

# STATISTICS

*for the Life Sciences*

FOURTH EDITION



SAMUELS | WITMER | SCHAFFNER

---

# STATISTICS FOR THE LIFE SCIENCES

Fourth Edition

MYRA L. SAMUELS

*Purdue University*

JEFFREY A. WITMER

*Oberlin College*

ANDREW A. SCHAFFNER

*California Polytechnic State University,  
San Luis Obispo*

**Prentice Hall**

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

Editor-in-Chief: *Deirdre Lynch*  
Acquisitions Editor: *Christopher Cummings*  
Senior Content Editor: *Joanne Dill*  
Associate Editor: *Christina Lepre*  
Senior Managing Editor: *Karen Wernholm*  
Production Project Manager: *Patty Bergin*  
Digital Assets Manager: *Marianne Groth*  
Production Coordinator: *Katherine Roz*  
Associate Media Producer: *Nathaniel Koven*  
Marketing Manager: *Alex Gay*  
Marketing Assistant: *Kathleen DeChavez*  
Senior Author Support/Technology Specialist: *Joe Vetere*  
Permissions Project Supervisor: *Michael Joyce*  
Senior Manufacturing Buyer: *Carol Melville*  
Design Manager: *Andrea Nix*  
Cover Designer: *Christina Gleason*  
Interior Designer: *Tamara Newnam*  
Production Management/Composition: *Prepare*  
Art Studio: *Laserwords*  
Cover image: © Rudchenko Liliia/Shutterstock

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 Education was aware of a trademark claim, the designations have been printed in initial caps or all caps.

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

Samuels, Myra L.

Statistics for the life sciences / Myra Samuels, Jeffrey Witmer. -- 4th ed. / Andrew Schaffner.

p. cm.

Includes bibliographical references and index.

ISBN 978-0-321-65280-5

1. Biometry--Textbooks. 2. Medical statistics--Textbooks. 3.

Agriculture--Statistics--Textbooks. I. Witmer, Jeffrey A. II.

Schaffner, Andrew. III. Title.

QH323.5.S23 2012

570.1'5195--dc22

2010003559

Copyright: © 2012, 2003, 1999 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 10

**Prentice Hall**  
is an imprint of



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

ISBN-10: 0-321-65280-0  
ISBN-13: 978-0-321-65280-5

# CONTENTS

---

## Preface vii

### 1 INTRODUCTION 1

---

- 1.1 Statistics and the Life Sciences 1
- 1.2 Types of Evidence 7
- 1.3 Random Sampling 15

### 2 DESCRIPTION OF SAMPLES AND POPULATIONS 26

---

- 2.1 Introduction 26
- 2.2 Frequency Distributions 28
- 2.3 Descriptive Statistics: Measures of Center 40
- 2.4 Boxplots 45
- 2.5 Relationships between Variables 52
- 2.6 Measures of Dispersion 59
- 2.7 Effect of Transformation of Variables (Optional) 68
- 2.8 Statistical Inference 73
- 2.9 Perspective 79

### 3 PROBABILITY AND THE BINOMIAL DISTRIBUTION 84

---

- 3.1 Probability and the Life Sciences 84
- 3.2 Introduction to Probability 84
- 3.3 Probability Rules (Optional) 94
- 3.4 Density Curves 99
- 3.5 Random Variables 102
- 3.6 The Binomial Distribution 107
- 3.7 Fitting a Binomial Distribution to Data (Optional) 116

### 4 THE NORMAL DISTRIBUTION 121

---

- 4.1 Introduction 121
- 4.2 The Normal Curves 123
- 4.3 Areas Under a Normal Curve 125
- 4.4 Assessing Normality 132
- 4.5 Perspective 142

## 5 SAMPLING DISTRIBUTIONS 145

---

- 5.1 Basic Ideas 145
- 5.2 The Sample Mean 149
- 5.3 Illustration of the Central Limit Theorem (Optional) 159
- 5.4 The Normal Approximation to the Binomial Distribution (Optional) 162
- 5.5 Perspective 167

## 6 CONFIDENCE INTERVALS 170

---

- 6.1 Statistical Estimation 170
- 6.2 Standard Error of the Mean 171
- 6.3 Confidence Interval for  $\mu$  177
- 6.4 Planning a Study to Estimate  $\mu$  187
- 6.5 Conditions for Validity of Estimation Methods 190
- 6.6 Comparing Two Means 199
- 6.7 Confidence Interval for  $(\mu_1 - \mu_2)$  206
- 6.8 Perspective and Summary 212

## 7 COMPARISON OF TWO INDEPENDENT SAMPLES 218

---

- 7.1 Hypothesis Testing: The Randomization Test 218
- 7.2 Hypothesis Testing: The  $t$  Test 223
- 7.3 Further Discussion of the  $t$  Test 234
- 7.4 Association and Causation 242
- 7.5 One-Tailed  $t$  Tests 250
- 7.6 More on Interpretation of Statistical Significance 260
- 7.7 Planning for Adequate Power (Optional) 267
- 7.8 Student's  $t$ : Conditions and Summary 273
- 7.9 More on Principles of Testing Hypotheses 277
- 7.10 The Wilcoxon-Mann-Whitney Test 282
- 7.11 Perspective 291

## 8 COMPARISON OF PAIRED SAMPLES 299

---

- 8.1 Introduction 299
- 8.2 The Paired-Sample  $t$  Test and Confidence Interval 300
- 8.3 The Paired Design 310

- 8.4 The Sign Test 315
- 8.5 The Wilcoxon Signed-Rank Test 321
- 8.6 Perspective 326

## 9 CATEGORICAL DATA: ONE-SAMPLE DISTRIBUTIONS 336

---

- 9.1 Dichotomous Observations 336
- 9.2 Confidence Interval for a Population Proportion 341
- 9.3 Other Confidence Levels (Optional) 347
- 9.4 Inference for Proportions: The Chi-Square Goodness-of-Fit Test 348
- 9.5 Perspective and Summary 359

## 10 CATEGORICAL DATA: RELATIONSHIPS 363

---

- 10.1 Introduction 363
- 10.2 The Chi-Square Test for the  $2 \times 2$  Contingency Table 365
- 10.3 Independence and Association in the  $2 \times 2$  Contingency Table 373
- 10.4 Fisher's Exact Test (Optional) 381
- 10.5 The  $r \times k$  Contingency Table 385
- 10.6 Applicability of Methods 391
- 10.7 Confidence Interval for Difference between Probabilities 395
- 10.8 Paired Data and  $2 \times 2$  Tables (Optional) 398
- 10.9 Relative Risk and the Odds Ratio (Optional) 401
- 10.10 Summary of Chi-Square Test 409

## 11 COMPARING THE MEANS OF MANY INDEPENDENT SAMPLES 414

---

- 11.1 Introduction 414
- 11.2 The Basic One-Way Analysis of Variance 418
- 11.3 The Analysis of Variance Model 427
- 11.4 The Global  $F$  Test 429
- 11.5 Applicability of Methods 433
- 11.6 One-Way Randomized Blocks Design 437
- 11.7 Two-Way ANOVA 449
- 11.8 Linear Combinations of Means (Optional) 456
- 11.9 Multiple Comparisons (Optional) 464
- 11.10 Perspective 475

## 12 LINEAR REGRESSION AND CORRELATION 480

---

- 12.1** Introduction 480
- 12.2** The Correlation Coefficient 482
- 12.3** The Fitted Regression Line 492
- 12.4** Parametric Interpretation of Regression: The Linear Model 505
- 12.5** Statistical Inference Concerning  $\beta_1$  511
- 12.6** Guidelines for Interpreting Regression and Correlation 516
- 12.7** Precision in Prediction (Optional) 527
- 12.8** Perspective 531
- 12.9** Summary of Formulas 542

## 13 A SUMMARY OF INFERENCE METHODS 550

---

- 13.1** Introduction 550
- 13.2** Data Analysis Examples 552

**Appendices** 566

**Chapter Notes** 583

**Statistical Tables** 610

**Answers to Selected Exercises** 639

**Index** 647

**Index of Examples** 655

# PREFACE

---

*Statistics for the Life Sciences* is an introductory text in statistics, specifically addressed to students specializing in the life sciences. Its primary aims are (1) to show students how statistical reasoning is used in biological, medical, and agricultural research; (2) to enable students confidently to carry out simple statistical analyses and to interpret the results; and (3) to raise students' awareness of basic statistical issues such as randomization, confounding, and the role of independent replication.

## Style and Approach

The style of *Statistics for the Life Sciences* is informal and uses only minimal mathematical notation. There are no prerequisites except elementary algebra; anyone who can read a biology or chemistry textbook can read this text. It is suitable for use by graduate or undergraduate students in biology, agronomy, medical and health sciences, nutrition, pharmacy, animal science, physical education, forestry, and other life sciences.

**Use of Real Data** Real examples are more interesting and often more enlightening than artificial ones. *Statistics for the Life Sciences* includes hundreds of examples and exercises that use real data, representing a wide variety of research in the life sciences. Each example has been chosen to illustrate a particular statistical issue. The exercises have been designed to reduce computational effort and focus students' attention on concepts and interpretations.

**Emphasis on Ideas** The text emphasizes statistical ideas rather than computations or mathematical formulations. Probability theory is included only to support statistics concepts. Throughout the discussion of descriptive and inferential statistics, interpretation is stressed. By means of salient examples, the student is shown why it is important that an analysis be appropriate for the research question to be answered, for the statistical design of the study, and for the nature of the underlying distributions. The student is warned against the common blunder of confusing statistical nonsignificance with practical insignificance and is encouraged to use confidence intervals to assess the magnitude of an effect. The student is led to recognize the impact on real research of design concepts such as random sampling, randomization, efficiency, and the control of extraneous variation by blocking or adjustment. Numerous exercises amplify and reinforce the student's grasp of these ideas.

**The Role of Technology** The analysis of research data is usually carried out with the aid of a computer. Computer-generated graphs are shown at several places in the text. However, in studying statistics it is desirable for the student to gain experience working directly with data, using paper and pencil and a hand-held calculator, as well as a computer. This experience will help the student appreciate the nature and purpose of the statistical computations. The student is thus prepared to make intelligent use of the computer—to give it appropriate instructions and properly interpret the output. Accordingly, most of the exercises

in this text are intended for hand calculation. However, electronic data files are provided for many of the exercises, so that a computer can be used if desired. Selected exercises are identified as being intended to be completed with use of a computer. (Typically, the computer exercises require calculations that would be unduly burdensome if carried out by hand.)

## Organization

This text is organized to permit coverage in one semester of the maximum number of important statistical ideas, including power, multiple inference, and the basic principles of design. By including or excluding optional sections, the instructor can also use the text for a one-quarter course or a two-quarter course. It is suitable for a terminal course or for the first course of a sequence.

The following is a brief outline of the text.

**Chapter 1: Introduction.** The nature and impact of variability in biological data. The hazards of observational studies, in contrast with experiments. Random sampling.

**Chapter 2: Description of distributions.** Frequency distributions, descriptive statistics, the concept of population versus sample.

**Chapters 3, 4, and 5: Theoretical preparation.** Probability, binomial and normal distributions, sampling distributions.

**Chapter 6:** Confidence intervals for a single mean and for a difference in means.

**Chapter 7: Hypothesis testing, with emphasis on the  $t$  test.** The randomization test, the Wilcoxon-Mann-Whitney test.

**Chapter 8: Inference for paired samples.** Confidence interval,  $t$  test, sign test, and Wilcoxon signed-rank test.

**Chapter 9: Inference for a single proportion.** Confidence intervals and the chi-square goodness-of-fit test.

**Chapter 10: Relationships in categorical data.** Conditional probability, contingency tables. Optional sections cover Fisher's exact test, McNemar's test, and odds ratios.

**Chapter 11: Analysis of variance.** One-way layout, multiple comparison procedures, one-way blocked ANOVA, two-way ANOVA. Contrasts and multiple comparisons are included in optional sections.

**Chapter 12: Correlation and regression.** Descriptive and inferential aspects of correlation and simple linear regression and the relationship between them.

**Chapter 13: A summary of inference methods.**

Statistical tables are provided at the back of the book. The tables of critical values are especially easy to use, because they follow mutually consistent layouts and so are used in essentially the same way.

Optional appendices at the back of the book give the interested student a deeper look into such matters as how the Wilcoxon-Mann-Whitney null distribution is calculated.

## Changes to the Fourth Edition

- Some of the material that was in Chapter 8, on statistical principles of design, is now found in Chapter 1. Other parts of old Chapter 8 are now found sprinkled throughout the book, in the hope that students will come to appreciate that all statistical studies involve issues of data collection and scope of inference (much as appropriate graphics are not to be studied and used in isolation but are a central part of statistical analysis and thus appear throughout the book).
- Several other chapters have been reorganized. Changes include the following:
  - Inference for a single proportion has been moved from Chapter 6 to new Chapter 9.
  - The confidence interval for a difference in means has been moved from Chapter 7 to Chapter 6.
  - A new chapter (9) presents inference procedures for a categorical variable observed on a single sample.
  - Chapter 11 provides deeper treatment of two-way ANOVA and of multiple comparison procedures in analysis of variance.
  - Chapter 12 now begins with correlation and then moves to regression, rather than the other way around.
- 25% of the problems in the book are new or revised. As before, the majority are based on real data and draw from a variety of subjects of interest to life science majors. Selected data sets that are used in the problems and exercises are available online.
- The tables used for the sign test, signed-rank test, and Wilcoxon-Mann-Whitney test have been reorganized.

## Instructor Supplements

### Online Instructor's Solutions Manual

Solutions to all exercises are provided in this manual. Careful attention has been paid to ensure that all methods of solution and notation are consistent with those used in the core text. Available for download from Pearson Education's online catalog at [www.pearsonhighered.com/irc](http://www.pearsonhighered.com/irc).

### PowerPoint Slides

Selected figures and tables from throughout the textbook are available on PowerPoint slides for use in creating custom PowerPoint Lecture presentations. These slides are available for download at [www.pearsonhighered.com/irc](http://www.pearsonhighered.com/irc).

## Student Supplements

**Student's Solutions Manual** (ISBN-13: 978-0-321-69307-5;  
ISBN-10: 0-321-69307-8)

Fully worked out solutions to selected exercises are provided in this manual. Careful attention has been paid to ensure that all methods of solution and notation are consistent with those used in the core text.

## Technology Supplements and Packaging Options

### Data Sets

The larger data sets used in problems and exercises in the book are available as .csv files on the Pearson Statistics Resources and Data Sets website:

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

### StatCrunch™ eText (ISBN-13: 978-0-321-73050-3; ISBN-10: 0-321-73050-X)

This interactive, online textbook includes StatCrunch, a powerful, web-based statistical software. Embedded StatCrunch buttons allow users to open all data sets and tables from the book with the click of a button and immediately perform an analysis using StatCrunch.

### The Student Edition of Minitab (ISBN-13: 978-0-321-11313-9; ISBN-10: 0-321-11313-6)

The Student Edition of Minitab is a condensed edition of the professional release of Minitab statistical software. It offers the full range of statistical methods and graphical capabilities, along with worksheets that can include up to 10,000 data points. Individual copies of the software can be bundled with the text.

### JMP Student Edition (ISBN-13: 978-0-321-67212-4; ISBN-10: 0-321-67212-7)

JMP Student Edition is an easy-to-use, streamlined version of JMP desktop statistical discovery software from SAS Institute, Inc., and is available for bundling with the text.

### SPSS, an IBM Company<sup>†</sup> (ISBN-13: 978-0-321-67537-8; ISBN-10: 0-321-67537-1)

SPSS, a statistical and data management software package, is also available for bundling with the text.

### StatCrunch™

StatCrunch™ is web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data. Users can upload their own data to StatCrunch, or search the library of over twelve thousand publicly shared data sets, covering almost any topic of interest. Interactive graphical outputs help users understand statistical concepts, and are available for export to enrich reports with visual representations of data. Additional features include:

- A full range of numerical and graphical methods that allow users to analyze and gain insights from any data set.
- Reporting options that help users create a wide variety of visually-appealing representations of their data.

---

<sup>†</sup>SPSS was acquired by IBM in October 2009.

- An online survey tool that allows users to quickly build and administer surveys via a web form.

StatCrunch is available to qualified adopters. For more information, visit our website at [www.statcrunch.com](http://www.statcrunch.com), or contact your Pearson representative.

Study Cards are also available for various technologies, including Minitab, SPSS, JMP, StatCrunch, R, Excel and the TI Graphing Calculator.

## Acknowledgments for the Fourth Edition

The fourth edition of *Statistics for the Life Science* retains the style and spirit of the writing of Myra Samuels. Prior to her tragic death from cancer, Myra wrote the first edition of the text, based on her experience both as a teacher of statistics and as a statistical consultant. Without her vision and efforts there never would have been a first edition, let alone a fourth.

Many researchers have contributed sets of data to the text, which have enriched the text considerably. We have benefited from countless conversations over the years with David Moore, Dick Scheaffer, Murray Clayton, Alan Agresti, Don Bentley, and many others who have our thanks.

We are grateful for the sound editorial guidance and encouragement of Chris Cummings and Joanne Dill and the careful reading and valuable comments provided by Soma Roy. We are also grateful for adopters of the third edition who pointed out errors of various kinds. In particular, Robert Wolf and Jeff May sent us many suggestions that have led to improvements in the current edition. Finally, we express our gratitude to the reviewers of this edition:

Marjorie E. Bond (Monmouth College), James Grover (University of Texas—Arlington), Leslie Hendrix (University of South Carolina), Yi Huang (University of Maryland, Baltimore County), Lawrence Kamin (Benedictine University), Tiantian Qin (Purdue University), Dimitre Stefanov (University of Akron)

## Special Thanks

To Merrilee, for enduring yet more meals and evenings alone while I was writing.  
JAW

To Michelle and my sons, Ganden and Tashi, for their patience with me and enthusiasm about this book.

AAS

*This page intentionally left blank*

# 1

## INTRODUCTION

---

### Objectives

In this chapter we will look at a series of examples of areas in the life sciences in which statistics is used, with the goal of understanding the scope of the field of statistics. We will also

- explain how experiments differ from observational studies.
- discuss the concepts of placebo effect, blinding, and confounding.
- discuss the role of random sampling in statistics.

### 1.1 Statistics and the Life Sciences

Researchers in the life sciences carry out investigations in various settings: in the clinic, in the laboratory, in the greenhouse, in the field. Generally, the resulting data exhibit some *variability*. For instance, patients given the same drug respond somewhat differently; cell cultures prepared identically develop somewhat differently; adjacent plots of genetically identical wheat plants yield somewhat different amounts of grain. Often the degree of variability is substantial even when experimental conditions are held as constant as possible.

The challenge to the life scientist is to discern the patterns that may be more or less obscured by the variability of responses in living systems. The scientist must try to distinguish the “signal” from the “noise.”

Statistics is the science of understanding data and of making decisions in the face of variability and uncertainty. The discipline of statistics has evolved in response to the needs of scientists and others whose data exhibit variability. The concepts and methods of statistics enable the investigator to describe variability and to plan research so as to take variability into account (i.e., to make the “signal” strong in comparison to the background “noise” in data that are collected). Statistical methods are used to analyze data so as to extract the maximum information and also to quantify the reliability of that information.

We begin with some examples that illustrate the degree of variability found in biological data and the ways in which variability poses a challenge to the biological researcher. We will briefly consider examples that illustrate some of the statistical issues that arise in life sciences research and indicate where in this book the issues are addressed.

The first two examples provide a contrast between an experiment that showed no variability and another that showed considerable variability.

**Example  
I.I.1**

**Vaccine for Anthrax** Anthrax is a serious disease of sheep and cattle. In 1881, Louis Pasteur conducted a famous experiment to demonstrate the effect of his vaccine against anthrax. A group of 24 sheep were vaccinated; another group of 24 unvaccinated sheep served as controls. Then, all 48 animals were inoculated with a virulent culture of anthrax bacillus. Table 1.1.1 shows the results.<sup>1</sup> The data of Table 1.1.1 show no variability; all the vaccinated animals survived and all the unvaccinated animals died.

**Table 1.1.1** Response of sheep to anthrax

Response	Treatment	
	Vaccinated	Not vaccinated
Died of anthrax	0	24
Survived	24	0
Total	24	24
Percent survival	100%	0%

**Example  
I.I.2**

**Bacteria and Cancer** To study the effect of bacteria on tumor development, researchers used a strain of mice with a naturally high incidence of liver tumors. One group of mice were maintained entirely germ free, while another group were exposed to the intestinal bacteria *Escherichia coli*. The incidence of liver tumors is shown in Table 1.1.2.<sup>2</sup>

**Table 1.1.2** Incidence of liver tumors in mice

Response	Treatment	
	<i>E. coli</i>	Germ free
Liver tumors	8	19
No liver tumors	5	30
Total	13	49
Percent with liver tumors	62%	39%

In contrast to Table 1.1.1, the data of Table 1.1.2 show variability; mice given the same treatment did not all respond the same way. Because of this variability, the results in Table 1.1.2 are equivocal; the data suggest that exposure to *E. coli* increases the risk of liver tumors, but the possibility remains that the observed difference in percentages (62% versus 39%) might reflect only chance variation rather than an effect of *E. coli*. If the experiment were replicated with different animals, the percentages might change substantially.

One way to explore what might happen if the experiment were replicated is to simulate the experiment, which could be done as follows. Take 62 cards and write “liver tumors” on 27 (= 8 + 19) of them and “no liver tumors” on the other 35 (= 5 + 30). Shuffle the cards and randomly deal 13 cards into one stack (to correspond to the *E. coli* mice) and 49 cards into a second stack. Next, count the number of cards in the “*E. coli* stack” that have the words “liver tumors” on them—to correspond to mice exposed to *E. coli* who develop liver tumors—and record whether this number is greater than or equal to 8. This process represents distributing 27 cases of liver tumors to two groups of mice (*E. coli* and germ free) randomly, with *E. coli* mice no more likely, nor any less likely, than germ-free mice to end up with liver tumors.

If we repeat this process many times (say, 10,000 times, with the aid of a computer in place of a physical deck of cards), it turns out that roughly 12% of the time we get 8 or more *E. coli* mice with liver tumors. Since something that happens 12% of the time is not terribly surprising, Table 1.1.2 does not provide significant evidence that exposure to *E. coli* increases the incidence of liver tumors. ■

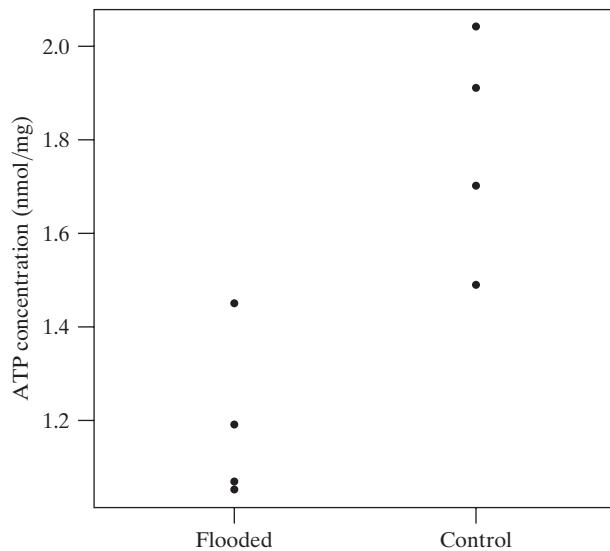
In Chapter 10 we will discuss statistical techniques for evaluating data such as those in Tables 1.1.1 and 1.1.2. Of course, in some experiments variability is minimal and the message in the data stands out clearly without any special statistical analysis. It is worth noting, however, that absence of variability is itself an experimental result that must be justified by sufficient data. For instance, because Pasteur's anthrax data (Table 1.1.1) show no variability at all, it is intuitively plausible to conclude that the data provide "solid" evidence for the efficacy of the vaccination. But note that this conclusion involves a judgment; consider how much *less* "solid" the evidence would be if Pasteur had included only 3 animals in each group, rather than 24. Statistical analyses can be used to make such a judgment, that is, to determine if the variability is indeed negligible. Thus, a statistical view can be helpful even in the absence of variability.

The next two examples illustrate additional questions that a statistical approach can help to answer.

### Example 1.1.3

**Flooding and ATP** In an experiment on root metabolism, a plant physiologist grew birch tree seedlings in the greenhouse. He flooded four seedlings with water for one day and kept four others as controls. He then harvested the seedlings and analyzed the roots for adenosine triphosphate (ATP). The measured amounts of ATP (nmoles per mg tissue) are given in Table 1.1.3 and displayed in Figure 1.1.1.<sup>3</sup>

<b>Table 1.1.3</b> ATP concentration in birch tree roots (nmol/mg)	
Flooded	Control
1.45	1.70
1.19	2.04
1.05	1.49
1.07	1.91



**Figure 1.1.1** ATP concentration in birch tree roots

The data of Table 1.1.3 raise several questions: How should one summarize the ATP values in each experimental condition? How much information do the data provide about the effect of flooding? How confident can one be that the reduced ATP in the flooded group is really a response to flooding rather than just random variation? What size experiment would be required in order to firmly corroborate the apparent effect seen in these data? ■

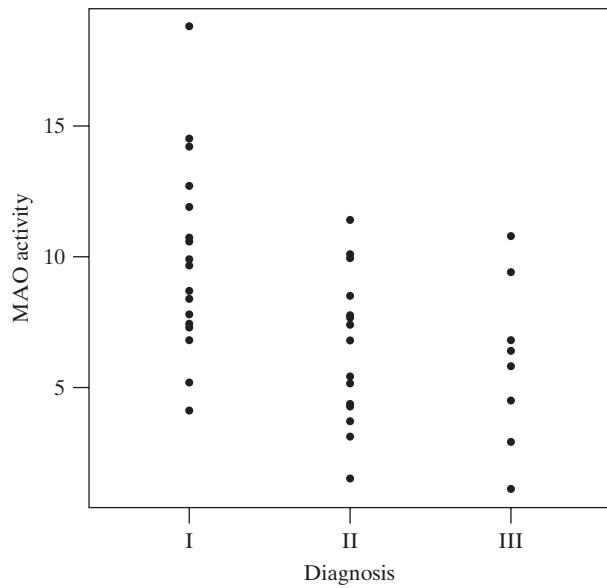
Chapters 2, 6, and 7 address questions like those posed in Example 1.1.3. One question that we can address here is whether the data in Table 1.1.3 are consistent with the claim that flooding has no effect on ATP concentration, or instead provide significant evidence that flooding affects ATP concentrations. If the claim of no effect is true, then should we be surprised to see that all four of the flooded observations are smaller than each of the control observations? Might this happen by chance alone? If we wrote each of the numbers 1.05, 1.07, 1.19, 1.45, 1.49, 1.91, 1.70, and 2.04 on cards, shuffled the eight cards, and randomly dealt them into two piles, what is the chance that the four smallest numbers would end up in one pile and the four largest numbers in the other pile? It turns out that we could expect this to happen 1 time in 35 random shufflings, so “chance alone” would only create the kind of imbalance seen in Figure 1.1.1 about 2.9% of the time (since  $1/35 = 0.029$ ). Thus, we have some evidence that flooding has an effect on ATP concentration. We will develop this idea more fully in Chapter 7.

### Example 1.1.4

**MAO and Schizophrenia** Monoamine oxidase (MAO) is an enzyme that is thought to play a role in the regulation of behavior. To see whether different categories of schizophrenic patients have different levels of MAO activity, researchers collected blood specimens from 42 patients and measured the MAO activity in the platelets. The results are given in Table 1.1.4 and displayed in Figure 1.1.2. (Values are expressed as nmol benzylaldehyde product per  $10^8$  platelets per hour.)<sup>4</sup> Note that it is much easier to get a feeling for the data by looking at the graph (Figure 1.1.2) than it is to read through the data in the table. The use of graphical displays of data is a very important part of data analysis. ■

**Table 1.1.4** MAO activity in schizophrenic patients

Diagnosis		MAO activity				
I:		6.8	4.1	7.3	14.2	18.8
Chronic undifferentiated		9.9	7.4	11.9	5.2	7.8
schizophrenic	(18 patients)	7.8	8.7	12.7	14.5	10.7
		8.4	9.7	10.6		
II:		7.8	4.4	11.4	3.1	4.3
Undifferentiated with		10.1	1.5	7.4	5.2	10.0
paranoid features	(16 patients)	3.7	5.5	8.5	7.7	6.8
		3.1				
III:		6.4	10.8	1.1	2.9	4.5
Paranoid schizophrenic	(8 patients)	5.8	9.4	6.8		



**Figure 1.1.2** MAO activity in schizophrenic patients

To analyze the MAO data, one would naturally want to make comparisons among the three groups of patients, to describe the reliability of those comparisons, and to characterize the variability within the groups. To go beyond the data to a biological interpretation, one must also consider more subtle issues, such as the following: How were the patients selected? Were they chosen from a common hospital

population, or were the three groups obtained at different times or places? Were precautions taken so that the person measuring the MAO was unaware of the patient's diagnosis? Did the investigators consider various ways of subdividing the patients before choosing the particular diagnostic categories used in Table 1.1.4? At first glance, these questions may seem irrelevant—can we not let the measurements speak for themselves? We will see, however, that the proper interpretation of data always requires careful consideration of how the data were obtained.

Chapters 2, 3, and 8 include discussions of selection of experimental subjects and of guarding against unconscious investigator bias. In Chapter 11 we will show how sifting through a data set in search of patterns can lead to serious misinterpretations and we will give guidelines for avoiding the pitfalls in such searches.

The next example shows how the effects of variability can distort the results of an experiment and how this distortion can be minimized by careful design of the experiment.

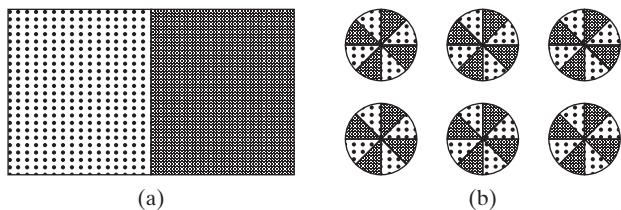
### Example 1.1.5

**Food Choice by Insect Larvae** The clover root curculio, *Sitona hispidulus*, is a root-feeding pest of alfalfa. An entomologist conducted an experiment to study food choice by *Sitona* larvae. She wished to investigate whether larvae would preferentially choose alfalfa roots that were nodulated (their natural state) over roots whose nodulation had been suppressed. Larvae were released in a dish where both nodulated and nonnodulated roots were available. After 24 hours, the investigator counted the larvae that had clearly made a choice between root types. The results are shown in Table 1.1.5.<sup>5</sup>

The data in Table 1.1.5 appear to suggest rather strongly that *Sitona* larvae prefer nodulated roots. But our description of the experiment has obscured an important point—we have not stated how the roots were arranged. To see the relevance of the arrangement, suppose the experimenter had used only one dish, placing all the nodulated roots on one side of the dish and all the nonnodulated roots on the other side, as shown in Figure 1.1.3(a), and had then released 120 larvae in the center of the dish. This experimental arrangement would be seriously deficient, because the data of Table 1.1.5 would then permit several competing interpretations—for instance, (a) perhaps the larvae really do prefer nodulated roots; or (b) perhaps the two sides of the dish were at slightly different temperatures and the larvae were responding to temperature rather than nodulation; or (c) perhaps one larva chose the nodulated roots just by chance and the other larvae followed its trail. Because of these possibilities the experimental arrangement shown in Figure 1.1.3(a) can yield only weak information about larval food preference.

**Table 1.1.5** Food choice by *Sitona* larvae

Choice	Number of larvae
Chose nodulated roots	46
Chose nonnodulated roots	12
Other (no choice, died, lost)	62
Total	120



**Figure 1.1.3** Possible arrangements of food choice experiment. The dark-shaded areas contain nodulated roots and the light-shaded areas contain nonnodulated roots.

- (a) A poor arrangement.
- (b) A good arrangement.

The experiment was actually arranged as in Figure 1.1.3(b), using six dishes with nodulated and nonnodulated roots arranged in a symmetric pattern. Twenty larvae were released into the center of each dish. This arrangement avoids the pitfalls of the arrangement in Figure 1.1.3(a). Because of the alternating regions of nodulated and nonnodulated roots, any fluctuation in environmental conditions (such as temperature) would tend to affect the two root types equally. By using several dishes, the experimenter has generated data that can be interpreted even if the larvae do tend to follow each other. To analyze the experiment properly, we would need to know the results in each dish; the condensed summary in Table 1.1.5 is not adequate.

In Chapter 11 we will describe various ways of arranging experimental material in space and time so as to yield the most informative experiment, as well as how to analyze the data to extract as much information as possible and yet resist the temptation to overinterpret patterns that may represent only random variation.

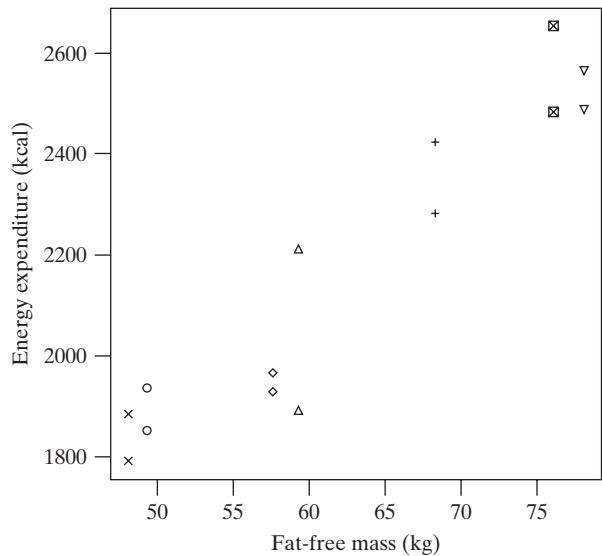
The following example is a study of the relationship between two measured quantities.

### Example 1.1.6

**Body Size and Energy Expenditure** How much food does a person need? To investigate the dependence of nutritional requirements on body size, researchers used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure during conditions of quiet sedentary activity; this was repeated twice for each subject. The results are shown in Table 1.1.6 and plotted in Figure 1.1.4.<sup>6</sup>

**Table 1.1.6** Fat-free mass and energy expenditure

Subject	Fat-free mass (kg)	24-hour energy expenditure (kcal)
1	49.3	1,851
2	59.3	2,209
3	68.3	2,283
4	48.1	1,885
5	57.6	1,929
6	78.1	2,490
7	76.1	2,484



**Figure 1.1.4** Fat-free mass and energy expenditure in seven men. Each man is represented by a different symbol.

A primary goal in the analysis of these data would be to describe the relationship between fat-free mass and energy expenditure—to characterize not only the overall trend of the relationship, but also the degree of scatter or variability in the relationship. (Note also that, to analyze the data, one needs to decide how to handle the duplicate observations on each subject.)

The focus of Example 1.1.6 is on the relationship between two variables: fat-free mass and energy expenditure. Chapter 12 deals with methods for describing such relationships, and also for quantifying the reliability of the descriptions.

## A Look Ahead

Where appropriate, statisticians make use of the computer as a tool in data analysis; computer-generated output and statistical graphics appear throughout this book. The computer is a powerful tool, but it must be used with caution. Using the computer to perform calculations allows us to concentrate on concepts. The danger when using a computer in statistics is that we will jump straight to the calculations without looking closely at the data and asking the right questions about the data. Our goal is to analyze, understand, and interpret data—which are numbers *in a specific context*—not just to perform calculations.

In order to understand a data set it is necessary to know how and why the data were collected. In addition to considering the most widely used methods in statistical inference, we will consider issues in data collection and experimental design. Together, these topics should provide the reader with the background needed to read the scientific literature and to design and analyze simple research projects.

The preceding examples illustrate the kind of data to be considered in this book. In fact, each of the examples will reappear as an exercise or example in an appropriate chapter. As the examples show, research in the life sciences is usually concerned with the comparison of two or more groups of observations, or with the relationship between two or more variables. We will begin our study of statistics by focusing on a simpler situation—observations of a *single* variable for a *single* group. Many of the basic ideas of statistics will be introduced in this oversimplified context. Two-group comparisons and more complicated analyses will then be discussed in Chapter 7 and later chapters.

## 1.2 Types of Evidence

Researchers gather information and make inferences about the state of nature in a variety of settings. Much of statistics deals with the *analysis* of data, but statistical considerations often play a key role in the planning and *design* of a scientific investigation. We begin with examples of the three major kinds of evidence that one encounters.

---

### Example 1.2.1

**Lightning and Deafness** On 15 July 1911, 65-year-old Mrs. Jane Decker was struck by lightning while in her house. She had been deaf since birth, but after being struck, she recovered her hearing, which led to a headline in the *New York Times*, “Lightning Cures Deafness.”<sup>7</sup> Is this compelling evidence that lightning is a cure for deafness? Could this event have been a coincidence? Are there other explanations for her cure? ■

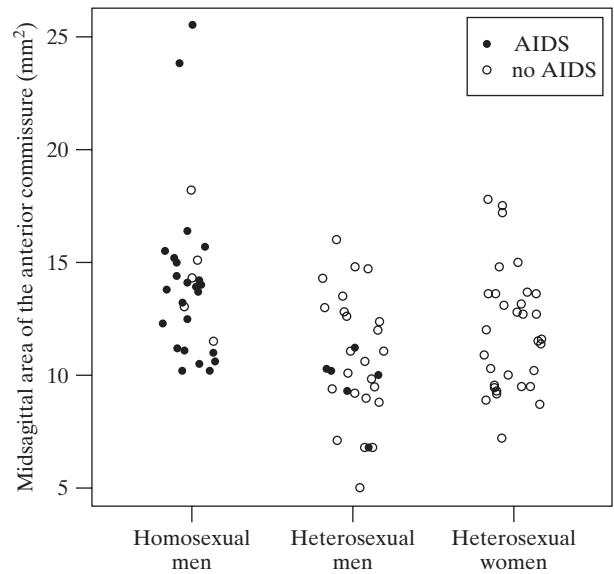
The evidence discussed in Example 1.2.1 is **anecdotal evidence**. An anecdote is a short story or an example of an interesting event, in this case, of lightning curing deafness. The accumulation of anecdotes often leads to conjecture and to scientific investigation, but it is predictable pattern, not anecdote, that establishes a scientific theory.

**Example  
1.2.2**

**Sexual Orientation** Some research has suggested that there is a genetic basis for sexual orientation. One such study involved measuring the midsagittal area of the anterior commissure (AC) of the brain for 30 homosexual men, 30 heterosexual men, and 30 heterosexual women. The researchers found that the AC tends to be larger in heterosexual women than in heterosexual men and that it is even larger in homosexual men. These data are summarized in Table 1.2.1 and are shown graphically in Figure 1.2.1.

**Table 1.2.1** Midsagittal area of the anterior commissure ( $\text{mm}^2$ )

Group	Average midsagittal area ( $\text{mm}^2$ ) of the anterior commissure
Homosexual men	14.20
Heterosexual men	10.61
Heterosexual women	12.03



**Figure 1.2.1** Midsagittal area of the anterior commissure ( $\text{mm}^2$ )

The data suggest that the size of the AC in homosexual men is more like that of heterosexual women than that of heterosexual men. When analyzing these data, we should take into account two things. (1) The measurements for two of the homosexual men were much larger than any of the other measurements; sometimes one or two such outliers can have a big impact on the conclusions of a study. (2) Twenty-four of the 30 homosexual men had AIDS, as opposed to 6 of the 30 heterosexual men; if AIDS affects the size of the anterior commissure, then this factor could account for some of the difference between the two groups of men.<sup>8</sup>

Example 1.2.2 presents an **observational study**. In an observational study the researcher systematically collects data from subjects, but only as an observer and not as someone who is manipulating conditions. By systematically examining all the data that arise in observational studies, one can guard against selectively viewing and reporting only evidence that supports a previous view. However, observational studies can be misleading due to *confounding variables*. In Example 1.2.2 we noted that having AIDS may affect the size of the anterior commissure. We would say that the effect of AIDS is confounded with the effect of sexual orientation in this study.

Note that the *context* in which the data arose is of central importance in statistics. This is quite clear in Example 1.2.2. The numbers themselves can be used to compute averages or to make graphs, like Figure 1.2.1, but if we are to understand what the data have to say, we must have an understanding of the context in which they arose. This context tells us to be on the alert for the effects that other factors, such as the impact of AIDS, may have on the size of the anterior commissure. Data analysis without reference to context is meaningless.

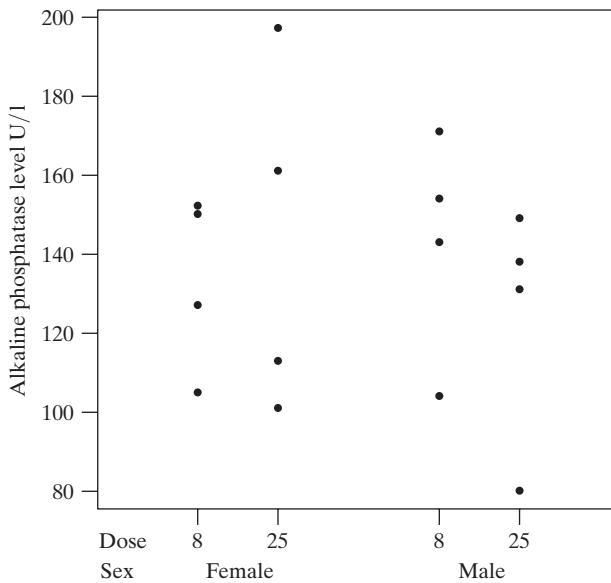
**Example 1.2.3**

**Health and Marriage** A study conducted in Finland found that people who were married at midlife were less likely to develop cognitive impairment (particularly Alzheimer's disease) later in life.<sup>9</sup> However, from an observational study such as this we don't know whether marriage *prevents* later problems or whether persons who are likely to develop cognitive problems are less likely to get married. ■

**Example 1.2.4**

**Toxicity in Dogs** Before new drugs are given to human subjects, it is common practice to first test them in dogs or other animals. In part of one study, a new investigational drug was given to eight male and eight female dogs at doses of 8 mg/kg and 25 mg/kg. Within each sex, the two doses were assigned at random to the eight dogs. Many "endpoints" were measured, such as cholesterol, sodium, glucose, and so on, from blood samples, in order to screen for toxicity problems in the dogs before starting studies on humans. One endpoint was alkaline phosphatase level (or APL, measured in U/l). The data are shown in Table 1.2.2 and plotted in Figure 1.2.2.<sup>10</sup>

<b>Table 1.2.2</b> Alkaline phosphatase level (U/l)		
Dose (mg/kg)	Male	Female
8	171	150
	154	127
	104	152
	143	105
Average	<b>143</b>	<b>133.5</b>
25	80	101
	149	113
	138	161
	131	197
Average	<b>124.5</b>	<b>143</b>



**Figure 1.2.2** Alkaline phosphatase level in dogs

The design of this experiment allows for the investigation of the interaction between two factors: sex of the dog and dose. These factors interacted in the following sense: For females, the effect of increasing the dose from 8 to 25 mg/kg was positive, although small (the average APL increased from 133.5 to 143 U/l), but for males the effect of increasing the dose from 8 to 25 mg/kg was negative (the average APL dropped from 143 to 124.5 U/l). Techniques for studying such interactions will be considered in Chapter 11. ■

Example 1.2.4 presents an **experiment**, in that the researchers imposed the conditions—in this case, doses of a drug—on the subjects (the dogs). By randomly assigning treatments (drug doses) to subjects (dogs), we can get around the problem of confounding that complicates observational studies and limits the conclusions that we can reach from them. Randomized experiments are considered the “gold standard” in scientific investigation, but they can also be plagued by difficulties.

Often human subjects in experiments are given a **placebo**—an inert substance, such as a sugar pill. It is well known that people often exhibit a *placebo response*; that is, they tend to respond favorably to *any* treatment, even if it is only inert. This psychological effect can be quite powerful. Research has shown that placebos are effective for roughly one-third of people who are in pain; that is, one-third of pain sufferers report their pain ending after being given a “painkiller” that is, in fact, an inert pill. For diseases such as bronchial asthma, angina pectoris (recurrent chest pain caused by decreased blood flow to the heart), and ulcers, the use of placebos has been shown to produce clinically beneficial results in over 60% of patients.<sup>11</sup> Of course, if a placebo control is used, then the subjects must not be told which group they are in—the group getting the active treatment or the group getting the placebo.

---

**Example**  
**1.2.5**

**Autism** Autism is a serious condition in which children withdraw from normal social interactions and sometimes engage in aggressive or repetitive behavior. In 1997, an autistic child responded remarkably well to the digestive enzyme secretin. This led to an experiment (a “clinical trial”) in which secretin was compared to a placebo. In this experiment, children who were given secretin improved considerably. However, the children given the placebo also improved considerably. There was no statistically significant difference between the two groups. Thus, the favorable response in the secretin group was considered to be only a “placebo response,” meaning, unfortunately, that secretin was not found to be beneficial (beyond inducing a positive response associated simply with taking a substance as part of an experiment).<sup>12</sup> ■

The word *placebo* means “I shall please.” The word *nocebo* (“I shall harm”) is sometimes used to describe adverse reactions to perceived, but nonexistent, risks. The following example illustrates the strength that psychological effects can have.

---

**Example**  
**1.2.6**

**Bronchial Asthma** A group of patients suffering from bronchial asthma were given a substance that they were told was a chest-constricting chemical. After being given this substance, several of the patients experienced bronchial spasms. However, during part of the experiment, the patients were given a substance that they were told would alleviate their symptoms. In this case, bronchial spasms were prevented. In reality, the second substance was identical to the first substance: Both were distilled water. It appears that it was the power of suggestion that brought on the bronchial spasms; the same power of suggestion prevented spasms.<sup>13</sup> ■

Similar to placebo treatment is *sham* treatment, which can be used on animals as well as humans. An example of sham treatment is injecting control animals with an inert substance such as saline. In some studies of surgical treatments, control animals (even, occasionally, humans) are given a “mock” surgery.

---

**Example**  
**1.2.7**

**Mammary Artery Ligation** In the 1950s, the surgical technique of internal mammary artery ligation became a popular treatment for patients suffering from angina pectoris. In this operation the surgeon would ligate (tie) the mammary artery, with the goal of increasing collateral blood flow to the heart. Doctors and patients alike enthusiastically endorsed this surgery as an effective treatment. In 1958, studies of internal mammary artery ligation in animals found that it was not effective and this raised doubts about its usefulness on humans. A study was conducted in which patients were randomly assigned to one of two groups. Patients in the treatment

group received the standard surgery. Patients in the control group received a sham operation in which an incision was made, the mammary artery was exposed as in the real operation, but the incision was closed *without* the artery being ligated. These patients had no way of knowing that their operation was a sham. The rates of improvement in the two groups of patients were nearly identical. (Patients who had the sham operation did slightly better than patients who had the real operation, but the difference was small.) A second randomized, controlled study also found that patients who received the sham surgery did as well as those who had the real operation. As a result of these studies, physicians stopped using internal mammary artery ligation.<sup>14</sup>

## Blinding

In experiments on humans, particularly those that involve the use of placebos, **blinding** is often used. This means that the treatment assignment is kept secret from the experimental subject. The purpose of blinding the subject is to minimize the extent to which his or her expectations influence the results of the experiment. If subjects exhibit a psychological reaction to getting a medication, that placebo response will tend to balance out between the two groups, so that any difference between the groups can be attributed to the effect of the active treatment.

In many experiments the persons who evaluate the responses of the subjects are also kept blind; that is, during the experiment they are kept ignorant of the treatment assignment. Consider, for instance, the following:

In a study to compare two treatments for lung cancer, a radiologist reads X-rays to evaluate each patient's progress. The X-ray films are coded so that the radiologist cannot tell which treatment each patient received.

Mice are fed one of three diets; the effects on their liver are assayed by a research assistant who does not know which diet each mouse received.

Of course, *someone* needs to keep track of which subject is in which group, but that person should not be the one who measures the response variable. The most obvious reason for blinding the person making the evaluations is to reduce the possibility of subjective bias influencing the observation process itself. Someone who *expects* or *wants* certain results may unconsciously influence those results. Such bias can enter even apparently “objective” measurements through subtle variation in dissection techniques, titration procedures, and so on.

