activities, you were shown a possible prior distribution for the proportion of brown or orange M&M's given in Table 11.2. Now consider another prior distribution for  $\pi$ , displayed in Table 11.8, that allows for four possible values of  $\pi$ .

- Compute the prior estimate of  $\pi$ ,  $E(\pi)$ , and the prior variance,  $\text{Var}(\pi)$ .
- Using your M&M's data or the MyMMs data, compute the posterior distribution for  $\pi$ .
- Compute the posterior estimate,  $p^*$ , and the posterior variance,  $\text{Var}(\pi | x)$ .
- Comment on the differences between the prior and posterior distributions for  $\pi$ . Include statements addressing the center and spread of the distributions.

**Table 11.8** Another proposal for the prior distribution for  $\pi$ .

$\pi$	0.25	0.30	0.35	0.40
$p(\pi)$	1/8	3/8	3/8	1/8

### E.2. M&M's

Data set: MyMMs or MMs

Construct another (discrete) prior distribution for the proportion of brown or orange M&M's,  $\pi$ . Use the same values of  $\pi$  given in Table 11.8 but use different probabilities,  $p(\pi)$ , so that the mean is the same but the variance is larger. Then based on this new prior and using your M&M's data or the MyMMs data, answer the following:

- Compute the posterior distribution for  $\pi$ , as well as its mean and variance.
- Compare the values of the posterior mean and variance to those calculated in Exercise 1. Is the change between the prior and posterior estimate of  $\pi$  more or less than the change between the prior and posterior estimate of  $\pi$  that you observed in Exercise 1? How did changing the variance of the prior distribution, but not the mean, affect the posterior mean and variance?

### E.3. Mammograms

Mammograms are extremely vital for early detection of breast cancer, yet for younger women under the age of 40 years, their performance is questionable.

Yankaskas and colleagues investigated the performance of first mammography examinations of women under the age of 40 years.<sup>10</sup> For women between the ages of 35 and 39, they found that the sensitivity of the exams was 76.1% and the specificity was 87.5%. The breast cancer rate among women between 35 and 39 years is 0.000293 (about 29.3 per 100,000 women).<sup>11</sup> Based on these data, if a mammography exam for a woman between the ages of 35 and 39 detects breast cancer, use Bayes' rule to find the probability that she does have breast cancer. What does your result indicate about the performance of mammograms for women between the ages of 35 and 39 years?

### E.4. A Sensitive Question

Bayes' Rule can be applied to situations involving the responses to sensitive questions where a respondent may not answer truthfully due to embarrassment or incrimination. For example, a university administrator may ask students whether they have ever cheated on a final exam for a college course. Instead of answering the question directly, each student rolls a die. If a 1 or 2 appears, then the student answers the question truthfully (a "yes" or "no" answer). If a 3, 4, 5, or 6 appears, then the student lies. Hence, if a student lies, then the student answers "yes" if s/he has never cheated, and answers "no" if s/he has ever cheated on a final exam. Since the administrator does not see the outcome of the roll of the die, he does not know if the student is answering the "cheating" question truthfully. Suppose the administrator uses this procedure to solicit responses from 500 randomly chosen students. Of those individuals, 225 of them answer "yes."

- If Chris is among those randomly selected students, what is the probability that he has ever cheated on a final exam? Hint: Use the law of total probability to find an equation for  $P(\text{yes})$  in terms of  $P(\text{cheated on final})$  and then solve for  $P(\text{cheated on final})$ .
- If Chris is one of those people who answered “yes,” what is the probability that he has ever cheated on a final exam?

#### E.5. ELISA

In Question 24 of the extended activities regarding the ELISA test data, you found that the probability that a randomly selected person who tests positive for HIV (the first time) actually has the virus is surprisingly low:  $P(\text{HIV} | + \text{ Test}) = 0.1175$ . Now let’s investigate how this probability changes depending on the *prevalence* rate of HIV—that is, the proportion of people who have HIV. For this problem, the probability of having HIV was believed to be 0.002 (i.e., the prevalence rate was 0.2%). Also, recall that the rate of correct positive test results for infected people (the *sensitivity* of the test) was 99.7% and the rate of correct negative test results for non-infected people (the *specificity* of the test) was 98.5%.

- If the prevalence rate was 0.1%, recalculate  $P(\text{HIV} | + \text{ Test})$ . Repeat the conditional probability calculation for a prevalence rate of 0.3%. What do you observe about  $P(\text{HIV} | + \text{ Test})$  as the prevalence rate increases? Briefly explain why this makes sense.
- Create an expression that relates the prevalence rate to  $P(\text{HIV} | + \text{ Test})$ . Then produce a plot that displays how  $P(\text{HIV} | + \text{ Test})$  changes as the prevalence rate increases from 0 to 1. Put the prevalence rate on the horizontal axis and  $P(\text{HIV} | + \text{ Test})$  on the vertical axis. Visually inspect the curve to determine an approximate prevalence rate that would be needed for  $P(\text{HIV} | + \text{ Test}) = 0.95$ .
- Now use the equation you found in Part B to find the prevalence rate that would be needed for  $P(\text{HIV} | + \text{ Test}) = 0.95$ . Does this prevalence rate seem high? What does it suggest about the population as a whole?
- To ensure a high probability that a person has HIV when the ELISA test result is positive, is high sensitivity (fixing the specificity level) or high specificity (fixing the sensitivity) more important for a given prevalence rate? Create graphs of  $P(\text{HIV} | + \text{ Test})$  versus values of the sensitivity and  $P(\text{HIV} | + \text{ Test})$  versus values of the specificity to help you answer this question.

#### E.6. M&M’s: Updating Our Model

The Bayesian model updating strategy used for the HIV testing example discussed in Section 11.6 can also be used to update the posterior estimate of the proportion of brown or orange M&M’s. In Section 11.8, we found that the posterior distribution for  $\pi$  (using the author’s data) was beta(24, 33), assuming a uniform prior on  $[0, 1]$  and a binomial likelihood of observing  $x = 23$  brown or orange M&M’s in a sample of  $n = 55$  candies. The posterior estimate was  $p^* = 0.421$ . Now suppose we take another sample of 51 M&M’s and find 19 brown or orange candies.

- Treat the posterior distribution as the new prior distribution for  $\pi$ —that is, the prior for  $\pi$  is beta(24, 33). If 19 brown or orange candies are observed in 51 M&M’s, find the updated posterior distribution for  $\pi$ .
- Find the revised posterior estimate for  $\pi$ . How does this value compare to the first posterior estimate  $p^* = 0.421$ ?
- The 95% Bayesian credible interval for  $\pi$  found in Section 11.8 was  $(0.297, 0.550)$ . Using the updated posterior distribution, construct a new 95% credible interval for  $\pi$ . What happened to the width of the interval when the new data were incorporated?

#### E.7. Examining the Likelihood Function

Recall that if  $x$  is the number of successes in  $n$  trials of a binomial setting with probability of success  $\pi$  on each trial, the likelihood function,  $p(x | \pi)$ , can be expressed as

$$p(x | \pi) = \pi^x (1 - \pi)^{n-x}$$

Consider four possible outcomes of a ganzfeld experiment:  $x = 9$  hits in  $n = 25$  sessions,  $x = 18$  hits in  $n = 50$  sessions,  $x = 36$  hits in  $n = 100$  sessions, and  $x = 72$  hits in  $n = 200$  sessions.

- For each of the four possible experimental outcomes, what is the observed hit rate?
- Construct a graph of the likelihood function corresponding to each of the four experimental outcomes. What do you observe about the shape of the likelihood function as the sample size

increases? Does it become more flat or more peaked? What do you observe about the likely values of  $\pi$  as the sample size increases? Does the *most* likely value of  $\pi$  appear to change for the different likelihood functions?

- c. Now let's consider parameter settings for a beta prior distribution for  $\pi$  for the skeptic:  $\alpha = 27$  and  $\beta = 81$ . Graph the posterior distribution of  $\pi$  for the skeptic. Now examine how the posterior estimate of  $\pi$  is related to the "peakedness" of the likelihood function. If the likelihood function has a sharp peak, does the prior for  $\pi$  have much of an effect on the posterior estimate  $p^*$ ? What does this tell you about the sensitivity of the posterior to the choice of prior if the likelihood function is very peaked? Very spread out?

#### E.8. Batting Averages

Through roughly the first week of the 2011 season, Miguel Montero of the Arizona Diamondbacks was the leading hitter in Major League Baseball. In his first 26 at-bats, he had 13 hits to his credit for a batting average of .500 (the batting average for a player is the number of hits divided by the total tries, called at-bats). We would like to estimate  $\pi$ , Montero's true batting average for the entire 2011 season, using Bayesian methods; that is, we want to compute the posterior estimate of  $\pi$ .

- a. Suppose you know nothing about baseball or a batting average beyond the fact that it is a number between 0 and 1. What prior distribution should you assign to  $\pi$ ? Briefly explain. What is the prior estimate for  $\pi$ ?
- b. Using the prior distribution you defined in Part A and assuming that the likelihood of observing 13 hits in 26 at-bats is binomial with the probability of a successful hit equal to  $\pi$ , find the posterior distribution of  $\pi$ . Calculate the posterior estimate of  $\pi$  and compare it to the prior estimate.
- c. Suppose you know that batting averages typically fall between .200 and .300 during a regular season. Use software to determine an appropriate beta prior distribution for  $\pi$ . Recompute the posterior distribution and posterior estimate for  $\pi$  after observing 13 hits in 26 at-bats. How does your posterior estimate compare to that found in Part B?
- d. Visit a sports website to find Montero's true 2011 season batting average, and compare it to your posterior estimates found in Parts B and C. Which posterior estimate was closer to his true batting average?

#### E.9. Ganzfeld Experiment

Consider the prior distributions assigned to  $\pi$ , the proportion of hits in a ganzfeld experiment, for the skeptic, the open-minded individual, and the believer. Assume that the beta prior distribution is assigned to  $\pi$ , and 18 hits were observed in 50 sessions in a ganzfeld experiment. Construct 95% Bayesian credible intervals for  $\pi$  corresponding to the skeptic, the open-minded individual, and the believer, and interpret the intervals.

#### E.10. Pass Completion Rate

In American football, it is important for the quarterback to throw the ball accurately to the receiver. The completion rate of a quarterback is the proportion of throws successfully caught by the receivers on the team; that is, it is the number of completed passes divided by the total number of attempted throws. One of the best quarterbacks of the 2010 season was Drew Brees of the New Orleans Saints. Let  $\pi$  represent Brees' true completion rate for the entire 2010 season, and suppose we want to estimate it using data from the first five games that he played during that season, using Bayesian methodology. After the first five games, Brees had completed 142 passes in 199 attempts. Suppose we assign a beta(60, 38) prior distribution for Brees' completion rate  $\pi$  (see Exercise 13 for insight into the values for the prior parameters).

- a. Assuming a binomial likelihood function for the 142 completed passes in 199 attempts and a beta(60, 38) prior distribution for  $\pi$ , compute the posterior distribution for  $\pi$  and the posterior estimate of  $\pi$ .
- b. Construct and interpret a 95% Bayesian credible interval for  $\pi$ .
- c. Construct a standard 95% confidence interval for  $\pi$ . Which interval for  $\pi$  is narrower?
- d. At the end of the 2010 season, Brees' completion rate of 68.1% was highest among all starting quarterbacks. How did your posterior estimate compare to his 2010 completion rate? Did his completion rate fall within the confidence interval and credible intervals?

### E.11. Another Approach for Finding a Prior Distribution

The following discussion provides an alternative approach to determine a prior distribution for  $\pi$ . Earlier it was mentioned that the larger the values of the parameters  $\alpha$  and  $\beta$ , the smaller the spread in the distribution. You can verify this by computing the variance of a beta random variable for small and large values of  $\alpha$  and  $\beta$ . The mean of the beta distribution is  $\alpha/(\alpha + \beta)$ , so fix  $E(\pi) = \alpha/(\alpha + \beta)$ , equal to a desired value, but set  $\alpha$  and  $\beta$  higher or lower depending on how much variability in  $\pi$  you are willing to allow. For example, perhaps you strongly believe that the prior estimate of the proportion of brown or orange M&M's is  $E(\pi) = 0.33$  and you are fairly certain that  $\pi$  does not vary too much. Then you might choose  $\alpha = 50$  and  $\beta = 100$ . However, if you are not too sure about the variability in the values of  $\pi$ , then you might select  $\alpha = 1$  and  $\beta = 2$ . Based on this approach, do the following:

- Determine some reasonable values for  $\alpha$  and  $\beta$  for the skeptic, the open-minded individual, and the believer in psychic abilities.
- Assume that in a ganzfeld experiment, 8 hits are observed in 40 sessions. For each type of individual, plot the prior and posterior distributions on the same graph. Compare the changes in the characteristics (mean and variance) of the prior and posterior distributions for the three types of individuals. For which type of person did the posterior estimate of  $\pi$  change the most from the prior estimate of  $\pi$ ?

### E.12. M&M's

Data set: MyMMs or MMs

We can take the approach to selecting a prior distribution for a population proportion  $\pi$  proposed in Exercise 11 one step further. Recall that the variance of a beta random variable,  $Y$ , is given by

$$\text{Var}(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

If we plan to use a beta prior for  $\pi$  and we have a prior estimate for  $\pi$ ,  $E(\pi)$ , as well as a prior variance for  $\pi$ ,  $\text{Var}(\pi)$ , then we can find the parameter values for the beta prior by solving two equations for two unknowns.

- Consider the M&M's activities. Suppose we believe that the prior mean for the proportion of brown or orange M&M's is  $E(\pi) = 0.33$  and we are fairly certain that the variability in the prior distribution of  $\pi$  is quite small, say 0.0001. Using the expressions for the mean and variance for a beta distribution, compute the parameter values for  $\alpha$  and  $\beta$ . Round the values to the nearest whole numbers.
- Use the values of  $\alpha$  and  $\beta$  that you found in Part A and your M&M's data or the MyMMs data to compute the posterior estimate of  $\pi$ .
- Construct a 90% Bayesian credible interval for  $\pi$ , given that you have observed your data.

### E.13. Pass Completion Rate: Empirical Bayes Methods

When observed data are used to develop the prior distribution in Bayesian procedures, the methods are often referred to as **empirical Bayes methods**. These methods are in contrast to standard Bayesian methods, which assume the prior distribution is fixed before any data are observed. In Exercise 10, a beta(60, 38) prior distribution was assigned to the 2010 completion rate for Drew Brees. The prior parameters were derived from pass completion data from other quarterbacks, making this an empirical Bayes problem. The average pass completion rate among all starting quarterbacks for the first five games was 61%, and the variance of the completion rates for the first five games (measured as a decimal) was 0.0024. We will assign a beta prior distribution for  $\pi$  with prior mean  $E(\pi) = 0.61$  and prior variance  $\text{Var}(\pi) = 0.0024$ . Use the approach for finding the prior distribution parameters discussed in Exercise 12 to show that the parameters for the beta prior assigned to Brees' 2010 completion rate are  $\alpha = 60$  and  $\beta = 38$  (rounding to the nearest integer values).

### E.14. A Normal Conjugate Prior Distribution

We have described the beta distribution for a population proportion  $\pi$  as an example of a conjugate prior when the likelihood function,  $p(x | \pi)$ , is binomial. Under certain conditions, the normal distribution also serves as a conjugate prior for the population mean  $\mu$ . It can be shown that if

- $p(\mu)$ , the prior distribution for  $\mu$ , is normal with mean  $\theta$  and variance  $\tau^2$  [i.e.,  $\mu \sim N(\theta, \tau^2)$ ] and
- $p(x_i | \mu)$ , the likelihood function for a single observation  $x_i$ , is  $N(\mu, \sigma^2)$ , for  $i = 1, 2, \dots, n$ ,

then the posterior distribution for  $\mu$ ,  $p(\mu | x_1, x_2, \dots, x_n)$  or simply  $p(\mu | \text{data})$ , is also normal, with posterior mean  $\mu^*$  and variance  $\sigma^{2*}$  given by

$$\mu^* = \frac{\sigma^2\theta + n\tau^2\bar{x}}{\sigma^2 + n\tau^2} \quad \text{and} \quad \sigma^{2*} = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}$$

where  $\bar{x}$  is the sample mean of the observations  $x_1, x_2, \dots, x_n$ .