In medical studies of human beings, blinding often serves additional purposes. For one thing, a patient must be asked whether he or she consents to participate in a medical study. If the physician who asks the question already knows which treatment the patient would receive, then by discouraging certain patients and encouraging others, the physician can (consciously or unconsciously) create noncomparable treatment groups. The effect of such biased assignment can be surprisingly large, and it has been noted that it generally favors the “new” or “experimental” treatment.<sup>15</sup> Another reason for blinding in medical studies is that a physician may (consciously or unconsciously) provide more psychological encouragement, or even better care, to the patients who are receiving the treatment that the physician regards as superior.

An experiment in which both the subjects and the persons making the evaluations of the response are blinded is called a **double-blind** experiment. The first mammary artery ligation experiment described in Example 1.2.7 was conducted as a double-blind experiment.

## The Need for Control Groups

### Example 1.2.8

**Clofibrate** An experiment was conducted in which subjects were given the drug clofibrate, which was intended to lower cholesterol and reduce the chance of death from coronary disease. The researchers noted that many of the subjects did not take all the medication that the experimental protocol called for them to take. They calculated the percentage of the prescribed capsules that each subject took and divided the subjects into two groups according to whether or not the subjects took at least 80% of the capsules they were given. Table 1.2.3 shows that the five-year mortality rate for those who took at least 80% of their capsules was much lower than the corresponding rate for subjects who did not adhere to the protocol. On the surface, this suggests that taking the medication lowers the chance of death. However, there was a placebo control group in the experiment and many of the placebo subjects took fewer than 80% of their capsules. The mortality rates for the two placebo groups—those who adhered to the protocol and those who did not—are quite similar to the rates for the clofibrate groups.

**Table 1.2.3** Mortality rates for the clofibrate experiment

	Clofibrate		Placebo	
Adherence	n	5-year mortality	n	5-year mortality
≥80%	708	15.0%	1813	15.1%
<80%	357	24.6%	882	28.2%

The clofibrate experiment seems to indicate that there are two kinds of subjects: those who adhere to the protocol and those who do not. The first group had a much lower mortality rate than the second group. This might be due simply to better health habits among people who are willing to follow a scientific protocol for five years than among people who don't adhere to the protocol. A further conclusion from the experiment is that clofibrate does not appear to be any more effective than placebo in reducing the death rate. Were it not for the presence of the placebo control group, the researchers might well have drawn the wrong conclusion from the study and attributed the lower death rate among adherers to clofibrate itself, rather than to other confounded effects that make the adherers different from the nonadherers.<sup>16</sup> ■

### Example 1.2.9

**The Common Cold** Many years ago, investigators invited university students who believed themselves to be particularly susceptible to the common cold to be part of an experiment. Volunteers were randomly assigned to either the treatment group, in which case they took capsules of an experimental vaccine, or to the control group, in which case they were told that they were taking a vaccine, but in fact were given a placebo—capsules that looked like the vaccine capsules but that contained lactose in place of the vaccine.<sup>17</sup> As shown in Table 1.2.4, both groups reported having dramatically fewer colds during the study than they had had in the previous year.

**Table 1.2.4** Number of colds in cold-vaccine experiment

	Vaccine	Placebo
n	201	203
Average number of colds		
Previous year (from memory)	5.6	5.2
Current year	1.7	1.6
% reduction	70%	69%

The average number of colds per person dropped 70% in the treatment group. This would have been startling evidence that the vaccine had an effect, except that the corresponding drop in the control group was 69%. ■

We can attribute much of the large drop in colds in Example 1.2.9 to the placebo effect. However, another statistical concern is **panel bias**, which is bias attributable to the study having influenced the behavior of the subjects—that is, people who know they are being studied often change their behavior. The students in this study reported from memory the number of colds they had suffered in the previous year. The fact that they were part of a study might have influenced their behavior, so that they were less likely to catch a cold during the study. Being in a study might also have affected the way in which they defined having a cold—during the study, they were “instructed to report to the health service whenever a cold developed”—so that some illness may have gone unreported during the study. (How sick do you have to be before you classify yourself as having a cold?)

## Historical Controls

Researchers may be particularly reluctant to use randomized allocation in medical experiments on human beings. Suppose, for instance, that researchers want to evaluate a promising new treatment for a certain illness. It can be argued that it would be unethical to withhold the treatment from any patients, and that therefore all current patients should receive the new treatment. But then who would serve as a control group? One possibility is to use historical controls—that is, previous patients with the same illness who were treated with another therapy. One difficulty with historical controls is that there is often a tendency for later patients to show a better response—even to the same therapy—than earlier patients with the same diagnosis. This tendency has been confirmed, for instance, by comparing experiments conducted at the same medical centers in different years.<sup>18</sup> One major reason for the tendency is that the overall characteristics of the patient population may change with time. For instance, because diagnostic techniques tend to improve, patients with a given diagnosis (say, breast cancer) in 2001 may have a better chance of recovery (even with the same treatment) than those with the same diagnosis in 1991, because they were diagnosed earlier in the course of the disease.

Medical researchers do not agree on the validity and value of historical controls. The following example illustrates the importance of this controversial issue.

---

### Example 1.2.10

**Coronary Artery Disease** Disease of the coronary arteries is often treated by surgery (such as bypass surgery), but it can also be treated with drugs only. Many studies have attempted to evaluate the effectiveness of surgical treatment for this common disease. In a review of 29 of these studies, each study was classified as to whether it used randomized controls or historical controls; the conclusions of the 29 studies are summarized in Table 1.2.5.<sup>19</sup>

**Table 1.2.5** Coronary artery disease studies

Type of controls	Conclusion about effectiveness of surgery		Total number of studies
	Effective	Not effective	
Randomized	1	7	8
Historical	16	5	21

It would appear from Table 1.2.5 that enthusiasm for surgery is much more common among researchers who use historical controls than among those who use randomized controls.

Proponents of the use of historical controls argue that statistical adjustment can provide meaningful comparison between a current group of patients and a group of historical controls; for instance, if the current patients are younger than the historical controls, then the data can be analyzed in a way that adjusts, or corrects, for the effect of age. Critics reply that such adjustment may be grossly inadequate.

The concept of historical controls is not limited to medical studies. The issue arises whenever a researcher compares current data with past data. Whether the data are from the lab, the field, or the clinic, the researcher must confront the question: Can the past and current results be meaningfully compared? One should always at least ask whether the experimental material, and/or the environmental conditions, may have changed enough over time to distort the comparison.

## Exercises 1.2.1–1.2.8

**1.2.1** Fluoridation of drinking water has long been a controversial issue in the United States. One of the first communities to add fluoride to their water was Newburgh, New York. In March 1944, a plan was announced to begin to add fluoride to the Newburgh water supply on April 1 of that year. During the month of April, citizens of Newburgh complained of digestive problems, which were attributed to the fluoridation of the water. However, there had been a delay in the installation of the fluoridation equipment, so that fluoridation did not begin until May 2.<sup>20</sup> Explain how the placebo effect/nocebo effect is related to this example.

**1.2.2** Olestra is a no-calorie, no-fat additive that is used in the production of some potato chips. After the Food and Drug Administration approved the use of olestra, some consumers complained that olestra caused stomach cramps and diarrhea. A randomized, double-blind experiment was conducted in which some subjects were given bags of potato chips made with olestra and other subjects were given ordinary potato chips. In the olestra group, 38% of the subjects reported having gastrointestinal symptoms. However, in the group given regular potato chips the corresponding percentage was 37%. (The two percentages are not statistically significantly different.)<sup>21</sup> Explain how the placebo effect/nocebo effect is related to this example. Also explain why it was important for this experiment to be double-blind.

**1.2.3 (Hypothetical)** In a study of acupuncture, patients with headaches are randomly divided into two groups. One group is given acupuncture and the other group is given aspirin. The acupuncturist evaluates the effectiveness of the acupuncture and compares it to the results from the aspirin group. Explain how lack of blinding biases the experiment in favor of acupuncture.

**1.2.4** Randomized, controlled experiments have found that vitamin C is not effective in treating terminal cancer

patients.<sup>22</sup> However, a 1976 research paper reported that terminal cancer patients given vitamin C survived much longer than did historical controls. The patients treated with vitamin C were selected by surgeons from a group of cancer patients in a hospital.<sup>23</sup> Explain how this experiment was biased in favor of vitamin C.

**1.2.5** On 3 November 2009, the blog lifehacker.com contained a posting by an individual with chronic toenail fungus. He remarked that after many years of suffering and trying all sorts of cures, he resorted to sanding his toenail as thin as he could tolerate, followed by daily application of vinegar and hydrogen-peroxide-soaked bandaids on his toenail. He repeated the vinegar peroxide bandaging for 100 days. After this time his nail grew out and the fungus was gone. Using the language of statistics, what kind of evidence is this? Is this convincing evidence that this procedure is an effective cure of toenail fungus?

- 1.2.6** For each of the following cases [(a), (b), and (c)],
- state whether the study should be observational or experimental.
  - state whether the study should be run blind, double-blind, or neither. If the study should be run blind or double-blind, who should be blinded?
    - An investigation of whether taking aspirin reduces one's chance of having a heart attack.
    - An investigation of whether babies born into poor families (family income below \$25,000) are more likely to weigh less than 5.5 pounds at birth than babies born into wealthy families (family income above \$65,000).
    - An investigation of whether the size of the mid-sagittal plane of the anterior commissure (a part of the brain) of a man is related to the sexual orientation of the man.

**I.2.7 (Hypothetical)** In order to assess the effectiveness of a new fertilizer, researchers applied the fertilizer to the tomato plants on the west side of a garden but did not fertilize the plants on the east side of the garden. They later measured the weights of the tomatoes produced by each plant and found that the fertilized plants grew larger tomatoes than did the nonfertilized plants. They concluded that the fertilizer works.

- Was this an experiment or an observational study? Why?
- This study is seriously flawed. Use the language of statistics to explain the flaw and how this affects the validity of the conclusion reached by the researchers.

- Could this study have used the concept of blinding (i.e., does the word “blind” apply to this study)? If so, how? Could it have been double-blind? If so, how?

**I.2.8** Researchers studied 1,718 persons over age 65 living in North Carolina. They found that those who attended religious services regularly were more likely to have strong immune systems (as determined by the blood levels of the protein interleukin-6) than those who didn't.<sup>24</sup> Does this mean that attending religious services improves one's health? Why or why not?

## 1.3 Random Sampling

In order to address research questions with data, we first must consider how those data are to be gathered. How we gather our data has tremendous implications on our choice of analysis methods and even on the validity of our studies. In this section we will examine some common types of data-gathering methods with special emphasis on the **simple random sample**.

### Samples and Populations

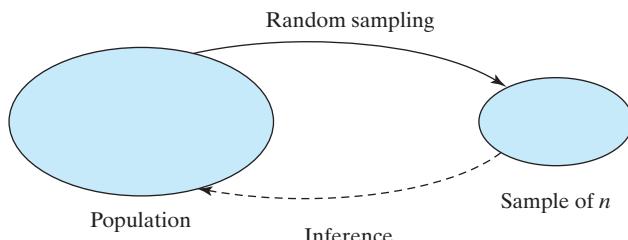
Before gathering data, we first consider the scope of our study by identifying the **population**. The population consists of all subjects/animals/specimens/plants, and so on, of interest. The following are all examples of populations:

- All birch tree seedlings in Florida
- All raccoons in Montaña de Oro State Park
- All people with schizophrenia in the United States
- All 100-ml water specimens in Chorro Creek

Typically we are unable to observe the entire population and therefore we must be content with gathering data from a subset of the population, a **sample** of size  $n$ . From this sample we make inferences about the population as a whole (see Figure 1.3.1). The following are all examples of samples:

- A selection of eight ( $n = 8$ ) Florida birch seedlings grown in a greenhouse.
- Thirteen ( $n = 13$ ) raccoons captured in traps at the Montaña de Oro campground.
- Forty-two ( $n = 42$ ) schizophrenic patients who respond to an advertisement in a U.S. newspaper.
- Ten ( $n = 10$ ) 100-ml vials of water collected one day at 10 locations along Chorro Creek.

**Figure 1.3.1** Sampling from a population



**Remark** There is some potential for confusion between the statistical meaning of the term *sample* and the sense in which this word is sometimes used in biology. If a biologist draws blood from 20 people and measures the glucose concentration in each, she might say she has 20 samples of blood. However, the statistician says she has *one* sample of 20 glucose measurements; the sample size is  $n = 20$ . In the interest of clarity, throughout this book we will use the term *specimen* where a biologist might prefer *sample*. So we would speak of glucose measurements on a sample of 20 specimens of blood.

Ideally our sample will be a representative subset of the population; however, unless we are careful, we may end up obtaining a **biased** sample. A biased sample systematically overestimates or systematically underestimates a characteristic of the population. For example, consider the raccoons from the sample described previously that are captured in traps at a campground. These raccoons may systematically differ from the population; they may be larger (from having ample access to food from dumpsters and campers), less timid (from being around people who feed them), and may be even longer lived than the general population of raccoons in the entire park.

One method to ensure that samples will be (in the long run) representative of the population is to use random sampling.

## Definition of a Simple Random Sample

Informally, the process of obtaining a simple random sample can be visualized in terms of labeled tickets, such as those used in a lottery or raffle. Suppose that each member of the population (e.g., raccoon, patient, plant) is represented by one ticket, and that the tickets are placed in a large box and thoroughly mixed. Then  $n$  tickets are drawn from the box by a blindfolded assistant, with new mixing after each ticket is removed. These  $n$  tickets constitute the sample. (Equivalently, we may visualize that  $n$  assistants reach in the box simultaneously, each assistant drawing one ticket.)

More abstractly, we may define random sampling as follows.

### A Simple Random Sample

A **simple random sample** of  $n$  items is a sample in which (a) every member of the population has the same chance of being included in the sample, and (b) the members of the sample are chosen independently of each other. [Requirement (b) means that the chance of a given member of the population being chosen does not depend on which other members are chosen.]\*

Simple random sampling can be thought of in other, equivalent, ways. We may envision the sample members being chosen one at a time from the population; under simple random sampling, at each stage of the drawing, every remaining member of the population is equally likely to be the next one chosen. Another view is to consider the totality of possible samples of size  $n$ . If all possible samples are equally likely to be obtained, then the process gives a simple random sample.

---

\*Technically, requirement (b) is that every pair of members of the population has the same chance of being selected for the sample, every group of 3 members of the population has the same chance of being selected for the sample, and so on. In contrast to this, suppose we had a population with 30 persons in it and we wrote the names of 3 persons on each of 10 tickets. We could then choose one ticket in order to get a sample of size  $n = 3$ , but this would not be a simple random sample, since the pair (1,2) could end up in the sample but the pair (1,4) could not. Here the selections of members of the sample are not independent of each other. [This kind of sampling is known as “cluster sampling,” with 10 clusters of size 3.] If the population is infinite, then the technical definition that all subsets of a given size are equally likely to be selected as part of the sample is equivalent to the requirement that the members of the sample are chosen independently.

## Employing Randomness

When conducting statistical investigations, we will need to make use of randomness. As previously discussed, we obtain simple random samples randomly—every member of the population has the same chance of being selected. In Chapter 7 we shall discuss experiments in which we wish to compare the effects of different treatments on members of a sample. To conduct these experiments we will have to assign the treatments to subjects randomly—so that every subject has the same chance of receiving treatment A as they do treatment B.

Unfortunately, as a practical matter, humans are not very capable of mentally employing randomness. We are unable to eliminate unconscious bias that often leads us to systematically excluding or including certain individuals in our sample (or at least decreasing or increasing the chance of choosing certain individuals). For this reason, we must use external resources for selecting individuals when we want a random sample: mechanical devices such as dice, coins, and lottery tickets; electronic devices that produce random digits such as computers and calculators; or tables of random digits such as Table 1 in the back of this book. Although straightforward, using mechanical devices such as tickets in a box is impractical, so we will focus on the use of random digits for sample selection.

## How to Choose a Random Sample

The following is a simple procedure for choosing a random sample of  $n$  items from a finite population of items.

- Create the **sampling frame**: a list of all members of the population with unique identification numbers for each member. All identification numbers must have the same number of digits; for instance, if the population contains 75 items, the identification numbers could be 01, 02, . . . , 75.
- Read numbers from Table 1, a calculator, or computer. Reject any numbers that do not correspond to any population member. (For example, if the population has 75 items that have been assigned identification numbers 01, 02, . . . , 75, then skip over the numbers 76, 77, . . . , 99 and 00.) Continue until  $n$  numbers have been acquired. (Ignore any repeated occurrence of the same number.)
- The population members with the chosen identification numbers constitute the sample.

The following example illustrates this procedure.

### Example 1.3.1

Suppose we are to choose a random sample of size 6 from a population of 75 members. Label the population members 01, 02, . . . , 75. Use Table 1, a calculator, or a computer to generate a string of random digits.\* For example, our calculator might produce the following string:

8 3 8 7 1 7 9 4 0 1 6 2 5 3 4 5 9 7 5 3 9 8 2 2

As we examine two-digit pairs of numbers, we ignore numbers greater than 75 as well as any pairs that identify a previously chosen individual.

8 3 8 7 1 7 9 4 0 1 6 2 5 3 4 5 9 7 5 3 9 8 2 2

Thus, the population members with the following identification numbers will constitute the sample: 17, 01, 62, 53, 45, 22. ■

---

\*Most calculators generate random numbers expressed as decimal numbers between 0 and 1; to convert these to random digits, simply ignore the leading zero and decimal and read the digits that follow the decimal. To generate a long string of random digits, simply call the random number function on the calculator repeatedly.

**Remark** In calling the digits in Table 1 or your calculator or computer *random* digits, we are using the term *random* loosely. Strictly speaking, random digits are digits produced by a random *process*—for example, tossing a 10-sided die. The digits in Table 1 or in your calculator or computer are actually *pseudorandom* digits; they are generated by a deterministic (although possibly very complex) process that is designed to produce sequences of digits that mimic randomly generated sequences.

**Remark** If the population is large, then computer software can be quite helpful in generating a sample. If you need a random sample of size 15 from a population with 2,500 members, have the computer (or calculator) generate 15 random numbers between 1 and 2,500. (If there are duplicates in the set of 15, then go back and get more random numbers.)

## Practical Concerns When Random Sampling

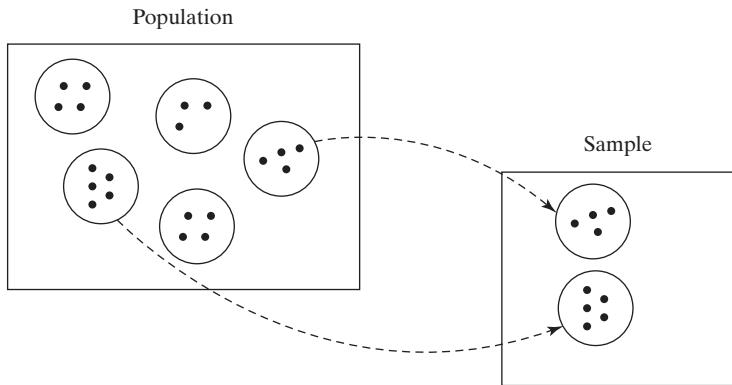
In many cases, obtaining a proper simple random sample is difficult or impossible. For example, to obtain a random sample of raccoons from Montaña de Oro State Park, one would first have to create the sampling frame, which provides a unique number for each raccoon in the park. Then, after generating the list of random numbers to identify our sample, one would have to capture those particular raccoons. This is likely an impossible task.

In practice, when it is possible to obtain a proper random sample, one should. When a proper random sample is impractical, it is important to take all precautions to ensure that the subjects in the study may be viewed *as if* they were obtained by random sampling from some population. That is, the sample should be comprised of individuals that all have the same chance of being selected from the population, and the individuals should be chosen independently. To do this, the first step is to define the population. The next step is to scrutinize the procedure by which the observational units are selected and to ask: Could the *observations* have been chosen at random? With the raccoon example, this might mean that we first define the population of raccoons by creating a sharp geographic boundary based on raccoon habitat and place traps at randomly chosen locations within the population habitat using a variety of baits and trap sizes. (We could use random numbers to generate latitude and longitude coordinates within the population habitat). While still less than ideal (some raccoons might be trap shy and baby raccoons may not enter the traps at all), this is certainly better than simply capturing raccoons at one nonrandomly chosen atypical location (e.g., the campground) within the park. Presumably, the vast majority of raccoons now have the same chance of being trapped (i.e., equally likely to be selected) and capturing one raccoon has little or no bearing on the capture of any other (i.e., they can be considered to be independently chosen). Thus, it seems reasonable to treat the observations as if they were chosen at random.

## Nonsimple Random Sampling Methods

There are other kinds of sampling that are random in a sense, but that are not simple. Two common nonsimple random sampling techniques are the **random cluster sample** and **stratified random sample**. To illustrate the concept of a cluster sample, consider a modification to the lottery method of generating a simple random sample. With cluster sampling, rather than assigning a unique ticket (or ID number) for each

**Figure 1.3.2** Random cluster sampling. The dots represent individuals within the population that are grouped into clusters (circles). Individuals in entire clusters are sampled from the population to form the sample.



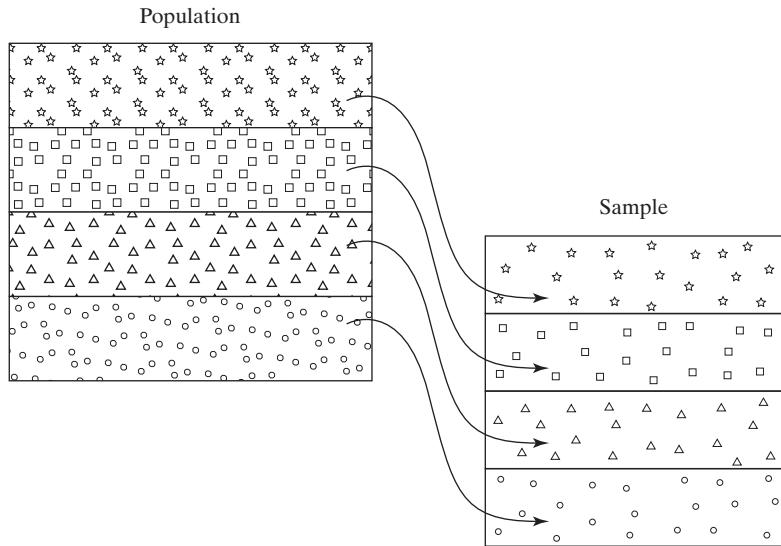
member of the population, IDs are assigned to entire groups of individuals. As tickets are drawn from the box, entire groups of individuals are selected for the sample as in the following example and Figure 1.3.2.

#### Example 1.3.2

**La Graciosa Thistle** The La Graciosa thistle (*Cirsium loncholepis*) is an endangered plant native to the Guadalupe Dunes on the central coast of California. In a seed germination study, 30 plants were randomly chosen from the population of plants in the Guadalupe dunes and all seeds from the 30 plants were harvested. The seeds form a cluster sample from the population of all La Graciosa thistle seeds in Guadalupe while the individual plants were used to identify the clusters.<sup>25</sup> ■

A stratified random sample is chosen by first dividing the population into **strata**—homogeneous collections of individuals. Then, many simple random samples are taken—one within each stratum—and combined to comprise the sample (see Figure 1.3.3). The following is an example of a stratified random sample.

**Figure 1.3.3** Stratified random sampling. The dots represent individuals within the population that are grouped into strata. Individuals from each stratum are randomly sampled and combined to form the sample.



#### Example 1.3.3

**Sand Crabs** In a study of parasitism of sand crabs (*Emerita analoga*), researchers obtained a stratified random sample of crabs by dividing a beach into 5-meter strips parallel to the water's edge. These strips were chosen as the strata because crab parasite loads may differ systematically based on the distance to the water's edge, thus making the parasite load for crabs within each stratum more similar than loads

across strata. The first stratum was the 5-meter strip of beach just under the water's edge parallel to the shoreline. The second stratum was the 5-meter strip of beach just above the shoreline, followed by the third and fourth strata—the next two 5-meter strips above the shoreline. Within each strata, 25 crabs were randomly sampled, yielding a total sample size of 100 crabs.<sup>26</sup>

The majority of statistical methods discussed in this textbook will assume we are working with data gathered from a simple random sample. A sample chosen by simple random sampling is often called a *random sample*. But note that it is actually the *process* of sampling rather than the sample itself that is defined as random; randomness is not a property of the particular sample that happens to be chosen.

## Sampling Error

How can we provide a rationale for inference from a limited sample to a much larger population? The approach of statistical theory is to refer to an idealized model of the sample–population relationship. In this model, which is called the **random sampling model**, the sample is chosen from the population by random sampling. The model is represented schematically in Figure 1.3.1.

The random sampling model is useful because it provides a basis for answering the question, How representative (of the population) is a sample likely to be? The model can be used to determine how much an inference might be influenced by chance, or “luck of the draw.” More explicitly, a randomly chosen sample will usually not exactly resemble the population from which it was drawn. The discrepancy between the sample and the population is called **chance error due to sampling** or **sampling error**. We will see in later chapters how statistical theory derived from the random sampling model enables us to set limits on the likely amount of error due to sampling in an experiment. The quantification of such error is a major contribution that statistical theory has made to scientific thinking.

Because our samples are chosen randomly, there will always be sampling error present. If we sample nonrandomly, however, we may exacerbate the sampling error in unpredictable ways such as by introducing **sampling bias**, which is a systematic tendency for some individuals of the population to be selected more readily than others. The following two examples illustrate sampling bias.

---

### Example 1.3.4

**Lengths of Fish** A biologist plans to study the distribution of body length in a certain population of fish in the Chesapeake Bay. The sample will be collected using a fishing net. Smaller fish can more easily slip through the holes in the net. Thus, smaller fish are less likely to be caught than larger ones, so that the sampling procedure is biased.

---

### Example 1.3.5

**Sizes of Nerve Cells** A neuroanatomist plans to measure the sizes of individual nerve cells in cat brain tissue. In examining a tissue specimen, the investigator must decide which of the hundreds of cells in the specimen should be selected for measurement. Some of the nerve cells are incomplete because the microtome cut through them when the tissue was sectioned. If the size measurement can be made only on complete cells, a bias arises because the smaller cells had a greater chance of being missed by the microtome blade.

When the sampling procedure is biased, the sample may not accurately represent the population, because it is systematically distorted. For instance, in Example 1.3.4

smaller fish will tend to be underrepresented in the sample, so that the length of the fish in the sample will tend to be larger than those in the population.

The following example illustrates a kind of nonrandomness that is different from bias.

#### Example 1.3.6

**Sucrose in Beet Roots** An agronomist plans to sample beet roots from a field in order to measure their sucrose content. Suppose she were to take all her specimens from a randomly selected small area of the field. This sampling procedure would not be biased but would tend to produce *too homogeneous* a sample, because environmental variation across the field would not be reflected in the sample. ■

Example 1.3.6 illustrates an important principle that is sometimes overlooked in the analysis of data: In order to check applicability of the random sampling model, one needs to ask not only whether the sampling procedure might be biased, but also whether the sampling procedure will adequately reflect the variability inherent in the population. Faulty information about variability can distort scientific conclusions just as seriously as bias can.

We now consider some examples where the random sampling model might reasonably be applied.

#### Example 1.3.7

**Fungus Resistance in Corn** A certain variety of corn is resistant to fungus disease. To study the inheritance of this resistance, an agronomist crossed the resistant variety with a nonresistant variety and measured the degree of resistance in the progeny plants. The actual progeny in the experiment can be regarded as a random sample from a conceptual population of all *potential* progeny of that particular cross. ■

When the purpose of a study is to *compare* two or more experimental conditions, a very narrow definition of the population may be satisfactory, as illustrated in the next example.

#### Example 1.3.8

**Nitrite Metabolism** To study the conversion of nitrite to nitrate in the blood, researchers injected four New Zealand White rabbits with a solution of radioactively labeled nitrite molecules. Ten minutes after injection, they measured for each rabbit the percentage of the nitrite that had been converted to nitrate.<sup>27</sup> Although the four animals were not literally chosen at random from a specified population, nevertheless it might be reasonable to view the measurements of nitrite metabolism as a random sample from similar measurements made on all New Zealand White rabbits. (This formulation assumes that age and sex are irrelevant to nitrite metabolism.) ■

#### Example 1.3.9

**Treatment of Ulcerative Colitis** A medical team conducted a study of two therapies, A and B, for treatment of ulcerative colitis. All the patients in the study were referral patients in a clinic in a large city. Each patient was observed for satisfactory “response” to therapy. In applying the random sampling model, the researchers might want to make an inference to the population of all ulcerative colitis patients in urban referral clinics. First, consider inference about the actual probabilities of response; such an inference would be valid if the probability of response to each therapy is the same at all urban referral clinics. However, this assumption might be somewhat questionable, and the investigators might believe that the population should be defined very narrowly—for instance, as “the type of ulcerative colitis patients who are referred to this clinic.” Even such a narrow population can be of interest in a comparative study. For instance, if treatment A is better than treatment B for the narrow population, it might be reasonable to infer that A would be better

than B for a broader population (even if the actual response probabilities might be different in the broader population). In fact, it might even be argued that the broad population should include all ulcerative colitis patients, not merely those in urban referral clinics. ■

It often happens in research that, for practical reasons, the population actually studied is narrower than the population that is of real interest. In order to apply the kind of rationale illustrated in Example 1.3.9, one must argue that the results in the narrowly defined population (or, at least, some aspects of those results) can meaningfully be extrapolated to the population of interest. This extrapolation is not a *statistical* inference; it must be defended on biological, not statistical, grounds.

In Section 2.8 we will say more about the connection between samples and populations as we further develop the concept of statistical inference.

## Nonsampling Errors

In addition to sampling errors, other concerns can arise in statistical studies. A **nonsampling error** is an error that is not caused by the sampling method; that is, a nonsampling error is one that would have arisen even if the researcher had a census of the entire population. For example, the way in which questions are worded can greatly influence how people answer them, as Example 1.3.10 shows.

---

### Example 1.3.10

**Abortion Funding** In 1991, the U.S. Supreme Court made a controversial ruling upholding a ban on abortion counseling in federally financed family-planning clinics. Shortly after the ruling, a sample of 1,000 people were asked, “As you may know, the U.S. Supreme Court recently ruled that the federal government is not required to use taxpayer funds for family planning programs to perform, counsel, or refer for abortion as a method of family planning. In general, do you favor or oppose this ruling?” In the sample, 48% favored the ruling, 48% were opposed, and 4% had no opinion.

A separate opinion poll conducted at nearly the same time, but by a different polling organization, asked over 1,200 people, “Do you favor or oppose that Supreme Court decision preventing clinic doctors and medical personnel from discussing abortion in family-planning clinics that receive federal funds?” In this sample, 33% favored the decision and 65% opposed it.<sup>28</sup> The difference in the percentages favoring the opinion is too large to be attributed to chance error in the sampling. It seems that the way in which the question was worded had a strong impact on the respondents. ■

Another type of nonsampling error is **nonresponse bias**, which is bias caused by persons not responding to some of the questions in a survey or not returning a written survey. It is common to have only one-third of those receiving a survey in the mail complete the survey and return it to the researchers. (We consider the people receiving the survey to be part of the sample, even if some of them don’t complete the entire survey, or even return the survey at all.) If the people who respond are unlike those who choose not to respond—and this is often the case, since people with strong feelings about an issue tend to complete a questionnaire, while others will ignore it—then the data collected will not accurately represent the population.

---

### Example 1.3.11

**HIV Testing** A sample of 949 men were asked if they would submit to an HIV test of their blood. Of the 782 who agreed to be tested, 8 (1.02%) were found to be HIV positive. However, some of the men refused to be tested. The health researchers

conducting the study had access to serum specimens that had been taken earlier from these 167 men and found that 9 of them (5.4%) were HIV positive.<sup>29</sup> Thus, those who refused to be tested were much more likely to have HIV than those who agreed to be tested. An estimate of the HIV rate based only on persons who agree to be tested is likely to substantially underestimate the true prevalence. ■

There are other cases in which an experimenter is faced with the vexing problem of **missing data**—that is, observations that were planned but could not be made. In addition to nonresponse, this can arise because experimental animals or plants die, because equipment malfunctions, or because human subjects fail to return for a follow-up observation.

A common approach to the problem of missing data is to simply use the remaining data and ignore the fact that some observations are missing. This approach is temptingly simple but must be used with extreme caution, because comparisons based on the remaining data may be seriously biased. For instance, if observations on some experimental mice are missing because the mice died of causes related to the treatment they received, it is obviously not valid to simply compare the mice that survived. As another example, if patients drop out of a medical study because they think their treatment is not working, then analysis of the remaining patients could produce a greatly distorted picture.

Naturally, it is best to make every effort to avoid missing data. But if data are missing, it is crucial that the possible reasons for the omissions be considered in interpreting and reporting the results.

Data can also be misleading if there is bias in how the data are collected. People have difficulty remembering the dates on which events happen and they tend to give unreliable answers if asked a question such as “How many times per week do you exercise?” They may also be biased as they make observations, as the following example shows.

### Example 1.3.12

**Sugar and Hyperactivity** Mothers who thought that their young sons were “sugar sensitive” were randomly divided into two groups. Those in the first group were told that their sons had been given a large dose of sugar, whereas those in the second group were told that their sons had been given a placebo. In fact, all the boys had been given the placebo. Nonetheless, the mothers in the first group rated their sons to be much more hyperactive during a 25-minute study period than did the mothers in the second group.<sup>30</sup> Neutral measurements found that boys in the first group were actually a bit *less* active than those in the second group. Numerous other studies have failed to find a link between sugar consumption and activity in children, despite the widespread belief that sugar causes hyperactive behavior. It seems that the expectations that these mothers had colored their observations.<sup>31</sup> ■

## Exercises 1.3.1–1.3.6

**1.3.1** In each of the following studies, identify which sampling technique best describes the way the data were collected (or could be treated as if they were collected): simple random sampling, random cluster sampling, or stratified random sampling. For cluster samples identify the clusters and for stratified samples identify the strata.

- (a) All 257 leukemia patients from three randomly chosen pediatric clinics in the United States were enrolled in a clinical trial for a new drug.

- (b) A total of twelve 10-g soil specimens were collected from random locations on a farm to study physical and chemical soil profiles.
- (c) In a pollution study three 100-ml air specimens were collected at each of four specific altitudes (100 m, 500 m, 1000 m, 2000 m) for a total of twelve 100-ml specimens.
- (d) A total of 20 individual grapes were picked from random vines in a vineyard to evaluate readiness for harvest.

- (e) Twenty-four dogs (eight randomly chosen small breed, eight randomly chosen medium breed, and eight randomly chosen large breed) were enrolled in an experiment to evaluate a new training program.

**1.3.2** For each of the following studies, identify the source(s) of sampling bias and describe (i) how it might affect the study conclusions and (ii) how you might alter the sampling method to avoid the bias.

- (a) Eight hundred volunteers were recruited from nightclubs to enroll in an experiment to evaluate a new treatment for social anxiety.
- (b) In a water pollution study, water specimens were collected from a stream on 15 rainy days.
- (c) To study the size (radius) distribution of scrub oaks (shrubby oak trees), 20 oak trees were selected by using random latitude/longitude coordinates. If the random coordinate fell within the canopy of a tree, the tree was selected; if not, another random location was generated.
- (d) To study the size distribution of rock cod (*Epinephelus puscus*) off the coast of southeastern Australia, the lengths and weights were recorded for all cod captured by a commercial fishing vessel on one day (using standard hook-and-line fishing methods).

**1.3.3 (A fun activity)** Write the digits 1, 2, 3, 4 in order on an index card. Bring this card to a busy place (e.g., dining hall, library, university union) and ask at least 30 people to look at the card and select one of the digits at random in their head. Record their responses.

- (a) If people can think “randomly,” about what fraction of the people should respond with the digit 1? 2? 3? 4?
- (b) What fraction of those surveyed responded with the digit 1? 2? 3? 4?
- (c) Do the results suggest anything about people’s ability to choose randomly?

**1.3.4** Consider a population consisting of 600 individuals with unique IDs: 001, 002, . . . , 600. Use the following string of random digits to select a simple random sample of 5 individuals. List the IDs of the individuals selected for your sample.

728121876442121593787803547216596851

**1.3.5 (Sampling exercise)** Refer to the collection of 100 ellipses shown in the accompanying figure, which can be thought of as representing a natural population of the mythical organism *C. ellipticus*. The ellipses have been given identification numbers 00, 01, . . . , 99 for convenience in sampling. Certain individuals of *C. ellipticus* are mutants and have two tail bristles.

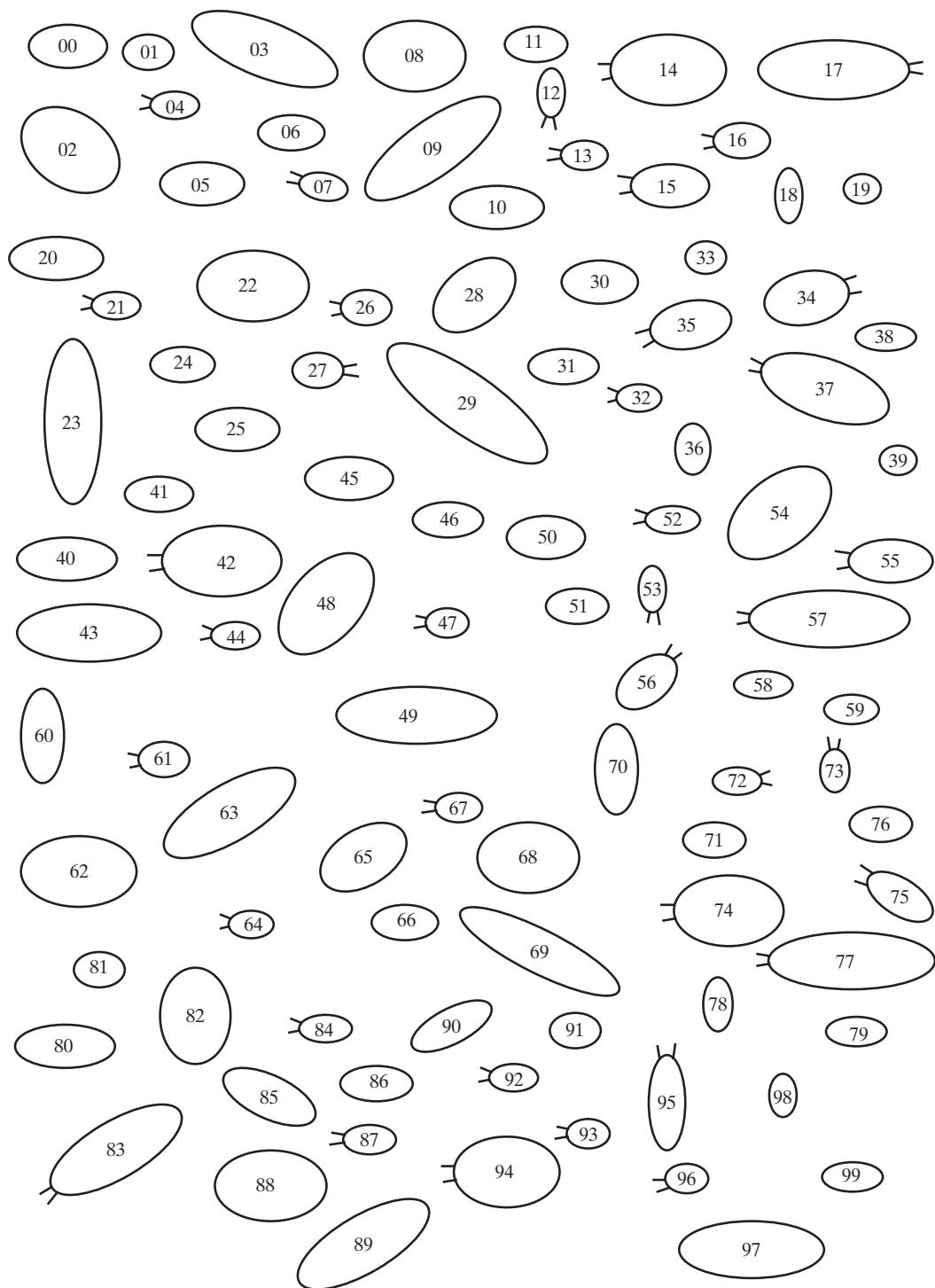
- (a) Use your *judgment* to choose a sample of size 10 from the population that you think is representative of the entire population. Note the number of mutants in the sample.
- (b) Use *random digits* (from Table 1 or your calculator or computer) to choose a random sample of size 10 from the population and note the number of mutants in the sample.

**1.3.6 (Sampling exercise)** Refer to the collection of 100 ellipses.

- (a) Use random digits (for Table 1 or your calculator or computer) to choose a random sample of size 5 from the population and note the number of mutants in the sample.
- (b) Repeat part (a) nine more times, for a total of 10 samples. (Some of the 10 samples may overlap.)

To facilitate pooling of results from the entire class, report your results in the following format:

NUMBER OF MUTANTS	FREQUENCY (NO. OF SAMPLES)
0	5
1	4
2	3
3	2
4	1
5	0
Total: 10	



# DESCRIPTION OF SAMPLES AND POPULATIONS

## Objectives

In this chapter we will study how to describe data. In particular, we will

- show how frequency distributions are used to make bar charts and histograms.
- compare the mean and median as measures of center.
- demonstrate how to construct and read a variety of graphics including dotplots, boxplots, and scatterplots.
- compare several measures of variability with emphasis on the standard deviation.
- examine how transformations of variables affect distributions.
- consider the relationship between populations and samples.

## 2.1 Introduction

Statistics is the science of analyzing and learning from data. In this section we introduce some terminology and notation for dealing with data.

### Variables

We begin with the concept of a **variable**. A variable is a characteristic of a person or a thing that can be assigned a number or a category. For example, blood type (A, B, AB, O) and age are two variables we might measure on a person.

Blood type is an example of a **categorical variable**: A categorical variable is a variable that records which of several categories a person or thing is in. Examples of categorical variables are

Blood type of a person: A, B, AB, O

Sex of a fish: male, female

Color of a flower: red, pink, white

Shape of a seed: wrinkled, smooth

For some categorical variables, the categories can be arrayed in a meaningful rank order. Such a variable is said to be **ordinal**. For example, the response of a patient to therapy might be none, partial, or complete.

Age is an example of a **numeric variable**. A numeric variable is a variable that records the amount of something. A **continuous variable** is a numeric variable that is measured on a continuous scale. Examples of continuous variables are

- Weight of a baby
- Cholesterol concentration in a blood specimen
- Optical density of a solution

A variable such as weight is continuous because, in principle, two weights can be arbitrarily close together. Some types of numeric variables are not continuous but fall on a discrete scale, with spaces between the possible values. A **discrete variable** is a numeric variable for which we can list the possible values. For example, the number of eggs in a bird's nest is a discrete variable because only the values 0, 1, 2, 3, . . . , are possible. Other examples of discrete variables are

- Number of bacteria colonies in a petri dish
- Number of cancerous lymph nodes detected in a patient
- Length of a DNA segment in basepairs

The distinction between continuous and discrete variables is not a rigid one. After all, physical measurements are always rounded off. We may measure the weight of a steer to the nearest kilogram, of a rat to the nearest gram, or of an insect to the nearest milligram. The scale of the actual measurements is always discrete, strictly speaking. The continuous scale can be thought of as an approximation to the actual scale of measurement.

## Observational Units

When we collect a sample of  $n$  persons or things and measure one or more variables on them, we call these persons or things **observational units** or cases. The following are some examples of samples.

Sample	Variable	Observational unit
150 babies born in a certain hospital	Birthweight (kg)	A baby
73 <i>Cecropia</i> moths caught in a trap	Sex	A moth
81 plants that are a progeny of a single parental cross	Flower color	A plant
Bacterial colonies in each of six petri dishes	Number of colonies	A petri dish

## Notation for Variables and Observations

We will adopt a notational convention to distinguish between a variable and an observed value of that variable. We will denote variables by uppercase letters such as  $Y$ . We will denote the observations themselves (that is, the data) by lowercase letters such as  $y$ . Thus, we distinguish, for example, between  $Y$  = birthweight (the variable) and  $y$  = 7.9 lb (the observation). This distinction will be helpful in explaining some fundamental ideas concerning variability.

## Exercises 2.1.1–2.1.4

For each of the following settings in Exercises 2.1.1–2.1.4, (i) identify the variable(s) in the study, (ii) for each variable tell the type of variable (e.g., categorical and ordinal, discrete, etc.), (iii) identify the observational unit (the thing sampled), and (iv) determine the sample size.

### 2.1.1

- (a) A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*.
- (b) The birthweight, date of birth, and the mother's race were recorded for each of 65 babies.

### 2.1.2

- (a) A physician measured the height and weight of each of 37 children.
- (b) During a blood drive, a blood bank offered to check the cholesterol of anyone who donated blood.

A total of 129 persons donated blood. For each of them, the blood type and cholesterol levels were recorded.

### 2.1.3

- (a) A biologist measured the number of leaves on each of 25 plants.
- (b) A physician recorded the number of seizures that each of 20 patients with severe epilepsy had during an eight-week period.

### 2.1.4

- (a) A conservationist recorded the weather (clear, partly cloudy, cloudy, rainy) and number of cars parked at noon at a trailhead on each of 18 days.
- (b) An enologist measured the pH and residual sugar content (g/l) of seven barrels of wine.

## 2.2 Frequency Distributions

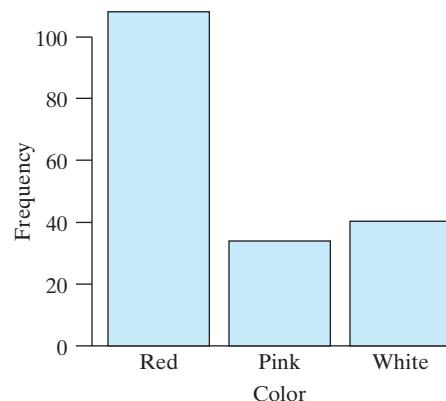
A first step toward understanding a set of data on a given variable is to explore the data and describe the data in summary form. In this chapter we discuss three mutually complementary aspects of summary data description: frequency distributions, measures of center, and measures of dispersion. These tell us about the shape, center, and spread of the data.

A **frequency distribution** is simply a display of the **frequency**, or number of occurrences, of each value in the data set. The information can be presented in tabular form or, more vividly, with a graph. A **bar chart** is a simple graphic showing the categories that a categorical variable takes on and the number of observations in each category for the data in the sample. Here are two examples of frequency distributions for categorical data.