Suppose the lengths (in inches) of a particular species of salmon are normally distributed with unknown mean  $\mu$  but known variance 36 inches<sup>2</sup>. The prior distribution of  $\mu$  is assumed to be normal with mean  $\theta = 32$  inches and variance  $\tau^2 = 25$  inches<sup>2</sup>. Ten salmon are caught, and their lengths are 19, 26, 18, 26, 31, 31, 32, 36, 33, and 21 inches.

- Find the posterior distribution of the mean salmon length,  $p(\mu | \text{data})$ , and provide the values of the posterior mean,  $\mu^*$ , and posterior variance,  $\sigma^{2*}$ .
- Use statistical software or a standard normal table to find the posterior probability that the average fish length is greater than 35 inches.

### E.15. Advertised Caloric Content

There is some debate regarding advertised nutrition and caloric content of restaurant foods. L. E. Urban and colleagues investigated the advertised caloric content of food items at 29 fast-food and sit-down restaurants and found that the measured caloric content of the food items in a laboratory setting tended to be higher than the reported or advertised values.<sup>12</sup> Suppose that the differences between the laboratory measured calories and the advertised calorie values are taken to be normal with unknown mean difference  $\mu$  calories and variance  $\sigma^2 = 1000$  calories<sup>2</sup>, and assign a normal prior distribution to  $\mu$  with mean 0 (no difference, on average, between the measured and advertised caloric content) and variance  $\sigma^2 = 10,000$ . The actual sample mean of the differences between the measured and advertised calorie values was 25.8 calories. Based on the data reported in the article, we'll use a Bayesian analysis to examine whether restaurants are systematically underreporting the caloric content of their food items.

- Describe the posterior distribution of  $\mu$ ,  $p(\mu | \text{data})$ , and provide the posterior mean and variance.
- Use your posterior distribution to find the probability that, on average, the measured caloric content is at least 10 calories more than the advertised caloric content of restaurant food items; that is, find  $p(\mu > 10 | \text{data})$ . Does your answer support the argument that restaurants are systematically underreporting caloric content?
- A 95% Bayesian credible interval can also be constructed for  $\mu$  by finding the limits  $L$  and  $U$  such that  $P(\mu < L | \text{data}) = 0.025$  and  $P(\mu > U | \text{data}) = 0.025$ . Construct a 95% credible interval for the true mean difference in caloric content, and provide an interpretation of the interval.

## Endnotes

- Blaise Pascal, *Pensées*, quoted in E. F. Keller, *A Feeling for the Organism* (San Francisco: Freeman, 1983), p. 197. Pascal (1623–1662), a mathematician, Christian theologian, and inventor, is well known for Pascal's triangle and his defense of the scientific method.
- Bayes' theorem was not published until after Bayes' death in 1761. His work was finally published in 1763 in T. Bayes, "An Essay Towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society of London*, 53 (1763): 370–418.
- For additional supporting comments regarding Bayesian statistical methods, see D. A. Berry, "Teaching Elementary Bayesian Statistics with Real Applications in Science," *American Statistician*, 51 (1997): 241–246; and M. Lavine, "What Is Bayesian Statistics and Why Everything Else Is Wrong," *Journal of Undergraduate Mathematics and Its Applications*, 20 (1999): 165–174.
- See, for example, M. H. DeGroot and M. J. Schervish, *Probability and Statistics* (Boston: Addison-Wesley, 2012).
- See R. Chou, L. H. Huffman, R. Fu, A. K. Smits, and P. T. Korthuis, "Screening for HIV: A Review of the Evidence for the U.S. Preventive Services Task Force," *Annals of Internal Medicine*, 143 (2005): 55–73.
- See M. P. Busch, S. A. Glynn, S. L. Stramer, D. M. Strong, S. Caglioti, D. J. Wright, B. Pappalardo, and S. H. Kleinman, "A New Strategy for Estimating Risks of Transfusion-Transmitted Viral Infections Based on Rates of Detection of Recently Infected Donors," *Transfusion*, 45 (2005): 254–265.

7. See the 2007 msnbc.com online article “Boo! One in Three People Believes in Ghosts: Survey Finds More People Believe in Specters Than Superstitions,” at <http://www.msnbc.com>.
8. For a complete discussion of the beta distribution, see any mathematical statistics textbook, such as M. H. DeGroot and M. J. Schervish, *Probability and Statistics* (Boston: Addison-Wesley, 2012).
9. For complete details and several examples of conjugate priors, see M. H. DeGroot and M. J. Schervish, *Probability and Statistics* (Boston: Addison-Wesley, 2012).
10. B. C. Yankaskas, S. Haneuse, J. M. Kapp, K. Kerlikowske, B. Geller, and D. S. M. Buist, “Performance of First Mammography Examination in Women Younger than 40 years,” *Journal of the National Cancer Institute*, 102 (2010): 668–669.
11. See L. A. G. Ries, D. Melbert, M. Krapcho, D. G. Stinchcomb, N. Howlader, M. J. Horner, A. Mariotto, B. A. Miller, E. I. Feuer, S. F. Altekruse, D. R. Lewis, L. Clegg, M. P. Eisner, M. Reichman, and B. K. Edwards, *SEER Cancer Statistics Review, 1975–2005*. (Bethesda, MD: National Cancer Institute, 2008), [http://seer.cancer.gov/csr/1975\\_2005/](http://seer.cancer.gov/csr/1975_2005/).
12. L. E. Urban, G. E. Dallal, L. M. Robinson, L. M. Ausman, E. Saltzman, and S. B. Roberts, “The Accuracy of Stated Energy Contents of Reduced-Energy, Commercially Prepared Foods,” *Journal of the American Dietetic Association*, 110 (2010): 116–123.
13. For a discussion of another application of Bayesian methods to home run data, see D. A. Berry and L. Chastain, “Inferences About Testosterone Abuse Among Athletes,” *Chance*, 17 (2004): 5–8.
14. The list of professional athletes identified as testing positive for illegal substances is extensive. We suggest performing a search on a sports website such as <http://espn.go.com> using keywords “banned substances” to read about various infractions.
15. See D. A. Berry and L. Chastain, “Inferences About Testosterone Abuse Among Athletes,” *Chance*, 17 (2004): 5–8; D. A. Berry, “The Science of Doping,” *Nature*, 454 (2008): 692–693.

# Research Project: Do You Believe in ESP?

Now that you have some experience with Bayesian statistics, you'll have the opportunity to conduct your own study to test for ESP abilities. Although it will not be possible to conduct an authentic ganzfeld study, you will be able to design and perform a simple experiment to test for ESP ability.

1. To learn more about Bayesian methods in parapsychology research, read J. Utts, M. Norris, E. Suess, and W. Johnson, "The Strength of Evidence versus the Power of Belief: Are We All Bayesians?" plenary paper presented at the 8th International Conference on Teaching Statistics, July 2010. (You can skip the section on Hierarchical Bayesian Models on page 6.) This paper can be found at the website [http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8\\_PL2\\_UTTS.pdf](http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_PL2_UTTS.pdf). Note that the approach in the paper to finding the parameters of the beta prior distribution is slightly different from the one discussed in this chapter, but the discussions of incorporating likely values of  $\pi$  into the prior distribution are quite similar. If there are any words that you do not understand, look them up and provide a short definition for each.

Before you begin designing the experiment and collecting data, determine the strength of your beliefs in psychic abilities—decide whether you are a skeptic, an open-minded individual, or a believer.

Based on your beliefs, determine two or three beta prior distributions for the proportion of hits. Try various strategies that have been discussed in the extended activities and the homework questions to determine the parameter values for the prior distribution. You will use these prior distributions in later questions.

## Designing the Study and Collecting the Data

2. Clearly define a problem and state the objectives of your experiment. What materials can you use to investigate the true proportion of hits?
3. Identify what other conditions need to be controlled during the experiment to eliminate potential biases. Identify how measurements, material, and process may involve unwanted variability. What conditions would be considered normal for this type of experiment? Are these conditions controllable? If this condition changed during the experiment, how might it impact the results? Explain how these conditions will be controlled throughout the experiment, even if they are simply held constant. Will subjects be allowed to practice transmitting targets before the actual experiment?
4. Choose an experimental design.
  - a. Keep the design and analysis as simple as possible. A straightforward design is usually better than a complex design. If the design is too complicated and the data are not collected properly, even the most advanced statistical techniques may not be able to draw appropriate conclusions from the experiment.
  - b. How many trials will be run? Can you completely randomize all the trials, or do you need to account for timing, subject variability, and other nuisance factors?
5. Explain how your experimental design builds on previous research. Identify relevant background such as theoretical relationships, expert knowledge/experience, or previous studies.
6. Discuss designs and decide on an experiment to be tested.
  - a. Prepare any questions you would like to ask cognitive psychologists or statisticians before you finalize your experimental design.
  - b. Write specific lab procedures that you will use while conducting the experiment. Determine who will collect the data at what time, how you will randomize the trials, how the data will be recorded, and exactly what will be measured.
  - c. Ensure that your group has received appropriate Institutional Review Board approval (if necessary).
7. Conduct the study.
  - a. Meet with your professor to discuss your experimental design and potential analysis.
  - b. Collect data. While conducting the study, did you identify any additional sources of variability that could be impacting the results?

8. Here is one possible set-up for your experiment:
  - a. Students get into pairs and determine who will be the sender and who will be the receiver by flipping a coin.
  - b. Students sit facing away from each other so that the receiver cannot see the card that the sender selects.
  - c. The sender shuffles the deck of cards and selects one target card from the deck. Through deep concentration, the sender attempts to transmit the image of the card to the receiver.
  - d. Once the transmission is complete, the sender mixes the target card with a few other randomly selected cards, which serve as decoys.
  - e. The sender places the target and decoy cards in front of the receiver, and the receiver attempts to pick the target card. If the receiver is correct, then the pick is considered a hit.
  - f. Repeat this process many times, and record the number of hits.

## Exploring the Data

Once you have collected your data, address the following.

9. For each potential prior distribution that you considered, construct the posterior distribution for  $\pi$  given the data you observed from the experiment you conducted. Plot the prior and posterior distributions on the same graph. Construct a 95% Bayesian credible interval for  $\pi$ , the probability of a correct guess.
10. How did your prior beliefs about ESP change after you conducted the experiment? Are you much less skeptical than you were before observing the data? Or are you more skeptical? Answer this question by comparing the characteristics of your prior and posterior distributions, as well as by examining the credible interval.
11. Repeat steps 8 and 9 using the entire class data set. Investigate which type of individual has been convinced the least and most in light of the larger data set.

## Presenting Your Results

Write a research paper. (See “How to Write a Scientific Paper or Poster” on the accompanying CD.) Bring three copies of your research paper to class. Submit one to your instructor. The other two will be randomly assigned to other students in your class to review. Use the “How to Write a Scientific Paper or Poster” checklist to review each other’s papers and provide comments.

## Final Revisions

Make final revisions to the research paper. Then submit the first draft, other students’ comments and checklists, the data set you used (in electronic format) along with descriptions of the variables in the data set, and your final paper.

## Other Project Ideas

Some of the end-of-chapter exercises might be used to develop your own project ideas. In addition, the following ideas can be developed into projects.

- **Home Runs.** What is the probability that your favorite baseball player will hit a home run during his next plate appearance? If you do not have a favorite player, then select any player (with the exception of pitchers in the American League, since they do not hit). Use Bayesian methodology to compute a posterior estimate for this probability,  $\pi$ , and create a 95% credible interval for  $\pi$ .<sup>13</sup>

First determine a prior distribution for  $\pi$  based on your knowledge or beliefs about the player (you might consider conducting some background research), and compute a prior estimate for  $\pi$ . You may consider discrete or continuous prior distributions for  $\pi$ . To gather data on your favorite player for the current season (or information from previous seasons if it is not baseball season),

you can visit the website [http://mlb.mlb.com/stats/sortable\\_player\\_stats.jsp](http://mlb.mlb.com/stats/sortable_player_stats.jsp) and locate the number of plate appearances (i.e., number of at-bats) in the column labeled “AB” and the number of home runs in the column labeled “HR.” The observed data are the current number of home runs hit ( $x$ ) and the number of at-bats ( $n$ ) for the season (or you can use data from the previous season).

Update your prior results using the sample data, and compute posterior quantities (e.g., mean and credible interval). If there are other students who share your favorite player, then compare your prior distributions and prior estimates of  $\pi$ , as well as your posterior distributions, posterior estimates, and credible intervals.

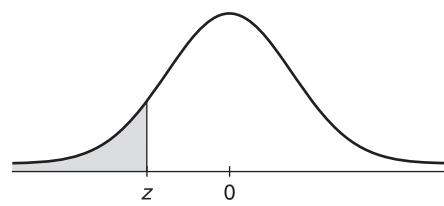
- **Drug Testing of Professional Athletes.** In professional sports, significant attention has focused on testing athletes for performance-enhancing drugs. Athletes have been suspended, banned, or stripped of their results in various sports such as cycling, track and field, and baseball because they tested positive for illegal substances including steroids and testosterone.<sup>14</sup> From a Bayesian perspective, the main question is “If the athlete tests positive for a drug, what is the probability that she or he is a user of the drug?”

In recent years, testing procedures have come under more scrutiny by statisticians who are concerned that the *sensitivity* and *specificity* of the test are being ignored.<sup>15</sup> Recall that the sensitivity is the proportion of times the test for a substance yields positive results when the person is actually using the substance, and the specificity is the rate of negative test results when the individual has not used the substance. Bayesian methods that take into account measures of sensitivity and specificity, as well as the prior probability that an athlete is a user, should be utilized when reaching conclusions about athletes’ test results.

For this project, you will investigate applications of Bayes’ rule to drug testing in a sport of your choice. You may want to begin by reading the articles by D. A. Berry “The Science of Doping,” [*Nature*, 454 (2008): 692–693] and D. A. Berry and L. Chastain “Inferences About Testosterone Abuse Among Athletes,” [*Chance*, 17 (2004): 5–8] and some of the references provided within those papers. Based on your readings, provide some hypothetical (but realistic) calculations for the probability that an athlete is a user of a drug given that the athlete tests positive for the drug. Change your prior probabilities (that an athlete is a user), sensitivity, and specificity rates and examine how the posterior probability of drug use changes.

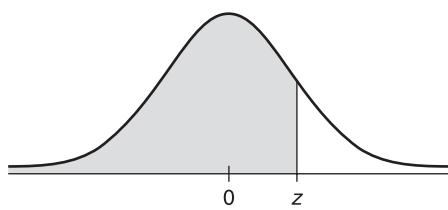
# Appendix of Tables

- 1. Standard Normal ( $z$ ) Distribution**
- 2.  $t$ -Distribution Critical Values**
- 3. Chi-Square Distribution Critical Values**

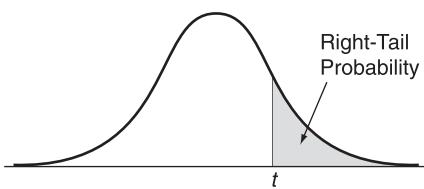


**Table 1** Standard normal ( $z$ ) distribution: Cumulative area from the LEFT.

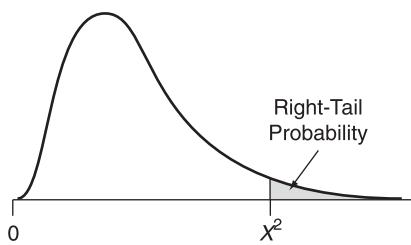
Negative $z$ Scores	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641



Standard normal ( $z$ ) distribution: Cumulative area from the LEFT.