### Example 2.2.1

**Color of Poinsettias** Poinsettias can be red, pink, or white. In one investigation of the hereditary mechanism controlling the color, 182 progeny of a certain parental cross were categorized by color.<sup>1</sup> The bar graph in Figure 2.2.1 is a visual display of the results given in Table 2.2.1. ■

**Figure 2.2.1** Bar chart of color of 182 poinsettias



**Table 2.2.1** Color of one hundred eighty-two poinsettias

Color	Frequency (number of plants)
Red	108
Pink	34
White	40
Total	182

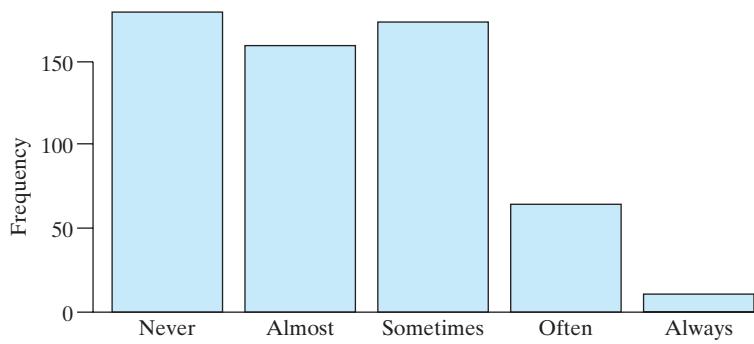
**Example 2.2.2**

**School Bags and Neck Pain** Physiologists in Australia were concerned that carrying a school bag loaded with heavy books was a cause of neck pain in adolescents, so they asked a sample of 585 teenage girls how often they get neck pain when carrying their school bag (e.g., never, almost never, sometimes, often, always). A summary of the results reported to them is given in Table 2.2.2 and displayed as a bar graph in Figure 2.2.2(a).<sup>2</sup> As the variable incidence is an ordinal categorical variable, our tables and graphs should respect the natural ordering. Figure 2.2.2(b) shows the same data but with the categories in alphabetical order (a default setting for much software), which obscures the information in the data. ■

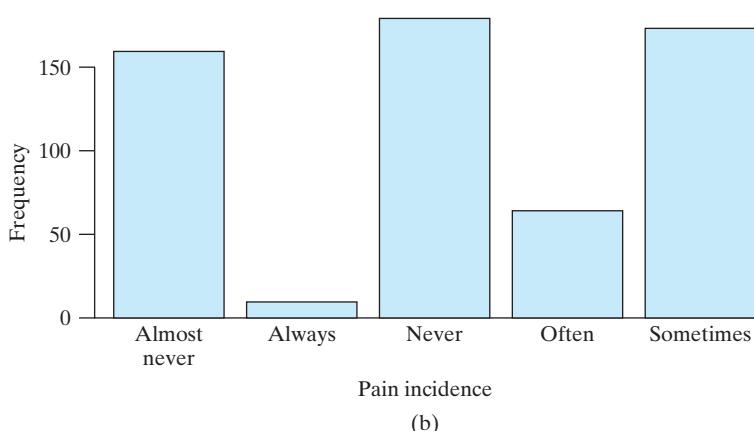
**Table 2.2.2** Neck pain associated with carrying a school bag

Incidence	Frequency (number of girls)
Never	179
Almost never	159
Sometimes	173
Often	64
Always	10
Total	585

**Figure 2.2.2** (a) Bar chart of incidence of neck pain reported by 585 adolescents; (b) the same data but with the categories in alphabetical order



(a)



(b)

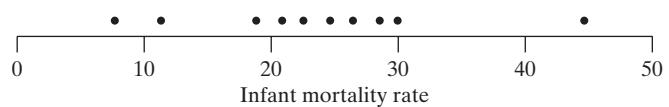
A **dotplot** is a simple graph that can be used to show the distribution of a numeric variable when the sample size is small. To make a dotplot, we draw a number line covering the range of the data and then put a dot above the number line for each observation, as the following example shows.

**Example  
2.2.3**

**Infant Mortality** Table 2.2.3 shows the infant mortality rate (infant deaths per 1,000 live births) in each of 12 countries in South America, as of 2009.<sup>3</sup> The distribution is shown in Figure 2.2.3. ■

**Table 2.2.3** Infant mortality in 12 South American countries

Country	Infant mortality rate
Argentina	11.4
Bolivia	44.7
Brazil	22.6
Chile	7.7
Colombia	18.9
Ecuador	20.9
Guyana	30.0
Paraguay	24.7
Peru	28.6
Suriname	18.8
Uruguay	11.3
Venezuela	26.5



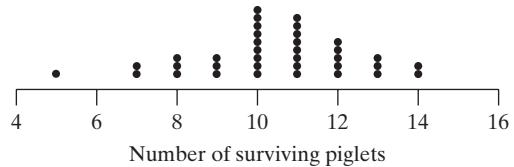
**Figure 2.2.3** Dotplot of infant mortality in 12 South American countries

When two or more observations take on the same value, we stack the dots in a dotplot on top of each other. This gives an effect similar to the effect of the bars in a bar chart. If we create bars, in place of the stacks of dots, we then have a **histogram**. A histogram is like a bar chart, except that a histogram displays a numeric variable, which means that there is a natural order and scale for the variable. In a bar chart the amount of space between the bars (if any) is arbitrary, since the data being displayed are categorical. In a histogram the scale of the variable determines the placement of the bars. The following example shows a dotplot and a histogram for a frequency distribution.

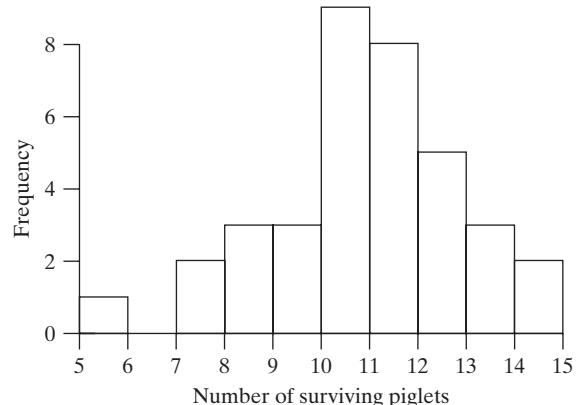
**Example  
2.2.4**

**Litter Size of Sows** A group of thirty-six 2-year-old sows of the same breed ( $\frac{3}{4}$  Duroc,  $\frac{1}{4}$  Yorkshire) were bred to Yorkshire boars. The number of piglets surviving to 21 days of age was recorded for each sow.<sup>4</sup> The results are given in Table 2.2.4 and displayed as a dotplot in Figure 2.2.4 and as a histogram in Figure 2.2.5. ■

<b>Table 2.2.4</b> Number of surviving piglets of 36 sows	
Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36



**Figure 2.2.4** Dotplot of number of surviving piglets of 36 sows



**Figure 2.2.5** Histogram of number of surviving piglets of 36 sows

## Relative Frequency

The frequency scale is often replaced by a **relative frequency** scale:

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

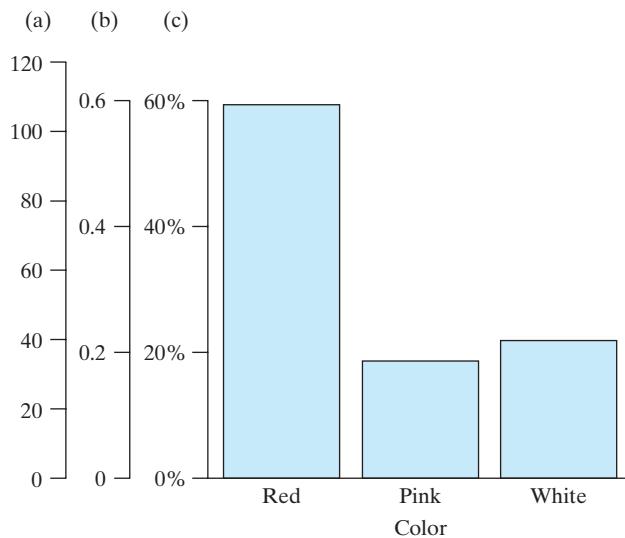
The relative frequency scale is useful if several data sets of different sizes ( $n$ 's) are to be displayed together for comparison. As another option, a relative frequency can be expressed as a percentage frequency. The shape of the display is not affected by the choice of frequency scale, as the following example shows.

### Example 2.2.5

**Color of Poinsettias** The poinsettia color distribution of Example 2.2.1 is expressed as frequency, relative frequency, and percent frequency in Table 2.2.5 and Figure 2.2.6. ■

<b>Table 2.2.5</b> Color of one hundred eighty-two poinsettias			
Color	Frequency	Relative frequency	Percent frequency
Red	108	.59	59
Pink	34	.19	19
White	40	.22	22
Total	182	1.00	100

**Figure 2.2.6** Bar chart of poinsettia colors on three scales:  
 (a) Frequency  
 (b) Relative frequency  
 (c) Percent frequency



## Grouped Frequency Distributions

In the preceding examples, simple ungrouped frequency distributions provided concise summaries of the data. For many data sets, it is necessary to group the data in order to condense the information adequately. (This is usually the case with continuous variables.) The following example shows a grouped frequency distribution.

### Example 2.2.6

**Serum CK** Creatine phosphokinase (CK) is an enzyme related to muscle and brain function. As part of a study to determine the natural variation in CK concentration, blood was drawn from 36 male volunteers. Their serum concentrations of CK (measured in U/l) are given in Table 2.2.6.<sup>5</sup> Table 2.2.7 shows these data grouped into **classes**. For instance, the frequency of the class [20,40) (all values in the interval  $20 \leq y < 40$ ) is 1, which means that one CK value fell in this range. The grouped frequency distribution is displayed as a histogram in Figure 2.2.7. ■

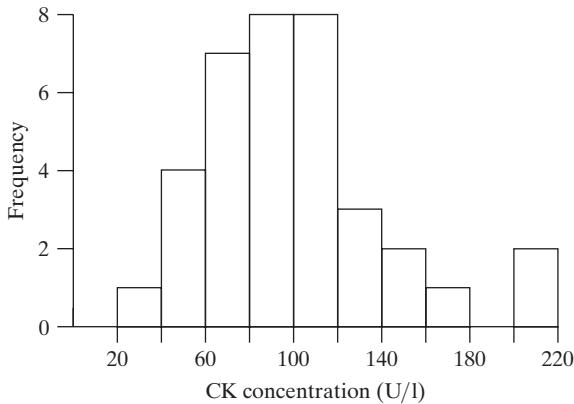
**Table 2.2.6** Serum CK values for 36 men

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

**Table 2.2.7** Frequency distribution of serum CK values for 36 men

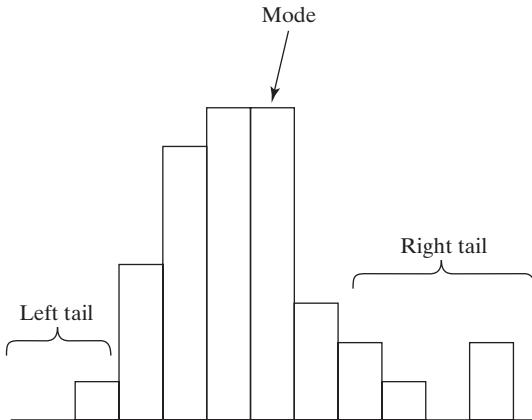
Serum CK (U/l)	Frequency (number of men)
[20,40)	1
[40,60)	4
[60,80)	7
[80,100)	8
[100,120)	8
[120,140)	3
[140,160)	2
[160,180)	1
[180,200)	0
[200,220)	2
Total	36

**Figure 2.2.7** Histogram of serum CK concentrations for 36 men



A grouped frequency distribution should display the essential features of the data. For instance, the histogram of Figure 2.2.7 shows that the average CK value is about 100 U/l, with the majority of the values falling between 60 and 140 U/l. In addition, the histogram shows the *shape* of the distribution. Note that the CK values are piled up around a central peak, or **mode**. On either side of this mode, the frequencies decline and ultimately form the **tails** of the distribution. These shape features are labeled in Figure 2.2.8. The CK distribution is not symmetric but is a bit **skewed to the right**, which means that the right tail is more stretched out than the left.\*

**Figure 2.2.8** Shape features of the CK distribution



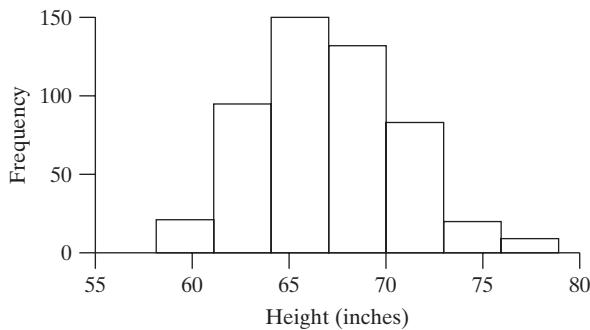
When making a histogram, we need to decide how many classes to have and how wide the classes should be. If we use computer software to generate a histogram, the program will choose the number of classes and the class width for us, but most software allows the user to change the number of classes and to specify the class width. If a data set is large and is quite spread out, it is a good idea to look at more than one histogram of the data, as is done in Example 2.2.7.

### Example 2.2.7

**Heights of Students** A sample of 510 college students were asked how tall they were. Note that they were not measured; rather, they just reported their heights.<sup>6</sup> Figure 2.2.9 shows the distribution of the self-reported values, using 7 classes and a

\*To help remember which tail of a skewed distribution is the longer tail, think of skew as stretch. Which side of the distribution is more stretched away from the center? A distribution that is skewed to the right is one in which the right tail stretches out more than the left.

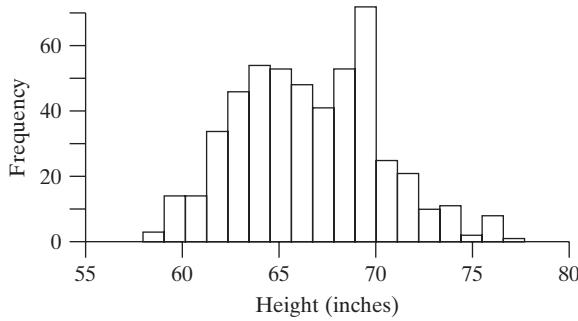
**Figure 2.2.9** Heights of students, using 7 classes (class width = 3)



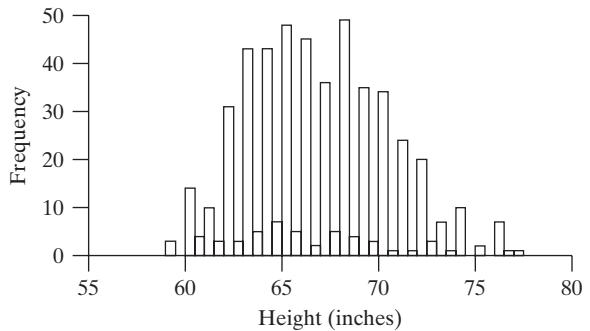
class width of 3 (inches). By using only 7 classes, the distribution appears to be reasonably symmetric, with a single peak around 66 inches.

Figure 2.2.10 shows the height data, but in a histogram that uses 18 classes and a class width of 1.1. This view of the data shows two modes—one for women and one for men.

Figure 2.2.11 shows the height data again, this time using 37 classes, each of width 0.5. Using such a large number of classes makes the distribution look jagged. In this case, we see an alternating pattern between classes with lots of observations and classes with few observations. In the middle of the distribution we see that there were many students who reported a height of 63 inches, few who reported a height of 63.5 inches, many who reported a height of 64 inches, and so on. It seems that most students round off to the nearest inch! ■



**Figure 2.2.10** Heights of students, using 18 classes (class width = 1.1)



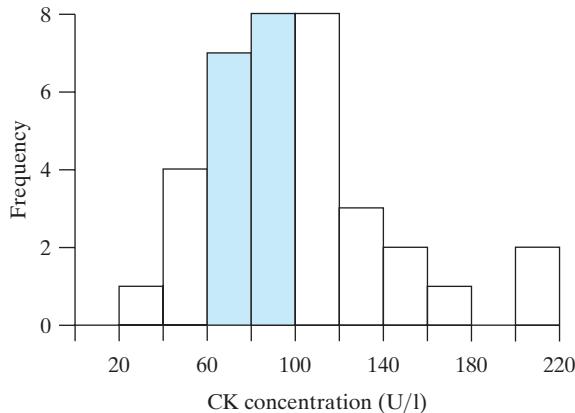
**Figure 2.2.11** Heights of students, using 37 classes (class width = 0.5)

## Interpreting Areas in a Histogram

A histogram can be looked at in two ways. The tops of the bars sketch out the shape of the distribution. But the *areas* within the bars also have a meaning. The area of each bar is proportional to the corresponding frequency. Consequently, the area of one or several bars can be interpreted as expressing the number of observations in the classes represented by the bars. For example, Figure 2.2.12 shows a histogram of the CK distribution of Example 2.2.6. The shaded area is 42% of the total area in all the bars. Accordingly, 42% of the CK values are in the corresponding classes; that is, 15 of 36 or 42% of the values are between 60 U/I and 100 U/I.\*

\*Strictly speaking, between 60 U/I and 99 U/I, inclusive.

**Figure 2.2.12** Histogram of CK distribution. The shaded area is 42% of the total area and represents 42% of the observations.

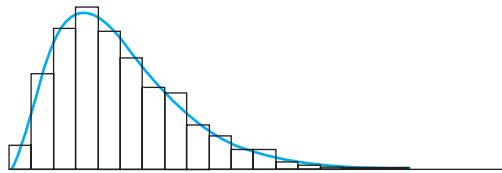


The area interpretation of histograms is a simple but important idea. In our later work with distributions we will find the idea to be indispensable.

## Shapes of Distributions

When discussing a set of data, we want to describe the shape, center, and spread of the distribution. In this section we concentrate on the shapes of frequency distributions and illustrate some of the diversity of distributions encountered in the life sciences. The shape of a distribution can be indicated by a smooth curve that approximates the histogram, as shown in Figure 2.2.13.

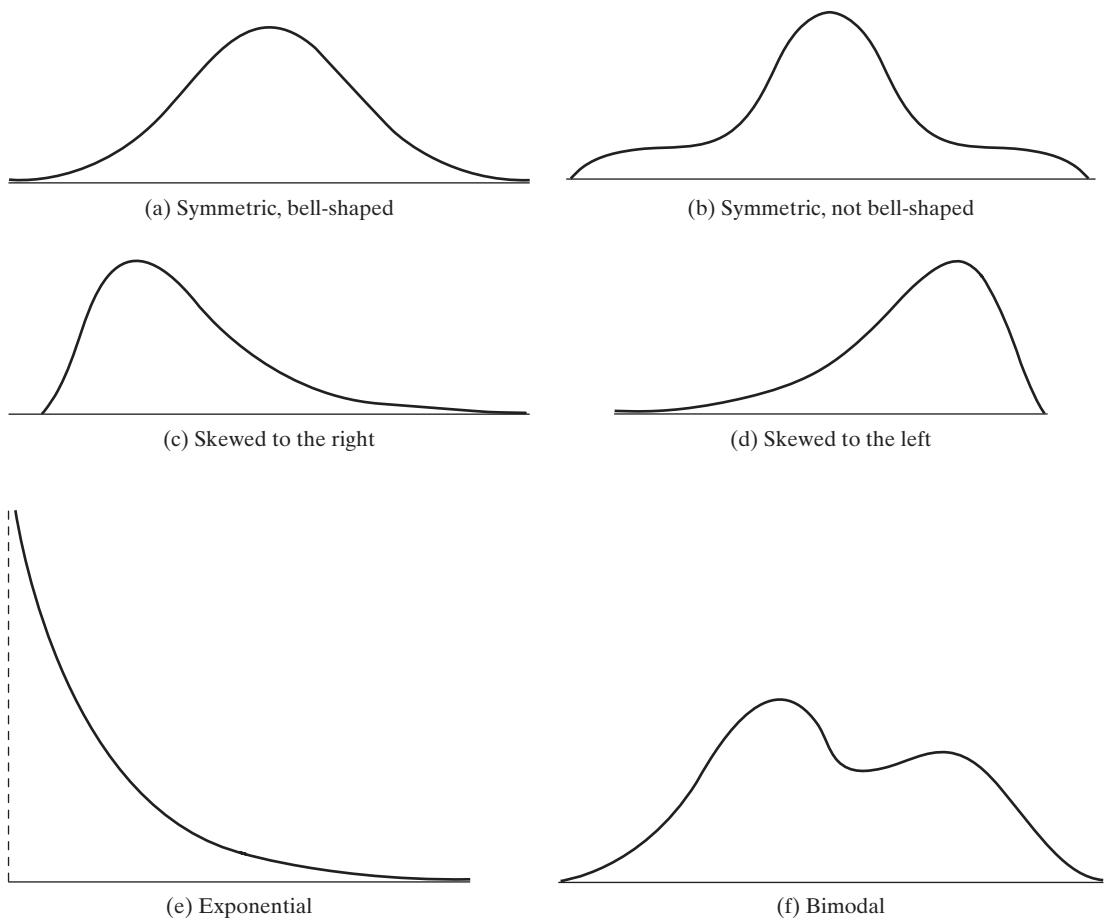
**Figure 2.2.13**  
Approximation of a histogram by a smooth curve



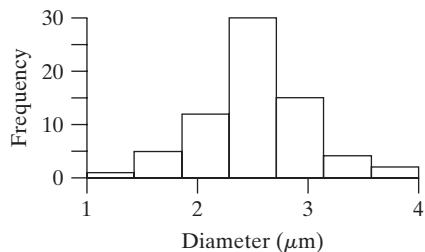
Some distributional shapes are shown in Figure 2.2.14. A common shape for biological data is **unimodal** (has one mode) and is somewhat skewed to the right, as in (c). Approximately bell-shaped distributions, as in (a), also occur. Sometimes a distribution is symmetric but differs from a bell in having long tails; an exaggerated version is shown in (b). Left-skewed (d) and exponential (e) shapes are less common. **Bimodality** (two modes), as in (f), can indicate the existence of two distinct subgroups of observational units.

Notice that the shape characteristics we are emphasizing, such as number of modes and degree of symmetry, are *scale free*; that is, they are not affected by the arbitrary choices of vertical and horizontal scale in plotting the distribution. By contrast, a characteristic such as whether the distribution appears short and fat, or tall and skinny, is affected by how the distribution is plotted and so is not an inherent feature of the biological variable.

The following three examples illustrate biological frequency distributions with various shapes. In the first example, the shape provides evidence that the distribution is in fact biological rather than nonbiological.

**Figure 2.2.14** Shapes of distributions**Example  
2.2.8**

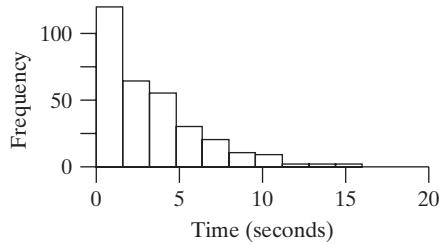
**Microfossils** In 1977, paleontologists discovered microscopic fossil structures, resembling algae, in rocks 3.5 billion years old. A central question was whether these structures were biological in origin. One line of argument focused on their size distribution, which is shown in Figure 2.2.15. This distribution, with its unimodal and rather symmetric shape, resembles that of known microbial populations, but not that of known nonbiological structures.<sup>7</sup> ■

**Figure 2.2.15** Sizes of microfossils

**Example  
2.2.9**

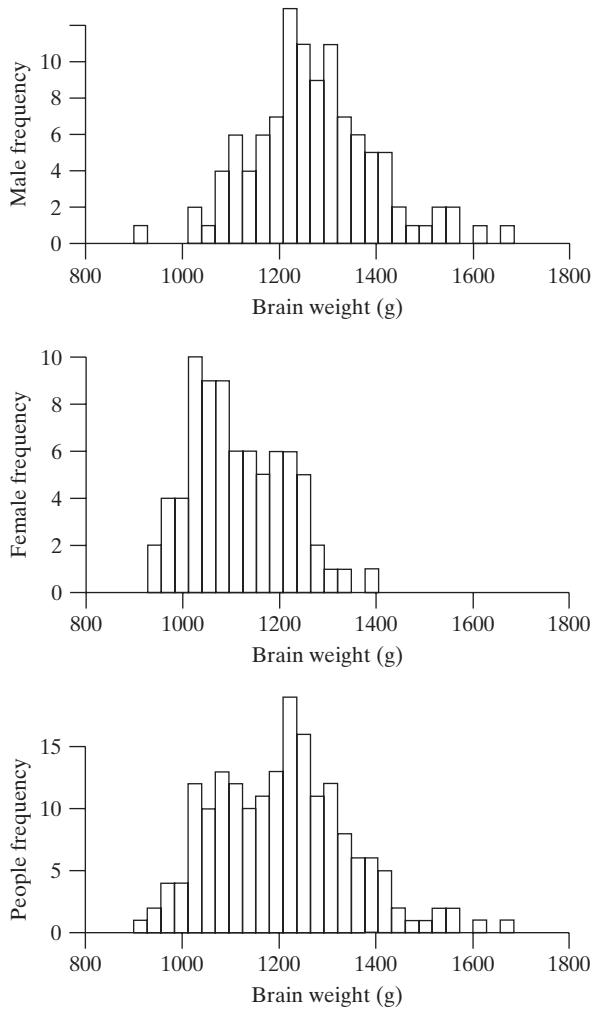
**Cell Firing Times** A neurobiologist observed discharges from rat muscle cells grown in culture together with nerve cells. The time intervals between 308 successive discharges were distributed as shown in Figure 2.2.16. Note the exponential shape of the distribution.<sup>8</sup> ■

**Figure 2.2.16** Time intervals between electrical discharges in rat muscle cells

**Example  
2.2.10**

**Brain Weight** In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The data for males and females are shown in Figure 2.2.17(a) and (b). The male distribution is fairly symmetric and bell shaped; the female distribution is somewhat skewed to the right. Part (c) of the figure shows the brain weight distribution for males and females combined. This combined distribution is slightly bimodal.<sup>9</sup> ■

**Figure 2.2.17** Brain weights



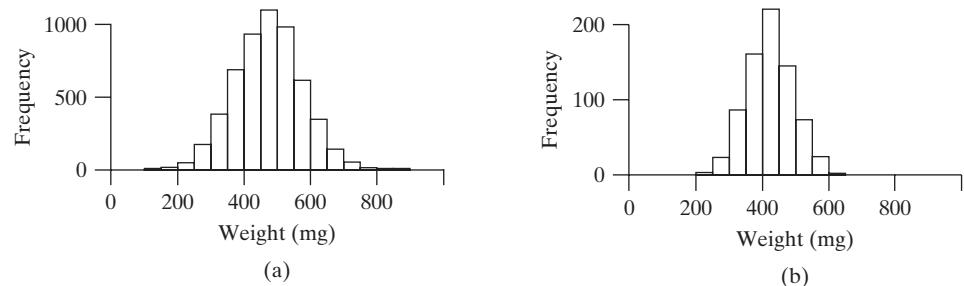
## Sources of Variation

In interpreting biological data, it is helpful to be aware of sources of variability. The variation among observations in a data set often reflects the combined effects of several underlying factors. The following two examples illustrate such situations.

### Example 2.2.11

**Weights of Seeds** In a classic experiment to distinguish environmental from genetic influence, a geneticist weighed seeds of the princess bean *Phaseolus vulgaris*. Figure 2.2.18 shows the weight distributions of (a) 5,494 seeds from a commercial seed lot, and (b) 712 seeds from a highly inbred line that was derived from a single seed from the original lot. The variability in (a) is due to both environmental and genetic factors; in (b), because the plants are nearly genetically identical, the variation in weights is due largely to environmental influence.<sup>10</sup> Thus, there is less variability in the inbred line. ■

**Figure 2.2.18** Weights of princess bean seeds:  
(a) from an open-bred population; (b) from an inbred line



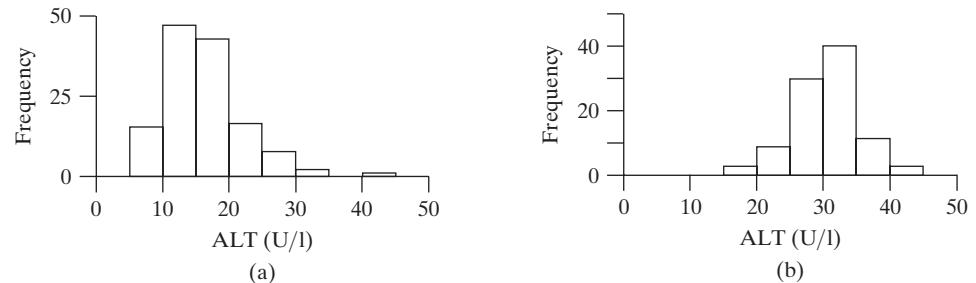
### Example 2.2.12

**Serum ALT** Alanine aminotransferase (ALT) is an enzyme found in most human tissues. Part (a) of Figure 2.2.19 shows the serum ALT concentrations for 129 adult volunteers. The following are potential sources of variability among the measurements:

1. Interindividual
  - (a) Genetic
  - (b) Environmental
2. Intraindividual
  - (a) Biological: changes over time
  - (b) Analytical: imprecision in assay

The effect of the last source—analytical variation—can be seen in part (b) of Figure 2.2.19, which shows the frequency distribution of 109 assays of the same specimen of serum; the figure shows that the ALT assay is fairly imprecise.<sup>11</sup> ■

**Figure 2.2.19** Distribution of serum ALT measurements (a) for 129 volunteers; (b) for 109 assays of the same specimen



## Exercises 2.2.1–2.2.9

**2.2.1** A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*. The results were as follows:<sup>12</sup>

6.1	5.7	6.0	6.5	6.0	5.7
6.1	5.8	5.9	6.1	6.2	6.0
6.3	6.2	6.1	6.2	6.0	5.7
6.2	5.8	5.7	6.3	6.2	5.7
6.2	6.1	5.9	6.5	5.4	6.7
5.9	6.1	5.9	5.9	6.1	6.1

- (a) Construct a frequency distribution and display it as a table and as a histogram.
- (b) Describe the shape of the distribution.

**2.2.2** In a study of schizophrenia, researchers measured the activity of the enzyme monoamine oxidase (MAO) in the blood platelets of 18 patients. The results (expressed as nmoles benzylaldehyde product per 108 platelets) were as follows:<sup>13</sup>

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

Construct a dotplot of the data.

**2.2.3** Consider the data presented in Exercise 2.2.2. Construct a frequency distribution and display it as a table and as a histogram.

**2.2.4** A dendritic tree is a branched structure that emanates from the body of a nerve cell. As part of a study of brain development, 36 nerve cells were taken from the brains of newborn guinea pigs. The investigators counted the number of dendritic branch segments emanating from each nerve cell. The numbers were as follows:<sup>14</sup>

23	30	54	28	31	29	34	35	30
27	21	43	51	35	51	49	35	24
26	29	21	29	37	27	28	33	33
23	37	27	40	48	41	20	30	57

Construct a dotplot of the data.

**2.2.5** Consider the data presented in Exercise 2.2.4. Construct a frequency distribution and display it as a table and as a histogram.

**2.2.6** The total amount of protein produced by a dairy cow can be estimated from periodic testing of her milk. The following are the total annual protein production values (lb) for twenty-eight 2-year-old Holstein cows. Diet, milking procedures, and other conditions were the same for all the animals.<sup>15</sup>

425	481	477	434	410	397	438
545	528	496	502	529	500	465
539	408	513	496	477	445	546
471	495	445	565	499	508	426

Construct a frequency distribution and display it as a table and as a histogram.

**2.2.7** For each of 31 healthy dogs, a veterinarian measured the glucose concentration in the anterior chamber of the right eye and also in the blood serum. The following data are the anterior chamber glucose measurements, expressed as a percentage of the blood glucose.<sup>16</sup>

81	85	93	93	99	76	75	84
78	84	81	82	89	81	96	82
74	70	84	86	80	70	131	75
88	102	115	89	82	79	106	

Construct a frequency distribution and display it as a table and as a histogram.

**2.2.8** Agronomists measured the yield of a variety of hybrid corn in 16 locations in Illinois. The data, in bushels per acre, were<sup>17</sup>

241	230	207	219	266	167
204	144	178	158	153	
187	181	196	149	183	

- (a) Construct a dotplot of the data.

- (b) Describe the shape of the distribution.

**2.2.9 (Computer problem)** Trypanosomes are parasites that cause disease in humans and animals. In an early study of trypanosome morphology, researchers measured the lengths of 500 individual trypanosomes taken from the blood of a rat. The results are summarized in the accompanying frequency distribution.<sup>18</sup>

LENGTH ( $\mu\text{m}$ )	FREQUENCY (NUMBER OF INDIVIDUALS)	LENGTH ( $\mu\text{m}$ )	FREQUENCY (NUMBER OF INDIVIDUALS)
15	1	27	36
16	3	28	41
17	21	29	48
18	27	30	28
19	23	31	43
20	15	32	27
21	10	33	23
22	15	34	10
23	19	35	4
24	21	36	5
25	34	37	1
26	44	38	1

- (a) Construct a histogram of the data using 24 classes (i.e., one class for each integer length, from 15 to 38).
- (b) What feature of the histogram suggests the interpretation that the 500 individuals are a mixture of two distinct types?
- (c) Construct a histogram of the data using only 6 classes. Discuss how this histogram gives a qualitatively different impression than the histogram from part (a).

## 2.3 Descriptive Statistics: Measures of Center

For categorical data, the frequency distribution provides a concise and complete summary of a sample. For numeric variables, the frequency distribution can usefully be supplemented by a few numerical measures. A numerical measure calculated from sample data is called a **statistic**.\* **Descriptive statistics** are statistics that describe a set of data. Usually the descriptive statistics for a sample are calculated in order to provide information about a population of interest (see Section 2.8). In this section we discuss measures of the center of the data. There are several different ways to define the “center” or “typical value” of the observations in a sample. We will consider the two most widely used measures of center: the median and the mean.

### The Median

Perhaps the simplest measure of the center of a data set is the sample **median**. The sample median is the value that most nearly lies in the middle of the sample—it is the data value that splits the ordered data into two equal halves. To find the median, first arrange the observations in increasing order. In the array of ordered observations, the median is the middle value (if  $n$  is odd) or midway between the two middle values (if  $n$  is even). We denote the median of the sample by the symbol  $\tilde{y}$  (read “y-tilde”). Example 2.3.1 illustrates these definitions.

**Example  
2.3.1**

**Weight Gain of Lambs** The following are the two-week weight gains (lb) of six young lambs of the same breed that had been raised on the same diet:<sup>19</sup>

11 13 19 2 10 1

The ordered observations are

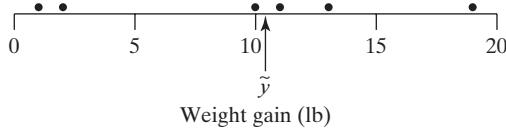
1 2 10 11 13 19

The median weight gain is

$$\tilde{y} = \frac{10 + 11}{2} = 10.5 \text{ lb}$$

The median divides the sorted data into two equal pieces (the same number of observations fall above and below the median). Figure 2.3.1 shows a dotplot of the lamb weight-gain data, along with the location of  $\tilde{y}$ .

**Figure 2.3.1** Plot of the lamb weight-gain data



\*Numerical measures based on the entire population are called **parameters**, which are discussed in greater detail in Section 2.8.

**Example  
2.3.2**

**Weight Gain of Lambs** Suppose the sample contained one more lamb, with the seven ranked observations as follows:

1 2 10 10 11 13 19

For this sample, the median weight gain is

$$\bar{y} = 10 \text{ lb}$$

(Notice that in this example there are two lambs whose weight gain is equal to the median. The fourth observation—the second 10—is the median.) ■

A more formal way to define the median is in terms of rank position in the ordered array (counting the smallest observation as rank 1, the next as 2, and so on). The rank position of the median is equal to

$$(0.5)(n + 1)$$

Thus, if  $n = 7$ , we calculate  $(0.5)(n + 1) = 4$ , so that the median is the fourth largest observation; if  $n = 6$ , we have  $(0.5)(n + 1) = 3.5$ , so that the median is midway between the third and fourth largest observations. Note that the formula  $(0.5)(n + 1)$  does not give the median, it gives the location of the median within the ordered list of the data.

## The Mean

The most familiar measure of center is the ordinary average or **mean** (sometimes called the arithmetic mean). The mean of a sample (or “the sample mean”) is the sum of the observations divided by the number of observations. If we denote a variable by  $Y$ , then we denote the observations in a sample by  $y_1, y_2, \dots, y_n$  and we denote the mean of the sample by the symbol  $\bar{y}$  (read “y-bar”). Example 2.3.3 illustrates this notation.

**Example  
2.3.3**

**Weight Gain of Lambs** The following are the data from Example 2.3.1:

11 13 19 2 10 1

Here  $y_1 = 11$ ,  $y_2 = 13$ , and so on, and  $y_6 = 1$ . The sum of the observations is  $11 + 13 + \dots + 1 = 56$ . We can write this using “summation notation” as  $\sum_{i=1}^n y_i = 56$ . The symbol  $\sum_{i=1}^n y_i$  means to “add up the  $y_i$ ’s.” Thus, when  $n = 6$ ,  $\sum_{i=1}^n y_i = y_1 + y_2 + y_3 + y_4 + y_5 + y_6$ . In this case we get  $\sum_{i=1}^n y_i = 11 + 13 + 19 + 2 + 10 + 1 = 56$ .

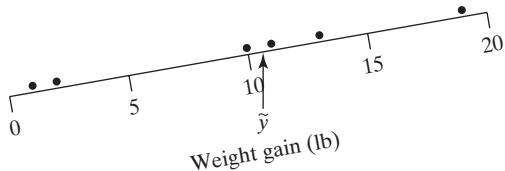
The mean weight gain of the six lambs in this sample is

$$\begin{aligned}\bar{y} &= \frac{11 + 13 + 19 + 2 + 10 + 1}{6} \\ &= \frac{56}{6} \\ &= 9.33 \text{ lb}\end{aligned}$$

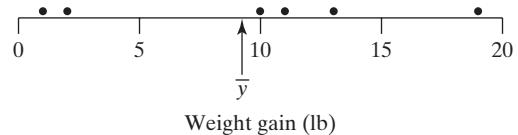
**The Sample Mean** The general definition of the sample mean is

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

where the  $y_i$ ’s are the observations in the sample and  $n$  is the sample size (that is, the number of  $y_i$ ’s).



**Figure 2.3.2** Plot of the lamb weight-gain data with the sample median as the fulcrum of a balance



**Figure 2.3.3** Plot of the lamb weight-gain data with the sample mean as the fulcrum of a balance

While the median divides the data into two equal pieces (i.e., the same number of observations above and below), the mean is the “point of balance” of the data. Figure 2.3.2 shows a dotplot of the lamb weight-gain data, along with the location of  $\tilde{y}$ . If the data points were children on a weightless seesaw, then the seesaw would tip if the fulcrum were placed at  $\tilde{y}$  despite there being the same number of children on either side. The children on the left side (below  $\tilde{y}$ ) tend to sit further from  $\tilde{y}$  than the children on the right (above  $\tilde{y}$ ) causing the seesaw to tip. However, if the fulcrum were placed at  $\bar{y}$ , the seesaw would exactly balance as in Figure 2.3.3. ■

The difference between a data point and the mean is called a **deviation**:  $\text{deviation}_i = y_i - \bar{y}$ . The mean has the property that the sum of the deviations from the mean is zero—that is,  $\sum_{i=1}^n (y_i - \bar{y}) = 0$ . In this sense, the mean is a center of the distribution—the positive deviations balance the negative deviations.

---

### Example 2.3.4

**Weight Gain of Lambs** For the lamb weight-gain data, the deviations are as follows:

$$\begin{aligned}\text{deviation}_1 &= y_1 - \bar{y} = 11 - 9.33 = 1.67 \\ \text{deviation}_2 &= y_2 - \bar{y} = 13 - 9.33 = 3.67 \\ \text{deviation}_3 &= y_3 - \bar{y} = 19 - 9.33 = 9.67 \\ \text{deviation}_4 &= y_4 - \bar{y} = 2 - 9.33 = -7.33 \\ \text{deviation}_5 &= y_5 - \bar{y} = 10 - 9.33 = 0.67 \\ \text{deviation}_6 &= y_6 - \bar{y} = 1 - 9.33 = -8.33\end{aligned}$$

The sum of the deviations is  $\sum_{i=1}^n (y_i - \bar{y}) = 1.67 + 3.67 + 9.67 - 7.33 + 0.67 - 8.33 = 0$ . ■

**Robustance** A statistic is said to be **robust** or **resistant** if the value of the statistic is relatively unaffected by changes in a small portion of the data, even if the changes are dramatic ones. The median is a robust statistic, but the mean is not robust because it can be greatly shifted by changes in even one observation. Example 2.3.5 illustrates this behavior.

---

### Example 2.3.5

**Weight Gain of Lambs** Recall that for the lamb weight-gain data

$$1 \ 2 \ 10 \ 11 \ 13 \ 19$$

we found

$$\bar{y} = 9.3 \text{ and } \tilde{y} = 10.5$$

Suppose now that the observation 19 is changed, or even omitted. How would the mean and median be affected? You can visualize the effect by imagining moving or removing the right-hand dot in Figure 2.3.3. Clearly the mean could change a great deal; the median would generally be less affected. For instance,

If the 19 is changed to 12, the mean becomes 8.2 and the median does not change.

If the 19 is omitted, the mean becomes 7.4 and the median becomes 10.

These changes are not wild ones; that is, the changed samples might well have arisen from the same feeding experiment. Of course, a huge change, such as changing the 19 to 100, would shift the mean very drastically. Note that it would not shift the median at all. ■

## Visualizing the Mean and Median

We can visualize the mean and the median in relation to the histogram of a distribution. The median divides the area under the histogram roughly in half because it divides the observations roughly in half [“roughly” because some observations may be tied at the median, as in Example 2.3.3(b), and because the observations within each class are not uniformly distributed across the class]. The mean can be visualized as the point of balance of the histogram: If the histogram were made out of plywood, it would balance if supported at the mean.

If the frequency distribution is symmetric, the mean and the median are equal and fall in the center of the distribution. If the frequency distribution is skewed, both measures are pulled toward the longer tail, but the mean is usually pulled farther than the median. The effect of skewness is illustrated by the following example.

### Example 2.3.6

**Cricket Singing Times** Male Mormon crickets (*Anabrus simplex*) sing to attract mates. A field researcher measured the duration of 51 unsuccessful songs—that is, the time until the singing male gave up and left his perch.<sup>20</sup> Figure 2.3.4 shows the histogram of the 51 singing times. Table 2.3.1 gives the raw data. The median is 3.7 min and the mean is 4.3 min. The discrepancy between these measures is due largely to the long straggly tail of the distribution; the few unusually long singing times influence the mean, but not the median. ■

Table 2.3.1 Fifty-one cricket singing times (min)								
4.3	3.9	17.4	2.3	0.8	1.5	0.7	3.7	
24.1	9.4	5.6	3.7	5.2	3.9	4.2	3.5	
6.6	6.2	2.0	0.8	2.0	3.7	4.7		
7.3	1.6	3.8	0.5	0.7	4.5	2.2		
4.0	6.5	1.2	4.5	1.7	1.8	1.4		
2.6	0.2	0.7	11.5	5.0	1.2	14.1		
4.0	2.7	1.6	3.5	2.8	0.7	8.6		

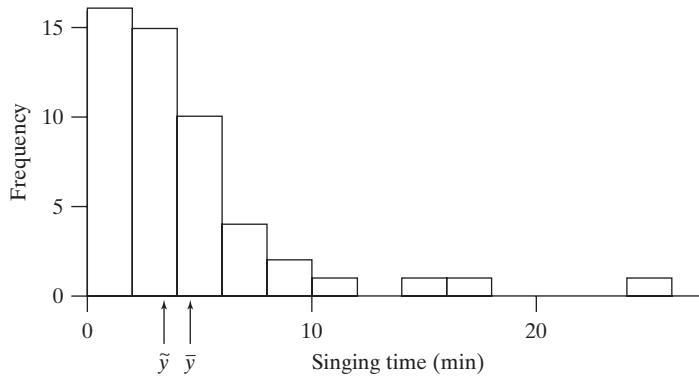


Figure 2.3.4 Histogram of cricket singing times

## Mean versus Median

Both the mean and the median are usually reasonable measures of the center of a data set. The mean is related to the sum; for example, if the mean weight gain of 100 lambs is 9 lb, then the total weight gain is 900 lb, and this total may be of primary interest since it translates more or less directly into profit for the farmer. In some

situations the mean makes very little sense. Suppose, for example, that the observations are survival times of cancer patients on a certain treatment protocol, and that most patients survive less than 1 year, while a few respond well and survive for 5 or even 10 years. In this case, the mean survival time might be greater than the survival time of most patients; the median would more nearly represent the experience of a “typical” patient. Note also that the mean survival time cannot be computed until the last patient has died; the median does not share this disadvantage. Situations in which the median can readily be computed, but the mean cannot, are not uncommon in bioassay, survival, and toxicity studies.

We have noted that the median is more resistant than the mean. If a data set contains a few observations rather distant from the main body of the data—that is, a long “straggly” tail—then the mean may be unduly influenced by these few unusual observations. Thus, the “tail” may “wag the dog”—an undesirable situation. In such cases, the resistance of the median may be advantageous.

An advantage of the mean is that in some circumstances it is more efficient than the median. Efficiency is a technical notion in statistical theory; roughly speaking, a method is efficient if it takes full advantage of all the information in the data. Partly because of its efficiency, the mean has played a major role in classical methods in statistics.

## Exercises 2.3.1–2.3.16

**2.3.1** Invent a sample of size 5 for which the sample mean is 20 and not all the observations are equal.

**2.3.2** Invent a sample of size 5 for which the sample mean is 20 and the sample median is 15.

**2.3.3** A researcher applied the carcinogenic (cancer-causing) compound benzo(a)pyrene to the skin of five mice, and measured the concentration in the liver tissue after 48 hours. The results (nmol/gm) were as follows:<sup>21</sup>

6.3 5.9 7.0 6.9 5.9

Determine the mean and the median.

**2.3.4** Consider the data from Exercise 2.3.3. Do the calculated mean and median support the claim that, in general, liver tissue concentration after 48 hours differs from 6.3 nmol/gm?

**2.3.5** Six men with high serum cholesterol participated in a study to evaluate the effects of diet on cholesterol level. At the beginning of the study their serum cholesterol levels (mg/dl) were as follows:<sup>22</sup>

366 327 274 292 274 230

Determine the mean and the median.

**2.3.6** Consider the data from Exercise 2.3.5. Suppose an additional observation equal to 400 were added to the sample. What would be the mean and the median of the seven observations?

**2.3.7** The weight gains of beef steers were measured over a 140-day test period. The average daily gains (lb/day) of 9 steers on the same diet were as follows:<sup>23</sup>

3.89 3.51 3.97 3.31 3.21  
3.36 3.67 3.24 3.27

Determine the mean and median.

**2.3.8** Consider the data from Exercise 2.3.7. Are the calculated mean and median consistent with the claim that, in general, steers gain 3.5 lb/day? Are they consistent with a claim of 4.0 lb/day?

**2.3.9** Consider the data from Exercise 2.3.7. Suppose an additional observation equal to 2.46 were added to the sample. What would be the mean and the median of the 10 observations?

**2.3.10** As part of a classic experiment on mutations, 10 aliquots of identical size were taken from the same culture of the bacterium *E. coli*. For each aliquot, the number of bacteria resistant to a certain virus was determined. The results were as follows:<sup>24</sup>

14 15 13 21 15  
14 26 16 20 13

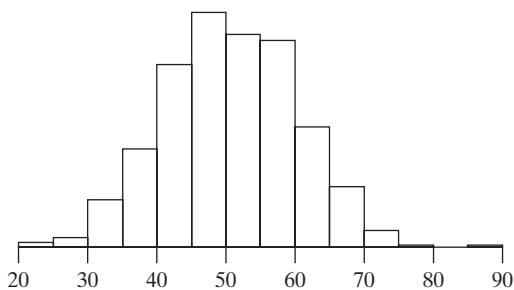
- Construct a frequency distribution of these data and display it as a histogram.
- Determine the mean and the median of the data and mark their locations on the histogram.

**2.3.11** The accompanying table gives the litter size (number of piglets surviving to 21 days) for each of 36 sows (as in Example 2.2.4). Determine the median litter size. (*Hint:* Note that there is one 5, but there are two 7's, three 8's, etc.)

NUMBER OF PIGLETS	FREQUENCY (NUMBER OF SOWS)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

**2.3.12** Consider the data from Exercise 2.3.11. Determine the mean of the 36 observations. (*Hint:* Note that there is one 5 but there are two 7's, three 8's, etc. Thus,  $\sum y_i = 5 + 7 + 7 + 8 + 8 + 8 + \dots = 5 + 2(7) + 3(8) + \dots$ )

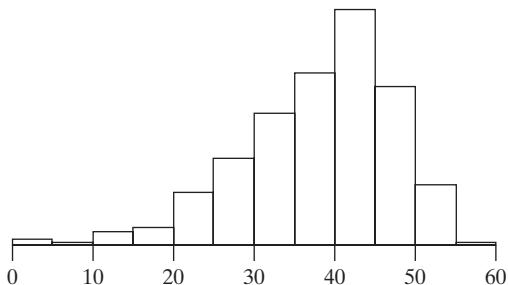
**2.3.13** Here is a histogram.



- (a) Estimate the median of the distribution.
- (b) Estimate the mean of the distribution.

**2.3.14** Consider the histogram from Exercise 2.3.13. By “reading” the histogram, estimate the percentage of observations that are less than 40. Is this percentage closest to 15%, 25%, 35%, or 45%? (*Note:* The frequency scale is not given for this histogram, because there is no need to calculate the number of observations in each class. Rather, the percentage of observations that are less than 40 can be estimated by looking at area.)

**2.3.15** Here is a histogram.



- (a) Estimate the median of the distribution.
- (b) Estimate the mean of the distribution.