**Table 2**  $t$ -distribution critical values.

df	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
$\infty$	1.282	1.645	1.960	2.326	2.576	3.091

**Table 3** Chi-square distribution critical values.

df	Right-Tail Probability						
	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.80	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.4

# Index

## A

Accelerated failure time models, 305  
Adequacy, overall test of model, 86  
Adjusted R<sup>2</sup>, 92  
Akaike's information criterion (AIC), 93, 228  
Algebra, matrix, 351–353  
Allocation, random, 8–9, 11, 41–43, 55, 139–140  
 $\alpha$ -level, 7, 22  
Alternative hypothesis, 4, 22  
Analysis of covariance (ANCOVA), 161–163  
Analysis of variance (ANOVA)  
to compare population means, 39–41  
Kruskal-Wallis test *vs.*, 18  
mathematical calculations for, 155–157  
model, 39–40  
model assumptions for, 40–41  
one-way, 124–125  
sums of squares and, 52, 120  
table, 53, 85, 86  
three-way, 114–115  
transformations for, 46–50  
two-way, 124–125  
variability in, 110  
ANCOVA. *See* Analysis of covariance (ANCOVA)  
ANOVA. *See* Analysis of variance (ANOVA)  
Arcsine transformation, 46  
Association, measures of, 228–229  
Assumption(s)  
equal variance definition of, 33  
two-sample *t* test with, 50–52  
for logistic regression models, 221  
model  
for analysis of variance (ANOVA), 40–41  
checking, 38, 150  
for linear regression model, 37–38  
for multiple regression, 73–80  
for two-sample *t*-test, 34–35  
Autocorrelation, 75, 94  
taxonomic, 76

## B

Back transformations, 47–48  
Backward stepwise regression, 71  
Balanced design, 107  
Bayesian approach, 370  
Bayesian credible intervals, 394–399  
Bayesian data analysis, 380–399  
Bayesian model updating, 382–383  
Bayes' information criterion (BIC), 93, 228  
Bayes' rule, 377  
Berra, Yogi, 30  
Best subsets techniques, 71–73, 94  
Beta distribution, 373, 387–389  
Between-block factors, 145  
Between-group variability, 110  
Between-level variability, 53  
Bias  
bootstrap estimate of, 14  
incomplete observations and, 287  
Binary variable, 178  
Binomial counts, logistic regression models for, 230–231  
Binomial model, for count data, 255–256  
Block design, 106, 142, 153  
Blocking, 139, 143, 163  
Bonferroni's method, 21, 125  
Bootstrap confidence intervals, 22  
Bootstrap distribution, 12–14, 22  
Bootstrap estimate of bias, 14  
Bootstrap percentile confidence interval, 15, 23  
Bootstrap *t* confidence interval, 14–15, 23  
Box, George E. P., 67

## C

Cards, simulation study with, 180  
Case-control studies, 190  
Categorical data, summarization of, 178–179  
Categorical explanatory variables, 81–83  
Categorical variable, 178  
Censoring  
interval, 317–318, 321  
left, 317–318, 321  
right, 287, 321  
truncation and, 319–320

Central limit theorem, 3  
Change-in-deviance test, 227  
Characteristic vector, 338  
Chi-square test, 183–186  
conducting, 185–186  
goodness-of-fit, 193–195, 233  
for homogeneity, 184  
for independence, 184  
simulation of statistic, 186  
Churchill, Winston, 102  
Coefficient of determination, 68, 71  
Cohort studies, 190  
Column variable, 178  
Comparative groups, 139  
Comparisonwise type I error, 21  
Complete event time, 287  
Completely randomized factorial designs, 106, 142  
Component, random, 218  
Computer simulation  
overview of, 180  
randomization test with, 5–7  
Concordant pair, 228  
Conditional failure rate, 306  
Conditional probability, 375  
Conditional proportion, 187  
Conditional test of independence, 182  
Confidence interval  
bootstrap, 22  
bootstrap percentile, 15, 23  
bootstrap *t*, 14–15, 23  
definition of, 14  
form, 54  
for regression coefficient, 54  
for survival probabilities, 297–300, 321  
Wald, 222  
Confidence level, 54  
Confirmation, as goal of multiple regression, 69  
Confounding variables, 142  
Conjugate priors, 389–390, 399  
Constant, normalizing, 387  
Contingency table, 178  
Contrasts, 124–126  
Correlation matrix, 337–338  
Count data  
binomial model for, 255–256  
building models for, 254–255  
comparing, for groups, 252–254  
Poisson model for, 256–260, 260–263  
Covariance matrix, 344

Covariate, for Poisson count model, 260–263  
Covariate pattern, 236  
Cox regression model, 305  
Credible intervals, Bayesian, 394–399  
Critical value  
choice of, 21–22  
definition of, 54  
Cross-classification studies, 189–190  
Crossed effects, 151–154  
Crossed factors, 145–146, 163  
Cubic terms, 90–91  
Cumulative hazard function, 311–317, 321  
Curve  
hazard, 308  
Kaplan-Meier, 292–293

## D

Data  
categorical, summarization of, 178–179  
failure-time, 286  
nominal, 178  
ordinal, 178  
survival, 286  
time series, 286  
time-to-event, 286  
Decomposition, of sums of squares, 83–85  
Degrees of freedom (df), 53, 121–124, 155–157  
Delayed entry, 319  
Delta beta, 236  
Delta chi-square, 236  
Delta deviance, 236  
Description, as goal of multiple regression, 69  
Descriptive statistics  
categorical data and, 178–179  
survival analysis and, 294–297  
Design  
experimental  
balanced, 107  
block, 106, 142, 153  
choice of, 105–106  
completely randomized factorial, 106, 142  
definition of, 104  
elements of good, 103–107, 139–141  
repeated measures, 106, 141

- split-plot, 106, 141, 144–145, 148–149, 154, 163  
 statistical analysis based on, 141  
 three-way factorial, 113–115  
 two-way factorial, 107–113  
 visualization of structures in, 146  
 sampling, 189–192
- Determination, coefficient of, 68, 71
- Deviation
- delta, 236
  - null, 223
  - pooled standard, 51
  - residual, 223, 268
  - statistic, 268, 277
  - test, 240
    - change-in, 227
    - drop-in, 226–228, 240
- Df. *See* Degrees of freedom (df)
- Diagnostic plots, 236–237
- Diagrams, Hasse, 157–161
- Direction of largest variability, 338–340
- Discordant pair, 228
- Dispersion parameter, 277
- Distribution
- beta, 373, 387–389
  - binomial, 258
  - bootstrap, 12–14, 13, 22
  - hypergeometric, 181
  - for  $\pi$ 
    - posterior, 375–378
    - prior, 374–375
  - Poisson, 258
  - posterior, 375, 378, 399
  - prior, 399
    - beta, for  $\pi$ , 396–398
    - non-informative, 385
    - objective, 385
    - for  $\pi$ , 390–394
    - uniform, 385–387
  - sampling, 12, 13
  - survival time, 296, 320
- Distribution-free test, 3
- Dotplots, 3, 10, 11
- Drop-in-deviance test, 226–228, 240
- Dummy variables, 36, 81
- E**
- Effect(s)
- calculation of, 117–120
  - crossed, 151–154
  - fixed, 106
  - interaction
- calculating, 119–120
- definition of, 111
- main, 39, 118–119
- nested, 151–154
- random, 106
- Eigenvalue, 338, 342, 351–353
- Eigenvector, 338, 339, 351–353
- Empirical Bayes methods, 403
- Empirical *p*-value, 5, 22
- Empirical randomization distribution, 6
- Empirical survival function, 288–289
- Entry, delayed, 319
- Equal variance assumption
- definition of, 33
  - two-sample *t* test with, 50–52
  - within-group variability and, 111
- Error
- mean square, 53, 111, 157
  - standard
    - of estimate, 54
    - of Kaplan-Meier estimator, 298–300
  - terms, 55, 117
  - type I, 21
  - type II, 21
- Error sum of squares ( $SS_{\text{Error}}$ ), 52, 121
- Estimate, 54
- Estimated hazard function, 307–309
- Estimated hazard rate, 309
- Estimator
- Kaplan-Meier, 320
    - standard error of, 298–300
    - in survival analysis, 289–292
  - Nelson-Aalen, 314–315
- Event times
- complete, 287
  - incomplete, 287
- Exact *p*-value, 7, 22
- Experiment(s)
- controls in, 8
  - definition of, 103
  - design
    - balanced, 107
    - block, 106, 142, 153
    - choice of, 105–106
    - completely randomized factorial, 106, 142
    - definition of, 104
    - repeated measures, 106
    - split-plot, 106, 141, 148–149, 154, 163
  - statistical analysis based on, 141
- three-way factorial, 113–115
- two-way factorial, 107–113
- visualization of structures in, 146
- elements of well-designed, 103–107, 139–141
- factor identification in, 105
- ganzfeld, 384–394
- observational study vs., 103
- response variable identification in, 104
- sample size determination for, 106–107
- Experimental unit, 105
- Explanatory variables
- in analysis of variance (ANOVA), 39
  - goodness-of-fit tests for continuous, 235
  - multiple
    - logistic regression with, 224–225
    - notation for, 108–110
- Exposure, 252, 276
- Extraneous variables, 41, 55, 105, 142
- Extra sum of squares *F*-test, 86–87, 95
- F**
- Factor(s)
- analysis, 349
  - between-block, 145
  - crossed, 145–146, 163
  - fixed, 127, 147–149, 163–164
  - identification of, 105
  - of interest, 105
  - level of, 39
  - loadings, 348
  - nested, 106, 145–146, 163
  - nuisance, 142
  - random, 147–149, 163–164
  - split-plot, 144
  - whole-plot, 144
  - within-block, 145
- Factorial design
- completely randomized, 106, 142
  - three-way, 113–115
  - two-way, 107–113
- Failure time, 286
- Failure-time data, 286
- Familywise type I error, 21
- Fisher's exact test, 181–182, 195
- Fixed effects, 106
- Fixed factors, 127, 147–149, 163–164
- Forward stepwise regression, 70
- Frame, sampling, 42
- Frequentist approach, 370
- F*-statistic, 53, 112, 157
- F*-tests
- extra sum of squares, 86–87, 95
  - for multiple regression, 83–87
- Full model, 86, 223, 267
- Function(s)
- hazard, 305–311, 321
  - cumulative, 311–317, 321
  - estimated, 307–309
  - population, 306–307
  - likelihood, 238
  - link, 218, 261, 276
  - monotonic, 46
  - probability density, 385
  - step, 292
  - survival, 288–294, 300–305
- Fundamental question for inference, 4–5
- G**
- Ganzfeld experiments, 384–394
- Generalized linear model (GLM), 218, 261, 270–271
- GLM. *See* Generalized linear model (GLM)
- Goodman-Kruskal gamma, 228, 240
- Goodness-of-fit test
- chi-square, 193–195, 233
  - for continuous explanatory variables, 235
  - for logistic regression models, 232–235
- Grand mean, 39
- Graph(s)
- descriptive, 178–179
  - of Kaplan-Meier curve, 292–293
- Group mean squares ( $MS_{\text{Group}}$ ), 53
- Groups
- comparative, 139
  - comparing count data for, 252–254
- Group sum of squares ( $SS_{\text{Group}}$ ), 52
- H**
- Hasse diagrams, 157–161
- Hazard curve, 308
- Hazard function, 305–311, 321
- cumulative, 311–317, 321
  - estimated, 307–309
  - population, 306–307

- Hazard rate, 306  
estimated, 308, 309
- Heteroskedasticity, 73–75, 94
- Homogeneity  
chi-square test for, 184  
independence *vs.*, tests for, 192–193  
of odds, test for, 191–192  
of proportions, test for, 196
- Homoskedasticity, 73–75
- Honest significant difference (HSD), 126
- Hooke, Robert, 176
- Hosmer-Lemeshow test, 233, 240
- HSD. *See* Honest significant difference (HSD)
- Hypergeometric distribution, 181
- Hypothesis  
alternative, 4, 22  
null, 4, 22
- I**
- Incidence rate, 252
- Incomplete event times, 287
- Independence  
chi-square test for, 184  
conditional test of, 182  
homogeneity *vs.*, tests for, 192–193
- Independent and identically distributed (iid) variables, 34
- Indicator variables, 36, 81–82
- Individual value plots, 3
- Inference  
fundamental question for, 4  
for Poisson regression models, 266–268  
through randomization test, 4–5
- Influential observation  
definition of, 76  
identification of, 237  
multiple regression and, 76
- Information criterion  
Akaike's, 93, 228  
Bayes', 93, 228
- Interaction  
definition of, 88  
effect  
calculating, 119–120  
definition of, 111
- Interaction plot, 113
- Interaction terms  
degrees of freedom for, 122–123  
interpretation of, 113  
mixed, 160
- multiple regression and, 88–90
- Interest, factors of, 105
- Interval censoring, 317–318, 321
- Iterative techniques, 94
- K**
- Kaplan-Meier curve, 292–293
- Kaplan-Meier estimator, 320  
standard error of, 298–300  
in survival analysis, 289–292
- Kendall's tau-a, 228, 240
- Kruskal-Wallis test, 18–20, 23
- L**
- Lack of normality, 77, 94
- Least-significant differences method (LSD), 21, 125
- Least squares estimation, 270
- Least squares regression  
building, 215  
equations, 85  
overview of, 215
- Left censoring, 317–318, 321
- Left truncation, 319, 321
- Level  
of factor, 39  
identification of, 105
- Likelihood estimates, maximum, 218–219
- Likelihood function, 238
- Likelihood ratio test (LRT), 223–224, 240, 267
- Linear combinations,  
interpretation of, 335–336
- Linear models, generalized, 218
- Linear regression model  
assumptions for, 37–38  
to compare population means, 36–38  
simple, 36
- Link function, 218, 261, 276
- Loadings, factor, 348
- Logistic regression  
assessing fit of, 232–235  
assumptions for, 221  
for binomial counts, 230–231  
inference for, 221–224  
interpretation of, 219–221  
with maximum likelihood estimates, 218–219, 238–240  
with multiple explanatory variables, 224–225  
overview of, 216–218  
residuals for, 231–232  
transformations and, 216
- Logit transformation, 46, 216
- Log-linear regression model, 261
- Log log transformation, 47
- Log-odds, 216, 240
- Log-rank test, 301–304
- Log-rank test statistic, 303–304
- Log transformation, 46
- LRT. *See* Likelihood ratio test (LRT)
- LSD. *See* Least-significant differences method (LSD)
- Lurking variables, 142
- M**
- Main effect, 39, 118–119
- Main effects plot, 40
- Mallows'  $C_p$ , 92–93
- Mann-Whitney test, 17–18, 23
- Marginal likelihood of data, 376, 399
- Marginal probability of data, 376
- Matched pairs design  
definition of, 10  
permutation tests for, 10–12  
randomization tests for, 10–12
- Matrix  
algebra, 351–353  
correlation, 337–338  
covariance, 344  
plot, 334  
scatterplot, 334
- Maximum likelihood estimates, 271, 276  
calculating, 239  
definition of, 238  
with logistic regression, 218–219, 238–240
- Mean(s)  
population  
analysis of variance (ANOVA) to compare, 39–41  
linear regression model to compare, 36–38  
posterior, 379
- Meaningful scale, 286
- Mean response, 117
- Mean square error (MSE), 53, 111, 157
- Mean squares (MS), 53, 110, 157
- Mean survival time, 295–296, 320
- Measurements  
of association, 228–229  
repeated, 144
- Median survival time, 296
- Minitab (computer software), 18
- Mixed interaction terms, 160
- Model(s)  
accelerated failure time, 305  
analysis of variance (ANOVA), 39–40  
Bayesian updating, 382–383  
to confirm theory, 87–88  
for count data  
binomial, 255–256  
building, 254–255  
Poisson, 256–260, 260–263  
full, 86, 223, 267  
linear  
generalized, 218, 270–271  
log, 261  
nested, 227  
nonparametric, 317  
null, 223, 267  
parametric, 317  
parametric regression, 305  
population, 16  
posterior, 372  
prior, 372  
randomization, 16  
reduced, 86, 223  
restricted, 223  
saturated, 269  
statistical, in two-sample *t*-test, 32–34  
unrestricted, 223  
validation, 93–94
- Model assumptions  
for analysis of variance (ANOVA), 40–41  
checking, 38, 150  
for linear regression model, 37–38  
for multiple regression, 73–80  
for two-sample *t*-test, 34–35
- Model coefficients, 68, 80–81
- Monotonic functions, 46
- Montgomery, Douglas, 1
- MS. *See* Mean squares (MS)
- MSE. *See* Mean square error (MSE)
- Multicollinearity, 78–79, 94, 274–275, 277, 351
- Multiple comparisons, 20–22, 124–126
- Multiple regression  
adjusted R<sup>2</sup> and, 92  
analysis of variance (ANOVA) table and, 85, 86  
autocorrelation and, 75–76, 94  
best subset techniques in, 71–73  
confirmation as goal of, 69

cubic terms and, 90–91  
 description as goal of, 69  
 $F$ -tests for, 83–87, 95  
 goals of, 69–70  
 heteroskedasticity and,  
   73–75, 94  
 homoskedasticity and, 73–75  
 indicator variables in, 81–82  
 influential observations and,  
   76–77  
 interaction terms and, 88–90  
 interpretation of model  
   coefficients in, 80–81  
 lack of normality and, 77, 94  
 Mallows'  $C_p$  and, 92–93  
 model assumptions for,  
   73–80  
 model validation in, 93–94  
 multicollinearity and,  
   78–79, 94  
 normally distributed residuals  
   and, 77–78  
 outliers and, 76–77, 94  
 prediction as goal of, 69  
 quadratic terms and, 90–91  
 stepwise regression in, 70–71  
 sums of squares decomposition in, 83–85  
 taxonomic autocorrelation  
   and, 76