**2.3.16** Consider the histogram from Exercise 2.3.15. By “reading” the histogram, estimate the percentage of observations that are greater than 45. Is this percentage closest to 15%, 25%, 35%, or 45%? (*Note:* The frequency scale is not given for this histogram, because there is no need to calculate the number of observations in each class. Rather, the percentage of observations that are greater than 45 can be estimated by looking at area.)

## 2.4 Boxplots

One of the most efficient graphics, both for examining a single distribution and for making comparisons between distributions, is known as a boxplot, which is the topic of this section. Before discussing boxplots, however, we need to discuss quartiles.

### Quartiles and the Interquartile Range

The median of a distribution splits the distribution into two parts, a lower part and an upper part. The **quartiles** of a distribution divide each of these parts in half, thereby dividing the distribution into four quarters. The **first quartile**, denoted by  $Q_1$ , is

the median of the data values in the lower half of the data set. The **third quartile**, denoted by  $Q_3$ , is the median of the data values in the upper half of the data set.\* The following example illustrates these definitions.

**Example  
2.4.1**

**Blood Pressure** The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:<sup>25</sup>

151 124 132 170 146 124 113

Putting these values in rank order, the sample is

113 124 124 132 146 151 170

The median is the fourth largest observation, which is 132. There are three data points in the lower part of the distribution: 113, 124, and 124. The median of these three values is 124. Thus, the first quartile,  $Q_1$ , is 124.

Likewise, there are three data points in the upper part of the distribution: 146, 151 and 170. The median of these three values is 151. Thus, the third quartile,  $Q_3$ , is 151.

$$\begin{array}{ccccccc} 113 & & 124 & & 124 & & 132 & & 146 & & 151 & & 170 \\ & \uparrow & & & & | & & & & & \uparrow & & \\ & \text{first quartile} & & & \text{median} & & & & & & \text{third quartile} & & \\ & Q_1 & & & & & & & & & & Q_3 & & \blacksquare \end{array}$$

Note that the median is not included in either the lower part or the upper part of the distribution. If the sample size,  $n$ , is even, then exactly one-half of the observations are in the lower part of the distribution and one-half are in the upper part.

The **interquartile range** is the difference between the first and third quartiles and is abbreviated as **IQR**:  $IQR = Q_3 - Q_1$ . For the blood pressure data in Example 2.4.1, the IQR is  $151 - 124 = 27$ .

**Example  
2.4.2**

**Pulse** The pulses of 12 college students were measured.<sup>26</sup> Here are the data, arranged in order, with the position of the median indicated by a dashed line:

62 64 68 70 70 74 | 74 76 76 78 78 80

The median is  $\frac{74 + 74}{2} = 74$ . There are six observations in the lower part of the distribution: 62, 64, 68, 70, 70, 74. Thus, the first quartile is the average of the third and fourth largest data values:

$$Q_1 = \frac{68 + 70}{2} = 69$$

There are six observations in the upper part of the distribution: 74, 76, 76, 78, 78, 80. Thus, the third quartile is the average of the ninth and tenth largest data values (the third and fourth values in the upper part of the distribution):

$$Q_3 = \frac{76 + 78}{2} = 77$$

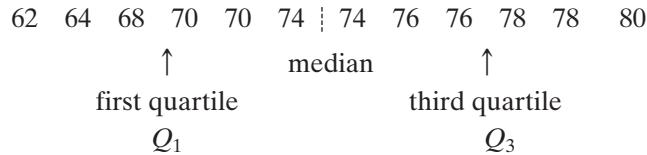
---

\*Some authors use other definitions of quartiles, as does some computer software. A common alternative definition is to say that the first quartile has rank position  $(.25)(n + 1)$  and that the third quartile has rank position  $(.75)(n + 1)$ . Thus, if  $n = 10$ , the first quartile would have rank position  $(.25)(11) = 2.75$ —that is, to find the first quartile we would have to interpolate between the second and third largest observations. If  $n$  is large, then there is little practical difference between the definitions that various authors use.

Thus, the interquartile range is

$$\text{IQR} = 77 - 69 = 8$$

We have

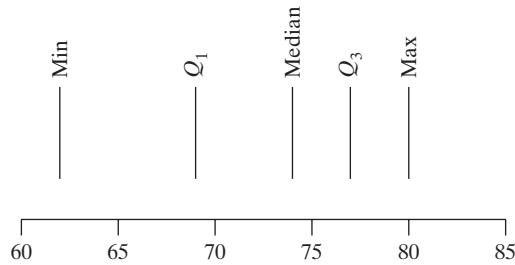


The minimum pulse value is 62 and the maximum is 80. ■

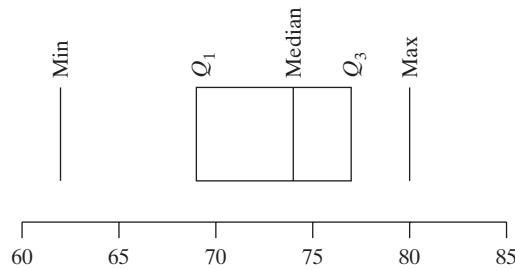
The minimum, the maximum, the median, and the quartiles, taken together, are referred to as the **five-number summary** of the data.

## Boxplots

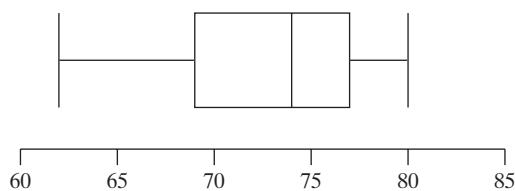
A **boxplot** is a visual representation of the five-number summary. To make a boxplot, we first make a number line; then we mark the positions minimum,  $Q_1$ , the median,  $Q_3$ , and the maximum:



Next, we make a box connecting the quartiles:



Note that the interquartile range is equal to the length of the box. Finally, we extend “whiskers” from  $Q_1$  down to the minimum and from  $Q_3$  up to the maximum:



A boxplot gives a quick visual summary of the distribution. We can immediately see where the center of the data is from the line within the box that locates the median. We see the spread of the total distribution, from the minimum up to the maximum, as well as the spread of the middle half of the distribution—the interquartile range—from the length of the box. The boxplot also gives an indication of the shape of the distribution; the preceding boxplot has a long lower whisker, indicating that the distribution is skewed to the left. Example 2.4.3 shows a boxplot for data from a radish growth experiment.\*

**Example  
2.4.3**

**Radish Growth** A common biology experiment involves growing radish seedlings under various conditions. In one version of this experiment, a moist paper towel is put into a plastic bag. Staples are put in the bag about one-third of the way from the bottom of the bag and then radish seeds are placed along the staple seam. One group of students kept their radish seed bags in total darkness for three days and then measured the length, in mm, of each radish shoot at the end of the three days. They collected 14 observations; the data are shown in Table 2.4.1.<sup>27</sup>

**Table 2.4.1** Radish growth, in mm, after three days in total darkness

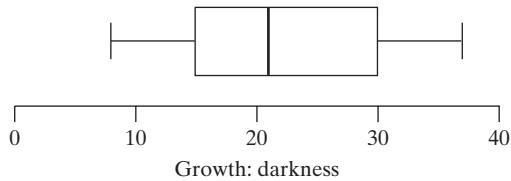
15	20	11	30	33
20	29	35	8	10
22	37	15	25	

Here are the data in order from smallest to largest:

8	10	11	<b>15</b>	15	20	<b>20</b>		<b>22</b>	25	29	<b>30</b>	33	35	37
↑				median				↑						
first quartile								third quartile						
$Q_1$								$Q_3$						

The quartiles are  $Q_1 = 15$  and  $Q_3 = 30$ . The median,  $\tilde{y} = 21$ , is the average of the two middle values of 20 and 22. Figure 2.4.1 shows a boxplot of the same data. ■

**Figure 2.4.1** Boxplot of data on radish growth in darkness



## Outliers

Sometimes a data point differs so much from the rest of the data that it doesn't seem to belong with the other data. Such a point is called an **outlier**. An outlier might occur because of a recording error or typographical error when the data are recorded, because of an equipment failure during an experiment, or for many other rea-

\*This and subsequent boxplots in our text are slightly stylized. Different computer packages present the plot somewhat differently, but all boxplots have the same basic five-number summary.

sons. Outliers are the most interesting points in a data set. Sometimes outliers tell us about a problem with the experimental protocol (e.g., an equipment failure or a failure of a patient to take his or her medication consistently during a medical trial). At other times an outlier might alert us to the fact that a special circumstance has happened (e.g., an abnormally high or low value on a medical test could indicate the presence of a disease in a patient).

People often use the term “outlier” informally. There is, however, a common definition of “outlier” in statistical practice. To give a definition of outlier, we first discuss what are known as fences. The **lower fence** of a distribution is

$$\text{lower fence} = Q_1 - 1.5 \times \text{IQR}$$

The **upper fence** of a distribution is

$$\text{upper fence} = Q_3 + 1.5 \times \text{IQR}$$

This means that the fences are located  $1.5 \times \text{IQR}$  (i.e.,  $1.5 \times$  the length of the box) beyond the end of the box in a boxplot.

Note that the fences need not be data values; indeed, there might be no data near the fences. The fences just locate limits within the sample distribution. These limits give us a way to define outliers. *An outlier is a data point that falls outside of the fences.* That is, if

$$\text{data point} < Q_1 - 1.5 \times \text{IQR}$$

or

$$\text{data point} > Q_3 + 1.5 \times \text{IQR}$$

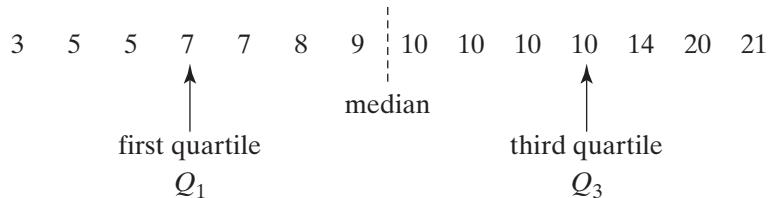
then we call the point an outlier.

#### Example 2.4.4

**Pulse** In Example 2.4.2 we saw that  $Q_1 = 69$ ,  $Q_3 = 77$ , and  $\text{IQR} = 8$ . Thus, the lower fence is  $69 - 1.5 \times 8 = 69 - 12 = 57$ . Any point less than 57 would be an outlier. The upper fence is  $77 + 1.5 \times 8 = 77 + 12 = 89$ . Any point greater than 89 would be an outlier. Since there are no points less than 57 or greater than 89, there are no outliers in this data set. ■

#### Example 2.4.5

**Radish Growth in Light** The data in Example 2.4.3 were for radish seedlings grown in total darkness. In another part of the experiment students grew 14 radish seedlings in constant light. The observations, in order, are



Thus, the median is  $\frac{9 + 10}{2} = 9.5$ ,  $Q_1$  is 7, and  $Q_3$  is 10. The interquartile range is

$\text{IQR} = 10 - 7 = 3$ . The lower fence is  $7 - 1.5 \times 3 = 7 - 4.5 = 2.5$ , so any point less than 2.5 would be an outlier. The upper fence is  $10 + 1.5 \times 3 = 10 + 4.5 = 14.5$ , so any point greater than 14.5 is an outlier. Thus, the two largest observations in this data set are outliers: 20 and 21. ■

The method we have defined for identifying outliers allows the bulk of the data to determine how extreme an observation must be before we consider it to be an outlier, since the quartiles and the IQR are determined from the data themselves. Thus, a point that is an outlier in one data set might not be an outlier in another data set. We label a point as an outlier if it is unusual relative to the inherent variability in the entire data set.

After an outlier has been identified, people are often tempted to remove the outlier from the data set. In general this is not a good idea. If we can identify that an outlier occurred due to an equipment error, for example, then we have good reason to remove the outlier before analyzing the rest of the data. However, quite often outliers appear in data sets without any identifiable, external reason for them. In such cases, we simply proceed with our analysis, aware that there is an outlier present. In some cases, we might want to calculate the mean, for example, with and without the outlier and then report both calculations, to show the effect of the outlier in the overall analysis. This is preferable to removing the outlier, which obscures the fact that there was an unusual data point present. In presenting data graphically, we can draw attention to outliers by using modified boxplots, which we now introduce.

## Modified Boxplot

A standard variation on the idea of a boxplot is what is known as a modified boxplot. A **modified boxplot** is a boxplot in which the outliers, if any, are graphed as separate points. The advantage of a modified boxplot is that it lets us quickly see where the outliers are, if there are any.

To make a modified boxplot, we proceed as we did when first making a boxplot, except for the last step. After drawing the box for the boxplot, we check to see if there are outliers. If there are no outliers, then we extend whiskers from the box out to the extremes (the minimum and the maximum). However, if there are outliers in the upper part of the distribution, then we identify them with a dot or other plotting symbol. We then extend a whisker from  $Q_3$  up to the largest data point that is not an outlier. Likewise, if there are outliers in the lower part of the distribution, we identify them with asterisks and extend a whisker from  $Q_1$  down to the smallest observation that is not an outlier. Figure 2.4.2 shows the distribution of radish seedlings grown under constant light. The area between the lower and upper fences is white, while the outlying region is blue.

**Figure 2.4.2** Dotplot and boxplot of data on radish growth in constant light. The points in the blue region are outliers.

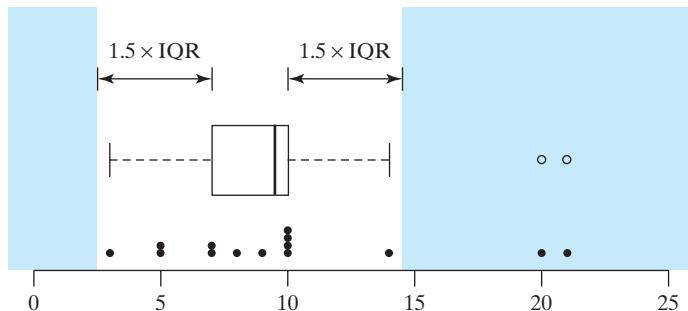
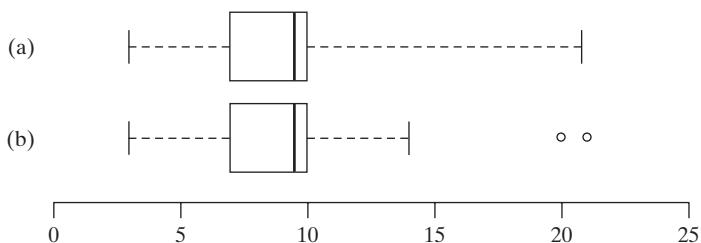


Figure 2.4.3 shows a boxplot and a modified boxplot of the data on radish seedlings grown in constant light.

**Figure 2.4.3** (a) Boxplot of data on radish growth in constant light; (b) modified boxplot of radish growth data



Most often, when people make boxplots, they make modified boxplots. Computer software is typically programmed to produce a modified boxplot when the user asks for a boxplot. Thus, we will use the term “boxplot” to mean “modified boxplot.”

### Exercises 2.4.1–2.4.8

**2.4.1** Here are the data from Exercise 2.3.10 on the number of virus-resistant bacteria in each of 10 aliquots:

14	15	13	21	15
14	26	16	20	13

- (a) Determine the median and the quartiles.
- (b) Determine the interquartile range.
- (c) How large would an observation in this data set have to be in order to be an outlier?

**2.4.2** Here are the 18 measurements of MAO activity reported in Exercise 2.2.2:

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

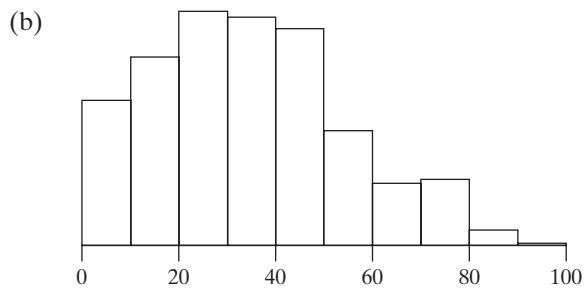
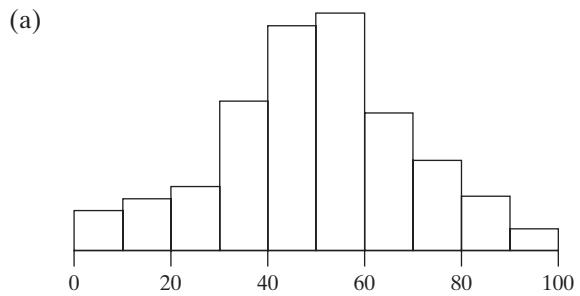
- (a) Determine the median and the quartiles.
- (b) Determine the interquartile range.
- (c) How large would an observation in this data set have to be in order to be an outlier?
- (d) Construct a (modified) boxplot of the data.

**2.4.3** In a study of milk production in sheep (for use in making cheese), a researcher measured the three-month milk yield for each of 11 ewes. The yields (liters) were as follows:<sup>28</sup>

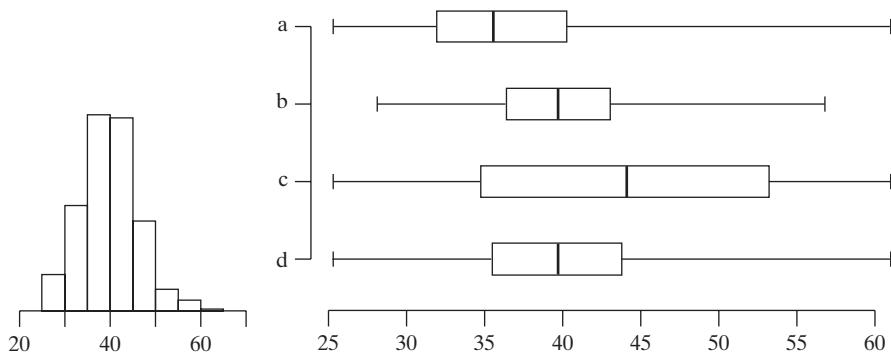
56.5	89.8	110.1	65.6	63.7	82.6
75.1	91.5	102.9	44.4	108.1	

- (a) Determine the median and the quartiles.
- (b) Determine the interquartile range.
- (c) Construct a (modified) boxplot of the data.

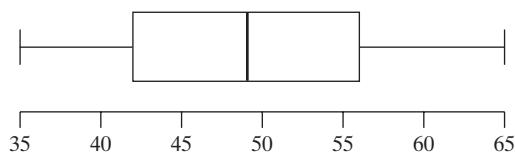
**2.4.4** For each of the following histograms, use the histogram to estimate the median and the quartiles; then construct a boxplot for the distribution.



**2.4.5** The following histogram shows the same data that are shown in one of the four boxplots. Which boxplot goes with the histogram? Explain your answer.



**2.4.6** The following boxplot shows the five-number summary for a data set. For these data the minimum is 35,  $Q_1$  is 42, the median is 49,  $Q_3$  is 56, and the maximum is 65. Is it possible that no observation in the data set equals 42? Explain your answer.



**2.4.7** Statistics software can be used to find the five-number summary of a data set. Here is an example of

MINITAB's descriptive statistics summary for a variable stored in column 1 (C1) of MINITAB's worksheet.

Variable	N	Mean	Median	TrMean	StDev	SEMean
C1	75	119.94	118.40	119.98	9.98	1.15

Variable	Min	Max	Q1	Q3
C1	95.16	145.11	113.59	127.42

- Use the MINITAB output to calculate the interquartile range.
- Are there any outliers in this set of data?

**2.4.8** Consider the data from Exercise 2.4.7. Use the five-number summary that is given to create a boxplot of the data.

## 2.5 Relationships between Variables

In the previous sections we have studied **univariate** summaries of both numeric and categorical variables. A univariate summary is a graphical or numeric summary of a single variable.

The histogram, boxplot, sample mean, and median are all examples of univariate summaries for numeric data. The bar chart, frequency, and relative frequency tables are examples of univariate summaries for categorical data. In this section we present some common **bivariate** graphical summaries used to examine the *relationship* between pairs of variables.

### Categorical–Categorical Relationships

To understand the relationship between two categorical variables, we first summarize the data in a **bivariate frequency table**. Unlike the frequency table presented in Section 2.2 (a univariate table), the bivariate frequency table has both rows and columns—one dimension for each variable. The choice of which variable to list with the rows and which to list with the columns is arbitrary. The following example considers the relationship between two categorical variables: *E. Coli* Source and Sampling Location.

**Example  
2.5.1**

**E. Coli Watershed Contamination** In an effort to determine if there are differences in the primary sources of fecal contamination at different locations in the Morro Bay watershed,  $n = 623$  water specimens were collected at three primary locations that feed into Morro Bay: Chorro Creek ( $n_1 = 241$ ), Los Osos Creek ( $n_2 = 256$ ), and Baywood Seeps ( $n_3 = 126$ ).<sup>29</sup> DNA fingerprinting techniques were used to determine the intestinal origin of the dominant *E. coli* strain in each water specimen. *E. coli* origins were classified into the following five categories: bird, domestic pet (e.g., cat or dog), farm animal (e.g., horse, cow, pig), human, or other terrestrial mammal (e.g., fox, mouse, coyote . . .). Thus, each water specimen had two categorical variables measured: location (Chorro, Los Osos, or Baywood) and *E. coli* source (bird, . . ., terrestrial mammal). Table 2.5.1 presents a frequency table of the data. ■

**Table 2.5.1** Frequency table of *E. coli* source by location

Location	E. Coli Source					Total
	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	
<b>Chorro Creek</b>	46	29	106	38	22	<b>241</b>
<b>Los Osos Creek</b>	79	56	32	63	26	<b>256</b>
<b>Baywood</b>	35	23	0	60	8	<b>126</b>
<b>Total</b>	<b>160</b>	<b>108</b>	<b>138</b>	<b>161</b>	<b>56</b>	<b>623</b>

While Table 2.5.1 provides a concise summary of the data, it is difficult to discover any patterns in the data. Examining relative frequencies (row or column proportions) often helps us make meaningful comparisons as seen in the following example.

**Example  
2.5.2**

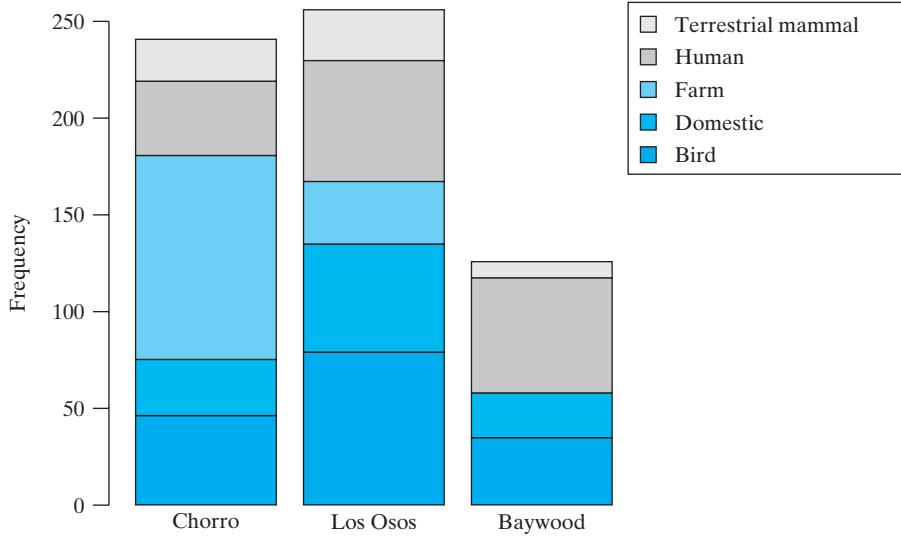
**E. Coli Watershed Contamination** Are domestic pets more of an *E. coli* problem (i.e., source) at Chorro Creek or Baywood? Table 2.5.1 shows that the domestic pet *E. coli* source count at Chorro (29) is higher than Baywood (23), so at first glance it seems that pets are more problematic at Chorro. However, as more water specimens were collected at Chorro ( $n_1 = 241$ ) than Baywood ( $n_2 = 126$ ), the relative frequency of domestic pet source *E. coli* is actually lower at Chorro ( $29/241 = 0.120$ ) than Baywood ( $23/126 = 0.183$ ). Table 2.5.2 displays row percentages and thus facilitates comparisons of *E. coli* sources among the locations. (Note that column percentages would not be meaningful in this context since the water was sampled by location and not by *E. coli* source.). ■

**Table 2.5.2** Bivariate relative frequency table (row percentages) of *E. coli* source by location

Location	E. Coli Source					Total
	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	
<b>Chorro Creek</b>	19.1	12.0	44.0	15.8	9.1	<b>100</b>
<b>Los Osos Creek</b>	30.9	21.9	12.5	24.6	10.2	<b>100</b>
<b>Baywood</b>	27.8	18.3	0.0	47.6	6.3	<b>100</b>
<b>Total</b>	<b>25.7</b>	<b>17.3</b>	<b>22.2</b>	<b>25.8</b>	<b>9.0</b>	<b>100</b>

To visualize the data in Tables 2.5.1 and 2.5.2 we can examine **stacked bar charts**. With a stacked frequency bar chart, the overall height of each bar reflects the sample size for a level of the  $X$  categorical variable (e.g., location) while the height or thickness of a slice that makes up a bar represents the count of the  $Y$  categorical variable (e.g., *E. coli* source) for that level of  $X$ . Figure 2.5.1 displays a stacked bar chart for the *E. coli* watershed count data in Table 2.5.1.

**Figure 2.5.1** Stacked frequency chart of *E. coli* source by location



Like the frequency table, the stacked frequency bar chart is not conducive to making comparisons across the three locations as the sample sizes differ for these locations. (This graph does help highlight the difference in sample sizes; for example, it is very clear that many fewer water specimens were collected at Baywood.) A chart that better displays the distribution of one categorical variable across levels of another is a **stacked relative frequency** (or percentage) bar chart, which graphs the summaries from a bivariate relative frequency table such as Table 2.5.2. Figure 2.5.2 provides an example using the *E. coli* watershed contamination data. This plot normalizes the bars of Figure 2.5.1 to have the same height (100%) to facilitate comparisons across the three locations.

**Figure 2.5.2** Stacked relative frequency (percentage) chart of *E. coli* source by location

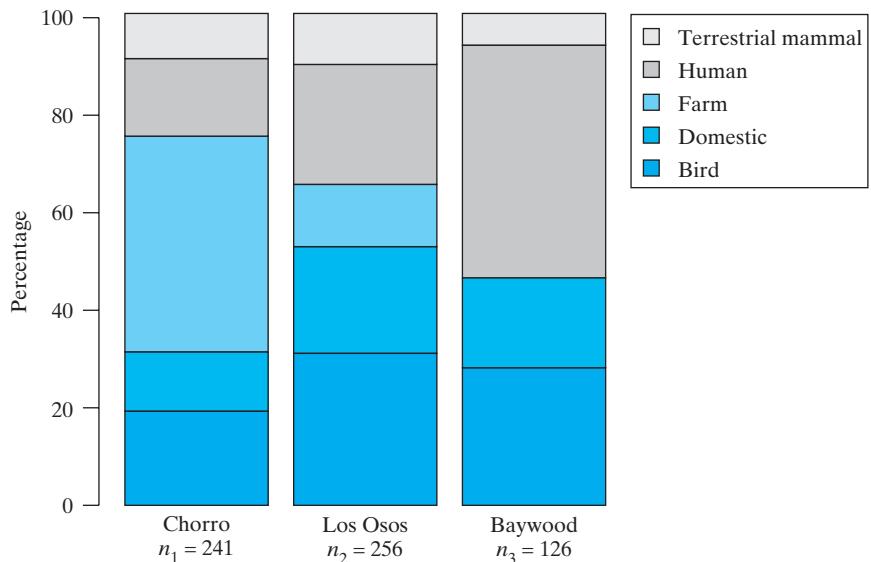


Figure 2.5.2 makes it very easy to see that farm animals are the largest contributors of *E. coli* to Chorro Creek while humans are primarily responsible for the pollution at Baywood. The distribution of the slices in the three bars appears quite different, suggesting that the distribution of *E. coli* sources is not the same at the three locations. In Chapter 10 we will learn how to determine if these apparent differences are large enough to be compelling evidence for real differences in the distribution of *E. coli* source by location, or whether they are likely due to chance variation.

## Numeric–Categorical Relationships

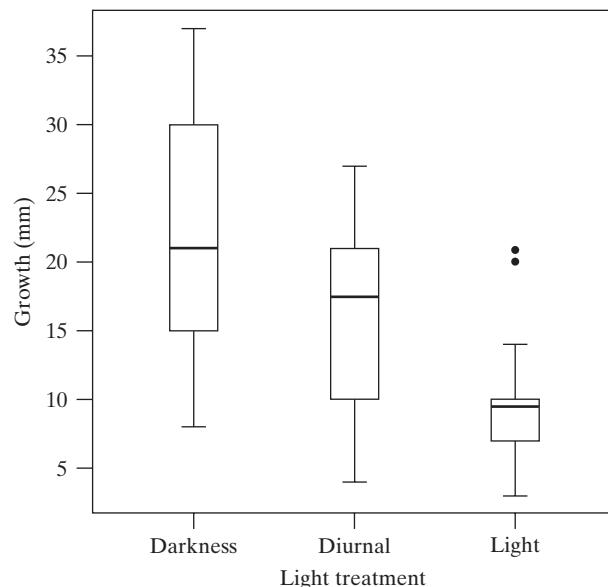
In Section 2.4 we learned that boxplots are graphs based on only five numbers: the minimum, first quartile, median, third quartile, and maximum. They are appealing plots because they are very simple and uncluttered, yet contain easy to read information about center, spread, skewness, and even outliers of a data set. By displaying **side-by-side boxplots** on the same graph, we are able to compare numeric data among several groups. We now consider an extension of the radish shoot growth problem in Example 2.4.3.

---

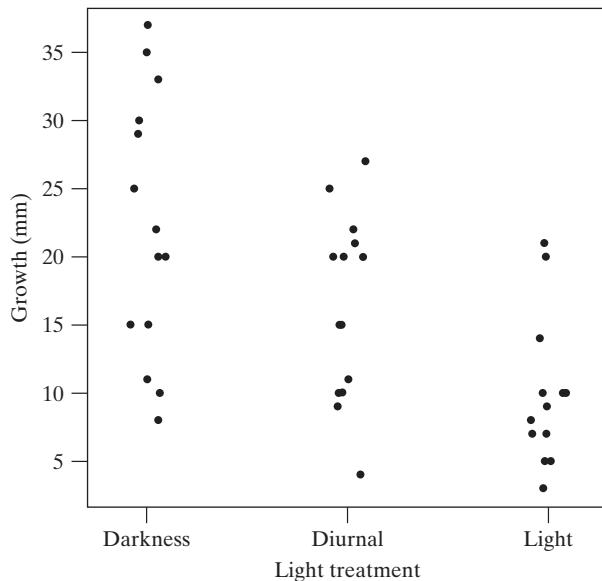
### Example 2.5.3

**Radish Growth** Does light exposure alter initial radish shoot growth? The complete radish growth experiment of Example 2.4.3 actually involved a total of 42 radish seeds randomly divided to receive one of three lighting conditions for germination (14 seeds in each lighting condition): 24-hour light, diurnal light (12 hours of light and 12 hours of darkness each day), and 24 hours of darkness. At the end of three days, shoot length was measured (mm). Thus, each shoot has two variables that are measured in this study: the categorical variable lighting condition (light, diurnal, dark) and the numeric variable sprout length (mm). Figure 2.5.3 displays side-by-side boxplots of the data. The boxplots make it very easy to compare the growth under the three conditions: It appears that light inhibits shoot growth. Are the observed differences in growth among the lighting conditions just due to chance variation, or is light really altering growth? We will learn how to numerically measure the strength of this evidence and answer this question in Chapters 7 and 11. ■

**Figure 2.5.3** Side-by-side boxplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light



**Figure 2.5.4** Side-by-side jittered dotplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light



For smaller data sets, we also may consider side-by-side dotplots of the data. Figure 2.5.4 displays a jittered side-by-side dotplot of the radish growth data of Example 2.5.3. The “jitter” is a common software option that adds horizontal scatter to the plot, helping to reduce the overlap of the dots. Choosing between side-by-side boxplots and dotplots is matter of personal preference. A good rule of thumb is to choose the plot that accurately reflects patterns in the data in the cleanest (least ink on the paper) way possible. For the radish growth example, the boxplot enables a very clean comparison of the growth under the three light treatments without hiding any information revealed by the dotplot.

## Numeric–Numeric Relationships

Each of the previous examples considered comparing the distribution of one variable (either categorical or numeric) among several groups (i.e., across levels of a categorical variable). In the next example we illustrate the **scatterplot** as a tool to examine the relationship between two numeric variables,  $X$  and  $Y$ . A scatterplot plots each observed  $(x,y)$  pair as a dot on the  $x$ - $y$  plane.

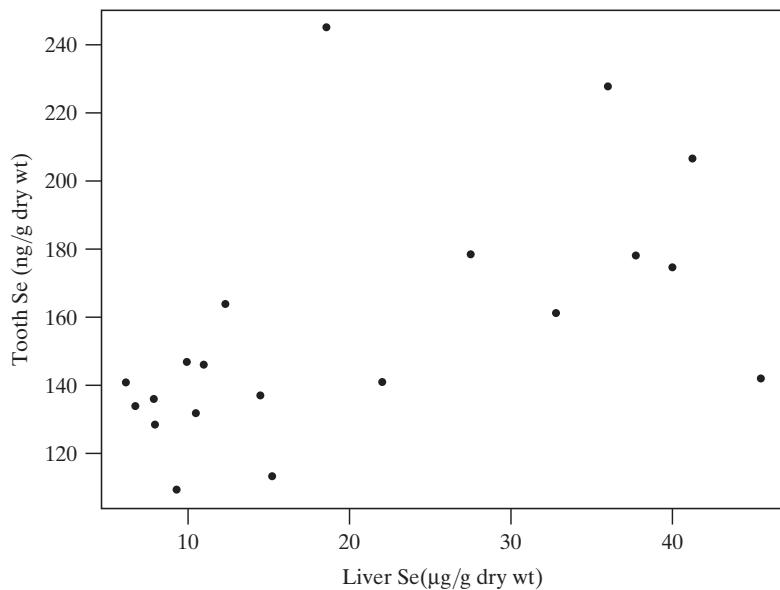
---

### Example 2.5.4

**Whale Selenium** Can metal concentration in marine mammal teeth be used as a bioindicator for body burden? Selenium (Se) is an essential element that has been shown to play an important role in protecting marine mammals against the toxic effects of mercury (Hg) and other metals. Twenty beluga whales (*Delphinapterus leucas*) were harvested from the Mackenzie Delta, Northwest Territories, as part of an annual traditional Inuit hunt.<sup>30</sup> Each whale yielded two numeric measurements: Tooth Se ( $\mu\text{g/g}$ ) and Liver Se (ng/g). Selenium concentrations for the whales are listed in Table 2.5.3. Tooth Se concentration ( $Y$ ) is graphed against Liver Se concentration ( $X$ ) in the scatterplot of Figure 2.5.5. ■

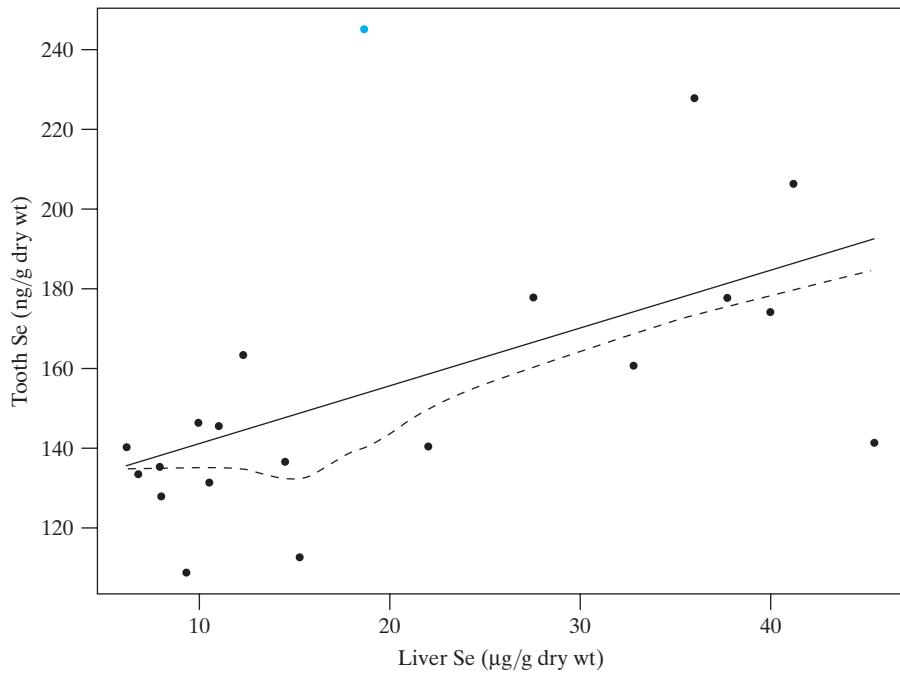
**Table 2.5.3** Liver and tooth selenium concentrations of twenty belugas

Whale	Liver Se ( $\mu\text{g/g}$ )	Tooth Se ( $\text{ng/g}$ )	Whale	Liver Se ( $\mu\text{g/g}$ )	Tooth Se ( $\text{ng/g}$ )
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

**Figure 2.5.5** Scatterplot of tooth selenium concentration against liver selenium concentration for 20 belugas

Scatterplots are helpful in revealing relationships between numeric variables. In Figure 2.5.6 two lines have been added to the whale selenium scatterplot of Figure 2.5.5 to highlight the increasing trend in the data: Tooth Se concentration tends to increase with liver Se concentration. The dashed line is called a **lowess smooth** whereas the straight solid line is called a **regression line**. Many software packages allow one to easily add these lines to a scatterplot. The lowess smooth is particularly helpful in visualizing curved or nonlinear relationships in data, while the regression line is used to highlight linear trend. Generally speaking, we would choose only one of these to display on our graph. In this case, since the pattern is fairly linear (the lowess smooth is fairly straight), we would choose the solid regression line. In Chapter 12 we will learn how to identify the equation of the regression line that best summarizes the data and determine if the apparent trend in the data is likely to be just due to chance or if there is evidence for a real relationship between  $X$  and  $Y$ .

**Figure 2.5.6** Scatterplot of tooth selenium concentration against liver selenium concentration for 20 belugas with regression (solid) and lowess (dashed) summary lines and outlier marked in blue



In addition to revealing relationships between two numeric variables, scatterplots also help reveal outliers that might otherwise be unnoticed in univariate plots (e.g., histograms, single boxplots, etc.). The colored point on Figure 2.5.6 falls far from the scatter of the other points. The  $X$  value of this point is not unusual in any way, and even the  $Y$  value, though large, doesn't appear extreme. The scatterplot, however, shows that the particular  $(x,y)$  pair for this whale is unusual.

## Exercises 2.5.1–2.5.3

**2.5.1** The two claws of the lobster (*Homarus americanus*) are identical in the juvenile stages. By adulthood, however, the two claws normally have differentiated into a stout claw called a “crusher” and a slender claw called a “cutter.” In a study of the differentiation process, 26 juvenile animals were reared in smooth plastic trays and 18 were reared in trays containing oyster chips (which they could use to exercise their claws). Another 23 animals were reared in trays containing only one oyster chip. The claw configurations of all the animals as adults are summarized in the table.<sup>31</sup>

TREATMENT	CLAW CONFIGURATION		
	RIGHT CRUSHER, LEFT CUTTER	RIGHT CUTTER, LEFT CRUSHER	RIGHT AND LEFT CUTTER (NO CRUSHER)
	8	9	1
Oyster chips	8	9	1
Smooth plastic	2	4	20
One oyster chip	7	9	7

- Create a stacked frequency bar chart to display these data.
- Create a stacked relative frequency bar chart to display these data.
- Of the two charts you created in parts (a) and (b), which is more useful for comparing the claw configurations across the three treatments? Why?

**2.5.2** Does the length (mm) of the golden mantled ground squirrel (*Spermophilus lateralis*) differ by latitude in California? A graduate student captured squirrels at four locations across California. Listed from south to north the locations are Hemet, Big Bear, Susanville, and Loop Hill.<sup>32</sup>

	HEMET	BIG BEAR	SUSANVILLE	LOOP HILL
	263	274	245	273
	256	256	272	291
	251	249	263	278
	242	264	260	281
	248		271	
	281			

- Create side-by-side dotplots of the data. Consider the geography of these four locations when making your plot. Is alphabetic order of the locations the most appropriate, or is there a better way to order the location categories?
- Create side-by-side boxplots of the data. Again, consider the geography of these four locations when making your plot.
- Of the two plots created in parts (a) and (b), which do you prefer and why?

**2.5.3** The rowan (*Sorbus aucuparia*) is a tree that grows in a wide range of altitudes. To study how the tree adapts to its varying habitats, researchers collected twigs with attached buds from 12 trees growing at various altitudes in North Angus, Scotland. The buds were brought back to the laboratory and measurements were made of the dark respiration rate. The accompanying table shows the altitude of origin (in meters) of each batch of buds and the dark respiration rate (expressed as  $\mu\text{l}$  of oxygen per hour per mg dry weight of tissue).<sup>33</sup>

TREE	ALTITUDE OF ORIGIN X (M)	RESPIRATION RATE Y ( $\mu\text{l}/\text{hr} \cdot \text{mg}$ )
1	90	0.11
2	230	0.20
3	240	0.13
4	260	0.15
5	330	0.18
6	400	0.16
7	410	0.23
8	550	0.18
9	590	0.23
10	610	0.26
11	700	0.32
12	790	0.37

- Create a scatterplot of the data.
- If your software allows, add a regression line to summarize the trend.
- If your software allows, create a scatterplot with a lowess smooth to summarize the trend.

## 2.6 Measures of Dispersion

We have considered the shapes and centers of distributions, but a good description of a distribution should also characterize how spread out the distribution is—are the observations in the sample all nearly equal, or do they differ substantially? In Section 2.4 we defined the interquartile range, which is one measure of dispersion. We will now consider other measures of dispersion: the range, the standard deviation, and the coefficient of variation.

### The Range

The sample **range** is the difference between the largest and smallest observations in a sample. Here is an example.

---

**Example  
2.6.1**

**Blood Pressure** The systolic blood pressures (mm Hg) of seven middle-aged men were given in Example 2.4.1 as follows:

113 124 124 132 146 151 170

For these data, the sample range is

$$170 - 113 = 57 \text{ mm Hg}$$



The range is easy to calculate, but it is very sensitive to extreme values; that is, it is not robust. If the maximum in the blood pressure sample had been 190 rather than 170, the range would have been changed from 57 to 77.

We defined the interquartile range (IQR) in Section 2.4 as the difference between the quartiles. Unlike the range, the IQR is robust. The IQR of the blood

pressure data is  $151 - 124 = 17$ . If the maximum in the blood pressure sample had been 190 rather than 170, the IQR would not have changed; it would still be 17.

## The Standard Deviation

The standard deviation is the classical and most widely used measure of dispersion. Recall that a *deviation* is the difference between an observation and the sample mean:

$$\text{deviation} = \text{observation} - \bar{y}$$

The standard deviation of the sample, or sample **standard deviation**, is determined by combining the deviations in a special way, as described in the following box.

**The Sample Standard Deviation** The sample standard deviation is denoted by  $s$  and is defined by the following formula:

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

In this formula, the expression  $\sum_{i=1}^n (y_i - \bar{y})^2$  denotes the sum of the squared deviations.

So, to find the standard deviation of a sample, first find the deviations. Then

1. square
2. add
3. divide by  $n - 1$
4. take the square root

To illustrate the use of the formula, we have chosen a data set that is especially simple to handle because the mean happens to be an integer.

**Example**  
**2.6.2**

**Growth of Chrysanthemums** In an experiment on chrysanthemums, a botanist measured the stem elongation (mm in 7 days) of five plants grown on the same greenhouse bench. The results were as follows.<sup>34</sup>

76 72 65 70 82

The data are tabulated in the first column of Table 2.6.1. The sample mean is

$$\bar{y} = \frac{365}{5} = 73 \text{ mm}$$

The deviations ( $y_i - \bar{y}$ ) are tabulated in the second column of Table 2.6.1; the first observation is 3 mm above the mean, the second is 1 mm below the mean, and so on.

The third column of Table 2.6.1 shows that the sum of the squared deviations is

$$= \sum_{i=1}^n (y_i - \bar{y})^2 = 164$$

**Table 2.6.1** Illustration of the formula for the sample standard deviation

Observation ( $y_i$ )	Deviation ( $y_i - \bar{y}$ )	Squared deviation ( $y_i - \bar{y}$ ) <sup>2</sup>
76	3	9
72	-1	1
65	-8	64
70	-3	9
82	9	81
Sum $365 = \sum_{i=1}^n y_i$	0	$164 = \sum_{i=1}^n (y_i - \bar{y})^2$

Since  $n = 5$ , the standard deviation is

$$\begin{aligned}s &= \sqrt{\frac{164}{4}} \\ &= \sqrt{41} \\ &= 6.4 \text{ mm}\end{aligned}$$

Note that the units of  $s$  (mm) are the same as the units of  $Y$ . This is because we have squared the deviations and then later taken the square root. ■

The sample **variance**, denoted by  $s^2$ , is simply the standard deviation squared: variance =  $s^2$ . Thus,  $s = \sqrt{\text{variance}}$ .

### Example 2.6.3

**Chrysanthemum Growth** The variance of the chrysanthemum growth data is

$$s^2 = 41 \text{ mm}^2$$

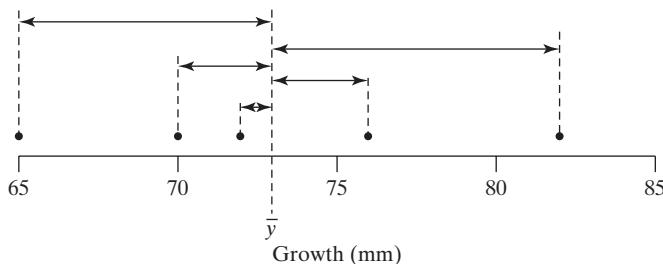
Note that the units of the variance ( $\text{mm}^2$ ) are not the same as the units of  $Y$ . ■

**An abbreviation** We will frequently abbreviate “standard deviation” as “SD”; the symbol “ $s$ ” will be used in formulas.

### Interpretation of the Definition of $s$

The magnitude (disregarding sign) of each deviation ( $y_i - \bar{y}$ ) can be interpreted as the *distance* of the corresponding observation from the sample mean  $\bar{y}$ . Figure 2.6.1 shows a plot of the chrysanthemum growth data (Example 2.6.2) with each distance marked.

**Figure 2.6.1** Plot of chrysanthemum growth data with deviations indicated as distances



From the formula for  $s$ , you can see that each deviation contributes to the SD. Thus, a sample of the same size but with less dispersion will have a smaller SD, as illustrated in the following example.

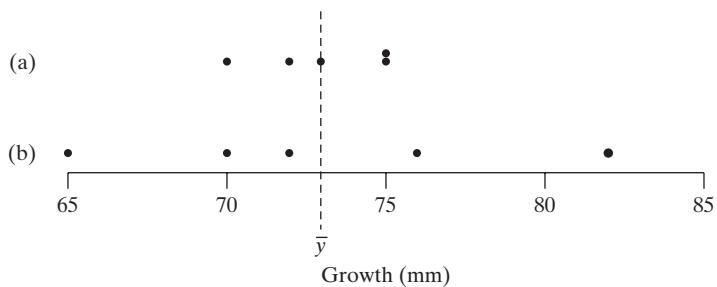
**Example  
2.6.4**

**Chrysanthemum Growth** If the chrysanthemum growth data of Example 2.6.2 are changed to

75 72 73 75 70

then the mean is the same ( $\bar{y} = 73$  mm), but the SD is smaller ( $s = 2.1$  mm), because the observations lie closer to the mean. The relative dispersion of the two samples can easily be seen from Figure 2.6.2. ■

**Figure 2.6.2** Two samples of chrysanthemum growth data with the same mean but different standard deviations: (a)  $s = 2.1$  mm; (b)  $s = 6.3$  mm



Let us look more closely at the way in which the deviations are combined to form the SD. The formula calls for dividing by  $(n - 1)$ . If the divisor were  $n$  instead of  $(n - 1)$ , then the quantity inside the square root sign would be the average (the mean) of the squared deviations. Unless  $n$  is very small, the inflation due to dividing by  $(n - 1)$  instead of  $n$  is not very great, so that the SD can be interpreted approximately as

$$s \approx \sqrt{\text{sample average value of } (y_i - \bar{y})^2}$$

Thus, it is roughly appropriate to think of the SD as a “typical” distance of the observations from their mean.

**Why  $n - 1$ ?** Since dividing by  $n$  seems more natural, you may wonder why the formula for the SD specifies dividing by  $(n - 1)$ . Note that the sum of the deviations  $y_i - \bar{y}$  is always zero. Thus, once the first  $n - 1$  deviations have been calculated, the last deviation is constrained. This means that in a sample with  $n$  observations there are only  $n - 1$  units of information concerning deviation from the average. The quantity  $n - 1$  is called the **degrees of freedom** of the standard deviation or variance. We can also give an intuitive justification of why  $n - 1$  is used by considering the extreme case when  $n = 1$ , as in the following example.

**Example  
2.6.5**

**Chrysanthemum Growth** Suppose the chrysanthemum growth experiment of Example 2.6.2 had included only one plant, so that the sample consisted of the single observation

73

For this sample,  $n = 1$  and  $\bar{y} = 73$ . However, the SD formula breaks down (giving  $\frac{0}{0}$ ), so the SD cannot be computed. This is reasonable, because the sample gives no information about variability in chrysanthemum growth under the experimental conditions. If the formula for the SD said to divide by  $n$ , we would obtain an SD of zero,

suggesting that there is little or no variability; such a conclusion hardly seems justified by observation of only one plant.

## The Coefficient of Variation

The **coefficient of variation** is the standard deviation expressed as a percentage of the mean: coefficient of variation =  $\frac{s}{\bar{y}} \times 100\%$ . Here is an example.

### Example 2.6.6

**Chrysanthemum Growth** For the chrysanthemum growth data of Example 2.6.2, we have  $\bar{y} = 73.0$  mm and  $s = 6.4$  mm. Thus,

$$\frac{s}{\bar{y}} \times 100\% = \frac{6.4}{73.0} \times 100\% = 0.088 \times 100\% = 8.8\%$$

The sample coefficient of variation is 8.8%. Thus, the standard deviation is 8.8% as large as the mean.

Note that the coefficient of variation is not affected by multiplicative changes of scale. For example, if the chrysanthemum data were expressed in inches instead of mm, then both  $\bar{y}$  and  $s$  would be in inches, and the coefficient of variation would be unchanged. Because of its imperviousness to scale change, the coefficient of variation is a useful measure for comparing the dispersions of two or more variables that are measured on different scales.

### Example 2.6.7

**Girls' Height and Weight** As part of the Berkeley Guidance Study,<sup>35</sup> the heights (in cm) and weights (in kg) of 13 girls were measured at age two. At age two, the average height was 86.6 cm and the SD was 2.9 cm. Thus, the coefficient of variation of height at age two is

$$\frac{s}{\bar{y}} \times 100\% = \frac{2.9}{86.6} \times 100\% = .033 \times 100\% = 3.3\%$$

For weight at age two the average was 12.6 kg and the SD was 1.4 kg. Thus, the coefficient of variation of weight at age two is

$$\frac{s}{\bar{y}} \times 100\% = \frac{1.4}{12.6} \times 100\% = .111 \times 100\% = 11.1\%$$

There is considerably more variability in weight than there is in height, when we express each measure of variability as a percentage of the mean. The SD of weight is a fairly large percentage of the average weight, but the SD of height is a rather small percentage of the average height.

## Visualizing Measures of Dispersion

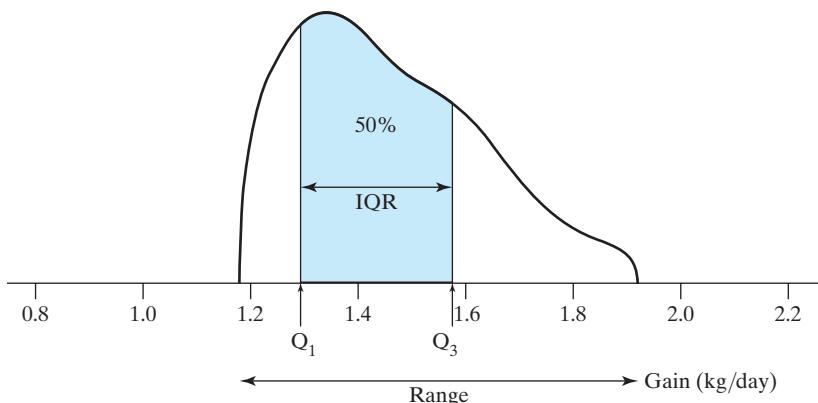
The range and the interquartile range are easy to interpret. The range is the spread of all the observations and the interquartile range is the spread of (roughly) the middle 50% of the observations. In terms of the histogram of a data set, the range can be visualized as (roughly) the width of the histogram. The quartiles are (roughly) the values that divide the area into four equal parts and the interquartile range is the distance between the first and third quartiles. The following example illustrates these ideas.

**Example  
2.6.8**

**Daily Gain of Cattle** The performance of beef cattle was evaluated by measuring their weight gain during a 140-day testing period on a standard diet. Table 2.6.2 gives the average daily gains (kg/day) for 39 bulls of the same breed (Charolais); the observations are listed in increasing order.<sup>36</sup> The values range from 1.18 kg/day to 1.92 kg/day. The quartiles are 1.29, 1.41, and 1.58 kg/day. Figure 2.6.3 shows a histogram of the data, the range, the quartiles, and the interquartile range (IQR). The shaded area represents the middle 50% (approximately) of the observations. ■

Table 2.6.2 Average daily gain (kg/day) of thirty-nine Charolais bulls							
1.18	1.24	1.29	1.37	1.41	1.51	1.58	1.72
1.20	1.26	1.33	1.37	1.41	1.53	1.59	1.76
1.23	1.27	1.34	1.38	1.44	1.55	1.64	1.83
1.23	1.29	1.36	1.40	1.48	1.57	1.64	1.92
1.23	1.29	1.36	1.41	1.50	1.58	1.64	1.65

**Figure 2.6.3** Smoothed histogram of 39 daily gain measurements, showing the range, the quartiles, and the interquartile range (IQR). The shaded area represents about 50% of the observations.



## Visualizing the Standard Deviation

We have seen that the SD is a combined measure of the distances of the observations from their mean. It is natural to ask how many of the observations are within  $\pm 1$  SD of the mean, within  $\pm 2$  SDs of the mean, and so on. The following example explores this question.

**Example  
2.6.9**

**Daily Gain of Cattle** For the daily-gain data of Example 2.6.8, the mean is  $\bar{y} = 1.445$  kg/day and the SD is  $s = 0.183$  kg/day. In Figure 2.6.4 the intervals  $\bar{y} \pm s$ ,  $\bar{y} \pm 2s$ , and  $\bar{y} \pm 3s$  have been marked on a histogram of the data. The interval  $\bar{y} \pm s$  is

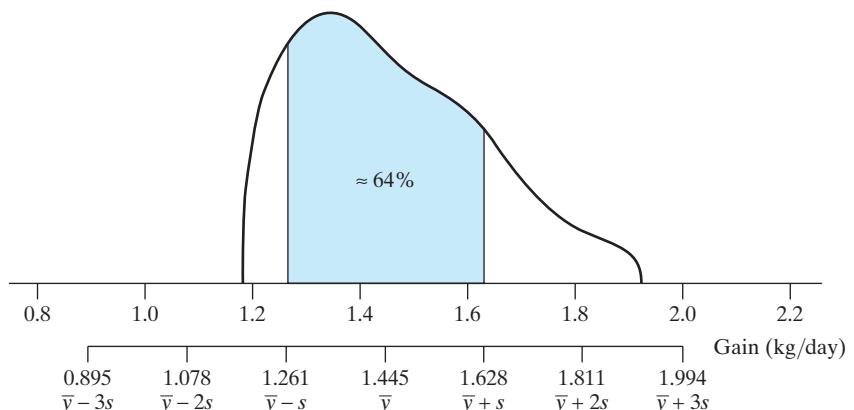
$$1.445 \pm 0.183 \text{ or } 1.262 \text{ to } 1.628$$

You can verify from Table 2.6.2 that this interval contains 25 of the 39 observations. Thus,  $\frac{25}{39}$  or 64% of the observations are within  $\pm 1$  SD of the mean; the corresponding area is shaded in Figure 2.6.4. The intervals  $\bar{y} \pm 2s$  is

$$1.445 \pm 0.366 \text{ or } 1.079 \text{ to } 1.811$$

This interval contains  $\frac{37}{39}$  or 95% of the observations. You may verify that the interval  $\bar{y} \pm 3s$  contains all the observations. ■

**Figure 2.6.4** Histogram of daily-gain data showing intervals 1, 2, and 3 standard deviations from the mean. The shaded area represents about 64% of the observations.



It turns out that the percentages found in Example 2.6.9 are fairly typical of distributions that are observed in the life sciences.

#### Typical Percentages: The Empirical Rule

For “nicely shaped” distributions—that is, unimodal distributions that are not too skewed and whose tails are not overly long or short—we usually expect to find

- about 68% of the observations within  $\pm 1$  SD of the mean.
- about 95% of the observations within  $\pm 2$  SDs of the mean.
- >99% of the observations within  $\pm 3$  SDs of the mean.

The typical percentages enable us to construct a rough mental image of a frequency distribution if we know just the mean and SD. (The value 68% may seem to come from nowhere. Its origin will become clear in Chapter 4.)

#### Estimating the SD from a Histogram

The empirical rule gives us a way to construct a rough mental image of a frequency distribution if we know just the mean and SD: We can envision a histogram centered at the mean and extending out a bit more than 2 SDs in either direction. Of course, the actual distribution might not be symmetric, but our rough mental image will often be fairly accurate.

Thinking about this the other way around, we can look at a histogram and estimate the SD. To do this, we need to estimate the endpoints of an interval that is centered at the mean and that contains about 95% of the data. The empirical rule implies that this interval is roughly the same as  $(\bar{y} - 2s, \bar{y} + 2s)$ , so the length of the interval should be about 4 times the SD:

$$(\bar{y} - 2s, \bar{y} + 2s) \text{ has length } 2s + 2s = 4s$$

This means

$$\text{length of interval} = 4s$$

so

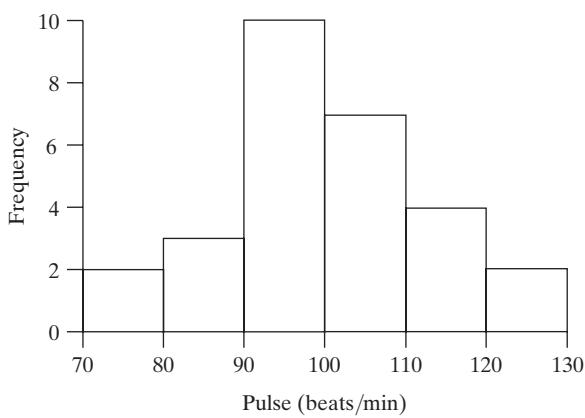
$$\text{estimate of } s = \frac{\text{length of interval}}{4}$$

Of course, our visual estimate of the interval that covers the middle 95% of the data could be off. Moreover, the empirical rule works best for distributions that are symmetric. Thus, this method of estimating the SD will give only a general estimate. The method works best when the distribution is fairly symmetric, but it works reasonably well even if the distribution is somewhat skewed.

**Example  
2.6.10**

**Pulse after Exercise** A group of 28 adults did some moderate exercise for five minutes and then measured their pulses. Figure 2.6.5 shows the distribution of the data.<sup>37</sup> We can see that about 95% of the observations are between about 75 and 125.\* Thus, an interval of length 50 (125 – 75) covers the middle 95% of the data. From this, we can estimate the SD to be  $\frac{50}{4} = 12.5$ . The actual SD is 13.4, which is not far off from our estimate. ■

**Figure 2.6.5** Pulse after moderate exercise for a group of adults



The typical percentages given by the empirical rule may be grossly wrong if the sample is small or if the shape of the frequency distribution is not “nice.” For instance, the cricket singing time data (Table 2.3.1 and Figure 2.3.4) has  $s = 4.4$  mm, and the interval  $\bar{y} \pm s$  contains 90% of the observations. This is much higher than the “typical” 68% because the SD has been inflated by the long straggly tail of the distribution.

## Comparison of Measures of Dispersion

The dispersion, or spread, of the data in a sample can be described by the standard deviation, the range, or the interquartile range. The range is simple to understand, but it can be a poor descriptive measure because it depends only on the extreme tails of the distribution. The interquartile range, by contrast, describes the spread in the central “body” of the distribution. The standard deviation takes account of all the observations and can roughly be interpreted in terms of the spread of the observations around their mean. However, the SD can be inflated by observations in the extreme tails. The interquartile range is a resistant measure, while the SD is nonresistant. Of course, the range is very highly nonresistant.

The descriptive interpretation of the SD is less straightforward than that of the range and the interquartile range. Nevertheless, the SD is the basis for most

\*It is difficult to visually assess exactly where the middle 95% of the data lay using a histogram, but as this is only a visual estimate, we need not concern ourselves with producing an exact value. Our visual estimates of the SD might differ from one another, but they should all be relatively close.

standard classical statistical methods. The SD enjoys this classic status for various technical reasons, including efficiency in certain situations.

The developments in later chapters will emphasize classical statistical methods, in which the mean and SD play a central role. Consequently, in this book we will rely primarily on the mean and SD rather than other descriptive measures.

## Exercises 2.6.1–2.6.16

**2.6.1** Calculate the standard deviation of each of the following fictitious samples:

- (a) 16, 13, 18, 13
- (b) 38, 30, 34, 38, 35
- (c) 1, -1, 5, -1
- (d) 4, 6, -1, 4, 2

**2.6.2** Calculate the standard deviation of each of the following fictitious samples:

- (a) 8, 6, 9, 4, 8
- (b) 4, 7, 5, 4
- (c) 9, 2, 6, 7, 6

**2.6.3**

- (a) Invent a sample of size 5 for which the deviations ( $y_i - \bar{y}$ ) are -3, -1, 0, 2, 2.
- (b) Compute the standard deviation of your sample.
- (c) Should everyone get the same answer for part (b)? Why?

**2.6.4** Four plots of land, each 346 square feet, were planted with the same variety ("Beau") of wheat. The plot yields (lb) were as follows:<sup>38</sup>

35.1 30.6 36.9 29.8

- (a) Calculate the mean and the standard deviation.
- (b) Calculate the coefficient of variation.

**2.6.5** A plant physiologist grew birch seedlings in the greenhouse and measured the ATP content of their roots. (See Example 1.1.3.) The results (nmol ATP/mg tissue) were as follows for four seedlings that had been handled identically.<sup>39</sup>

1.45 1.19 1.05 1.07

- (a) Calculate the mean and the standard deviation.
- (b) Calculate the coefficient of variation.

**2.6.6** Ten patients with high blood pressure participated in a study to evaluate the effectiveness of the drug Timolol in reducing their blood pressure. The accompanying table shows systolic blood pressure measurements taken before and after two weeks of treatment with Timolol.<sup>40</sup> Calculate the mean and standard deviation of the *change* in blood pressure (note that some values are negative).

PATIENT	BLOOD PRESSURE (mm HG)		
	BEFORE	AFTER	CHANGE
1	172	159	-13
2	186	157	-29
3	170	163	-7
4	205	207	2
5	174	164	-10
6	184	141	-43
7	178	182	4
8	156	171	15
9	190	177	-13
10	168	138	-30

**2.6.7** Dopamine is a chemical that plays a role in the transmission of signals in the brain. A pharmacologist measured the amount of dopamine in the brain of each of seven rats. The dopamine levels (nmoles/g) were as follows:<sup>41</sup>

6.8 5.3 6.0 5.9 6.8 7.4 6.2

- (a) Calculate the mean and standard deviation.
- (b) Determine the median and the interquartile range.
- (c) Calculate the coefficient of variation.
- (d) Replace the observation 7.4 by 10.4 and repeat parts (a) and (b). Which of the descriptive measures display resistance and which do not?

**2.6.8** In a study of the lizard *Sceloporus occidentalis*, biologists measured the distance (m) run in two minutes for each of 15 animals. The results (listed in increasing order) were as follows:<sup>42</sup>

18.4 22.2 24.5 26.4 27.5 28.7 30.6 32.9  
32.9 34.0 34.8 37.5 42.1 45.5 45.5

- (a) Determine the quartiles and the interquartile range.
- (b) Determine the range.

**2.6.9** Refer to the running-distance data of Exercise 2.6.8. The sample mean is 32.23 m and the SD is 8.07 m. What percentage of the observations are within

- (a) 1 SD of the mean?
- (b) 2 SDs of the mean?

**2.6.10** Compare the results of Exercise 2.6.9 with the predictions of the empirical rule.

**2.6.11** Listed in increasing order are the serum creatine phosphokinase (CK) levels (U/l) of 36 healthy men (these are the data of Example 2.2.6):

25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/l and the SD is 40.4 U/l. What percentage of the observations are within

- (a) 1 SD of the mean?
- (b) 2 SDs of the mean?
- (c) 3 SDs of the mean?

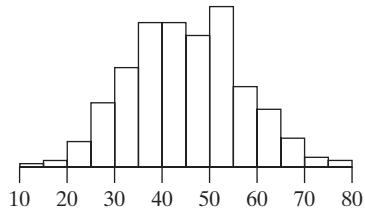
**2.6.12** Compare the results of Exercise 2.6.11 with the predictions of the empirical rule.

**2.6.13** The girls in the Berkeley Guidance Study (Example 2.6.7) who were measured at age two were measured again at age nine. Of course, the average height and weight were much greater at age nine than at age two. Likewise, the SDs of height and of weight were much greater at age nine, than they were at age two. But what

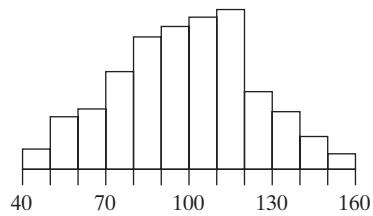
about the coefficient of variation of height and the coefficient of variation of weight? It turns out that one of these went up a moderate amount from age two to age nine, but for the other variable the increase in the coefficient of variation was fairly large. For which variable, height or weight, would you expect the coefficient of variation to change more between age two and age nine? Why? (*Hint:* Think about how genetic factors influence height and weight and how environmental factors influence height and weight.)

**2.6.14** Consider the 13 girls mentioned in Example 2.6.7. At age 18 their average height was 166.3 cm and the SD of their heights was 6.8 cm. Calculate the coefficient of variation.

**2.6.15** Here is a histogram. Estimate the mean and the SD of the distribution.



**2.6.16** Here is a histogram. Estimate the mean and the SD of the distribution.



## 2.7 Effect of Transformation of Variables (Optional)

Sometimes when we are working with a data set, we find it convenient to transform a variable. For example, we might convert from inches to centimeters or from °F to °C. Transformation, or reexpression, of a variable  $Y$  means replacing  $Y$  by a new variable, say  $Y'$ . To be more comfortable working with data, it is helpful to know how the features of a distribution are affected if the observed variable is transformed.

The simplest transformations are **linear** transformations, so called because a graph of  $Y$  against  $Y'$  would be a straight line. A familiar reason for linear transformation is a change in the scale of measurement, as illustrated in the following two examples.

**Example 2.7.1**

**Weight** Suppose  $Y$  represents the weight of an animal in kg, and we decide to reexpress the weight in lb. Then

$$\begin{aligned} Y &= \text{Weight in kg} \\ Y' &= \text{Weight in lb} \end{aligned}$$

so

$$Y' = 2.2Y$$

This is a **multiplicative** transformation, because  $Y'$  is calculated from  $Y$  by multiplying by the constant value 2.2. ■

**Example 2.7.2**

**Body Temperature** Measurements of basal body temperature (temperature on waking) were made on 47 women.<sup>43</sup>

Typical observations  $Y$ , in °C, were

$$Y: 36.23, 36.41, 36.77, 36.15, \dots$$

Suppose we convert these data from °C to °F, and call the new variable  $Y'$ :

$$Y': 97.21, 97.54, 98.19, 97.07, \dots$$

The relation between  $Y$  and  $Y'$  is

$$Y' = 1.8Y + 32$$

The combination of **additive** (+32) and multiplicative ( $\times 1.8$ ) changes indicates a linear relationship. ■

Another reason for linear transformation is **coding**, which means transforming the data for convenience in handling the numbers. The following is an example.

**Example 2.7.3**

**Body Temperature** Consider the temperature data of Example 2.7.2. If we subtract 36 from each observation, the data become

$$0.23, 0.41, 0.77, 0.15, \dots$$

This is additive coding, since we added a constant value (-36) to each observation. Now suppose we further transform the data to the form

$$23, 41, 77, 15, \dots$$

This step of the coding is multiplicative, since each observation is multiplied by a constant value (100). ■

As the foregoing examples illustrate, a linear transformation consists of (1) multiplying all the observations by a constant, or (2) adding a constant to all the observations, or (3) both.

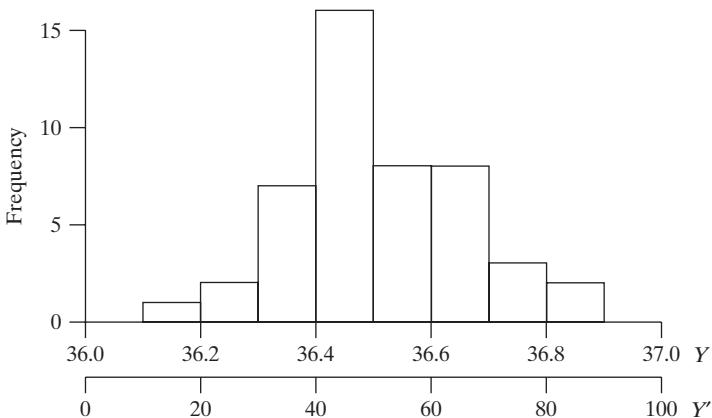
## How Linear Transformations Affect the Frequency Distribution

A linear transformation of the data does not change the essential shape of its frequency distribution; by suitably scaling the horizontal axis, you can make the transformed histogram identical to the original histogram. Example 2.7.4 illustrates this idea.

**Example  
2.7.4**

**Body Temperature** Figure 2.7.1 shows the distribution of 47 temperature measurements that have been transformed by first subtracting 36 from each observation and then multiplying by 100 (as in Examples 2.7.2 and 2.7.3). That is,  $Y' = (Y - 36) \times 100$ . The figure shows that the two distributions can be represented by the same histogram with different horizontal scales. ■

**Figure 2.7.1** Distribution of 47 temperature measurements showing original and linearly transformed scales



## How Linear Transformations Affect $\bar{y}$ and $s$

The effect of a linear transformation on  $\bar{y}$  is “natural”; that is, **under a linear transformation**,  $\bar{y}$  changes like  $Y$ . For instance, if temperatures are converted from  $^{\circ}\text{C}$  to  $^{\circ}\text{F}$ , then the mean is similarly converted:

$$Y' = 1.8Y + 32 \quad \text{so} \quad \bar{y}' = 1.8\bar{y} + 32$$

The effect of multiplying  $Y$  by a positive constant on  $s$  is “natural”; if  $Y' = c \times Y$ , with  $c > 0$ , then  $s' = c \times s$ . For instance, if weights are converted from kg to lb, the SD is similarly converted:  $s' = 2.2s$ . If  $Y' = c \times Y$  and  $c < 0$ , then  $s' = -c \times s$ . In general, if  $Y' = c \times Y$  then  $s' = |c| \times s$ .

However, an additive transformation does not affect  $s$ . If we add or subtract a constant, we do not change how spread out the distribution is, so  $s$  does not change. Thus, for example, we would *not* convert the SD of temperature data from  $^{\circ}\text{C}$  to  $^{\circ}\text{F}$  in the same way as we convert each observation; we would multiply the SD by 1.8, but we would *not* add 32. The fact that the SD is unchanged by additive transformation will appear less surprising if you recall (from the definition) that  $s$  depends only on the deviations  $(y_i - \bar{y})$ , and these are not changed by an additive transformation. The following example illustrates this idea.

**Example  
2.7.5**

**Additive Transformation** Consider a simple set of fictitious data, coded by subtracting 20 from each observation. The original and transformed observations are shown in Table 2.7.1.

The SD for the original observations is

$$\begin{aligned}s &= \sqrt{\frac{(-1)^2 + (0)^2 + (2)^2 + (-1)^2}{3}} \\ &= 1.4\end{aligned}$$

**Table 2.7.1** Effect of additive transformation

Original observations ( $y$ )	Deviations ( $y_i - \bar{y}$ )	Transformed observations ( $y'$ )	Deviations ( $y'_i - \bar{y}$ )
25	-1	5	-1
26	0	6	0
28	2	8	2
25	-1	5	-1
Mean	26	6	

Because the deviations are unaffected by the transformation, the SD for the transformed observations is the same:

$$s' = 1.4$$

An additive transformation effectively picks up the histogram of a distribution and moves it to the left or to the right on the number line. The shape of the histogram does not change and the deviations do not change, so the SD does not change. A multiplicative transformation, on the other hand, stretches or shrinks the distribution, so the SD gets larger or smaller accordingly.

**Other Statistics** Under linear transformations, other measures of center (for instance, the median) change like  $\bar{y}$ , and other measures of dispersion (for instance, the interquartile range) change like  $s$ . The quartiles themselves change like  $\bar{y}$ .

## Nonlinear Transformations

Data are sometimes reexpressed in a nonlinear way. Examples of nonlinear transformations are

$$Y' = \sqrt{Y}$$

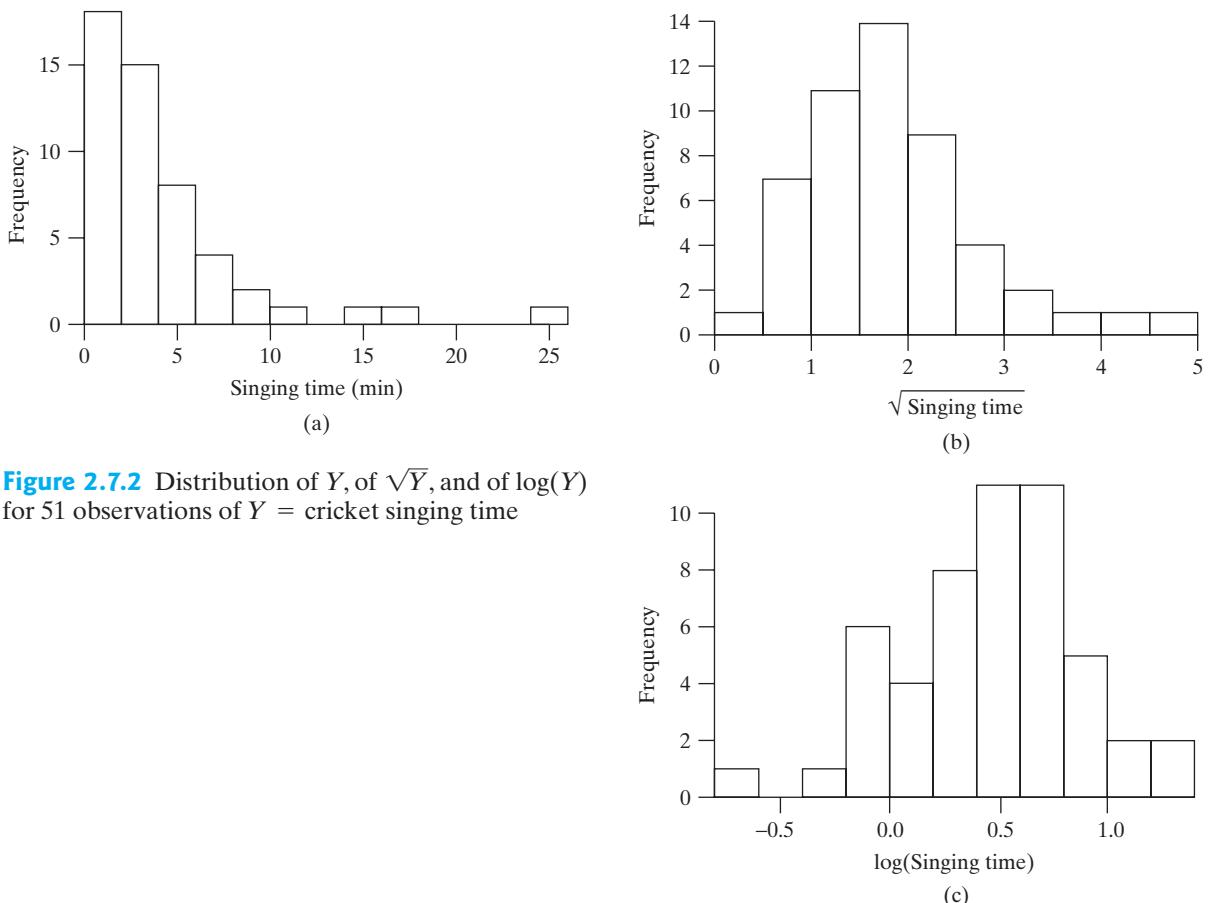
$$Y' = \log(Y)$$

$$Y' = \frac{1}{Y}$$

$$Y' = Y^2$$

These transformations are termed “nonlinear” because a graph of  $Y'$  against  $Y$  would be a curve rather than a straight line. Computers make it easy to use nonlinear transformations. The logarithmic transformation is especially common in biology because many important relationships can be simply expressed in terms of logs. For instance, there is a phase in the growth of a bacterial colony when  $\log(\text{colony size})$  increases at a constant rate with time. [Note that logarithms are used in some familiar scales of measurement, such as pH measurement or earthquake magnitude (Richter scale).]

Nonlinear transformations can affect data in complex ways. For example, the mean does not change “naturally” under a log transformation; the log of the mean is *not* the same as the mean of the logs. Furthermore, nonlinear transformations (unlike linear ones) *do* change the essential shape of a frequency distribution.



**Figure 2.7.2** Distribution of  $Y$ , of  $\sqrt{Y}$ , and of  $\log(Y)$  for 51 observations of  $Y$  = cricket singing time

In future chapters we will see that if a distribution is skewed to the right, such as the cricket singing-time distribution shown in Figure 2.7.2, then we may wish to apply a transformation that makes the distribution more symmetric, by pulling in the right-hand tail. Using  $Y' = \sqrt{Y}$  will pull in the right-hand tail of a distribution and push out the left-hand tail. The transformation  $Y' = \log(Y)$  is more severe than  $\sqrt{Y}$  in this regard. The following example shows the effect of these transformations.

### Example 2.7.6

**Cricket Singing Times** Figure 2.7.2(a) shows the distribution of the cricket singing-time data of Table 2.3.1. If we transform these data by taking square roots, the transformed data have the distribution shown in Figure 2.7.2(b). Taking logs (base 10) yields the distribution shown in Figure 2.7.2(c). Notice that the transformations have the effect of “pulling in” the straggly upper tail and “stretching out” the clumped values on the lower end of the original distribution. ■

## Exercises 2.7.1–2.7.6

**2.7.1** A biologist made a certain pH measurement in each of 24 frogs; typical values were<sup>44</sup>

$$7.43, \quad 7.16, \quad 7.51, \dots$$

She calculated a mean of 7.373 and a standard deviation of 0.129 for these original pH measurements. Next, she

transformed the data by subtracting 7 from each observation and then multiplying by 100. For example, 7.43 was transformed to 43. The transformed data are

$$43, \quad 16, \quad 51, \quad \dots$$

What are the mean and standard deviation of the transformed data?

**2.7.2** The mean and SD of a set of 47 body temperature measurements were as follows:<sup>45</sup>

$$\bar{y} = 36.497 \text{ } ^\circ\text{C} \quad s = 0.172 \text{ } ^\circ\text{C}$$

If the 47 measurements were converted to  $^\circ\text{F}$ ,

- (a) What would be the new mean and SD?
- (b) What would be the new coefficient of variation?

**2.7.3** A researcher measured the average daily gains (in kg/day) of 20 beef cattle; typical values were<sup>46</sup>

$$1.39, \quad 1.57, \quad 1.44, \quad \dots$$

The mean of the data was 1.461 and the standard deviation was 0.178.

- (a) Express the mean and standard deviation in lb/day.  
(Hint: 1 kg = 2.20 lb.)
- (b) Calculate the coefficient of variation when the data are expressed (i) in kg/day; (ii) in lb/day.

**2.7.4** Consider the data from Exercise 2.7.3. The mean and SD were 1.461 and 0.178. Suppose we transformed the data from

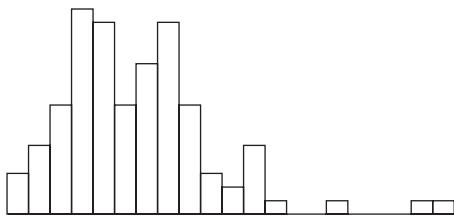
$$1.39, \quad 1.57, \quad 1.44, \quad \dots$$

to

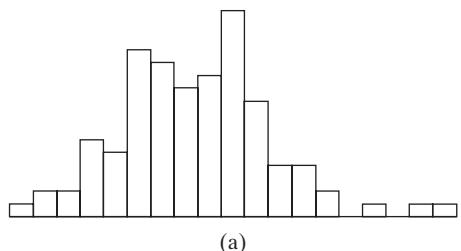
$$39, \quad 57, \quad 44, \quad \dots$$

What would be the mean and standard deviation of the transformed data?

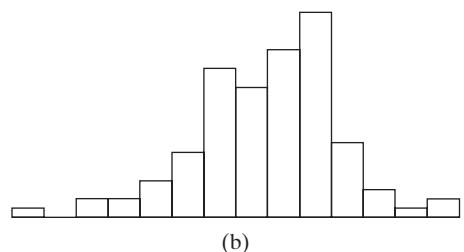
**2.7.5** The following histogram shows the distribution for a sample of data:



One of the following histograms is the result of applying a square root transformation and the other is the result of applying a log transformation. Which is which? How do you know?



(a)



(b)

**2.7.6 (Computer problem)** The file ‘Exer2.7.6.csv’ is included on the data disk packaged with this text. This file contains 36 observations on the number of dendritic branch segments emanating from nerve cells taken from the brains of newborn guinea pigs. (These data were used in Exercise 2.2.4.) Open the file and enter the data into a statistics package. Make a histogram of the data, which are skewed to the right. Now consider the following possible transformations:  $\text{sqrt}(Y)$ ,  $\log(Y)$ , and  $1/\text{sqrt}(Y)$ . Which of these transformations does the best job of meeting the goal of making the resulting distribution reasonably symmetric?

## 2.8 Statistical Inference

The description of a data set is sometimes of interest for its own sake. Usually, however, the researcher hopes to generalize, to extend the findings beyond the limited scope of the particular group of animals, plants, or other units that were actually observed. Statistical theory provides a rational basis for this process of generalization, building on the random sampling model from Section 1.3 and taking into account the variability of the data. The key idea of the statistical approach is to view the particular data in a study as a sample from a larger population; the population is the real focus of scientific and/or practical interest. The following example illustrates this idea.

**Example  
2.8.1**

**Blood Types** In an early study of the ABO blood-typing system, researchers determined blood types of 3,696 persons in England. The results are given in Table 2.8.1.<sup>47</sup>

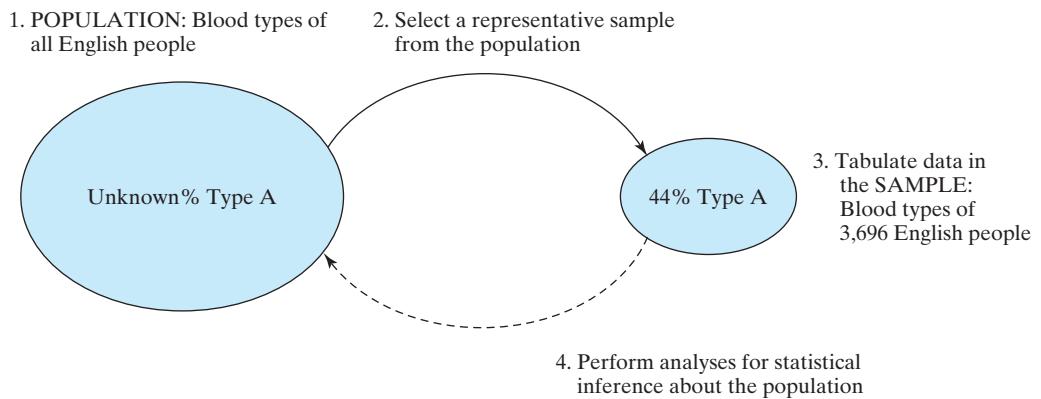
<b>Table 2.8.1</b> Blood types of 3,696 persons	
Blood type	Frequency
A	1,634
B	327
AB	119
O	1,616
Total	3,696

These data were not collected for the purpose of learning about the blood types of those particular 3,696 people. Rather, they were collected for their scientific value as a source of information about the distribution of blood types in a larger population. For instance, one might presume that the blood type distribution of all English people should resemble the distribution for these 3,696 people. In particular, the observed relative frequency of type A blood was

$$\frac{1634}{3696} \text{ or } 44\% \text{ type A}$$

One might conclude from this that approximately 44% of the people in England have type A blood. ■

The process of drawing conclusions about a population, based on observations in a sample from that population, is called **statistical inference**. For instance, in Example 2.8.1 the conclusion that approximately 44% of the people in England have type A blood would be a statistical inference. The inference is shown schematically in Figure 2.8.1. Of course, such an inference might be entirely wrong—perhaps the 3,696 people are not at all representative of English people in general. We might be worried about two possible sources of difficulty: (1) the 3,696 people might have been selected in a way that was systematically biased for (or against) type A people, and (2) the number of people examined might have been too small to permit generalization to a population of many millions. In general, it turns out that the population size being in the millions is *not* a problem, but bias in the way people are selected is a big concern.



**Figure 2.8.1** Schematic representation of inference from sample to population regarding prevalence of blood type A

In making a statistical inference, we hope that the sample resembles the population closely—that the sample is *representative* of the population. In Section 1.3 we saw how sampling errors and nonsampling errors can lead to nonrepresentative samples. However, even in the absence of bias we must ask how likely it is that a particular sample will provide a good representation of the population. The important question is: *How representative (of the population) is a sample likely to be?* We will see in Chapter 5 how statistical theory can help to answer this question.

## Specifying the Population

In Section 1.3 we emphasized that the collection of individuals that comprise a sample should be representative of the population. In fact, this requirement is a bit stronger than what is actually necessary. Ultimately, what matters is that the measurements that we obtain on the variable of interest are representative of the values present in the population. The following provides an example of a case where the sample members might not be representative of the population, but one could argue that the measurements taken from this sample could be viewed as representative of the larger population.

---

### Example 2.8.2

**Blood Types** How were the 3,696 English people of Example 2.8.1 actually chosen? It appears from the original paper that this was a “sample of convenience,” that is, friends of the investigators, employees, and sundry unspecified sources. There is little basis for believing that the *people* themselves would be representative of the entire English population. Nevertheless, one might argue that their *blood types* might be (more or less) representative of the population. The argument would be that the biases that entered into the selection of those particular people were probably not related to blood type. [Nonetheless, an objection to this argument might be made on the basis of race. For example, the racial distribution of the sample could differ substantially from the racial distribution of England (the population) and there are known differences in blood type distributions among races.] The argument for representativeness would be much less plausible if the observed variable were blood pressure rather than blood type; we know that blood pressure tends to increase with age, and the selection procedure was undoubtedly biased against certain age groups (for example, elderly people). ■

As Example 2.8.2 shows, whether the measurements obtained from a sample are likely to be representative of the measurements from a population depends not only on how the observational units (in this case people) were chosen, but also on the variable that was observed. Ideally we would always work with random samples, but we have noted that in some instances random samples are not possible or convenient. However, by turning our attention to the measurements themselves rather than the individuals from which they came, we can often make an argument for the generalizability (or lack of generalizability) of our results to a larger population. We do this by thinking of the population as consisting of observations or a collection of values from a measurement process, rather than of people or other observational units. The following is another example.

---

### Example 2.8.3

**Alcohol and MOPEG** The biochemical MOPEG plays a role in brain function. Seven healthy male volunteers participated in a study to determine whether drinking alcohol might elevate the concentration of MOPEG in the cerebrospinal fluid. The MOPEG concentration was measured twice for each man—once at the start of the experiment, and again after he drank 80 gm of ethanol. The results (in pmol/ml) are given in Table 2.8.2.<sup>48</sup>

Let us focus on the rightmost column, which shows the change in MOPEG concentration (that is, the difference between the “after” and the “before” measurements). In thinking of these values as a sample from a population, we need to specify all the details of the experimental conditions—how the cerebrospinal specimens were obtained, the exact timing of the measurements and the alcohol

**Table 2.8.2** Effect of alcohol on MOPEG

Volunteer	MOPEG concentration		
	Before	After	Change
1	46	56	10
2	47	52	5
3	41	47	6
4	45	48	3
5	37	37	0
6	48	51	3
7	58	62	4

consumption, and so on—as well as relevant characteristics of the volunteers themselves. Thus, the definition of the population might be something like this:

**Population** Change in cerebrospinal MOPEG concentration in healthy young men when measured before and after drinking 80 gm of ethanol, both measurements being made at 8:00 A.M., . . . (other relevant experimental conditions are specified here).

There is no single “correct” definition of a population for an experiment like this. A scientist reading a report of the experiment might find this definition too narrow (for instance, perhaps it does not matter that the volunteers were measured at 8:00 A.M.) or too broad. She might use her knowledge of alcohol and brain chemistry to formulate her own definition, and she would then use that definition as a basis for interpreting these seven observations. ■

## Describing a Population

Because observations are made only on a sample, characteristics of biological populations are almost never known exactly. Typically, our knowledge of a population characteristic comes from a sample. In statistical language, we say that the sample characteristic is an estimate of the corresponding population characteristic. Thus, estimation is a type of statistical inference.

Just as each sample has a distribution, a mean, and an SD, so also we can envision a population distribution, a population mean, and a population SD. In order to discuss inference from a sample to a population, we will need a language for describing the population. This language parallels the language that describes the sample. A sample characteristic is called a **statistic**; a population characteristic is called a **parameter**.

## Proportions

For a categorical variable, we can describe a population by simply stating the proportion, or relative frequency, of the population in each category. The following is a simple example.

---

**Example**  
**2.8.4**

**Oat Plants** In a certain population of oat plants, resistance to crown rust disease is distributed as shown in Table 2.8.3.<sup>49</sup> ■

Table 2.8.3 Disease resistance in oats	
Resistance	Proportion of plants
Resistant	0.47
Intermediate	0.43
Susceptible	0.10
Total	1.00

**Remark** The population described in Example 2.8.4 is realistic, but it is not a specific real population; the exact proportions for any real population are not known. For similar reasons, we will use fictitious but realistic populations in several other examples, here and in Chapters 3, 4, and 5.

For categorical data, the sample proportion of a category is an estimate of the corresponding population proportion. Because these two proportions are not necessarily the same, it is essential to have a notation that distinguishes between them. We denote the population proportion of a category by  $p$  and the sample proportion by  $\hat{p}$  (read “ $p$ -hat”):

$$\begin{aligned} p &= \text{Population proportion} \\ \hat{p} &= \text{Sample proportion} \end{aligned}$$

The symbol “ $\hat{\phantom{x}}$ ” can be interpreted as “estimate of.” Thus,

$$\hat{p} \text{ is an estimate of } p.$$

We illustrate this notation with an example.

---

**Example  
2.8.5**

**Lung Cancer** Eleven patients suffering from adenocarcinoma (a type of lung cancer) were treated with the chemotherapeutic agent Mitomycin. Three of the patients showed a positive response (defined as shrinkage of the tumor by at least 50%).<sup>50</sup> Suppose we define the population for this study as “responses of all adenocarcinoma patients.” Then we can represent the sample and population proportions of the category “positive response” as follows:

$p$  = Proportion of positive responders among all adenocarcinoma patients

$\hat{p}$  = Proportion of positive responders among the 11 patients in the study

$$\hat{p} = \frac{3}{11} = 0.27$$

Note that  $p$  is unknown, and  $\hat{p}$ , which is known, is an estimate of  $p$ . ■

We should emphasize that an “estimate,” as we are using the term, may or may not be a *good* estimate. For instance, the estimate  $\hat{p}$  in Example 2.8.5 is based on very few patients; estimates based on a small number of observations are subject to considerable uncertainty. Of course, the question of whether an estimation procedure is good or poor is an important one, and we will show in later chapters how this question can be answered.

## Other Descriptive Measures

If the observed variable is quantitative, one can consider descriptive measures other than proportions—the mean, the quartiles, the SD, and so on. Each of these quantities can be computed for a sample of data, and each is an estimate of its corresponding

population analog. For instance, the sample median is an estimate of the population median. In later chapters, we will focus especially on the mean and the SD, and so we will need a special notation for the population mean and SD. **The population mean is denoted by  $\mu$  (mu), and the population SD is denoted by  $\sigma$  (sigma).** We may define these as follows for a quantitative variable  $Y$ :

$$\mu = \text{Population average value of } Y$$

$$\sigma = \sqrt{\text{Population average value of } (Y - \mu)^2}$$

The following example illustrates this notation.

**Example  
2.8.6**

**Tobacco Leaves** An agronomist counted the number of leaves on each of 150 tobacco plants of the same strain (Havana). The results are shown in Table 2.8.4.<sup>51</sup>

The sample mean is

$$\bar{y} = 19.78 = \text{Mean number of leaves on the 150 plants}$$

**Table 2.8.4** Number of leaves on tobacco plants

Number of leaves	Frequency (number of plants)
17	3
18	22
19	44
20	42
21	22
22	10
23	6
24	1
Total	150

The population mean is

$$\mu = \text{Mean number of leaves on Havana tobacco plants grown under these conditions}$$

We do not know  $\mu$ , but we can regard  $\bar{y} = 19.78$  as an estimate of  $\mu$ . The sample SD is

$$s = 1.38 = \text{SD of number of leaves on the 150 plants}$$

The population SD is

$$\sigma = \text{SD of number of leaves on Havana tobacco plants grown under these conditions}$$

We do not know  $\sigma$ , but we can regard  $s = 1.38$  as an estimate of  $\sigma$ .\*

---

\*You may wonder why we use  $\bar{y}$  and  $s$  instead of  $\hat{\mu}$  and  $\hat{\sigma}$ . One answer is tradition. Another answer is that since “ $\wedge$ ” means estimate, you might have other estimates in mind.

## 2.9 Perspective

In this chapter we have considered various ways of describing a set of data. We have also introduced the notion of regarding features of a sample as estimates of corresponding features of a suitably defined population.

### Parameters and Statistics

Some features of a distribution—for instance, the mean—can be represented by a single number, while some—for instance, the shape—cannot. We have noted that a numerical measure that describes a sample is called a statistic. Correspondingly, a numerical measure that describes a population is called a parameter. For the most important numerical measures, we have defined notations to distinguish between the statistic and the parameter. These notations are summarized in Table 2.9.1 for convenient reference.

**Table 2.9.1** Notation for some important statistics and parameters

Measure	Sample value (statistic)	Population value (parameter)
Proportion	$\hat{p}$	$p$
Mean	$\bar{y}$	$\mu$
Standard deviation	$s$	$\sigma$

### A Look Ahead

It is natural to view a sample characteristic (for instance,  $\bar{y}$ ) as an estimate of the corresponding population characteristic (for instance,  $\mu$ ). But in taking such a view, one must guard against unjustified optimism. Of course, if the sample were perfectly representative of the population, then the estimate would be perfectly accurate. But this raises the central question: How representative (of the population) is a sample likely to be? Intuition suggests that, if the observational units are appropriately selected, then the sample should be more or less representative of the population. Intuition also suggests that larger samples should tend to be more representative than smaller samples. These intuitions are basically correct, but they are too vague to provide practical guidance for research in the life sciences. Practical questions that need to be answered are

1. How can an investigator judge whether a sample can be viewed as “more or less” representative of a population?
2. How can an investigator quantify “more or less” in a specific case?