**N**

Nelson-Aalen estimator,  
   314–315  
 Nested effects, 151–154  
 Nested factors, 106,  
   145–146, 163  
 Nested model, 227  
 Nominal data, 178  
 Non-informative prior  
   distribution, 385  
 Nonparametric models, 317  
 Nonparametric test, 3, 22  
 Normality, lack of, 77, 94  
 Normalization, of vector, 339  
 Normalizing constant, 387  
 Normally distributed residuals,  
   77–78  
 Normal probability plots,  
   44–46, 56  
 Notation  
   for multiple explanatory  
 variables, 108–110  
   vector, 336–337  
 Nuisance factors, 142  
 Null deviance, 223  
 Null hypothesis, 4, 22  
 Null model, 223, 267

**O**

Observation  
   incomplete, 287  
 influential  
   definition of, 76  
   identification of, 237  
   multiple regression and, 76  
 Observational studies, 9, 103  
 Odds  
   definition of, 188  
   homogeneity of, tests for,  
 191–192  
   log, 216, 240  
   ratio, 188, 195, 240  
 One-way analysis of variance,  
   124–125  
 Order  
   randomization, 11  
   residual plots across, 75  
   standard, 107  
 Ordered complete times, 291  
 Ordinal data, 178  
 Orthogonal contrast, 125  
 Outliers  
   definition of, 76  
   identification of, 237  
   multiple regression and,  
 76–77, 94  
 Overall test of model  
   adequacy, 86  
 Overdispersion, 275–276, 277

**P**

Pair  
   concordant, 228  
   discordant, 228  
   tied, 228  
 Parameters, definition of, 33  
 Parametric models, 317  
 Parametric regression  
   models, 305  
 Parametric tests, 22  
 Pascal, Blaise, 369  
 PCA. *See* Principal component  
   analysis (PCA)  
 Pearson chi-square goodness-of-fit test, 233, 240  
 Pearson residuals, 272  
 Percentiles, of survival time  
   distribution, 296, 320  
 Permutation tests  
   definition of, 22  
   for matched pairs designs,  
 10–12  
   randomization tests *vs.*, 9–10  
 $\pi$   
   beta prior distribution for,  
 396–398

combining prior information  
   about, with data,  
 371–373  
 posterior distribution for,  
   375–378  
 posterior probabilities  
   for, 377  
 prior distributions for,  
   374–375, 390–394  
 relative frequency estimate  
   of, 371  
 subjective estimate of,  
   370–371  
 Planned variability, 41  
 Plot(s)  
   diagnostic, 236–237  
   dotplots, 3, 10, 11  
   individual value, 3  
   interaction, 113  
   main effects, 40  
   matrix, 334  
   normal probability,  
 44–46, 56  
   residual, across time/order, 75  
   scatterplot, in principal  
 component analysis,  
 333–334  
   scree, 354  
   time series, principal  
 component analysis in,  
 335–336  
 Poisson model for count data,  
   256–260, 260–263  
 Poisson regression model, 261  
   assessing fit of, 268–270  
   inference for, 266–268  
   with more than one covariate,  
 264–266  
   multicollinearity in, 274–275  
 Polya, George, 251  
 Pooled standard deviation, 51  
 Population hazard function,  
   306–307  
 Population means  
   analysis of variance  
 (ANOVA) to compare,  
 39–41  
   linear regression model to  
 compare, 36–38  
 Population model, 16  
 Posterior distribution,  
   375, 378, 399  
 Posterior estimate, 372, 373  
 Posterior mean, 379  
 Posterior model, 372  
 Posterior probabilities, 375,  
   377–378  
 Potential, 305

Prediction, as goal of multiple  
 regression, 69  
 Principal component analysis  
   (PCA)  
   calculation of principal com-  
 ponents in, 336–338  
   correlation matrix in, 337–338  
   covariance matrix in, 344–345  
   creation of principal compo-  
 nents in, 340–342  
 definition of, 333  
 determination of number of  
   components to retain  
   in, 346–347  
 direction of largest variability  
   in, 338–340  
 eigenvalues in, 342  
 interpretation of principal  
   components in,  
 347–350  
 linear combinations and,  
   335–336  
 in other statistical meth-  
   ods, 351  
 regression and, 350  
 scatterplots for, 333–334  
 three-dimensional example,  
   342–343  
 time series plots for, 335–336  
 variable standardization in,  
   344–345  
 vector notation and, 336–337  
 visual interpretation of,  
   333–336  
 Prior distribution, 399  
   beta, for  $\pi$ , 396–398  
   non-informative, 385  
   objective, 385  
   for  $\pi$ , 390–394  
   uniform, 385–387  
 Prior estimate, 375  
 Prior model, 372  
 Prior probability, 372  
 Priors, conjugate, 389–390, 399  
 Probability density function, 385  
 Probability plots, normal,  
   44–46, 56  
 Proportion  
   conditional, 187  
   equivalence of two, 195  
   homogeneity of, test for, 196  
 Proportional hazards model, 305  
 Proxy measurement, 359  
 $P$ -value  
   definition of, 5  
   empirical, 5, 22  
   exact, 7, 22  
   *F*-statistic, 112

**Q**

Quadratic terms, 90–91

**R**

Random allocation, 8, 11, 41–43, 55, 139–140

Random component, 218

Random effects, 106

Random error term, 117

Random factors, 147–149, 163–164

Randomization hypothesis, 3

Randomization of order, 11

Randomization test

- with computer simulation, 5–7

- by hand, 4–5

- for matched pairs designs, 10–12

- permutation test *vs.*, 9–10

- statistical inference through, 4–5

- steps in, 22

- two-sample *t*-test and, 16–17

Randomization tests, 9–10

Random sampling, 8–9, 11, 41–43, 55, 140

Random variability, 41

Rate

- conditional failure, 306

- hazard, 306

- estimated, 308, 309

- incidence, 252

Ratio, odds, 188, 195, 240

R (computer software), 18

Reciprocal transformation, 46

Reduced model, 86, 223

Regression

- best subsets, 71–73, 94

- Cox, 305

- forward stepwise, 70

- least squares

  - building, 215

  - equations, 85

  - overview of, 215

- linear

  - assumptions for, 37–38
  - to compare population means, 36–38

  - log, 261

  - simple, 36

- logistic

  - assessing fit of, 232–235

  - assumptions for, 221

  - for binomial counts, 230–231

  - inference for, 221–224

  - interpretation of, 219–221

with maximum likelihood estimates, 218–219, 238–240

with multiple explanatory variables, 224–225

overview of, 216–218

residuals for, 231–232

transformations and, 216

multiple

- adjusted  $R^2$  and, 92

- analysis of variance (ANOVA) table and, 85, 86

- autocorrelation and, 75–76, 94

- best subsets techniques in, 71–73

- best subset techniques in, 71–73

- confirmation as goal of, 69

- cubic terms and, 90–91

- description as goal of, 69

- F*-tests for, 83–87, 95

- goals of, 69–70

- heteroskedasticity and, 73–75, 94

- homoskedasticity and, 73–75

- indicator variables in, 81–82

- influential observations and, 76–77

- interaction terms and, 88–90

- interpretation of model coefficients in, 80–81

- lack of normality and, 77, 94

- Mallows'  $C_p$ , and, 92–93

- model assumptions for, 73–80

- model validation in, 93–94

- multicollinearity and, 78–79, 94

- normally distributed residuals and, 77–78

- outliers and, 76–77, 94

- prediction as goal of, 69

- quadratic terms and, 90–91

- stepwise regression in, 70–71

- sums of squares decomposition in, 83–85

- taxonomic autocorrelation and, 76

- parametric, 305

- Poisson, 261

- principal components and, 350

stepwise, 70–71

- backward, 71

- forward, 70

- use of, 94

for survival data, 305

transformations for, 48

Relative risk reduction, 189

Repeated measurements, 144

Repeated measures design, 106, 141

- adjusted  $R^2$  and, 92

Replacement, sampling with, 12

Replications, 144

Resampling techniques, 12

Residual(s), 33, 56

- deviance, 223, 268

- in generalized linear models, 270

- hand calculation of, 233–235

- in logistic models with binomial counts, 231–232

- normally distributed, 77–78

- Pearson, 272

- plots, across time/order, 75

Residual deviance, 223

Response variable, identification of, 104

Restricted model, 223

Retrospective studies, 190–191

Right censoring, 287, 321

Right truncation, 319, 321

Risk reduction, relative, 189

Row variable, 178

$R^2$  value, 68, 71, 92

**S**

Sample size, 106–107, 140–141

Sampling

- designs, 189–192

- distribution, 12, 13

- frame, 42

- random, 8–9, 11, 41–43,

  - 55, 140

  - with replacement, 12

  - resampling and, 12

  - variability, 297

SAS (computer software), 18

Saturated model, 269

Scale, meaningful, 286

Scatterplot

- matrices, 334

- for principal component

  - analysis, 333–334

Scheffe's method, 126

Scree plot, 354

Sensitivity, 380

Sensitivity analysis, 398–399

Significance, of regression

- model, test of, 86

Significance level, 7, 22

Simple linear regression model, 36

Simulation study

- with cards, 180

- computer

  - overview of, 180

  - randomization test with, 5–7

  - overview of, 179–180

Somers'  $D$ , 228, 240

Specificity, 380

Split-plot design, 106, 141, 144–145, 148–149, 154, 163

Split-plot factor, 144

Split-plot units, 144

S-plus (computer software), 18

Square-root transformation, 46

SS. *See* Sums of squares (SS)

Standard deviation, pooled, 51

Standard error

- of the estimate, 54

  - of Kaplan-Meier estimator, 298–300

Standard order, 107

Statistic(s)

- chi-square, simulation of, 186

- definition of, 33

- descriptive

  - categorical data and, 178–179

  - survival analysis and, 294–297

- deviance, 268, 277

- $F$ -, 53, 112, 157

- holistic view of, 102

- log-rank test, 303–304

- Wald, 221–222, 276

Statistically significant at level  $\alpha$ , 7, 22

Statistical model, in two-sample *t*-test, 32–34

Step function, 292

Stepwise regression, 70–71

- backward, 71

- forward, 70

- use of, 94

Study(ies)

- case-control, 190

- cohort, 190

- cross-classification, 189–190

- observational, 9, 103

- retrospective, 190–191

simulation

- with cards, 180

- computer

  - overview of, 180

- randomization test  
with, 5–7  
overview of, 179–180
- S**ubjective definition of probability, 370
- Sums of squares (SS), 52  
in analysis of variance (ANOVA), 120  
calculating, 155  
decomposition of, 83–85  
error, 52, 121  
group, 52  
total, 52, 121
- S**urvival analysis  
confidence intervals in, 297–300, 321  
definition of, 286  
descriptive statistics in, 294–297  
hazard function in, 305–311, 321  
cumulative, 311–317, 321  
estimated, 307–309  
population, 306–307  
interval censoring in, 317–318  
Kaplan-Meier curve and, 292–293  
Kaplan-Meier estimator and, 289–292, 320  
left censoring in, 317–318  
log-rank test in, 301–304  
mean survival time in, 295–296, 320  
median survival time in, 296  
Nelson-Aalen estimator  
in, 314  
overview of, 286–287  
percentiles in, 296  
regression in, 305  
truncation mechanisms in, 319–320  
Wilcoxon test in, 304
- S**urvival data, 286
- S**urvival function, 288–294, 300–305
- S**urvival time, 286
- T**
- T**able  
analysis of variance  
(ANOVA), 53, 85, 86  
contingency, 178  
 $2 \times 2$ , 178  
two-way contingency, 178
- Taxonomic autocorrelation, 76
- T**erms  
cubic, 90–91  
interaction
- degrees of freedom for, 122–123  
interpretation of, 113  
mixed, 160  
multiple regression and, 88–90  
quadratic, 90–91  
random error, 117
- T**est(s)  
chi-square, 183–186  
conducting, 185–186  
goodness-of-fit, 193–195, 233  
for homogeneity, 184  
for independence, 184  
simulation of statistic, 186
- deviance, 240  
change-in, 227  
drop-in, 226–228, 240
- distribution-free, 3
- F**-  
extra sum of squares, 86–87, 95  
for multiple regression, 83–87
- Fisher's exact, 181–182, 195
- for homogeneity of odds, 191–192
- Hosmer-Lemeshow, 233, 240
- of independence  
conditional, 182  
vs. homogeneity, 192–193
- Kruskal-Wallis, 18–20, 23
- likelihood ratio, 223–224, 240, 267
- log-rank, 301–304
- Mann-Whitney, 17–18, 23
- nonparametric, 3, 22
- overall, of model adequacy, 86
- parametric, 22
- permutation  
definition of, 22  
for matched pairs designs, 10–12  
randomization tests vs., 9–10
- randomization  
with computer simulation, 5–7  
by hand, 4–5  
for matched pairs designs, 10–12  
permutation test vs., 9–10  
statistical inference through, 4–5
- steps in, 22
- two-sample *t*-test and, 16–17
- of significance of regression model, 86
- two-sample *t*  
with equal variance assumption, 50–52  
model assumptions for, 34–35  
randomization test and, 16–17  
statistical model for, 32–34
- two-sided, 7–8
- two-sided hypothesis, 182–183
- Wald's, 221, 240, 266–267
- Wilcoxon rank sum, 17–18, 23, 304
- T**heory, developing model to confirm, 87–88
- T**hree-way factorial design, 113–115
- T**ied pairs, 228
- T**ime  
beginning of, 286  
event  
complete, 286  
incomplete, 286  
ordered complete, 291
- Time series data, 286
- Time series plot, for principal component analysis, 335–336
- Time-to-event data, 286
- Time-to-event random variable, 286
- Total estimated hazard, 314
- Total sum of squares ( $SS_{Total}$ ), 52, 121
- T**ransformation(s)  
for analysis of variance (ANOVA), 46–50  
arcsine, 46  
back, 47–48  
choice of, 49–50  
definition of, 46  
log, 46  
logistic regression model and, 216  
logit, 46, 216  
log log, 47  
reciprocal, 46  
for regression, 48  
square-root, 46  
use of, 56
- T**runcation  
censoring and, 319–320  
definition of, 319  
left, 319, 321  
mechanisms, 319–320
- right, 319, 321
- Tukey, John, 211, 332
- Tukey's honest significant difference, 126
- $2 \times 2$  table, 178
- Two-sample *t*-test  
with equal variance assumption, 50–52  
model assumptions for, 34–35  
randomization test and, 16–17  
statistical model in, 32–34
- Two-sided hypothesis tests, 182–183
- Two-sided tests, 7–8
- T**wo-way analysis of variance, 124–125
- Two-way contingency table, 178
- Type I error, 21
- Type II error, 21
- U**
- Uniform prior distribution, 385–387
- Unit, experimental, 105
- Unplanned systematic variability, 41
- Unrestricted model, 223
- V**
- Validation, model, 93–94
- V**ariability  
in analysis of variance (ANOVA), 110  
between-group, 110  
between-level, 53  
comparison of, 110–113  
direction of largest, 338–340  
planned, 41  
Poisson, 275–276  
of principal components, 353  
random, 41  
in random samples, 45–46  
sampling, 297  
unplanned systematic, 41  
within-group, 111
- V**ariable(s)  
binary, 178  
categorical, 178  
categorical explanatory, 81–83  
column, 178  
confounding, 142  
dummy, 36, 81  
explanatory  
in analysis of variance (ANOVA), 39

- Variable(s) (*Continued*)  
goodness-of-fit tests for  
continuous, 235  
multiple  
logistic regression,  
224–225  
notation for, 108–110  
extraneous, 41, 55, 105, 142  
independent and identically  
distributed, 34  
indicator, 36, 81–82  
lurking, 142  
response, identification  
of, 104  
row, 178  
selection criteria, 92–94,  
226, 240  
standardization of, 344–345  
time-to-event random, 286  
Variance inflation factor (VIF), 79  
Vector  
characteristic, 338  
eigenvector, 338, 339  
normalization of, 339  
notation, 336–337  
VIF. *See* Variance inflation factor  
(VIF)
- W**
- Wald confidence intervals, 222  
Wald statistic, 221–222, 276  
Wald's test, 221, 240, 266–267  
Watt, William Whyte, 137  
Whole-plot factor, 144  
Wilcoxon rank sum test, 17–18,  
23, 304  
Within-block factors, 145  
Within-group variability, 111