In Section 1.3 we described a theoretical probability model based on random sampling that provides a framework for the judgment in question (1), and in Chapter 6 we will see how this model can provide a concrete answer to question (2). Specifically, in Chapter 6 we will see how to analyze a set of data so as to quantify how closely the sample mean ( $\bar{y}$ ) estimates the population mean ( $\mu$ ). But before returning to data analysis in Chapter 6, we will need to lay some groundwork in Chapters 3, 4, and 5; the developments in these chapters are an essential prelude to understanding the techniques of statistical inference.

## Supplementary Exercises 2.S.1–2.S.20

**2.S.1** A sample of four students had the following heights (in cm): 180, 182, 179, 176. Suppose a fifth student were added to the group. How tall would that student have to be to make the mean height of the group equal 181?

**2.S.2** A botanist grew 15 pepper plants on the same greenhouse bench. After 21 days, she measured the total stem length (cm) of each plant, and obtained the following values:<sup>52</sup>

12.4	12.2	13.4
10.9	12.2	12.1
11.8	13.5	12.0
14.1	12.7	13.2
12.6	11.9	13.1

- (a) Construct a dotplot for these data, and mark the positions of the quartiles.
- (b) Calculate the interquartile range.

**2.S.3** In a behavioral study of the fruitfly *Drosophila melanogaster*, a biologist measured, for individual flies, the total time spent preening during a six-minute observation period. The following are the preening times (sec) for 20 flies:<sup>53</sup>

34	24	10	16	52
76	33	31	46	24
18	26	57	32	25
48	22	48	29	19

- (a) Determine the median and the quartiles.
- (b) Determine the interquartile range.
- (c) Construct a (modified) boxplot of the data.

**2.S.4** To calibrate a standard curve for assaying protein concentrations, a plant pathologist used a spectrophotometer to measure the absorbance of light (wavelength 500 nm) by a protein solution. The results of 27 replicate assays of a standard solution containing 60 µg protein per ml water were as follows:<sup>54</sup>

0.111	0.115	0.115	0.110	0.099
0.121	0.107	0.107	0.100	0.110
0.106	0.116	0.098	0.116	0.108
0.098	0.120	0.123	0.124	0.122
0.116	0.130	0.114	0.100	0.123
0.119	0.107			

Construct a frequency distribution and display it as a table and as a histogram.

**2.S.5** Refer to the absorbance data of Exercise 2.S.4.

- (a) Determine the median, the quartiles, and the interquartile range.
- (b) How large must an observation be to be an outlier?

**2.S.6** The midrange is defined as the average of the minimum and maximum of a distribution. Is the midrange a robust statistic? Why or why not?

**2.S.7** Twenty patients with severe epilepsy were observed for eight weeks. The following are the numbers of major seizures suffered by each patient during the observation period:<sup>55</sup>

5	0	9	6	0	0	5	0	6	1
5	0	0	0	7	0	0	4	7	

- (a) Determine the median number of seizures.
- (b) Determine the mean number of seizures.
- (c) Construct a histogram of the data. Mark the positions of the mean and the median on the histogram.
- (d) What feature of the frequency distribution suggests that neither the mean nor the median is a meaningful summary of the experience of these patients?

**2.S.8** Calculate the standard deviation of each of the following fictitious samples:

- (a) 11, 8, 4, 10, 7
- (b) 23, 29, 24, 21, 23
- (c) 6, 0, -3, 2, 5

**2.S.9** To study the spatial distribution of Japanese beetle larvae in the soil, researchers divided a 12- × 12-foot section of a cornfield into 144 one-foot squares. They counted the number of larvae  $Y$  in each square, with the results shown in the following table.<sup>56</sup>

NUMBER OF LARVAE	FREQUENCY (NUMBER OF SQUARES)
0	13
1	34
2	50
3	18
4	16
5	10
6	2
7	1
Total	144

- (a) The mean and standard deviation of  $Y$  are  $\bar{y} = 2.23$  and  $s = 1.47$ . What percentage of the observations are within

- (i) 1 standard deviation of the mean?
- (ii) 2 standard deviations of the mean?
- (b) Determine the total number of larvae in all 144 squares. How is this number related to  $\bar{y}$ ?
- (c) Determine the median value of the distribution.

**2.S.10** One measure of physical fitness is maximal oxygen uptake, which is the maximum rate at which a person can consume oxygen. A treadmill test was used to determine the maximal oxygen uptake of nine college women before and after participation in a 10-week program of vigorous exercise. The accompanying table shows the before and after measurements and the change (after-before); all values are in ml O<sub>2</sub> per mm per kg body weight.<sup>57</sup>

PARTICIPANT	MAXIMAL OXYGEN UPTAKE		
	BEFORE	AFTER	CHANGE
1	48.6	38.8	-9.8
2	38.0	40.7	2.7
3	31.2	32.0	0.8
4	45.5	45.4	-0.1
5	41.7	43.2	1.5
6	41.8	45.3	3.5
7	37.9	38.9	1.0
8	39.2	43.5	4.3
9	47.2	45.0	-2.2

The following computations are to be done on the *change* in maximal oxygen uptake (the right-hand column).

- (a) Calculate the mean and the standard deviation.
- (b) Determine the median.
- (c) Eliminate participant 1 from the data and repeat parts (a) and (b). Which of the descriptive measures display resistance and which do not?

**2.S.11** A veterinary anatomist investigated the spatial arrangement of the nerve cells in the intestine of a pony. He removed a block of tissue from the intestinal wall, cut the block into many equal sections, and counted the number of nerve cells in each of 23 randomly selected sections. The counts were as follows.<sup>58</sup>

35	19	33	34	17	26	16	40
28	30	23	12	27	33	22	31
28	28	35	23	23	19	29	

- (a) Determine the median, the quartiles, and the interquartile range.
- (b) Construct a boxplot of the data.

**2.S.12** Exercise 2.S.11 asks for a boxplot of the nerve-cell data. Does this graphic support the claim that the data came from a reasonably symmetric distribution?

**2.S.13** A geneticist counted the number of bristles on a certain region of the abdomen of the fruitfly *Drosophila melanogaster*. The results for 119 individuals were as shown in the table.<sup>59</sup>

NUMBER OF BRISTLES	NUMBER OF FLIES	NUMBER OF BRISTLES	NUMBER OF FLIES
29	1	38	18
30	0	39	13
31	1	40	10
32	2	41	15
33	2	42	10
34	6	43	2
35	9	44	2
36	11	45	3
37	12	46	2

- (a) Find the median number of bristles.
- (b) Find the first and third quartiles of the sample.
- (c) Make a boxplot of the data.
- (d) The sample mean is 38.45 and the standard deviation is 3.20. What percentage of the observations fall within 1 standard deviation of the mean?

**2.S.14** The carbon monoxide in cigarettes is thought to be hazardous to the fetus of a pregnant woman who smokes. In a study of this hypothesis, blood was drawn from pregnant women before and after smoking a cigarette. Measurements were made of the percent of blood hemoglobin bound to carbon monoxide as carboxyhemoglobin (COHb). The results for 10 women are shown in the table.<sup>60</sup>

SUBJECT	BLOOD COHB (%)		
	BEFORE	AFTER	INCREASE
1	1.2	7.6	6.4
2	1.4	4.0	2.6
3	1.5	5.0	3.5
4	2.4	6.3	3.9
5	3.6	5.8	2.2
6	0.5	6.0	5.5
7	2.0	6.4	4.4
8	1.5	5.0	3.5
9	1.0	4.2	3.2
10	1.7	5.2	3.5

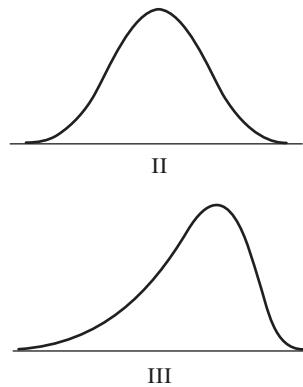
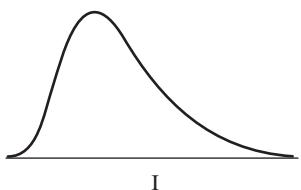
- Calculate the mean and standard deviation of the increase in COHb.
- Calculate the mean COHb before and the mean after. Is the mean increase equal to the increase in means?
- Determine the median increase in COHb.
- Repeat part (c) for the before measurements and for the after measurements. Is the median increase equal to the increase in medians?

**2.S.15 (Computer problem)** A medical researcher in India obtained blood specimens from 31 young children, all of whom were infected with malaria. The following data, listed in increasing order, are the numbers of malarial parasites found in 1 ml of blood from each child.<sup>61</sup>

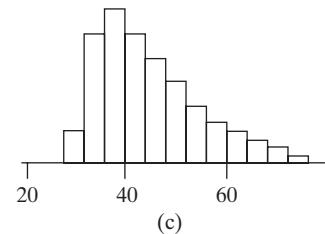
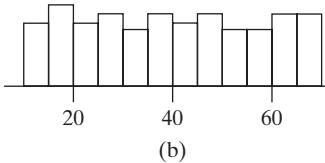
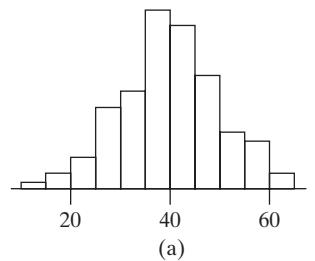
100	140	140	271	400	435	455	770
826	1,400	1,540	1,640	1,920	2,280	2,340	3,672
4,914	6,160	6,560	6,741	7,609	8,547	9,560	10,516
14,960	16,855	18,600	22,995	29,800	83,200	134,232	

- Construct a frequency distribution of the data, using a class width of 10,000; display the distribution as a histogram.
- Transform the data by taking the logarithm (base 10) of each observation. Construct a frequency distribution of the transformed data and display it as a histogram. How does the log transformation affect the shape of the frequency distribution?
- Determine the mean of the original data and the mean of the log-transformed data. Is the mean of the logs equal to the log of the mean?
- Determine the median of the original data and the median of the log-transformed data. Is the median of the logs equal to the log of the median?

**2.S.16** Rainfall, measured in inches, for the month of June in Cleveland, Ohio, was recorded for each of 41 years.<sup>62</sup> The values had a minimum of 1.2, an average of 3.6, and a standard deviation of 1.6. Which of the following is a rough histogram for the data? How do you know?



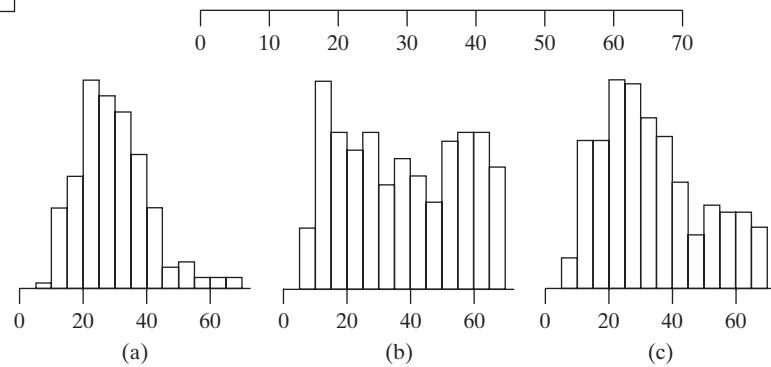
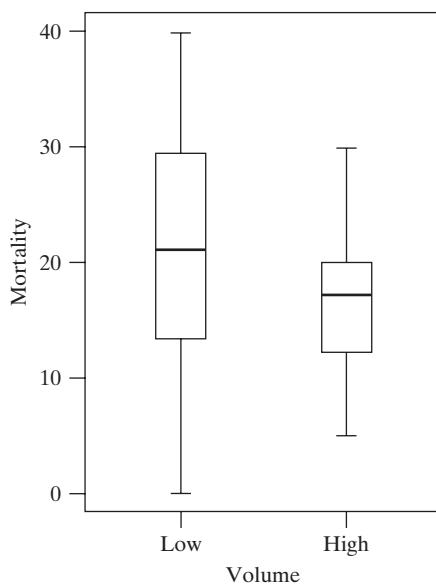
**2.S.17** The following histograms (a), (b), and (c) show three distributions.



The accompanying computer output shows the mean, median, and standard deviation of the three distributions, plus the mean, median, and standard deviation for a fourth distribution. Match the histograms with the statistics. Explain your reasoning. (One set of statistics will not be used.)

1. Count	100	2. Count	100
Mean	41.3522	Mean	39.6761
Median	39.5585	Median	39.5377
StdDev	13.0136	StdDev	10.0476
3. Count	100	4. Count	100
Mean	37.7522	Mean	39.6493
Median	39.5585	Median	39.5448
StdDev	13.0136	StdDev	17.5126

**2.S.18** The following boxplots show mortality rates (deaths within one year per 100 patients) for heart transplant patients at various hospitals. The low-volume hospitals are those that perform between 5 and 9 transplants per year. The high-volume hospitals perform 10 or more transplants per year.<sup>63</sup> Describe the distributions, paying special attention to how they compare to one another. Be sure to note the shape, center, and spread of each distribution.

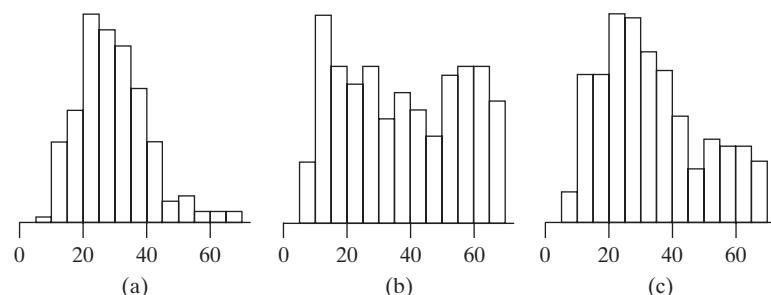


**2.S.19 (Computer problem)** Physicians measured the concentration of calcium (nM) in blood samples from 38 healthy persons. The data are listed as follows:<sup>64</sup>

95	110	135	120	88	125
112	100	130	107	86	130
122	122	127	107	107	107
88	126	125	112	78	115
78	102	103	93	88	110
104	122	112	80	121	126
90	96				

Calculate appropriate measures of the center and spread of the distribution. Describe the shape of the distribution and any unusual features in the data.

**2.S.20** The following boxplot shows the same data that are shown in one of the three histograms. Which histogram goes with the boxplot? Explain your answer.



# PROBABILITY AND THE BINOMIAL DISTRIBUTION

## Objectives

In this chapter we will study the basic ideas of probability, including

- the “limiting frequency” definition of probability.
- the use of probability trees.
- the concept of a random variable.
- rules for finding means and standard deviations of random variables.
- the use of the binomial distribution.

### 3.1 Probability and the Life Sciences

Probability, or chance, plays an important role in scientific thinking about living systems. Some biological processes are affected directly by chance. A familiar example is the segregation of chromosomes in the formation of gametes; another example is the occurrence of mutations.

Even when the biological process itself does not involve chance, the results of an experiment are always somewhat affected by chance: chance fluctuations in environmental conditions, chance variation in the genetic makeup of experimental animals, and so on. Often, chance also enters directly through the design of an experiment; for instance, varieties of wheat may be randomly allocated to plots in a field. (Random allocation will be discussed in Chapter 11.)

The conclusions of a statistical data analysis are often stated in terms of probability. Probability enters statistical analysis not only because chance influences the results of an experiment, but also because probability models allow us to quantify how likely, or unlikely, an experimental result is, given certain modeling assumptions. In this chapter we will introduce the language of probability and develop some simple tools for manipulating probabilities.

### 3.2 Introduction to Probability

In this section we introduce the language of probability and its interpretation.

#### Basic Concepts

A **probability** is a numerical quantity that expresses the likelihood of an event. The probability of an event  $E$  is written as

$$\Pr\{E\}$$

The probability  $\Pr\{E\}$  is always a number between 0 and 1, inclusive.

We can speak meaningfully about a probability  $\Pr\{E\}$  only in the context of a chance operation—that is, an operation whose outcome is determined at least partially by chance. The chance operation must be defined in such a way that *each time the chance operation is performed, the event E either occurs or does not occur*. The following two examples illustrate these ideas.

---

**Example  
3.2.1**

**Coin Tossing** Consider the familiar chance operation of tossing a coin, and define the event

$$E: \text{Heads}$$

Each time the coin is tossed, either it falls heads or it does not. If the coin is equally likely to fall heads or tails, then

$$\Pr\{E\} = \frac{1}{2} = 0.5$$

Such an ideal coin is called a “fair” coin. If the coin is not fair (perhaps because it is slightly bent), then  $\Pr\{E\}$  will be some value other than 0.5, for instance,

$$\Pr\{E\} = 0.6$$

■

---

**Example  
3.2.2**

**Coin Tossing** Consider the event

$$E: 3 \text{ heads in a row}$$

The chance operation “toss a coin” is *not* adequate for this event, because we cannot tell from one toss whether  $E$  has occurred. A chance operation that would be adequate is

*Chance operation:* Toss a coin 3 times.

Another chance operation that would be adequate is

*Chance operation:* Toss a coin 100 times

with the understanding that  $E$  occurs if there is a run of 3 heads anywhere in the 100 tosses. Intuition suggests that  $E$  would be more likely with the second definition of the chance operation (100 tosses) than with the first (3 tosses). This intuition is correct and serves to underscore the importance of the chance operation in interpreting a probability.

■

The language of probability can be used to describe the results of random sampling from a population. The simplest application of this idea is a sample of size  $n = 1$ ; that is, choosing one member at random from a population. The following is an illustration.

---

**Example  
3.2.3**

**Sampling Fruitflies** A large population of the fruitfly *Drosophila melanogaster* is maintained in a lab. In the population, 30% of the individuals are black because of a mutation, while 70% of the individuals have the normal gray body color. Suppose one fly is chosen at random from the population. Then the probability that a black fly is chosen is 0.3. More formally, define

$$E: \text{Sampled fly is black}$$

Then

$$\Pr\{E\} = 0.3$$

■

The preceding example illustrates the basic relationship between probability and random sampling: *The probability that a randomly chosen individual has a certain characteristic is equal to the proportion of population members with the characteristic.*

## Frequency Interpretation of Probability

The **frequency interpretation** of probability provides a link between probability and the real world by relating the probability of an event to a measurable quantity, namely, the long-run relative frequency of occurrence of the event.\*

According to the frequency interpretation, the probability of an event  $E$  is meaningful only in relation to a chance operation that can in principle be repeated indefinitely often. Each time the chance operation is repeated, the event  $E$  either occurs or does not occur. *The probability  $\Pr\{E\}$  is interpreted as the relative frequency of occurrence of  $E$  in an indefinitely long series of repetitions of the chance operation.*

Specifically, suppose that the chance operation is repeated a large number of times, and that for each repetition the occurrence or nonoccurrence of  $E$  is noted. Then we may write

$$\Pr\{E\} \leftrightarrow \frac{\text{\# of times } E \text{ occurs}}{\text{\# of times chance operation is repeated}}$$

The arrow in the preceding expression indicates “approximate equality in the long run”; that is, if the chance operation is repeated many times, the two sides of the expression will be approximately equal. Here is a simple example.

---

**Example  
3.2.4**

**Coin Tossing** Consider again the chance operation of tossing a coin, and the event

$$E: \text{Heads}$$

If the coin is fair, then

$$\Pr\{E\} = 0.5 \leftrightarrow \frac{\text{\# of heads}}{\text{\# of tosses}}$$

The arrow in the preceding expression indicates that, in a long series of tosses of a fair coin, we expect to get heads about 50% of the time. ■

The following two examples illustrate the relative frequency interpretation for more complex events.

---

**Example  
3.2.5**

**Coin Tossing** Suppose that a fair coin is tossed twice. For reasons that will be explained later in this section, the probability of getting heads both times is 0.25. This probability has the following relative frequency interpretation.

---

\*Some statisticians prefer a different view, namely that the probability of an event is a subjective quantity expressing a person’s “degree of belief” that the event will happen. Statistical methods based on this “subjectivist” interpretation are rather different from those presented in this book.

*Chance operation:* Toss a coin twice

*E:* Both tosses are heads

$$\Pr\{E\} = 0.25 \leftrightarrow \frac{\text{\# of times both tosses are heads}}{\text{\# of pairs of tosses}}$$

**Example  
3.2.6**

**Sampling Fruitflies** In the *Drosophila* population of Example 3.2.3, 30% of the flies are black and 70% are gray. Suppose that two flies are randomly chosen from the population. We will see later in this section that the probability that both flies are the same color is 0.58. This probability can be interpreted as follows:

*Chance operation:* Choose a random sample of size  $n = 2$

*E:* Both flies in the sample are the same color

$$\Pr\{E\} = 0.58 \leftrightarrow \frac{\text{\# of times both flies are same color}}{\text{\# of times a sample of } n = 2 \text{ is chosen}}$$

We can relate this interpretation to a concrete sampling experiment. Suppose that the *Drosophila* population is in a very large container, and that we have some mechanism for choosing a fly at random from the container. We choose one fly at random, and then another; these two constitute the first sample of  $n = 2$ . After recording their colors, we put the two flies back into the container, and we are ready to repeat the sampling operation once again. Such a sampling experiment would be tedious to carry out physically, but it can readily be simulated using a computer. Table 3.2.1 shows a partial record of the results of choosing 10,000 random samples of size  $n = 2$  from a simulated *Drosophila* population. After each repetition of the chance operation (that is, after each sample of  $n = 2$ ), the cumulative relative frequency of occurrence of the event *E* was updated, as shown in the rightmost column of the table.

Figure 3.2.1 shows the cumulative relative frequency plotted against the number of samples. Notice that, as the number of samples becomes large, the relative frequency of occurrence of *E* approaches 0.58 (which is  $\Pr\{E\}$ ). In other words, the percentage of color-homogeneous samples among all the samples approaches 58% as the number of samples increases. It should be emphasized, however, that the *absolute* number of color-homogeneous samples generally does *not* tend to get closer to 58% of the total number. For instance, if we compare the results shown in Table 3.2.1 for the first 100 samples and the first 1,000 samples, we find the following:

	<b>Color-Homogeneous</b>			<b>Deviation from 58% of Total</b>		
First 100 samples:	54	or	54 %	-4	or	-4 %
First 1,000 samples:	596	or	59.6%	+16	or	+1.6%

Note that the deviation from 58% is larger in absolute terms, but smaller in relative terms (i.e., in percentage terms), for 1,000 samples than for 100 samples. Likewise, for 10,000 samples the deviation from 58% is rather larger (a deviation of -30),

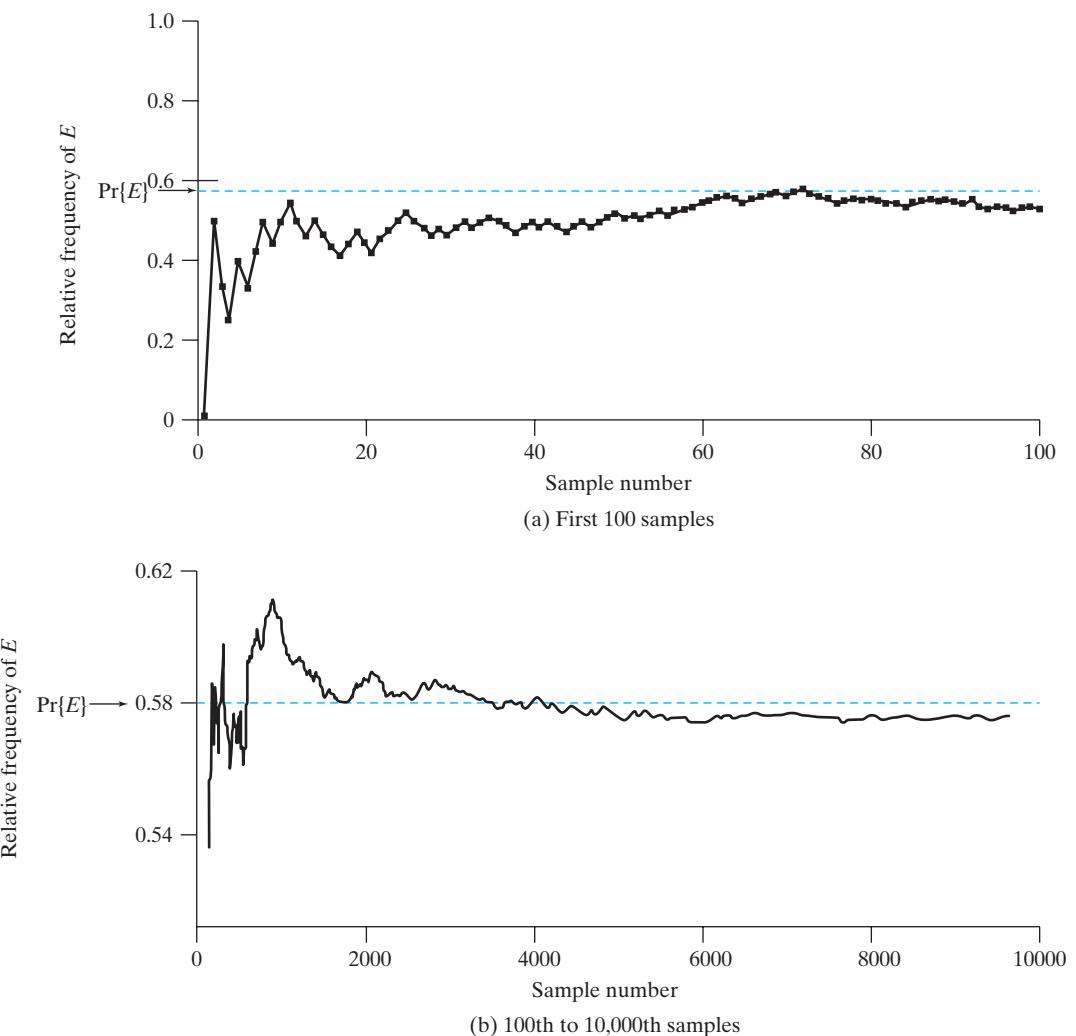
**Table 3.2.1** Partial results of simulated sampling from a *Drosophila* population

Sample number	Color		Did E occur?	Relative frequency of E (cumulative)
	1st Fly	2nd Fly		
1	G	B	No	0.000
2	B	B	Yes	0.500
3	B	G	No	0.333
4	G	B	No	0.250
5	G	G	Yes	0.400
6	G	B	No	0.333
7	B	B	Yes	0.429
8	G	G	Yes	0.500
9	G	B	No	0.444
10	B	B	Yes	0.500
·	·	·	·	·
·	·	·	·	·
·	·	·	·	·
20	G	B	No	0.450
·	·	·	·	·
·	·	·	·	·
·	·	·	·	·
100	G	B	No	0.540
·	·	·	·	·
·	·	·	·	·
·	·	·	·	·
1,000	G	G	Yes	0.596
·	·	·	·	·
·	·	·	·	·
·	·	·	·	·
10,000	B	B	Yes	0.577

but the percentage deviation is quite small ( $30/10,000$  is 0.3%). The deficit of 4 color-homogeneous samples among the first 100 samples is not *canceled* by a corresponding excess in later samples but rather is *swamped*, or overwhelmed, by a larger denominator.

## Probability Trees

Often it is helpful to use a **probability tree** to analyze a probability problem. A probability tree provides a convenient way to break a problem into parts and to organize the information available. The following examples show some applications of this idea.

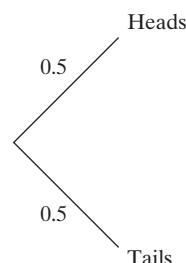


**Figure 3.2.1** Results of sampling from fruitfly population. Note that the axes are scaled differently in (a) and (b).

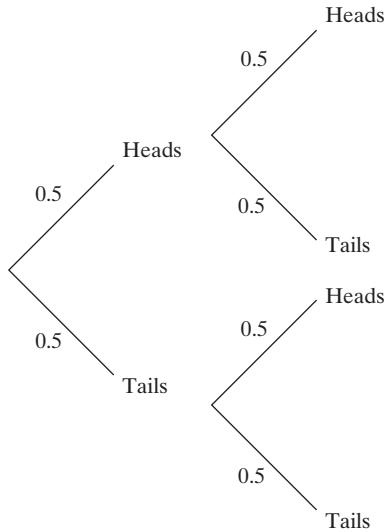
---

**Example  
3.2.7**

**Coin Tossing** If a fair coin is tossed twice, then the probability of heads is 0.5 on each toss. The first part of a probability tree for this scenario shows that there are two possible outcomes for the first toss and that they have probability 0.5 each.



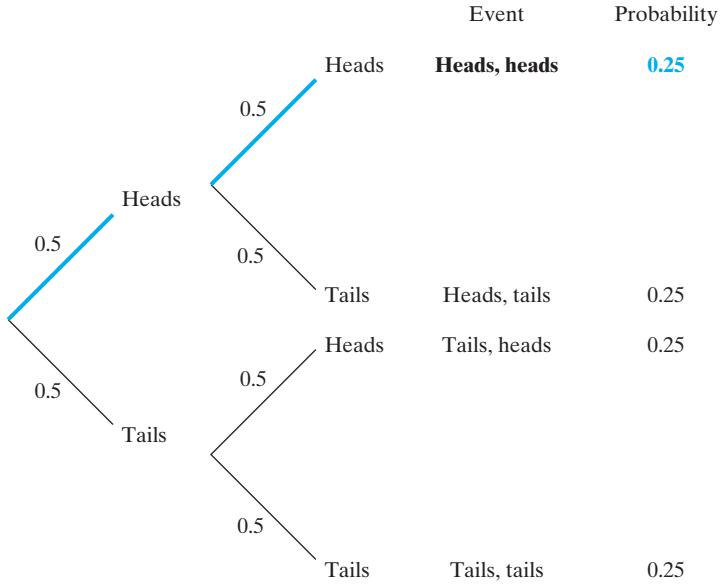
Then the tree shows that, for either outcome of the first toss, the second toss can be either heads or tails, again with probabilities 0.5 each.



To find the probability of getting heads on both tosses, we consider the path through the tree that produces this event. We multiply together the probabilities that we encounter along the path. Figure 3.2.2 summarizes this example and shows that

$$\Pr \{ \text{heads on both tosses} \} = 0.5 \times 0.5 = 0.25. \quad \blacksquare$$

**Figure 3.2.2** Probability tree for two coin tosses



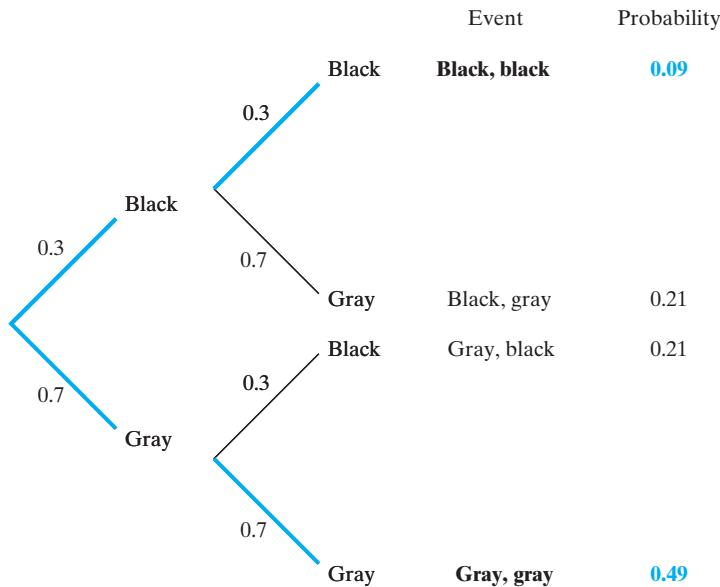
## Combination of Probabilities

If an event can happen in more than one way, the relative frequency interpretation of probability can be a guide to appropriate combinations of the probabilities of subevents. The following example illustrates this idea.

**Example  
3.2.8**

**Sampling Fruitflies** In the *Drosophila* population of Examples 3.2.3 and 3.2.6, 30% of the flies are black and 70% are gray. Suppose that two flies are randomly chosen from the population. Suppose we wish to find the probability that both flies are the same color. The probability tree displayed in Figure 3.2.3 shows the four possible outcomes from sampling two flies. From the tree, we can see that the probability of getting two black flies is  $0.3 \times 0.3 = 0.09$ . Likewise, the probability of getting two gray flies is  $0.7 \times 0.7 = 0.49$ .

**Figure 3.2.3** Probability tree for sampling two flies



To find the probability of the event

$E$ : Both flies in the sample are the same color

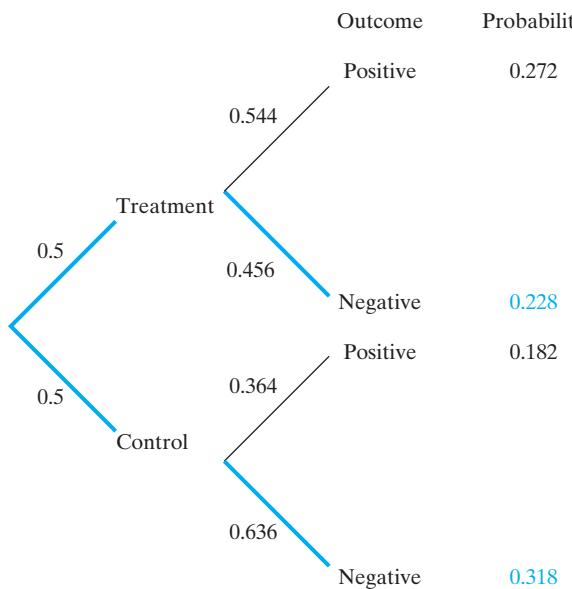
we add the probability of black, black to the probability of gray, gray to get  $0.09 + 0.49 = 0.58$ . ■

In the coin tossing setting of Example 3.2.7, the second part of the probability tree had the same structure as the first part—namely, a 0.5 chance of heads and a 0.5 chance of tails—because the outcome of the first toss does not affect the probability of heads on the second toss. Likewise, in Example 3.2.8 the probability of the second fly being black was 0.3, regardless of the color of the first fly, because the population was assumed to be very large, so that removing one fly from the population would not affect the proportion of flies that are black. However, in some situations we need to treat the second part of the probability tree differently than the first part.

**Example  
3.2.9**

**Nitric Oxide** Hypoxic respiratory failure is a serious condition that affects some newborns. If a newborn has this condition, it is often necessary to use extracorporeal membrane oxygenation (ECMO) to save the life of the child. However, ECMO is an invasive procedure that involves inserting a tube into a vein or artery near the heart, so physicians hope to avoid the need for it. One treatment for hypoxic respiratory failure is to have the newborn inhale nitric oxide. To test the effectiveness of this treatment, newborns suffering hypoxic respiratory failure were assigned at

**Figure 3.2.4** Probability tree for nitric oxide example



random to either be given nitric oxide or a control group.<sup>1</sup> In the treatment group 45.6% of the newborns had a negative outcome, meaning that either they needed ECMO or that they died. In the control group, 63.6% of the newborns had a negative outcome. Figure 3.2.4 shows a probability tree for this experiment.

If we choose a newborn at random from this group, there is a 0.5 probability that the newborn will be in the treatment group and, if so, a probability of 0.456 of getting a negative outcome. Likewise, there is a 0.5 probability that the newborn will be in the control group and, if so, a probability of 0.636 of getting a negative outcome. Thus, the probability of a negative outcome is

$$0.5 \times 0.456 + 0.5 \times 0.636 = 0.228 + 0.318 = 0.546.$$

### Example 3.2.10

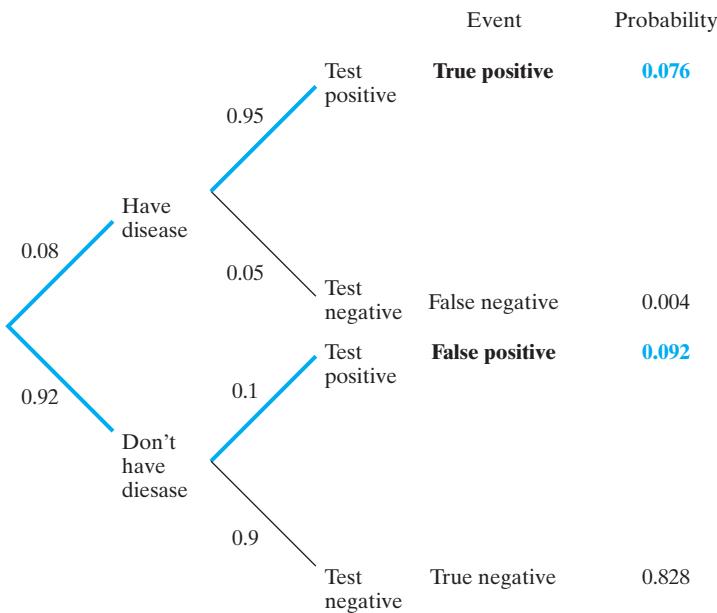
**Medical Testing** Suppose a medical test is conducted on someone to try to determine whether or not the person has a particular disease. If the test indicates that the disease is present, we say the person has “tested positive.” If the test indicates that the disease is not present, we say the person has “tested negative.” However, there are two types of mistakes that can be made. It is possible that the test indicates that the disease is present, but the person does not really have the disease; this is known as a false positive. It is also possible that the person has the disease, but the test does not detect it; this is known as a false negative.

Suppose that a particular test has a 95% chance of detecting the disease if the person has it (this is called the sensitivity of the test) and a 90% chance of correctly indicating that the disease is absent if the person really does not have the disease (this is called the specificity of the test). Suppose 8% of the population has the disease. What is the probability that a randomly chosen person will test positive?

Figure 3.2.5 shows a probability tree for this situation. The first split in the tree shows the division between those who have the disease and those who don’t. If someone has the disease, then we use 0.95 as the chance of the person testing positive. If the person doesn’t have the disease, then we use 0.10 as the chance of the person testing positive. Thus, the probability of a randomly chosen person testing positive is

$$0.08 \times 0.95 + 0.92 \times 0.10 = 0.076 + 0.092 = 0.168.$$

**Figure 3.2.5** Probability tree for medical testing example



**Example  
3.2.11**

**False Positives** Consider the medical testing scenario of Example 3.2.10. If someone tests positive, what is the chance the person really has the disease? In Example 3.2.10 we found that 0.168 (16.8%) of the population will test positive, so if 1,000 persons are tested, we would expect 168 to test positive. The probability of a true positive is 0.076, so we would expect 76 “true positives” out of 1,000 persons tested. Thus, we expect 76 true positives out of 168 total positives, which is to say that the probability that someone really has the disease, given that the person tests positive, is  $\frac{76}{168} = \frac{0.076}{0.168} \approx 0.452$ . This probability is quite a bit smaller than most people expect it to be, given that the sensitivity and specificity of the test are 0.95 and 0.90. ■

## Exercises 3.2.1–3.2.7

- 3.2.1** In a certain population of the freshwater sculpin, *Cottus rotheus*, the distribution of the number of tail vertebrae is as shown in the table.<sup>2</sup>

NO. OF VERTEBRAE	PERCENT OF FISH
20	3
21	51
22	40
23	6
Total	100

Find the probability that the number of tail vertebrae in a fish randomly chosen from the population

- (a) equals 21.
- (b) is less than or equal to 22.

- (c) is greater than 21.

- (d) is no more than 21.

- 3.2.2** In a certain college, 55% of the students are women. Suppose we take a sample of two students. Use a probability tree to find the probability

- (a) that both chosen students are women.
- (b) that at least one of the two students is a woman.

- 3.2.3** Suppose that a disease is inherited via a sex-linked mode of inheritance, so that a male offspring has a 50% chance of inheriting the disease, but a female offspring has no chance of inheriting the disease. Further suppose that 51.3% of births are male. What is the probability that a randomly chosen child will be affected by the disease?

- 3.2.4** Suppose that a student who is about to take a multiple choice test has only learned 40% of the material covered by the exam. Thus, there is a 40% chance that she

will know the answer to a question. However, even if she does not know the answer to a question, she still has a 20% chance of getting the right answer by guessing. If we choose a question at random from the exam, what is the probability that she will get it right?

**3.2.5** If a woman takes an early pregnancy test, she will either test positive, meaning that the test says she is pregnant, or test negative, meaning that the test says she is not pregnant. Suppose that if a woman really is pregnant, there is a 98% chance that she will test positive. Also, suppose that if a woman really is *not* pregnant, there is a 99% chance that she will test negative.

- (a) Suppose that 1,000 women take early pregnancy tests and that 100 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive?
- (b) Suppose that 1,000 women take early pregnancy tests and that 50 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive?

### 3.2.6

(a) Consider the setting of Exercise 3.2.5, part (a). Suppose that a woman tests positive. What is the probability that she really is pregnant?

(b) Consider the setting of Exercise 3.2.5, part (b). Suppose that a woman tests positive. What is the probability that she really is pregnant?

**3.2.7** Suppose that a medical test has a 92% chance of detecting a disease if the person has it (i.e., 92% sensitivity) and a 94% chance of correctly indicating that the disease is absent if the person really does not have the disease (i.e., 94% specificity). Suppose 10% of the population has the disease.

- (a) What is the probability that a randomly chosen person will test positive?
- (b) Suppose that a randomly chosen person does test positive. What is the probability that this person really has the disease?

## 3.3 Probability Rules (Optional)

We have defined the probability of an event,  $\Pr\{E\}$ , as the long-run relative frequency with which the event occurs. In this section we will briefly consider a few rules that help determine probabilities. We begin with three basic rules.

### Basic Rules

Rule (1) The probability of an event  $E$  is always between 0 and 1. That is,  $0 \leq \Pr\{E\} \leq 1$ .

Rule (2) The sum of the probabilities of all possible events equals 1. That is, if the set of possible events is  $E_1, E_2, \dots, E_k$ , then  $\sum_{i=1}^k \Pr\{E_i\} = 1$ .

Rule (3) The probability that an event  $E$  does not happen, denoted by  $E^C$ , is one minus the probability that the event happens. That is,  $\Pr\{E^C\} = 1 - \Pr\{E\}$ . (We refer to  $E^C$  as the complement of  $E$ .)

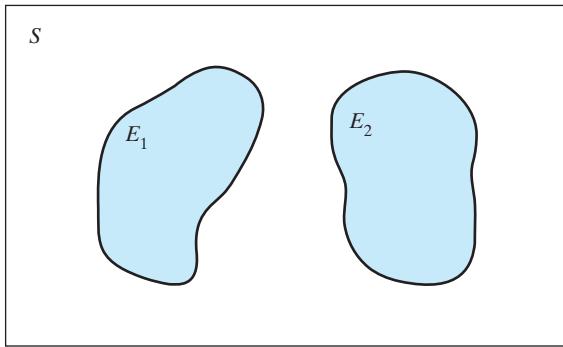
We illustrate these rules with an example.

---

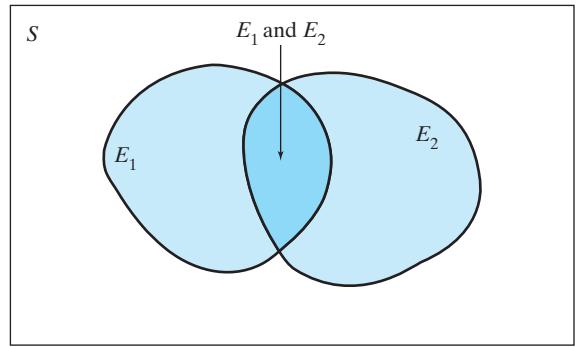
#### Example 3.3.1

**Blood Type** In the United States, 44% of the population has type O blood, 42% has type A, 10% has type B, and 4% has type AB.<sup>3</sup> Consider choosing someone at random and determining the person's blood type. The probability of a given blood type will correspond to the population percentage.

- (a) The probability that the person will have type O blood =  $\Pr\{O\} = 0.44$ .
- (b)  $\Pr\{O\} + \Pr\{A\} + \Pr\{B\} + \Pr\{AB\} = 0.44 + 0.42 + 0.10 + 0.04 = 1$ .



**Figure 3.3.1** Venn diagram showing two disjoint events



**Figure 3.3.2** Venn diagram showing union (total shaded area) and intersection (middle area) of two events

- (c) The probability that the person will *not* have type O blood =  $\Pr\{O^C\} = 1 - 0.44 = 0.56$ . This could also be found by adding the probabilities of the other blood types:  $\Pr\{O^C\} = \Pr\{A\} + \Pr\{B\} + \Pr\{AB\} = 0.42 + 0.10 + 0.04 = 0.56$ . ■

We often want to discuss two or more events at once; to do this we will find some terminology to be helpful. We say that two events are *disjoint*\* if they cannot occur simultaneously. Figure 3.3.1 is a *Venn diagram* that depicts a *sample space*  $S$  of all possible outcomes as a rectangle with two disjoint events depicted as nonoverlapping regions.

The *union* of two events is the event that one or the other occurs or both occur. The *intersection* of two events is the event that they both occur. Figure 3.3.2 is a Venn diagram that shows the union of two events as the total shaded area, with the intersection of the events being the overlapping region in the middle.

If two events are disjoint, then the probability of their union is the sum of their individual probabilities. If the events are not disjoint, then to find the probability of their union we take the sum of their individual probabilities and subtract the probability of their intersection (the part that was “counted twice”).

## Addition Rules

Rule (4) If two events  $E_1$  and  $E_2$  are disjoint, then  

$$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\}.$$

Rule (5) For any two events  $E_1$  and  $E_2$ ,  

$$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \text{ and } E_2\}.$$

We illustrate these rules with an example.

### Example 3.3.2

**Hair Color and Eye Color** Table 3.3.1 shows the relationship between hair color and eye color for a group of 1,770 German men.<sup>4</sup>

\*Another term for disjoint events is “mutually exclusive” events.

**Table 3.3.1** Hair color and eye color

		Hair color			
		Brown	Black	Red	Total
Eye color	Brown	400	300	20	720
	Blue	800	200	50	1,050
	Total	1,200	500	70	1,770

- (a) Because events “black hair” and “red hair” are disjoint, if we choose someone at random from this group then  $\Pr\{\text{black hair or red hair}\} = \Pr\{\text{black hair}\} + \Pr\{\text{red hair}\} = 500/1,770 + 70/1,770 = 570/1,770$ .
- (b) If we choose someone at random from this group, then  $\Pr\{\text{black hair}\} = 500/1,770$ .
- (c) If we choose someone at random from this group, then  $\Pr\{\text{blue eyes}\} = 1,050/1,770$ .
- (d) The events “black hair” and “blue eyes” are not disjoint, since there are 200 men with both black hair and blue eyes. Thus,  $\Pr\{\text{black hair or blue eyes}\} = \Pr\{\text{black hair}\} + \Pr\{\text{blue eyes}\} - \Pr\{\text{black hair and blue eyes}\} = 500/1,770 + 1,050/1,770 - 200/1,770 = 1,350/1,770$ . ■

Two events are said to be *independent* if knowing that one of them occurred does not change the probability of the other one occurring. For example, if a coin is tossed twice, the outcome of the second toss is independent of the outcome of the first toss, since knowing whether the first toss resulted in heads or in tails does not change the probability of getting heads on the second toss.

Events that are not independent are said to be *dependent*. When events are dependent, we need to consider the *conditional probability* of one event, given that the other event has happened. We use the notation

$$\Pr\{E_2|E_1\}$$

to represent the probability of  $E_2$  happening, given that  $E_1$  happened.

**Example  
3.3.3**

**Hair Color and Eye Color** Consider choosing a man at random from the group shown in Table 3.3.1. Overall, the probability of blue eyes is  $1,050/1,770$ , or about 59.3%. However, if the man has black hair, then the conditional probability of blue eyes is only  $200/500$ , or 40%; that is,  $\Pr\{\text{blue eyes}|\text{black hair}\} = 0.40$ . Because the probability of blue eyes depends on hair color, the events “black hair” and “blue eyes” are dependent. ■

Refer again to Figure 3.3.2, which shows the intersection of two regions (for  $E_1$  and  $E_2$ ). If we know that the event  $E_1$  has happened, then we can restrict our attention to the  $E_1$  region in the Venn diagram. If we now want to find the chance that  $E_2$  will happen, we need to consider the intersection of  $E_1$  and  $E_2$  relative to the entire  $E_1$  region. In the case of Example 3.3.3, this corresponds to knowing that a randomly chosen man has black hair, so that we restrict our attention to the 500 men (out of 1,770 total in the group) with black hair. Of these men, 200 have blue eyes. The 200 are in the intersection of “black hair” and “blue eyes.” The fraction  $200/500$  is the conditional probability of having blue eyes, given that the man has black hair.

This leads to the following formal definition of the conditional probability of  $E_2$  given  $E_1$ :

**Definition** The conditional probability of  $E_2$ , given  $E_1$ , is

$$\Pr\{E_2|E_1\} = \frac{\Pr\{E_1 \text{ and } E_2\}}{\Pr\{E_1\}}$$

provided that  $\Pr\{E_1\} > 0$ .

**Example  
3.3.4**

**Hair Color and Eye Color** Consider choosing a man at random from the group shown in Table 3.3.1. The probability of the man having blue eyes given that he has black hair is

$$\begin{aligned}\Pr\{\text{blue eyes}|\text{black hair}\} &= \Pr\{\text{black hair and blue eyes}\}/\Pr\{\text{black hair}\} \\ &= \frac{200}{500} = \frac{200}{1,770} = 0.40.\end{aligned}$$

In Section 3.2 we used probability trees to study compound events. In doing so, we implicitly used multiplication rules that we now make explicit.

## Multiplication Rules

Rule (6) If two events  $E_1$  and  $E_2$  are independent then

$$\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2\}.$$

Rule (7) For any two events  $E_1$  and  $E_2$ ,  $\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2|E_1\}$ .

**Example  
3.3.5**

**Coin Tossing** If a fair coin is tossed twice, the two tosses are independent of each other. Thus, the probability of getting heads on both tosses is

$$\begin{aligned}\Pr\{\text{heads twice}\} &= \Pr\{\text{heads on first toss}\} \times \Pr\{\text{heads on second toss}\} \\ &= 0.5 \times 0.5 = 0.25.\end{aligned}$$

**Example  
3.3.6**

**Blood Type** In Example 3.3.1 we stated that 44% of the U.S. population has type O blood. It is also true that 15% of the population is Rh negative and that this is independent of blood group. Thus, if someone is chosen at random, the probability that the person has type O, Rh negative blood is

$$\begin{aligned}\Pr\{\text{group O and Rh negative}\} &= \Pr\{\text{group O}\} \times \Pr\{\text{Rh negative}\} \\ &= 0.44 \times 0.15 = 0.066.\end{aligned}$$

**Example  
3.3.7**

**Hair Color and Eye Color** Consider choosing a man at random from the group shown in Table 3.3.1. What is the probability that the man will have red hair and brown eyes? Hair color and eye color are dependent, so finding this probability involves using a conditional probability. The probability that the man will have red hair is 70/1,770. Given that the man has red hair, the conditional probability of brown eyes is 20/70. Thus,

$$\begin{aligned}\Pr\{\text{red hair and brown eyes}\} &= \Pr\{\text{red hair}\} \times \Pr\{\text{brown eyes}|\text{red hair}\} \\ &= 70/1,770 \times 20/70 = 20/1,770.\end{aligned}$$

Sometimes a probability problem can be broken into two conditional “parts” that are solved separately and the answers combined.

### Rule of Total Probability

Rule (8) For any two events  $E_1$  and  $E_2$ ,

$$\Pr\{E_1\} = \Pr\{E_2\} \times \Pr\{E_1|E_2\} + \Pr\{E_2^C\} \times \Pr\{E_1|E_2^C\}.$$

**Example**  
**3.3.8**

**Hand Size** Consider choosing someone at random from a population that is 60% female and 40% male. Suppose that for a woman the probability of having a hand size smaller than  $100 \text{ cm}^2$  is 0.31.<sup>5</sup> Suppose that for a man the probability of having a hand size smaller than  $100 \text{ cm}^2$  is 0.08. What is the probability that the randomly chosen person will have a hand size smaller than  $100 \text{ cm}^2$ ?

We are given that if the person is a woman, then the probability of a “small” hand size is 0.31 and that if the person is a man, then the probability of a “small” hand size is 0.08.

Thus,

$$\begin{aligned}\Pr\{\text{hand size} < 100\} &= \Pr\{\text{woman}\} \times \Pr\{\text{hand size} < 100|\text{woman}\} \\ &\quad + \Pr\{\text{man}\} \times \Pr\{\text{hand size} < 100|\text{man}\} \\ &= 0.6 \times 0.31 + 0.4 \times 0.08 \\ &= 0.186 + 0.032 \\ &= 0.218.\end{aligned}$$

### Exercises 3.3.1–3.3.5

**3.3.1** In a study of the relationship between health risk and income, a large group of people living in Massachusetts were asked a series of questions.<sup>6</sup> Some of the results are shown in the following table.

		INCOME			TOTAL
		LOW	MEDIUM	HIGH	
Smoke	Smoke	634	332	247	1,213
	Don't smoke	1,846	1,622	1,868	5,336
Total		2,480	1,954	2,115	6,549

- (a) What is the probability that someone in this study smokes?
- (b) What is the conditional probability that someone in this study smokes, given that the person has high income?
- (c) Is being a smoker independent of having a high income? Why or why not?

**3.3.2** Consider the data table reported in Exercise 3.3.1.

- (a) What is the probability that someone in this study is from the low income group and smokes?
- (b) What is the probability that someone in this study is not from the low income group?
- (c) What is the probability that someone in this study is from the medium income group?
- (d) What is the probability that someone in this study is from the low income group or from the medium income group?

**3.3.3** The following data table is taken from the study reported in Exercise 3.3.1. Here “stressed” means that the person reported that most days are extremely stress-

ful or quite stressful; “not stressed” means that the person reported that most days are a bit stressful, not very stressful, or not at all stressful.

	INCOME			TOTAL
	LOW	MEDIUM	HIGH	
Stressed	526	274	216	1,016
Not stressed	1,954	1,680	1,899	5,533
Total	2,480	1,954	2,115	6,549

- (a) What is the probability that someone in this study is stressed?
- (b) Given that someone in this study is from the high income group, what is the probability that the person is stressed?
- (c) Compare your answers to parts (a) and (b). Is being stressed independent of having high income? Why or why not?

**3.3.4** Consider the data table reported in Exercise 3.3.3.

- (a) What is the probability that someone in this study has low income?
- (b) What is the probability that someone in this study either is stressed or has low income (or both)?
- (c) What is the probability that someone in this study either is stressed and has low income?

**3.3.5** Suppose that in a certain population of married couples 30% of the husbands smoke, 20% of the wives smoke, and in 8% of the couples both the husband and the wife smoke. Is the smoking status (smoker or nonsmoker) of the husband independent of that of the wife? Why or why not?

## 3.4 Density Curves

The examples presented in Section 3.2 dealt with probabilities for discrete variables. In this section we will consider probability when the variable is continuous.

### Relative Frequency Histograms and Density Curves

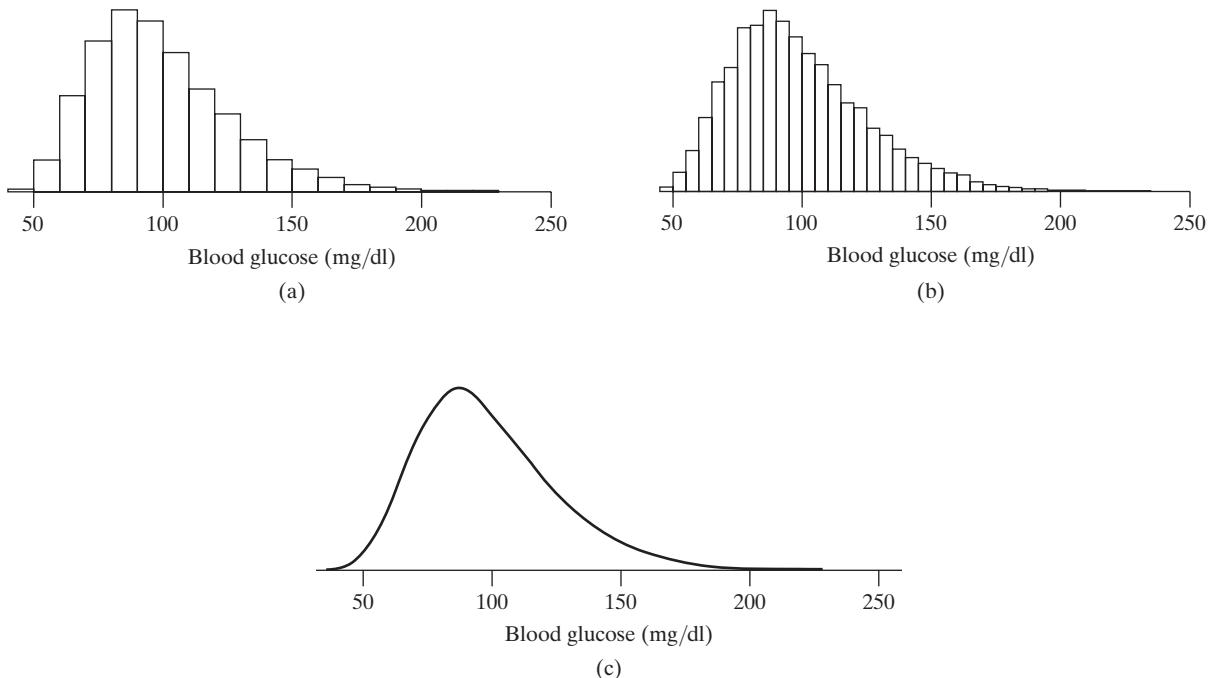
In Chapter 2 we discussed the use of a histogram to represent a frequency distribution for a variable. A relative frequency histogram is a histogram in which we indicate the proportion (i.e., the relative frequency) of observations in each category, rather than the count of observations in the category. We can think of the relative frequency histogram as an approximation of the underlying true population distribution from which the data came.

It is often desirable, especially when the observed variable is continuous, to describe a population frequency distribution by a smooth curve. We may visualize the curve as an idealization of a relative frequency histogram with very narrow classes. The following example illustrates this idea.

---

**Example**  
3.4.1

**Blood Glucose** A glucose tolerance test can be useful in diagnosing diabetes. The blood level of glucose is measured one hour after the subject has drunk 50 mg of glucose dissolved in water. Figure 3.4.1 shows the distribution of responses to this test for a certain population of women.<sup>7</sup> The distribution is represented by histograms with class widths equal to (a) 10 and (b) 5, and by (c) a smooth curve.



**Figure 3.4.1** Different representations of the distribution of blood glucose levels in a population of women

A smooth curve representing a frequency distribution is called a **density curve**. The vertical coordinates of a density curve are plotted on a scale called a **density scale**. When the density scale is used, relative frequencies are represented as areas under the curve. Formally, the relation is as follows:

### Interpretation of Density

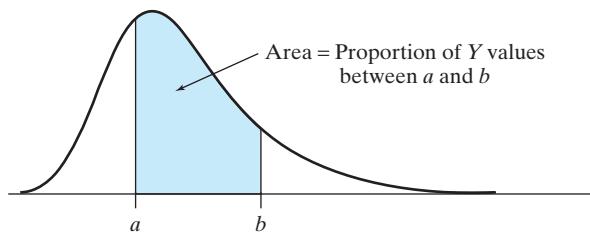
For any two numbers  $a$  and  $b$ ,

$$\text{Area under density curve between } a \text{ and } b = \text{Proportion of } Y \text{ values between } a \text{ and } b$$

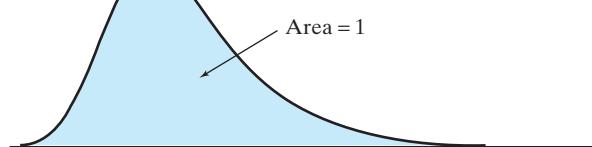
This relation is indicated in Figure 3.4.2 for an arbitrary distribution

Because of the way the density curve is interpreted, the density curve is entirely above (or equal to) the  $x$ -axis and the area under the entire curve must be equal to 1, as shown in Figure 3.4.3.

The interpretation of density curves in terms of areas is illustrated concretely in the following example.



**Figure 3.4.2** Interpretation of area under a density curve



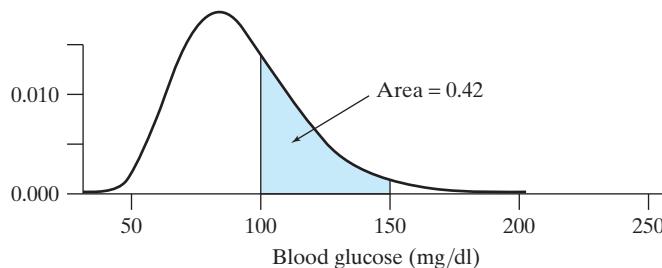
**Figure 3.4.3** The area under an entire density curve must be 1

#### Example 3.4.2

**Blood Glucose** Figure 3.4.4 shows the density curve for the blood glucose distribution of Example 3.4.1, with the vertical scale explicitly shown. The shaded area is equal to 0.42, which indicates that about 42% of the glucose levels are between 100 mg/dl and 150 mg/dl. The area under the density curve to the left of 100 mg/dl is equal to 0.50; this indicates that the population median glucose level is 100 mg/dl. The area under the entire curve is 1. ■

**Figure 3.4.4**

Interpretation of an area under the blood glucose density curve



**The Continuum Paradox** The area interpretation of a density curve has a paradoxical element. If we ask for the relative frequency of a single specific  $Y$  value, the answer is zero. For example, suppose we want to determine from Figure 3.4.4 the relative frequency of blood glucose levels *equal* to 150. The area interpretation gives an answer of zero. This seems to be nonsense—how can every value of  $Y$  have a relative frequency of zero? Let us look more closely at the question. If blood glucose is measured to the nearest mg/dl, then we are really asking for the relative frequency of glucose levels between 149.5 and 150.5 mg/dl, and the corresponding area is not zero. On the other hand, if we are thinking of blood glucose as an *idealized* continuous variable, then the relative frequency of any particular value (such as 150) *is* zero. This is admittedly a paradoxical situation. It is similar to the paradoxical fact that an idealized straight line can be 1 centimeter long, and yet each of the idealized points of which the line is composed has length equal to zero. In practice, the continuum paradox does not cause any trouble; we simply do not discuss the relative frequency of a single  $Y$  value (just as we do not discuss the length of a single point).

## Probabilities and Density Curves

If a variable has a continuous distribution, then we find probabilities by using the density curve for the variable. A probability for a continuous variable equals the area under the density curve for the variable between two points.

### Example 3.4.3

**Blood Glucose** Consider the blood glucose level, in mg/dl, of a randomly chosen subject from the population described in Example 3.4.2. We saw in Example 3.4.2 that 42% of the population glucose levels are between 100 mg/dl and 150 mg/dl. Thus,  $\Pr\{100 \leq \text{glucose level} \leq 150\} = 0.42$ .

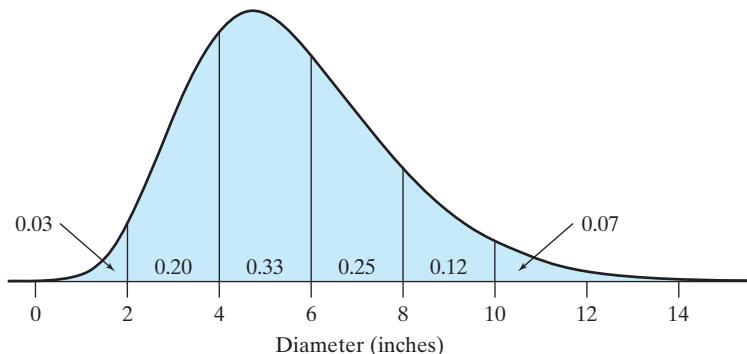
We are modeling blood glucose level as being a continuous variable, which means that  $\Pr\{\text{glucose level} = 100\} = 0$ , as we noted above. Thus,

$$\Pr\{100 \leq \text{glucose level} \leq 150\} = \Pr\{100 < \text{glucose level} < 150\} = 0.42. \quad \blacksquare$$

### Example 3.4.4

**Tree Diameters** The diameter of a tree trunk is an important variable in forestry. The density curve shown in Figure 3.4.5 represents the distribution of diameters (measured 4.5 feet above the ground) in a population of 30-year-old Douglas fir trees; areas under the curve are shown in the figure.<sup>8</sup> Consider the diameter, in inches, of a randomly chosen tree. Then, for example,  $\Pr\{4 < \text{diameter} < 6\} = 0.33$ . If we want to find the probability that a randomly chosen tree has a diameter greater than 8 inches, we must add the last two areas under the curve in Figure 3.4.3:  $\Pr\{\text{diameter} > 8\} = 0.12 + 0.07 = 0.19$ . ■

**Figure 3.4.5** Diameters of 30-year-old Douglas fir trees



### Exercises 3.4.1–3.4.4

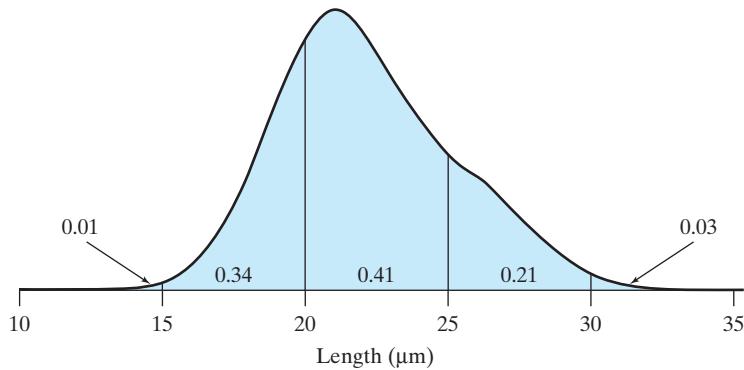
**3.4.1** Consider the density curve shown in Figure 3.4.5, which represents the distribution of diameters (measured 4.5 feet above the ground) in a population of 30-year-old Douglas fir trees. Areas under the curve are shown in the figure. What percentage of the trees have diameters

- (a) between 4 inches and 10 inches?
- (b) less than 4 inches?
- (c) more than 6 inches?

**3.4.2** Consider the diameter of a Douglas fir tree drawn at random from the population that is represented by the density curve shown in Figure 3.4.5. Find

- (a)  $\Pr\{\text{diameter} < 10\}$
- (b)  $\Pr\{\text{diameter} > 4\}$
- (c)  $\Pr\{2 < \text{diameter} < 8\}$

**3.4.3** In a certain population of the parasite *Trypanosoma*, the lengths of individuals are distributed as indicated by the density curve shown here. Areas under the curve are shown in the figure.<sup>9</sup>



Consider the length of an individual trypanosome chosen at random from the population. Find

- (a)  $\Pr\{20 < \text{length} < 30\}$
- (b)  $\Pr\{\text{length} > 20\}$
- (c)  $\Pr\{\text{length} < 20\}$

**3.4.4** Consider the distribution of *Trypanosoma* lengths shown by the density curve in Exercise 3.4.3. Suppose we take a sample of two trypanosomes. What is the probability that

- (a) both trypanosomes will be shorter than  $20\ \mu\text{m}$ ?
- (b) the first trypanosome will be shorter than  $20\ \mu\text{m}$  and the second trypanosome will be longer than  $25\ \mu\text{m}$ ?
- (c) exactly one of the trypanosomes will be shorter than  $20\ \mu\text{m}$  and one trypanosome will be longer than  $25\ \mu\text{m}$ ?

## 3.5 Random Variables

A **random variable** is simply a variable that takes on numerical values that depend on the outcome of a chance operation. The following examples illustrate this idea.

### Example 3.5.1

**Dice** Consider the chance operation of tossing a die. Let the random variable  $Y$  represent the number of spots showing. The possible values of  $Y$  are  $Y = 1, 2, 3, 4, 5$ , or  $6$ . We do not know the value of  $Y$  until we have tossed the die. If we know how the die is weighted, then we can specify the probability that  $Y$  has a particular value, say  $\Pr\{Y = 4\}$ , or a particular set of values, say  $\Pr\{2 \leq Y \leq 4\}$ . For instance, if the die is perfectly balanced so that each of the six faces is equally likely, then

$$\Pr\{Y = 4\} = \frac{1}{6} \approx 0.17$$

and

$$\Pr\{2 \leq Y \leq 4\} = \frac{3}{6} = 0.5$$

**Example  
3.5.2**

**Family Size** Suppose a family is chosen at random from a certain population, and let the random variable  $Y$  denote the number of children in the chosen family. The possible values of  $Y$  are  $0, 1, 2, 3, \dots$ . The probability that  $Y$  has a particular value is equal to the percentage of families with that many children. For instance, if 23% of the families have 2 children, then

$$\Pr\{Y = 2\} = 0.23$$

**Example  
3.5.3**

**Medications** After someone has heart surgery, the person is usually given several medications. Let the random variable  $Y$  denote the number of medications that a patient is given following cardiac surgery. If we know the distribution of the number of medications per patient for the entire population, then we can specify the probability that  $Y$  has a certain value or falls within a certain interval of values. For instance, if 52% of all patients are given 2, 3, 4, or 5 medications, then

$$\Pr\{2 \leq Y \leq 5\} = 0.52$$

**Example  
3.5.4**

**Heights of Men** Let the random variable  $Y$  denote the height of a man chosen at random from a certain population. If we know the distribution of heights in the population, then we can specify the probability that  $Y$  falls in a certain range. For instance, if 46% of the men are between 65.2 and 70.4 inches tall, then

$$\Pr\{65.2 \leq Y \leq 70.4\} = 0.46$$

Each of the variables in Examples 3.5.1–3.5.3 is a *discrete random variable*, because in each case we can list the possible values that the variable can take on. In contrast, the variable in Example 3.5.4, height, is a *continuous random variable*: Height, at least in theory, can take on any of an infinite number of values in an interval. Of course, when we measure and record a person's height, we generally measure to the nearest inch or half inch. Nonetheless, we can think of true height as being a continuous variable. We use density curves to model the distributions of continuous random variables, such as blood glucose level or tree diameter as discussed in Section 3.4.

## Mean and Variance of a Random Variable

In Chapter 2 we briefly considered the concepts of population mean and population standard deviation. For the case of a discrete random variable, we can calculate the population mean and standard deviation if we know the probability distribution for the random variable. We begin with the mean.

The mean of a discrete random variable  $Y$  is defined as

$$\mu_Y = \sum y_i \Pr(Y = y_i)$$

where the  $y_i$ 's are the values that the variable takes on and the sum is taken over all possible values.

The mean of a random variable is also known as the *expected value* and is often written as  $E(Y)$ ; that is,  $E(Y) = \mu_Y$ .

**Example  
3.5.5**

**Fish Vertebrae** In a certain population of the freshwater sculpin, *Cottus rotheus*, the distribution of the number of tail vertebrae,  $Y$ , is as shown in Table 3.5.1.<sup>2</sup>

<b>Table 3.5.1</b> Distribution of vertebrae	
No. of vertebrae	Percent of fish
20	3
21	51
22	40
23	6
Total	100

The mean of  $Y$  is

$$\begin{aligned}\mu_Y &= 20 \times \Pr\{Y = 20\} + 21 \times \Pr\{Y = 21\} + 22 \times \Pr\{Y = 22\} + 23 \times \Pr\{Y = 23\} \\ &= 20 \times .03 \quad + 21 \times .51 \quad + 22 \times .40 \quad + 23 \times .06 \\ &= 0.6 \quad + 10.71 \quad + 8.8 \quad + 1.38 \\ &= 21.49.\end{aligned}$$

---

**Example  
3.5.6**

**Dice** Consider rolling a die that is perfectly balanced so that each of the six faces is equally likely to come up and let the random variable  $Y$  represent the number of spots showing. The expected value, or mean, of  $Y$  is

$$E(Y) = \mu_Y = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{21}{6} = 3.5.$$

To find the standard deviation of a random variable, we first find the variance,  $\sigma^2$ , of the random variable and then take the square root of the variance to get the standard deviation,  $\sigma$ .

The variance of a discrete random variable  $Y$  is defined as

$$\sigma_Y^2 = \sum (y_i - \mu_Y)^2 \Pr(Y = y_i)$$

where the  $y_i$ 's are the values that the variable takes on and the sum is taken over all possible values.

We often write  $\text{VAR}(Y)$  to denote the variance of  $Y$ .

---

**Example  
3.5.7**

**Fish Vertebrae** Consider the distribution of vertebrae given in Table 3.5.1. In Example 3.5.5 we found that the mean of  $Y$  is  $\mu_Y = 21.49$ . The variance of  $Y$  is

$$\begin{aligned}\text{VAR}(Y) = \sigma_Y^2 &= (20 - 21.49)^2 \times \Pr\{Y = 20\} \\ &\quad + (21 - 21.49)^2 \times \Pr\{Y = 21\} \\ &\quad + (22 - 21.49)^2 \times \Pr\{Y = 22\} \\ &\quad + (23 - 21.49)^2 \times \Pr\{Y = 23\} \\ &= (-1.49)^2 \times 0.03 + (-.49)^2 \times 0.51 \\ &\quad + (0.51)^2 \times 0.40 + (1.51)^2 \times 0.06 \\ &= 2.2201 \times 0.03 + .2401 \times 0.51 + .2601 \times 0.40 + 2.2801 \times 0.06 \\ &= 0.066603 + 0.122451 + 0.10404 + 0.136806 \\ &= 0.4299.\end{aligned}$$

The standard deviation of  $Y$  is  $\sigma_Y = \sqrt{0.4299} \approx 0.6557$ .

**Example  
3.5.8**

**Dice** In Example 3.5.6 we found that the mean number obtained from rolling a fair die is 3.5 (i.e.,  $\mu_Y = 3.5$ ). The variance of the number obtained from rolling a fair die is

$$\begin{aligned}\sigma_Y^2 &= (1 - 3.5)^2 \times \Pr\{Y = 1\} + (2 - 3.5)^2 \times \Pr\{Y = 2\} \\&\quad + (3 - 3.5)^2 \times \Pr\{Y = 3\} + (4 - 3.5)^2 \times \Pr\{Y = 4\} \\&\quad + (5 - 3.5)^2 \times \Pr\{Y = 5\} + (6 - 3.5)^2 \times \Pr\{Y = 6\} \\&= (-2.5)^2 \times \frac{1}{6} + (-1.5)^2 \times \frac{1}{6} + (-0.5)^2 \times \frac{1}{6} + (0.5)^2 \times \frac{1}{6} \\&\quad + (1.5)^2 \times \frac{1}{6} + (2.5)^2 \times \frac{1}{6} \\&= (6.25) \times \frac{1}{6} + (2.25) \times \frac{1}{6} + (0.25) \times \frac{1}{6} + (0.25) \times \frac{1}{6} \\&\quad + (2.25) \times \frac{1}{6} + (6.25) \times \frac{1}{6} \\&= 17.5 \times \frac{1}{6} \\&\approx 2.9167.\end{aligned}$$

The standard deviation of  $Y$  is  $\sigma_Y = \sqrt{2.9167} \approx 1.708$ . ■

The preceding definitions are appropriate for discrete random variables. There are analogous definitions for continuous random variables, but they involve integral calculus and won't be presented here.

## Adding and Subtracting Random Variables (Optional)

If we add two random variables, it makes sense that we add their means. Likewise, if we create a new random variable by subtracting two random variables, then we subtract the individual means to get the mean of the new random variable. If we multiply a random variable by a constant (for example, if we are converting feet to inches, so that we are multiplying by 12), then we multiply the mean of the random variable by the same constant. If we add a constant to a random variable, then we add that constant to the mean.

The following rules summarize the situation:

### Rules for Means of Random Variables

Rule (1) If  $X$  and  $Y$  are two random variables, then  $\mu_{X+Y} = \mu_X + \mu_Y$ .

$$\mu_{X-Y} = \mu_X - \mu_Y$$

Rule (2) If  $Y$  is a random variable and  $a$  and  $b$  constants, then

$$\mu_{a+bY} = a + b\mu_Y.$$

**Example  
3.5.9**

**Temperature** The average summer temperature,  $\mu_Y$ , in a city is 81°F. To convert °F to °C, we use the formula  $^{\circ}\text{C} = (\text{°F} - 32) \times (5/9)$  or  $^{\circ}\text{C} = (5/9) \times ^{\circ}\text{F} - (5/9) \times 32$ . Thus, the mean in degrees Celsius is  $(5/9) \times (81) - (5/9) \times 32 = 45 - 17.78 = 27.22$ . ■

Dealing with standard deviations of functions of random variables is a bit more complicated. We work with the variance first and then take the square root, at the

end, to get the standard deviation we want. If we *multiply* a random variable by a constant (for example, if we are converting inches to centimeters by multiplying by 2.54), then we multiply the variance by the square of the constant. This has the effect of multiplying the standard deviation by the constant. If we *add* a constant to a random variable, then we are not changing the relative spread of the distribution, so the variance does not change.

**Example  
3.5.10**

**Feet to Inches** Let  $Y$  denote the height, in feet, of a person in a given population; suppose the standard deviation of  $Y$  is  $\sigma_Y = 0.35$  (feet). If we wish to convert from feet to inches, we can define a new variable  $X$  as  $X = 12Y$ . The variance of  $Y$  is  $0.35^2$  (the square of the standard deviation). The variance of  $X$  is  $12^2 \times 0.35^2$ , which means that the standard deviation of  $X$  is  $\sigma_X = 12 \times 0.35 = 4.2$  (inches). ■

If we add two random variables *that are independent of one another*, then we add their variances.\* Moreover, if we subtract two random variables *that are independent of one another*, then we *add* their variances. If we want to find the standard deviation of the sum (or difference) of two independent random variables, we first find the variance of the sum (or difference) and then take the square root to get the standard deviation of the sum (or difference).

**Example  
3.5.11**

**Mass** Consider finding the mass of a 10-ml graduated cylinder. If several measurements are made, using an analytical balance, then in theory we would expect the measurements to all be the same. In reality, however, the readings will vary from one measurement to the next. Suppose that a given balance produces readings that have a standard deviation of 0.03g; let  $X$  denote the value of a reading made using this balance. Suppose that a second balance produces readings that have a standard deviation of 0.04g; let  $Y$  denote the value of a reading made using this second balance.<sup>10</sup>

If we use each balance to measure the mass of a graduated cylinder, we might be interested in the difference,  $X - Y$ , of the two measurements. The standard deviation of  $X - Y$  is positive. To find the standard deviation of  $X - Y$ , we first find the variance of the difference. The variance of  $X$  is  $0.03^2$  and the variance of  $Y$  is  $0.04^2$ . The variance of the difference is  $0.03^2 + 0.04^2 = 0.0025$ . The standard deviation of  $X - Y$  is the square root of 0.0025, which is 0.05. ■

The following rules summarize the situation for variances:

**Rules for Variances of Random Variables**

Rule (3) If  $Y$  is a random variable and  $a$  and  $b$  constants, then  $\sigma_{a+bY}^2 = b^2\sigma_Y^2$ .

Rule (4) If  $X$  and  $Y$  are two *independent* random variables, then

$$\begin{aligned}\sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 \\ \sigma_{X-Y}^2 &= \sigma_X^2 + \sigma_Y^2\end{aligned}$$

---

\*If we add two random variables that are not independent of one another, then the variance of the sum depends on the degree of dependence between the variables. To take an extreme case, suppose that one of the random variables is the negative of the other. Then the sum of the two random variables will always be zero, so that the variance of the sum will be zero. This is quite different from what we would get by adding the two variances together. As another example, suppose  $Y$  is the number of questions correct on a 20-question exam and  $X$  is the number of questions wrong. Then  $Y + X$  is always equal to 20, so that there is no variability at all. Hence, the variance of  $Y + X$  is zero, even though the variance of  $Y$  is positive, as is the variance of  $X$ .

## Exercises 3.5.1–3.5.8

**3.5.1** In a certain population of the European starling, there are 5,000 nests with young. The distribution of brood size (number of young in a nest) is given in the accompanying table.<sup>11</sup>

BROOD SIZE	FREQUENCY (NO. OF BROODS)
1	90
2	230
3	610
4	1,400
5	1,760
6	750
7	130
8	26
9	3
10	1
Total	5,000

Suppose one of the 5,000 broods is to be chosen at random, and let  $Y$  be the size of the chosen brood. Find



**3.5.2** In the starling population of Exercise 3.5.1, there are 22,435 young in all the broods taken together. (There are 90 young from broods of size 1, there are 460 from broods of size 2, etc.) Suppose one of the young is to be chosen at random, and let  $Y'$  be the size of the chosen individual's brood.

- (a) Find  $\Pr\{Y' = 3\}$ .      (b) Find  $\Pr\{Y' \geq 7\}$ .

(c) Explain why choosing a young at random and then observing its brood is not equivalent to choosing a brood at random. Your explanation should show why the answer to part (b) is greater than the answer to part (b) of Exercise 3.5.1.

**3.5.3** Calculate the mean,  $\mu_Y$ , of the random variable  $Y$  from Exercise 3.5.1.

**3.5.4** Consider a population of the fruitfly *Drosophila melanogaster* in which 30% of the individuals are black because of a mutation, while 70% of the individuals have the normal gray body color. Suppose three flies are chosen at random from the population; let  $Y$  denote the number of black flies out of the three. Then the probability distribution for  $Y$  is given by the following table:

$Y$ (NO. BLACK)	PROBABILITY
0	0.343
1	0.441
2	0.189
3	0.027
Total	1.000

- (a) Find  $\Pr\{Y \geq 2\}$       (b) Find  $\Pr\{Y \leq 2\}$

**3.5.5** Calculate the mean,  $\mu_Y$ , of the random variable  $Y$  from Exercise 3.5.4.

**3.5.6** Calculate the standard deviation,  $\sigma_Y$ , of the random variable  $Y$  from Exercise 3.5.4.

**3.5.7** A group of college students were surveyed to learn how many times they had visited a dentist in the previous year.<sup>12</sup> The probability distribution for  $Y$ , the number of visits, is given by the following table:

Y (NO. VISITS)	PROBABILITY
0	0.15
1	0.50
2	0.35
Total	1.00

Calculate the mean,  $\mu_Y$ , of the number of visits.

**3.5.8** Calculate the standard deviation,  $\sigma_Y$ , of the random variable  $Y$  from Exercise 3.5.7.

### 3.6 The Binomial Distribution

To add some depth to the notion of probability and random variables, we now consider a special type of random variable, the **binomial**. The distribution of a binomial random variable is a probability distribution associated with a special kind of

chance operation. The chance operation is defined in terms of a set of conditions called the independent-trials model.

## The Independent-Trials Model

The **independent-trials model** relates to a sequence of chance “trials.” Each trial is assumed to have two possible outcomes, which are arbitrarily labeled “success” and “failure.” The probability of success on each individual trial is denoted by the letter  $p$  and is assumed to be constant from one trial to the next. In addition, the trials are required to be independent, which means that the chance of success or failure on each trial does not depend on the outcome of any other trial. The total number of trials is denoted by  $n$ . These conditions are summarized in the following definition of the model.

### Independent-Trials Model

A series of  $n$  independent trials is conducted. Each trial results in success or failure. The probability of success is equal to the same quantity,  $p$ , for each trial, regardless of the outcomes of the other trials.

The following examples illustrate situations that can be described by the independent-trials model.

#### Example 3.6.1

**Albinism** If two carriers of the gene for albinism marry, each of their children has probability  $1/4$  of being albino. The chance that the second child is albino is the same ( $1/4$ ) whether or not the first child is albino; similarly, the outcome for the third child is independent of the first two, and so on. Using the labels “success” for albino and “failure” for nonalbino, the independent-trials model applies with  $p = 1/4$  and  $n =$  the number of children in the family. ■

#### Example 3.6.2

**Mutant Cats** A study of cats in Omaha, Nebraska, found that 37% of them have a certain mutant trait.<sup>13</sup> Suppose that 37% of all cats have this mutant trait and that a random sample of cats is chosen from the population. As each cat is chosen for the sample, the probability is 0.37 that it will be mutant. This probability is the same as each cat is chosen, regardless of the results of the other cats, because the percentage of mutants in the large population remains equal to 0.37 even when a few individual cats have been removed. Using the labels “success” for mutant and “failure” for nonmutant, the independent-trials model applies with  $p = 0.37$  and  $n =$  the sample size. ■

## An Example of the Binomial Distribution

The binomial distribution specifies the probabilities of various numbers of successes and failures when the basic chance operation consists of  $n$  independent trials. Before giving the general formula for the binomial distribution, we consider a simple example.

**Example  
3.6.3**

**Albinism** Suppose two carriers of the gene for albinism marry (see Example 3.6.1) and have two children. Then the probability that both of their children are albino is

$$\Pr\{\text{both children are albino}\} = \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{16}$$

The reason for this probability can be seen by considering the relative frequency interpretation of probability. Of a great many such families with two children,  $\frac{1}{4}$  would have the first child albino; furthermore,  $\frac{1}{4}$  of these would have the second child albino; thus,  $\frac{1}{4}$  of  $\frac{1}{4}$ , or  $\frac{1}{16}$  of all the couples would have both albino children. A similar kind of reasoning shows that the probability that both children are not albino is

$$\Pr\{\text{both children are not albino}\} = \left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{16}$$

A new twist enters if we consider the probability that one child is albino and the other is not. There are two possible ways this can happen:

$$\Pr\{\text{first child is albino, second is not}\} = \left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{3}{16}$$

$$\Pr\{\text{first child is not albino, second is}\} = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) = \frac{3}{16}$$

To see how to combine these possibilities, we again consider the relative frequency interpretation of probability. Of a great many such families with two children, the fraction of families with one albino and one nonalbino child would be the total of the two possibilities, or

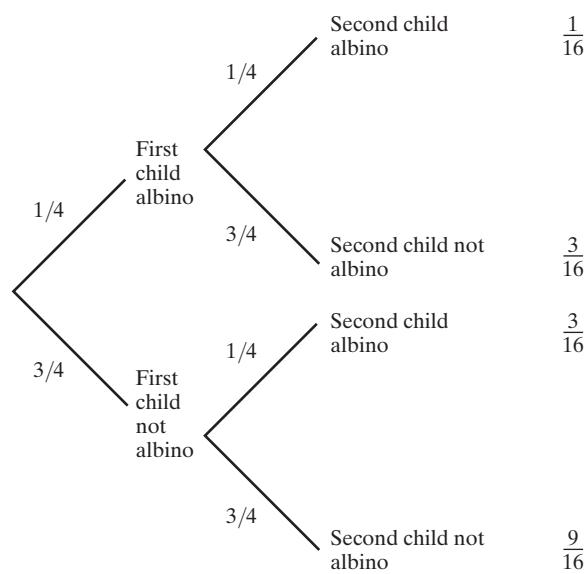
$$\left(\frac{3}{16}\right) + \left(\frac{3}{16}\right) = \frac{6}{16}$$

Thus, the corresponding probability is

$$\Pr\{\text{one child is albino, the other is not}\} = \frac{6}{16}$$

Another way to see this is to consider a probability tree. The first split in the tree represents the birth of the first child; the second split represents the birth of the second child. The four possible outcomes and their associated probabilities are shown in Figure 3.6.1. These probabilities are collected in Table 3.6.1. ■

The probability distribution in Table 3.6.1 is called the binomial distribution with  $p = \frac{1}{4}$  and  $n = 2$ . Note that the probabilities add to 1. This makes sense because all possibilities have been accounted for: We expect  $\frac{9}{16}$  of the families to have no albino children,  $\frac{6}{16}$  to have one albino child, and  $\frac{1}{16}$  to have two albino children; there are no other possible compositions for a two-child family. The number of albino children, out of the two children, is an example of a binomial random variable. A **binomial random variable** is a random variable that satisfies the following four conditions, abbreviated as **BInS**:

**Table 3.6.1** Probability distribution for number of albino children

Number of		Probability
Albino	Nonalbino	
0	2	$\frac{9}{16}$
1	1	$\frac{6}{16}$
2	0	$\frac{1}{16}$
		1

**Figure 3.6.1** Probability tree for albinism among two children of carriers of the gene for albinism

**Binary outcomes:** There are two possible outcomes for each trial (success and failure).

**Independent trials:** The outcomes of the trials are independent of each other.

**$n$  is fixed:** The number of trials,  $n$ , is fixed in advance.

**Same value of  $p$ :** The probability of a success on a single trial is the same for all trials.

## The Binomial Distribution Formula

A general formula is available that can be used to calculate probabilities associated with a binomial random variable for any values of  $n$  and  $p$ . This formula can be proved using logic similar to that in Example 3.6.3. (The formula is discussed further in Appendix 3.1.) The formula is given in the accompanying box.

### The Binomial Distribution Formula

For a binomial random variable  $Y$ , the probability that the  $n$  trials result in  $j$  successes (and  $n - j$  failures) is given by the following formula:

$$\Pr\{j \text{ successes}\} = \Pr\{Y = j\} = {}_nC_j p^j (1 - p)^{n-j}$$

The quantity  ${}_nC_j$  appearing in the formula is called a **binomial coefficient**. Each binomial coefficient is an integer depending on  $n$  and on  $j$ . Values of binomial coefficients are given in Table 2 at the end of this book and can be found by the formula

$${}_nC_j = \frac{n!}{j!(n - j)!}$$

where  $x!$  (“ $x$ -factorial”) is defined for any positive integer  $x$  as

$$x! = x(x - 1)(x - 2) \dots (2)(1)$$

and  $0! = 1$ . For more details, see Appendix 3.1.

For example, for  $n = 5$  the binomial coefficients are as follows:

$j:$	0	1	2	3	4	5
$_5C_j$ :	1	5	10	10	5	1

Thus, for  $n = 5$  the binomial probabilities are as indicated in Table 3.6.2. Notice the pattern in Table 3.6.2: The powers of  $p$  ascend  $(0, 1, 2, 3, 4, 5)$  and the powers of  $(1 - p)$  descend  $(5, 4, 3, 2, 1, 0)$ . (In using the binomial distribution formula, remember that  $x^0 = 1$  for any nonzero  $x$ .)

**Table 3.6.2** Binomial probabilities for  $n = 5$

Number of		
Successes $j$	Failures $n - j$	Probability
0	5	$1p^0(1 - p)^5$
1	4	$5p^1(1 - p)^4$
2	3	$10p^2(1 - p)^3$
3	2	$10p^3(1 - p)^2$
4	1	$5p^4(1 - p)^1$
5	0	$1p^5(1 - p)^0$

The following example shows a specific application of the binomial distribution with  $n = 5$ .

---

**Example  
3.6.4**

**Mutant Cats** Suppose we draw a random sample of five individuals from a large population in which 37% of the individuals are mutants (as in Example 3.6.2). The probabilities of the various possible samples are then given by the binomial distribution formula with  $n = 5$  and  $p = 0.37$ ; the results are displayed in Table 3.6.3. For instance, the probability of a sample containing 2 mutants and 3 nonmutants is

$$10(0.37)^2(0.63)^3 \approx 0.34$$

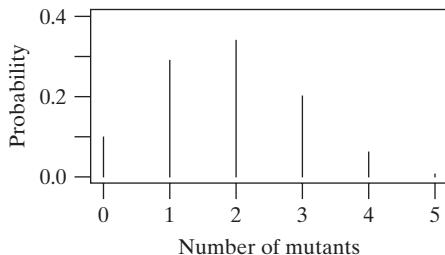
**Table 3.6.3** Binomial distribution with  $n = 5$  and  $p = 0.37$

Number of		
Mutants	Nonmutants	Probability
0	5	0.10
1	4	0.29
2	3	0.34
3	2	0.20
4	1	0.06
5	0	0.01
		1.00

Thus,  $\Pr\{Y = 3\} \approx 0.34$ . This means that about 34% of random samples of size 5 will contain two mutants and three nonmutants.

Notice that the probabilities in Table 3.6.3 add to 1. The probabilities in a probability distribution must always add to 1, because they account for 100% of the possibilities. ■

**Figure 3.6.2** Binomial distribution with  $n = 5$  and  $p = 0.37$



The binomial distribution of Table 3.6.3 is pictured graphically in Figure 3.6.2. The spikes in the graph emphasize that the probability distribution is discrete.

**Remark** In applying the independent-trials model and the binomial distribution, we assign the labels “success” and “failure” arbitrarily. For instance, in Example 3.6.4, we could say “success” = “mutant” and  $p = 0.37$ ; or, alternatively, we could say “success” = “nonmutant” and  $p = 0.63$ . Either assignment of labels is all right; it is only necessary to be consistent.

**Notes on Table 2** The following features in Table 2 are worth noting:

- (a) The first and last entries in each row are equal to 1. This will be true for any row; that is,  ${}_nC_0 = 1$  and  ${}_nC_n = 1$  for any value of  $n$ .
- (b) Each row of the table is symmetric; that is,  ${}_nC_j$  and  ${}_nC_{n-j}$  are equal.
- (c) The bottom rows of the table are left incomplete to save space, but you can easily complete them using the symmetry of the  ${}_nC_j$ 's; if you need to know  ${}_nC_j$  you can look up  ${}_nC_{n-j}$  in Table 2. For instance, consider  $n = 18$ ; if you want to know  ${}_{18}C_{15}$  you just look up  ${}_{18}C_3$ ; both  ${}_{18}C_3$  and  ${}_{18}C_{15}$  are equal to 816.

**Computational note** Computer and calculator technology makes it fairly easy to handle the binomial distribution formula for small or moderate values of  $n$ . For large values of  $n$ , the use of the binomial formula gets to be tedious and even a computer will balk at being asked to calculate a binomial probability. However, the binomial formula can be approximated by other methods. One of these will be discussed in the optional Section 5.5.

Sometimes a binomial probability question involves combining two or more possible outcomes. The following example illustrates this idea.

**Example  
3.6.5**

**Sampling Fruitflies** In a large *Drosophila* population, 30% of the flies are black (B) and 70% are gray (G). Suppose two flies are randomly chosen from the population (as in Example 3.2.3). The binomial distribution with  $n = 2$  and  $p = 0.3$  gives probabilities for the possible outcomes as shown in Table 3.6.4. (Using the binomial formula agrees with the results given by the probability tree shown in Figure 3.2.3.)

<b>Table 3.6.4</b>		
Sample composition	$Y$	Probability
Both G	0	0.49
One B, one G	1	0.42
Both B	2	0.09
		1.00

Let  $E$  be the event that both flies are the same color. Then  $E$  can happen in two ways: Both flies are gray or both are black. To find the probability of  $E$ , consider what would happen if we repeated the sampling procedure many times: Forty-nine

percent of the samples would have both flies gray, and 9% would have both flies black. Consequently, the percentage of samples with both flies the same color would be  $49\% + 9\% = 58\%$ . Thus, we have shown that the probability of  $E$  is

$$\Pr\{E\} = 0.58$$

as we claimed in Example 3.2.3. ■

Whenever an event  $E$  can happen in two or more mutually exclusive ways, a rationale such as that of Example 3.6.5 can be used to find  $\Pr\{E\}$ .

---

**Example  
3.6.6**

**Blood Type** In the United States, 85% of the population has Rh positive blood. Suppose we take a random sample of 6 persons and count the number with Rh positive blood. The binomial model can be applied here, since the BIInS conditions are met: There is a binary outcome on each trial (Rh positive or Rh negative blood), the trials are independent (due to the random sampling),  $n$  is fixed at 6, and the same probability of Rh positive blood applies to each person ( $p = 0.85$ ).

Let  $Y$  denote the number of persons, out of 6, with Rh positive blood. The probabilities of the possible values of  $Y$  are given by the binomial distribution formula with  $n = 6$  and  $p = 0.85$ ; the results are displayed in Table 3.6.5. For instance, the probability that  $Y = 4$  is

$${}_6C_4(0.85)^4(0.15)^2 \approx 15(0.522)(0.0225) \approx 0.1762$$

If we want to find the probability that at least 4 persons (out of the 6 sampled) will have Rh positive blood, we need to find  $\Pr\{Y \geq 4\} = \Pr\{Y = 4\} + \Pr\{Y = 5\} + \Pr\{Y = 6\} = 0.1762 + 0.3993 + 0.3771 = 0.9526$ . This means that the probability of getting at least 4 persons with Rh positive blood in a sample of size 6 is 0.9526. ■

**Table 3.6.5** Binomial distribution with  $n = 6$  and  $p = 0.85$

Number of successes	Probability
0	<0.0001
1	0.0004
2	0.0055
3	0.0415
4	0.1762
5	0.3993
6	0.3771
	1

In some problems, it is easier to find the probability that an event *does not happen* rather than finding the probability of the event happening. To solve such problems we use the fact that the probability of an event happening is 1 minus the probability that the event does not happen:  $\Pr\{E\} = 1 - \Pr\{E \text{ does not happen}\}$ . The following is an example.

---

**Example  
3.6.7**

**Blood Type** As in Example 3.6.6, let  $Y$  denote the number of persons, out of 6, with Rh positive blood. Suppose we want to find the probability that  $Y$  is less than 6 (i.e., the probability that there is *at least 1* person in the sample who has Rh *negative* blood). We could find this directly as  $\Pr\{Y = 0\} + \Pr\{Y = 1\} + \dots + \Pr\{Y = 5\}$ . However, it is easier to find  $\Pr\{Y \neq 6\}$  and subtract this from 1:

$$\Pr\{Y < 6\} = 1 - \Pr\{Y = 6\} = 1 - 0.3771 = 0.6229. ■$$

## Mean and Standard Deviation of a Binomial

If we toss a fair coin 10 times, then we expect to get 5 heads, on average. This is an example of a general rule: *For a binomial random variable, the mean (that is, the average number of successes) is equal to  $np$ .* This is an intuitive fact: The probability of success on each trial is  $p$ , so if we conduct  $n$  trials, then  $np$  is the expected number of successes. In Appendix 3.2 we show that this result is consistent with the rule given in Section 3.5 for finding the mean of the sum of random variables. *The standard deviation for a binomial random variable is given by  $\sqrt{np(1 - p)}$ .* This formula is not intuitively clear; a derivation of the result is given in Appendix 3.2. For the example of tossing a coin 10 times, the standard deviation of the number of heads is

$$\sqrt{10 \times 0.5 \times 0.5} = \sqrt{2.5} \approx 1.58.$$

### Example 3.6.8

**Blood Type** As discussed in Example 3.6.6, if  $Y$  denotes the number of persons with Rh positive blood in a sample of size 6, then a binomial model can be used to find probabilities associated with  $Y$ . The single most likely value of  $Y$  is 5 (which has probability 0.3993). The average value of  $Y$  is  $6 \times 0.85 = 5.1$ , which means that if we take many samples, each of size 6, and count the number of Rh positive persons in each sample, and then average those counts, we expect to get 5.1. The standard deviation of those counts is  $\sqrt{6 \times 0.85 \times .015} \approx 0.87$ . ■

## Applicability of the Binomial Distribution

A number of statistical procedures are based on the binomial distribution. We will study some of these procedures in later chapters. Of course, the binomial distribution is applicable only in experiments where the BInS conditions are satisfied in the real biological situation. We briefly discuss some aspects of these conditions.

**Application to Sampling** The most important application of the independent-trials model and the binomial distribution is to describe random sampling from a population when the observed variable is dichotomous—that is, a categorical variable with two categories (for instance, black and gray in Example 3.6.5). This application is valid if the sample size is a negligible fraction of the population size, so that the population composition is not altered appreciably by the removal of the individuals in the sample (so that the S part of BInS is satisfied: The probability of a success remains the same from trial to trial). However, if the sample is *not* a negligibly small part of the population, then the population composition may be altered by the sampling process, so that the “trials” involved in composing the sample are not independent and the probability of a success changes as the sampling progresses. In this case, the probabilities given by the binomial formula are not correct. In most biological studies, the population is so large that this kind of difficulty does not arise.

**Contagion** In some applications the phenomenon of contagion can invalidate the condition of independence between trials. The following is an example.

### Example 3.6.9

**Chickenpox** Consider the occurrence of chickenpox in children. Each child in a family can be categorized according to whether he had chickenpox during a certain year. One can say that each child constitutes a “trial” and that “success” is having chickenpox during the year, but the trials are *not* independent because the chance of a particular child catching chickenpox depends on whether his sibling caught chickenpox. As a specific example, consider a family with five children, and suppose that the

chance of an individual child catching chickenpox during the year is equal to 0.10. The binomial distribution gives the chance of all five children getting chickenpox as

$$\Pr\{5 \text{ children get chickenpox}\} = (0.10)^5 = 0.00001$$

However, this answer is not correct; because of contagion, the correct probability would be much larger. There would be many families in which one child caught chickenpox and then the other four children got chickenpox from the first child, so that all five children would get chickenpox. ■

## Exercises 3.6.1–3.6.10

**3.6.1** The seeds of the garden pea (*Pisum sativum*) are either yellow or green. A certain cross between pea plants produces progeny in the ratio 3 yellow:1 green.<sup>14</sup> If four randomly chosen progeny of such a cross are examined, what is the probability that

- (a) three are yellow and one is green?
- (b) all four are yellow?
- (c) all four are the same color?

**3.6.2** In the United States, 42% of the population has type A blood. Consider taking a sample of size 4. Let  $Y$  denote the number of persons in the sample with type A blood. Find

- (a)  $\Pr\{Y = 0\}$ .
- (b)  $\Pr\{Y = 1\}$ .
- (c)  $\Pr\{Y = 2\}$ .
- (d)  $\Pr\{0 \leq Y \leq 2\}$ .
- (e)  $\Pr\{0 < Y \leq 2\}$ .

**3.6.3** A certain drug treatment cures 90% of cases of hookworm in children.<sup>15</sup> Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that

- (a) all 20 will be cured.
- (b) all but 1 will be cured.
- (c) exactly 18 will be cured.
- (d) exactly 90% will be cured.

**3.6.4** The shell of the land snail *Limocolaria martensiana* has two possible color forms: streaked and pallid. In a certain population of these snails, 60% of the individuals have streaked shells.<sup>16</sup> Suppose that a random sample of 10 snails is to be chosen from this population. Find the probability that the percentage of streaked-shelled snails in the sample will be

- (a) 50%.      (b) 60%.      (c) 70%.

**3.6.5** Consider taking a sample of size 10 from the snail population in Exercise 3.6.4.

- (a) What is the mean number of streaked-shelled snails?
- (b) What is the standard deviation of the number of streaked-shelled snails?

**3.6.6** The sex ratio of newborn human infants is about 105 males:100 females.<sup>17</sup> If four infants are chosen at random, what is the probability that

- (a) two are male and two are female?
- (b) all four are male?
- (c) all four are the same sex?

**3.6.7** Construct a binomial setting (different from any examples presented in this book) and a problem for which the following is the answer:  ${}_7C_3(0.8)^3(0.2)^5$ .

**3.6.8** Neuroblastoma is a rare, serious, but treatable disease. A urine test, the VMA test, has been developed that gives a positive diagnosis in about 70% of cases of neuroblastoma.<sup>18</sup> It has been proposed that this test be used for large-scale screening of children. Assume that 300,000 children are to be tested, of whom 8 have the disease. We are interested in whether or not the test detects the disease in the 8 children who have the disease. Find the probability that

- (a) all eight cases will be detected.
- (b) only one case will be missed.
- (c) two or more cases will be missed. [Hint: Use parts (a) and (b) to answer part (c).]

**3.6.9** If two carriers of the gene for albinism marry, each of their children has probability  $\frac{1}{4}$  of being albino (see Example 3.6.1). If such a couple has six children, what is the probability that

- (a) none will be albino?
- (b) at least one will be albino? [Hint: Use part (a) to answer part (b); note that “at least one” means “one or more.”]

**3.6.10** Childhood lead poisoning is a public health concern in the United States. In a certain population, 1 child in 8 has a high blood lead level (defined as 30 µg/dl or more).<sup>19</sup> In a randomly chosen group of 16 children from the population, what is the probability that

- (a) none has high blood lead?
- (b) 1 has high blood lead?
- (c) 2 have high blood lead?
- (d) 3 or more have high blood lead? [Hint: Use parts (a)–(c) to answer part (d).]

### 3.7 Fitting a Binomial Distribution to Data (Optional)

Occasionally it is possible to obtain data that permit a direct check of the applicability of the binomial distribution. One such case is described in the next example.

**Example  
3.7.1**

**Sexes of Children** In a classic study of the human sex ratio, families were categorized according to the sexes of the children. The data were collected in Germany in the nineteenth century, when large families were common. Table 3.7.1 shows the results for 6,115 families with 12 children.<sup>20</sup>

It is interesting to consider whether the observed variation among families can be explained by the independent-trials model. We will explore this question by fitting a binomial distribution to the data.

**Table 3.7.1** Sex ratios in 6,115 families with twelve children

Boys	Number of Girls	Observed frequency (number of families)
0	12	3
1	11	24
2	10	104
3	9	286
4	8	670
5	7	1,033
6	6	1,343
7	5	1,112
8	4	829
9	3	478
10	2	181
11	1	45
12	0	7
		6,115

The first step in fitting the binomial distribution is to determine a value for  $p = \Pr\{\text{boy}\}$ . One possibility would be to assume that  $p = 0.50$ . However, since it is known that the human sex ratio at birth is not exactly 1:1 (in fact, it favors boys slightly), we will not make this assumption. Rather, we will “fit”  $p$  to the data; that is, we will determine a value for  $p$  that fits the data best. We observe that the total number of children in all the families is

$$(12)(6,115) = 73,380 \text{ children}$$

Among these children, the number of boys is

$$(3)(0) + (24)(1) + \dots + (12)(7) = 38,100 \text{ boys}$$

Therefore, the value of  $p$  that fits the data best is

$$p = \frac{38,100}{73,380} = 0.519215$$

The next step is to compute probabilities from the binomial distribution formula with  $n = 12$  and  $p = 0.519215$ . For instance, the probability of 3 boys and 9 girls is computed as

$$\begin{aligned} {}_{12}C_3(p)^3(1 - p)^9 &= 220(0.519215)^3(0.480785)^9 \\ &\approx 0.042269 \end{aligned}$$

For comparison with the observed data, we convert each probability to a theoretical or “expected” frequency by multiplying by 6,115 (the total number of families). For instance, the expected number of families with 3 boys and 9 girls is

$$(6,115)(0.042269) \approx 258.5$$

The expected and observed frequencies are displayed together in Table 3.7.2. Table 3.7.2 shows reasonable agreement between the observed frequencies and the predictions of the binomial distribution. But a closer look reveals that the discrepancies, although not large, follow a definite pattern. The data contain more unisexual, or preponderantly unisexual, sibships than expected. In fact, the observed frequencies are higher than the expected frequencies for nine types of families in which one sex or the other predominates, while the observed frequencies are lower than the expected frequencies for four types of more “balanced” families. This pattern is clearly revealed by the last column of Table 3.7.2, which shows the sign of the difference between the observed frequency and the expected frequency. Thus, the observed distribution of sex ratios has heavier “tails” and a lighter “middle” than the best-fitting binomial distribution.

The systematic pattern of deviations from the binomial distribution suggests that the observed variation among families cannot be entirely explained by the independent-trials model.\* What factors might account for the discrepancy?

**Table 3.7.2** Sex-ratio data and binomial expected frequencies

Number of Boys	Number of Girls	Observed frequency	Expected frequency	Sign of (Obs. – Exp.)
0	12	3	0.9	+
1	11	24	12.1	+
2	10	104	71.8	+
3	9	286	258.5	+
4	8	670	628.1	+
5	7	1,033	1,085.2	-
6	6	1,343	1,367.3	-
7	5	1,112	1,265.6	-
8	4	829	854.3	-
9	3	478	410.0	+
10	2	181	132.8	+
11	1	45	26.1	+
12	0	7	2.3	+
		6,115	6,115.0	

\*A chi-square goodness-of-fit test of the binomial model shows that there is strong evidence that the differences between the observed and expected frequencies did not happen due to chance error in the sampling process. We will explore the topic of goodness-of-fit tests in Chapter 9.

This intriguing question has stimulated several researchers to undertake more detailed analysis of these data. We briefly discuss some of the issues.

One explanation for the excess of predominantly unisexual families is that the probability of producing a boy may vary among families. If  $p$  varies from one family to another, then sex will appear to “run” in families in the sense that the number of predominantly unisexual families will be inflated. In order to clearly visualize this effect, consider the fictitious data set shown in Table 3.7.3.

**Table 3.7.3** Fictitious sex-ratio data and binomial expected frequencies

Boys	Girls	Observed frequency	Expected frequency	Sign of (Obs. – Exp.)
0	12	2,940	0.9	+
1	11	0	12.1	-
2	10	0	71.8	-
3	9	0	258.5	-
4	8	0	628.1	-
5	7	0	1,085.2	-
6	6	0	1,367.3	-
7	5	0	1,265.6	-
8	4	0	854.3	-
9	3	0	410.0	-
10	2	0	132.8	-
11	1	0	26.1	-
12	0	3,175	2.3	+
		6,115	6,115.0	

In the fictitious data set, there are  $(3,175)(12) = 38,100$  males among 73,380 children, just as there are in the real data set. Consequently, the best-fitting  $p$  is the same ( $p = 0.519215$ ) and the expected binomial frequencies are the same as in Table 3.7.2. The fictitious data set contains only unisexual sibships and so is an extreme example of sex “running” in families. The real data set exhibits the same phenomenon more weakly. One explanation of the fictitious data set would be that some families can have only boys ( $p = 1$ ) and other families can have only girls ( $p = 0$ ). In a parallel way, one explanation of the real data set would be that  $p$  varies slightly among families. Variation in  $p$  is biologically plausible, even though the mechanism causing the variation has not yet been discovered.

An alternative explanation for the inflated number of sexually homogeneous families would be that the sexes of the children in a family are literally dependent on one another, in the sense that the determination of an individual child’s sex is somehow influenced by the sexes of the previous children. This explanation is implausible on biological grounds because it is difficult to imagine how the biological system could “remember” the sexes of previous offspring. ■

Example 3.7.1 shows that poorness of fit to the independent-trials model can be biologically interesting. We should emphasize, however, that most statistical applications of the binomial distribution proceed from the assumption that the independent-trials model is applicable. In a typical application, the data are regarded as resulting from a *single* set of  $n$  trials. Data such as the family sex-ratio data, which refer to *many* sets of  $n = 12$  trials, are not often encountered.

## Exercises 3.7.1–3.7.3

**3.7.1** The accompanying data on families with 6 children are taken from the same study as the families with 12 children in Example 3.7.1. Fit a binomial distribution to the data. (Round the expected frequencies to one decimal place.) Compare with the results in Example 3.7.1. What features do the two data sets share?

NUMBER OF		NUMBER OF FAMILIES
BOYS	GIRLS	
0	6	1,096
1	5	6,233
2	4	15,700
3	3	22,221
4	2	17,332
5	1	7,908
6	0	1,579
		72,069

**3.7.2** An important method for studying mutation-causing substances involves killing female mice 17 days after mating and examining their uteri for living and dead embryos. The classical method of analysis of such data assumes that the survival or death of each embryo constitutes an independent binomial trial. The accompanying table, which is extracted from a larger study, gives data for 310 females, all of whose uteri contained 9 embryos; all of the animals were treated alike (as controls).<sup>21</sup>

NUMBER OF EMBRYOS		NUMBER OF FEMALE MICE
DEAD	LIVING	
0	9	136
1	8	103
2	7	50
3	6	13
4	5	6
5	4	1
6	3	1
7	2	0
8	1	0
9	0	0
		310

- (a) Fit a binomial distribution to the observed data. (Round the expected frequencies to one decimal place.)
- (b) Interpret the relationship between the observed and expected frequencies. Do the data cast suspicion on the classical assumption?

**3.7.3** Students in a large botany class conducted an experiment on the germination of seeds of the Saguaro cactus. As part of the experiment, each student planted five seeds in a small cup, kept the cup near a window, and checked every day for germination (sprouting). The class results on the seventh day after planting were as displayed in the table.<sup>22</sup>

GERMINATED	NUMBER OF SEEDS		NUMBER OF STUDENTS
	NOT GERMINATED		
0	5		17
1	4		53
2	3		94
3	2		79
4	1		33
5	0		4
			280

- (a) Fit a binomial distribution to the data. (Round the expected frequencies to one decimal place.)
- (b) Two students, Fran and Bob, were talking before class. All of Fran's seeds had germinated by the seventh day, whereas none of Bob's had. Bob wondered whether he had done something wrong. With the perspective gained from seeing all 280 students' results, what would you say to Bob? (*Hint:* Can the variation among the students be explained by the hypothesis that some of the seeds were good and some were poor, with each student receiving a randomly chosen five seeds?)
- (c) Invent a fictitious set of data for 280 students, with the same overall percentage germination as the observed data given in the table, but with all the students getting either Fran's results (perfect) or Bob's results (nothing). How would your answer to Bob differ if the actual data had looked like this fictitious data set?

## Supplementary Exercises 3.S.1–3.S.10

**3.S.1** In the United States, 10% of adolescent girls have iron deficiency.<sup>23</sup> Suppose two adolescent girls are chosen at random. Find the probability that

- (a) both girls have iron deficiency.
- (b) one girl has iron deficiency and the other does not.

**3.S.2** In preparation for an ecological study of centipedes, the floor of a beech woods is divided into a large number of 1-foot squares.<sup>24</sup> At a certain moment, the distribution of centipedes in the squares is as shown in the table.

NUMBER OF CENTIPEDES	PERCENT FREQUENCY (% OF SQUARES)
0	45
1	36
2	14
3	4
4	1
	100

Suppose that a square is chosen at random, and let  $Y$  be the number of centipedes in the chosen square. Find

$$(a) \Pr\{Y = 1\} \quad (b) \Pr\{Y \geq 2\}$$

**3.S.3** Refer to the distribution of centipedes given in Exercise 3.S.2. Suppose five squares are chosen at random. Find the probability that three of the squares contain centipedes and two do not.

**3.S.4** Refer to the distribution of centipedes given in Exercise 3.S.2. Suppose five squares are chosen at random. Find the expected value (i.e., the mean) of the number of squares that contain at least one centipede.

**3.S.5** Wavy hair in mice is a recessive genetic trait. If mice with wavy hair are mated with straight-haired (heterozygous) mice, each offspring has probability  $\frac{1}{2}$  of having wavy hair.<sup>25</sup> Consider a large number of such matings, each producing a litter of five offspring. What percentage of the litters will consist of

- (a) two wavy-haired and three straight-haired offspring?
- (b) three or more straight-haired offspring?
- (c) all the same type (either all wavy- or all straight-haired) offspring?

**3.S.6** A certain drug causes kidney damage in 1% of patients. Suppose the drug is to be tested on 50 patients. Find the probability that

- (a) none of the patients will experience kidney damage.
- (b) one or more of the patients will experience kidney damage. [Hint: Use part (a) to answer part (b).]

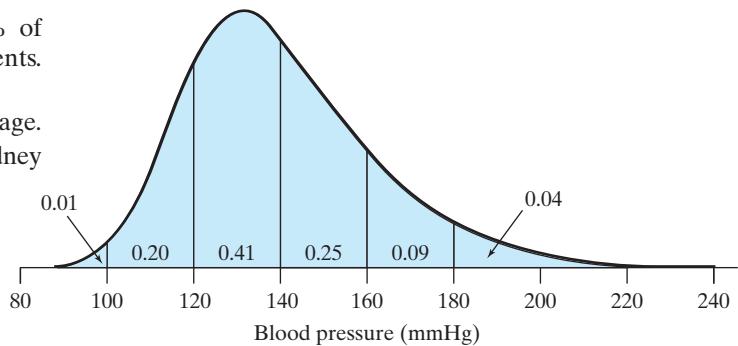
**3.S.7** Refer to Exercise 3.S.6. Suppose now that the drug is to be tested on  $n$  patients, and let  $E$  represent the event that kidney damage occurs in one or more of the patients. The probability  $\Pr\{E\}$  is useful in establishing criteria for drug safety.

- (a) Find  $\Pr\{E\}$  for  $n = 100$ .
- (b) How large must  $n$  be in order for  $\Pr\{E\}$  to exceed 0.95?

**3.S.8** To study people's ability to deceive lie detectors, researchers sometimes use the "guilty knowledge" technique.<sup>26</sup> Certain subjects memorize six common words; other subjects memorize no words. Each subject is then tested on a polygraph machine (lie detector), as follows. The experimenter reads, in random order, 24 words: the six "critical" words (the memorized list) and, for each critical word, three "control" words with similar or related meanings. If the subject has memorized the six words, he or she tries to conceal that fact. The subject is scored a "failure" on a critical word if his or her electrodermal response is higher on the critical word than on any of the three control words. Thus, on each of the six critical words, even an innocent subject would have a 25% chance of failing. Suppose a subject is labeled "guilty" if the subject fails on four or more of the six critical words. If an innocent subject is tested, what is the probability that he or she will be labeled "guilty"?

**3.S.9** The density curve shown here represents the distribution of systolic blood pressures in a population of middle-aged men.<sup>27</sup> Areas under the curve are shown in the figure. Suppose a man is selected at random from the population, and let  $Y$  be his blood pressure. Find

- (a)  $\Pr\{120 < Y < 160\}$ .
- (b)  $\Pr\{Y < 120\}$ .
- (c)  $\Pr\{Y > 140\}$ .



**3.S.10** Refer to the blood pressure distribution of Exercise 3.S.9. Suppose four men are selected at random from the population. Find the probability that

- (a) all four have blood pressures higher than 140 mm Hg.
- (b) three have blood pressures higher than 140, and one has blood pressure 140 or less.

# THE NORMAL DISTRIBUTION

## Objectives

In this chapter we will study the normal distribution, including

- the use of the normal curve in modeling distributions.
- finding probabilities using the normal curve.
- assessing normality of data sets with the use of normal probability plots.

### 4.1 Introduction

In Chapter 2, we introduced the idea of regarding a set of data as a sample from a population. In Section 3.4 we saw that the population distribution of a quantitative variable  $Y$  can be described by its mean  $\mu$  and its standard deviation  $\sigma$  and also by a density curve, which represents relative frequencies as areas under the curve. In this chapter we study the most important type of density curve: the **normal curve**. The normal curve is a symmetric “bell-shaped” curve whose exact form we will describe next. A distribution represented by a normal curve is called a **normal distribution**.

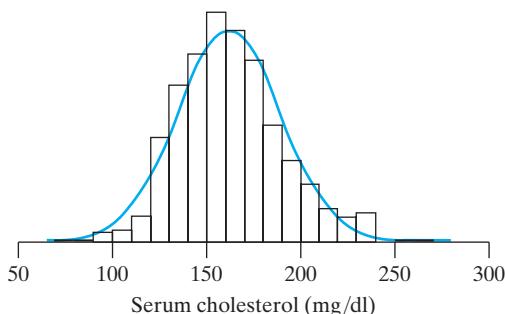
The family of normal distributions plays two roles in statistical applications. Its more straightforward use is as a convenient approximation to the distribution of an observed variable  $Y$ . The second role of the normal distribution is more theoretical and will be explored in Chapter 5.

An example of a natural population distribution that can be approximated by a normal distribution follows.

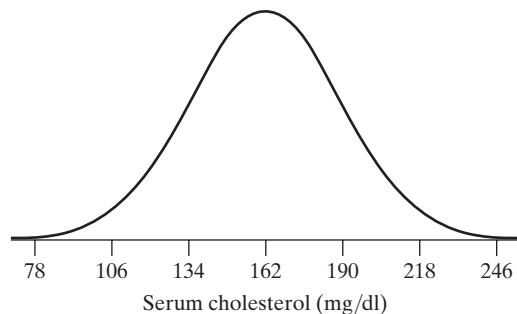
#### Example 4.1.1

**Serum Cholesterol** The relationship between the concentration of cholesterol in the blood and the occurrence of heart disease has been the subject of much research. As part of a government health survey, researchers measured serum cholesterol levels for a large sample of Americans including children. The distribution for children between 12 and 14 years of age can be fairly well approximated by a normal curve with mean  $\mu = 162$  mg/dl and standard deviation  $\sigma = 28$  mg/dl. Figure 4.1.1 shows a histogram based on a sample of 727 children between 12 and 14 years old, with the normal curve superimposed.<sup>1</sup> ■

To indicate how the mean  $\mu$  and standard deviation  $\sigma$  relate to the normal curve, Figure 4.1.2 shows the normal curve for the serum cholesterol distribution of Example 4.1.1, with tick marks at 1, 2, and 3 standard deviations from the mean.



**Figure 4.1.1** Distribution of serum cholesterol in 727 12- to 14-year-old children



**Figure 4.1.2** Normal distribution of serum cholesterol, with  $\mu = 162$  mg/dl and  $\sigma = 28$  mg/dl

The normal curve can be used to describe the distribution of an observed variable  $Y$  in two ways: (1) as a smooth approximation to a histogram based on a sample of  $Y$  values; and (2) as an idealized representation of the population distribution of  $Y$ . The normal curves in Figures 4.1.1 and 4.1.2 could be interpreted either way. For simplicity, in the remainder of this chapter we will consider the normal curve as representing a population distribution.

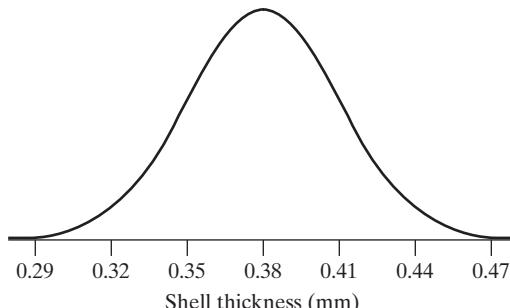
## Further Examples

We now give three more examples of normal curves that approximately describe real populations. In each figure, the horizontal axis is scaled with tick marks centered at the mean and one standard deviation apart.

---

### Example 4.1.2

**Eggshell Thickness** In the commercial production of eggs, breakage is a major problem. Consequently, the thickness of the eggshell is an important variable. In one study, the shell thicknesses of the eggs produced by a large flock of White Leghorn hens were observed to follow approximately a normal distribution with mean  $\mu = 0.38$  mm and standard deviation  $\sigma = 0.03$  mm. This distribution is pictured in Figure 4.1.3.<sup>2</sup>

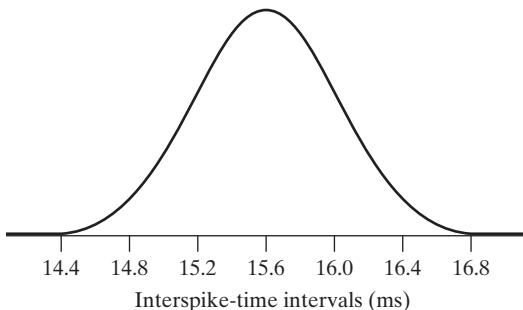


**Figure 4.1.3** Normal distribution of eggshell thickness, with  $\mu = 0.38$  mm and  $\sigma = 0.03$  mm

---

### Example 4.1.3

**Interspike Times in Nerve Cells** In certain nerve cells, spontaneous electrical discharges are observed that are so rhythmically repetitive that they are called “clock-spikes.” The timing of these spikes, even though remarkably regular, does exhibit variation. In one study, the interspike-time intervals (in milliseconds) for a single housefly (*Musca domestica*) were observed to follow approximately a normal distribution with mean  $\mu = 15.6$  ms and standard deviation  $\sigma = 0.4$  ms; this distribution is shown in Figure 4.1.4.<sup>3</sup>

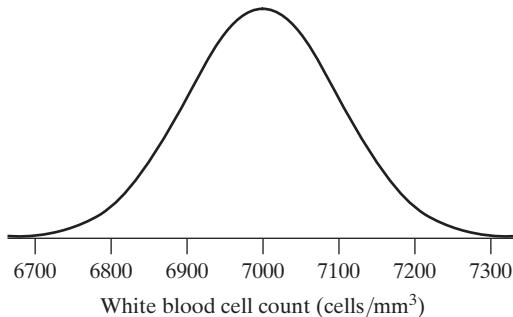


**Figure 4.1.4** Normal distribution of interspike-time intervals, with  $\mu = 15.6$  ms and  $\sigma = 0.4$  ms

The preceding examples have illustrated very different kinds of populations. In Example 4.1.3, the entire population consists of measurements on only one fly. Still another type of population is a *measurement error* population, consisting of repeated measurements of exactly the same quantity. The deviation of an individual measurement from the “correct” value is called measurement error. Measurement error is not the result of a mistake but rather is due to lack of perfect precision in the measuring process or measuring instrument. Measurement error distributions are often approximately normal; in this case the mean of the distribution of repeated measurements of the same quantity is the true value of the quantity (assuming that the measuring instrument is correctly calibrated), and the standard deviation of the distribution indicates the precision of the instrument. One measurement error distribution was described in Example 2.2.12. The following is another example.

**Example 4.1.4**

**Measurement Error** When a certain electronic instrument is used for counting particles such as white blood cells, the measurement error distribution is approximately normal. For white blood cells, the standard deviation of repeated counts based on the same blood specimen is about 1.4% of the true count. Thus, if the true count of a certain blood specimen were  $7,000 \text{ cells/mm}^3$ , then the standard deviation would be about  $100 \text{ cells/mm}^3$  and the distribution of repeated counts on that specimen would resemble Figure 4.1.5. ■



**Figure 4.1.5** Normal distribution of repeated white blood cell counts of a blood specimen whose true value is  $\mu = 7000 \text{ cells/mm}^3$ . The standard deviation is  $\sigma = 100 \text{ cells/mm}^3$ .

## 4.2 The Normal Curves

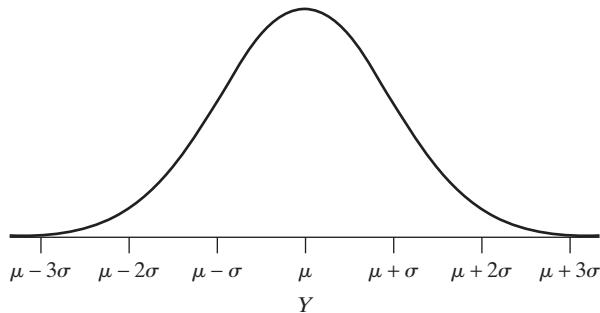
As the examples in Section 4.1 show, there are many normal curves; each particular normal curve is characterized by its mean and standard deviation. If a variable  $Y$  follows a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then it is common to write  $Y \sim N(\mu, \sigma)$ . All the normal curves can be described by a single formula. Even though we will not make any direct use of the formula in this book, we present it here, both as a matter of interest and also to emphasize that a normal curve is not just any symmetric curve, but rather a *specific* kind of symmetric curve.

If a variable  $Y$  follows a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then the density curve of the distribution of  $Y$  is given by the following formula:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

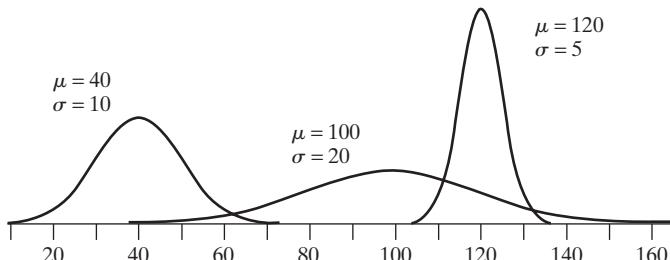
This function,  $f(y)$ , is called the *density function* of the distribution and expresses the height of the curve as a function of the position  $y$  along the  $y$ -axis. The quantities  $e$  and  $\pi$  that appear in the formula are constants, with  $e$  approximately equal to 2.71 and  $\pi$  approximately equal to 3.14.

Figure 4.2.1 shows a graph of a normal curve. The shape of the curve is like a symmetric bell, centered at  $y = \mu$ . The direction of curvature is downward (like an inverted bowl) in the central portion of the curve, and upward in the tail portions. The points of inflection (i.e., where the curvature changes direction) are  $y = \mu - \sigma$  and  $y = \mu + \sigma$ ; notice that the curve is almost linear near these points. In principle the curve extends to  $+\infty$  and  $-\infty$ , never actually reaching the  $y$ -axis; however, the height of the curve is very small for  $y$  values more than three standard deviations from the mean. The area under the curve is exactly equal to 1. (Note: It may seem paradoxical that a curve can enclose a finite area, even though it never descends to touch the  $y$ -axis. This apparent paradox is clarified in Appendix 4.1.)



**Figure 4.2.1** A normal curve with mean  $\mu$  and standard deviation  $\sigma$

All normal curves have the same essential shape, in the sense that they can be made to look identical by suitable choice of the vertical and horizontal scales for each. (For instance, notice that the curves in Figures 4.1.2–4.1.5 look identical.) But normal curves with different values of  $\mu$  and  $\sigma$  will not look identical if they are all plotted to the same scale, as illustrated by Figure 4.2.2. The location of the normal curve along the  $y$ -axis is governed by  $\mu$  since the curve is centered at  $y = \mu$ ; the width of the curve is governed by  $\sigma$ . The height of the curve is also determined by  $\sigma$ . Since the area under each curve must be equal to 1, a curve with a smaller value of  $\sigma$  must be taller. This reflects the fact that the values of  $Y$  are more highly concentrated near the mean when the standard deviation is smaller.



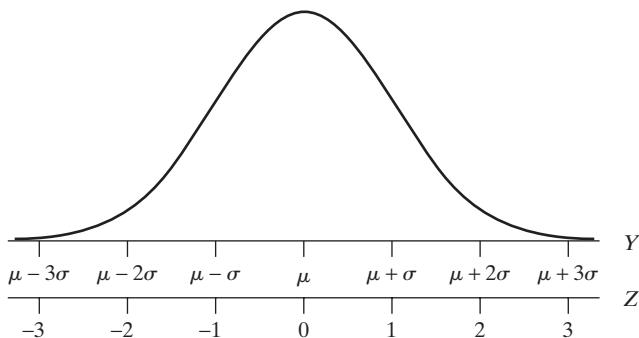
**Figure 4.2.2** Three normal curves with different means and standard deviations

## 4.3 Areas under a Normal Curve

As explained in Section 3.4, a density curve can be quantitatively interpreted in terms of areas under the curve. While areas can be roughly estimated by eye, for some purposes it is desirable to have fairly precise information about areas.

### The Standardized Scale

The areas under a normal curve have been computed mathematically and are tabulated here for practical use. The use of this tabulated information is much simplified by the fact that all normal curves can be made equivalent with respect to areas under them by suitable rescaling of the horizontal axis. The rescaled variable is denoted by  $Z$ ; the relationship between the two scales is shown in Figure 4.3.1.



**Figure 4.3.1** A normal curve, showing the relationship between the natural scale ( $Y$ ) and the standardized scale ( $Z$ )

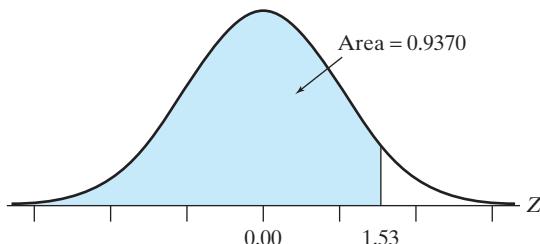
As Figure 4.3.1 indicates, the  $Z$  scale measures standard deviations from the mean:  $z = 1.0$  corresponds to 1.0 standard deviation above the mean;  $z = -2.5$  corresponds to 2.5 standard deviations below the mean, and so on. The  $Z$  scale is referred to as a **standardized scale**.

The correspondence between the  $Z$  scale and the  $Y$  scale can be expressed by the formula given in the following box.

#### Standardization Formula

$$Z = \frac{Y - \mu}{\sigma}$$

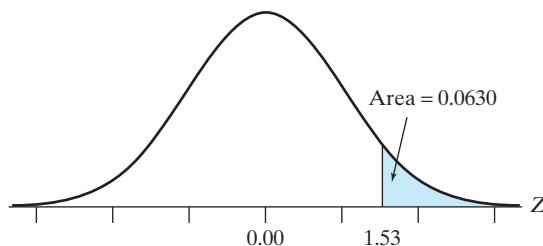
The variable  $Z$  is referred to as the **standard normal** and its distribution follows a normal curve with mean zero and standard deviation one. Table 3 at the end of this book gives areas under the standard normal curve, with distances along the horizontal axis measured in the  $Z$  scale. Each area tabled in Table 3 is the area under the standard normal curve below a specified value of  $z$ . For example, for  $z = 1.53$ , the tabled area is 0.9370; this area is shaded in Figure 4.3.2.



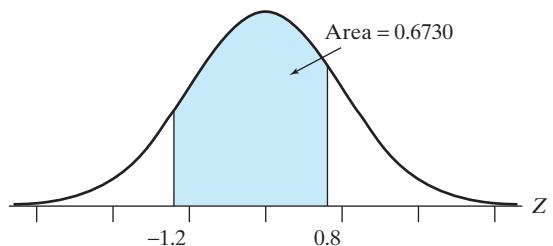
**Figure 4.3.2** Illustration of the use of Table 3

If we want to find the area above a given value of  $z$ , we subtract the tabulated area from 1. For example, the area above  $z = 1.53$  is  $1.0000 - 0.9370 = 0.0630$  (Figure 4.3.3).

To find the area between two  $z$  values (also commonly called  **$z$  scores**) we can subtract the areas given in Table 3. For example, to find the area under the  $Z$  curve between  $z = -1.2$  and  $z = 0.8$  (Figure 4.3.4), we take the area below 0.8, which is 0.7881, and subtract the area below  $-1.2$ , which is 0.1151, to get  $0.7881 - 0.1151 = 0.6730$ .

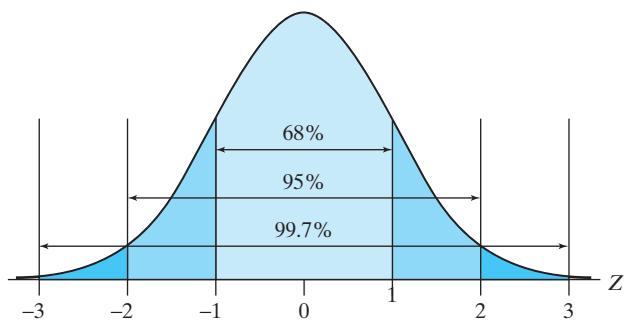


**Figure 4.3.3** Area under a standard normal curve above 1.53

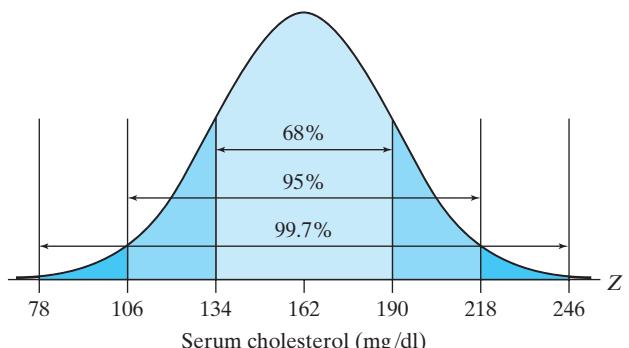


**Figure 4.3.4** Area under a standard normal curve between  $-1.2$  and  $0.8$

Using Table 3, we see that the area under the normal curve between  $z = -1$  and  $z = +1$  is  $0.8413 - 0.1578 = 0.6826$ . Thus, for any normal distribution, about 68% of the observations are within  $\pm 1$  standard deviation of the mean. Likewise, the area under the normal curve between  $z = -2$  and  $z = +2$  is  $0.9772 - 0.0228 = 0.9544$  and the area under the normal curve between  $z = -3$  and  $z = +3$  is  $0.9987 - 0.0013 = 0.9974$ . This means that for any normal distribution about 95% of the observations are within  $\pm 2$  standard deviations of the mean and about 99.7% of the observations are within  $\pm 3$  standard deviations of the mean. (See Figure 4.3.5.) For example, about 68% of the serum cholesterol values in the idealized distribution of Figure 4.1.2 are between 134 mg/dl and 190 mg/dl, about 95% are between 106 mg/dl and 218 mg/dl, and virtually all are between 78 mg/dl and 246 mg/dl. Figure 4.3.6 shows these percentages.



**Figure 4.3.5** Areas under a standard normal curve between  $-1$  and  $+1$ , between  $-2$  and  $+2$ , and between  $-3$  and  $+3$



**Figure 4.3.6** The 68/95/99.7 rule and the serum cholesterol distribution

- If the variable  $Y$  follows a normal distribution, then
- about 68% of the  $y$ 's are within  $\pm 1$  SD of the mean.
  - about 95% of the  $y$ 's are within  $\pm 2$  SDs of the mean.
  - about 99.7% of the  $y$ 's are within  $\pm 3$  SDs of the mean.

These statements provide a very definite interpretation of the standard deviation in cases where a distribution is approximately normal. (In fact, the statements are often approximately true for moderately nonnormal distributions; that is why, in Section 2.6, these percentages—68%, 95%, and >99%—were described as “typical” for “nicely shaped” distributions.)

## Determining Areas for a Normal Curve

By taking advantage of the standardized scale, we can use Table 3 to answer detailed questions about any normal population when the population mean and standard deviation are specified. The following example illustrates the use of Table 3. (Of course, the population described in the example is an idealized one, since no actual population follows a normal distribution *exactly*.)

### Example 4.3.1

**Lengths of Fish** In a certain population of the herring *Pomolobus aestivalis*, the lengths of the individual fish follow a normal distribution. The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm.<sup>5</sup> We will use Table 3 to answer various questions about the population.

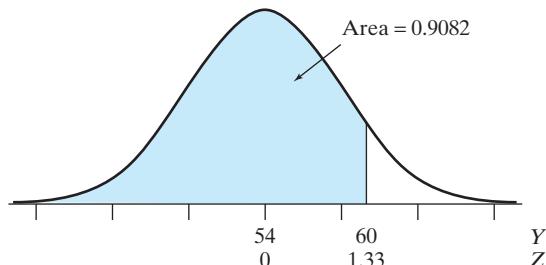
- (a) What percentage of the fish are less than 60 mm long?

Figure 4.3.7 shows the population density curve, with the desired area indicated by shading. In order to use Table 3, we convert the limits of the area from the  $Y$  scale to the  $Z$  scale, as follows:

For  $y = 60$ , the  $z$  score is

$$z = \frac{y - \mu}{\sigma} = \frac{60 - 54}{4.5} = 1.33$$

Thus, the question “What percentage of the fish are less than 60 mm long?” is equivalent to the question “What is the area under the standard normal curve below the  $z$  value of 1.33?” Looking up  $z = 1.33$  in Table 3, we find that the area is 0.9082; thus, 90.82% of the fish are less than 60 mm long.



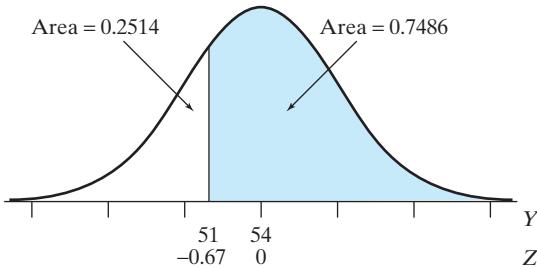
**Figure 4.3.7** Area under the normal curve in Example 4.3.1(a)

- (b) What percentage of the fish are more than 51 mm long?

The standardized value for  $y = 51$  is

$$z = \frac{y - \mu}{\sigma} = \frac{51 - 54}{4.5} = -0.67$$

Thus, the question “What percentage of the fish are more than 51 mm long?” is equivalent to the question “What is the area under the standard normal curve above the  $z$  value of -0.67?” Figure 4.3.8 shows this relationship. Look-



**Figure 4.3.8** Area under the normal curve in Example 4.3.1(b)

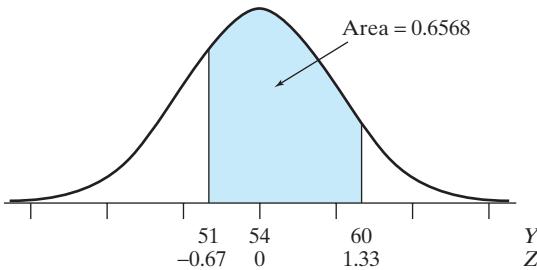
ing up  $z = -0.67$  in Table 3, we find that the area below  $z = -0.67$  is 0.2514. This means that the area above  $z = -0.67$  is  $1 - 0.2514 = 0.7486$ . Thus, 74.86% of the fish are more than 51 mm long.

- (c) What percentage of the fish are between 51 and 60 mm long?

Figure 4.3.9 shows the desired area. This area can be expressed as a difference of two areas found from Table 3. The area below  $y = 60$  is 0.9082, as found in part (a), and the area below  $y = 51$  is 0.2514, as found in part (b). Consequently, the desired area is computed as

$$0.9082 - 0.2514 = 0.6568$$

Thus, 65.68% of the fish are between 51 and 60 mm long.



**Figure 4.3.9** Area under the normal curve in Example 4.3.1(c)

- (d) What percentage of the fish are between 58 and 60 mm long?

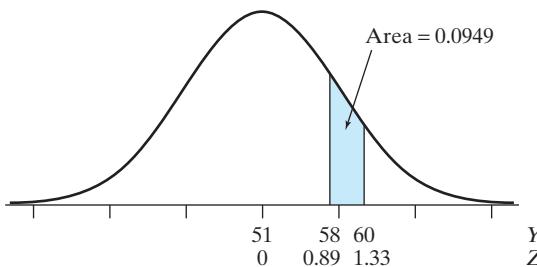
Figure 4.3.10 shows the desired area. This area can be expressed as a difference of two areas found from Table 3. The area below  $y = 60$  is 0.9082, as was found in part (a). To find the area below  $y = 58$ , we first calculate the  $z$  value that corresponds to  $y = 58$ :

$$z = \frac{y - \mu}{\sigma} = \frac{58 - 54}{4.5} = 0.89$$

The area under the  $Z$  curve below  $z = 0.89$  is 0.8133. Consequently, the desired area is computed as

$$0.9082 - 0.8133 = 0.0949$$

Thus, 9.49% of the fish are between 58 and 60 mm long. ■



**Figure 4.3.10** Area under the normal curve in Example 4.3.1(d)

Each of the percentages found in Example 4.3.1 can also be interpreted in terms of probability. Let the random variable  $Y$  represent the length of a fish randomly chosen from the population. Then the results in Example 4.3.1 imply that

$$\Pr\{Y < 60\} = 0.9082$$

$$\Pr\{Y > 51\} = 0.7486$$

$$\Pr\{51 < Y < 60\} = 0.6568$$

and

$$\Pr\{58 < Y < 60\} = 0.0949$$

Thus, the normal distribution can be interpreted as a continuous probability distribution.

Note that because the idealized normal distribution is perfectly continuous, probabilities such as

$$\Pr\{Y > 48\} \text{ and } \Pr\{Y \geq 48\}$$

are equal (see Section 3.4). That is,

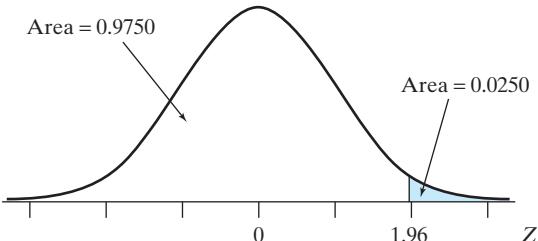
$$\begin{aligned}\Pr\{Y \geq 48\} &= \Pr\{Y > 48\} + \Pr\{Y = 48\} \\ &= \Pr\{Y > 48\} + 0 \text{ (since } Y \text{ is taken to be continuous)} \\ &= \Pr\{Y > 48\}\end{aligned}$$

If, however, the length were measured only to the nearest mm, then the measured variable would actually be discrete, so that  $\Pr\{Y > 48\}$  and  $\Pr\{Y \geq 48\}$  would differ somewhat from each other. In cases where this discrepancy is important, the computation can be refined to take into account the discontinuity of the measured distribution (we will later see such an example in Section 5.4).

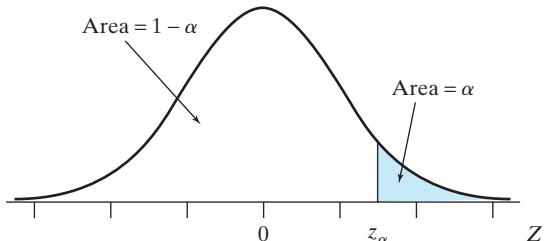
### Inverse Reading of Table 3

In determining facts about a normal distribution, it is sometimes necessary to read Table 3 in an “inverse” way—that is, to find the value of  $z$  corresponding to a given area rather than the other way around. For example, suppose we want to find the value on the  $Z$  scale that cuts off the top 2.5% of the distribution. This number is 1.96, as shown in Figure 4.3.11.

We will find it helpful, for future reference, to introduce some notation. We will use the notation  $z_\alpha$  to denote the number such that  $\Pr\{Z < z_\alpha\} = 1 - \alpha$  and  $\Pr\{Z > z_\alpha\} = \alpha$ , as shown in Figure 4.3.12. Thus,  $z_{0.025} = 1.96$ .



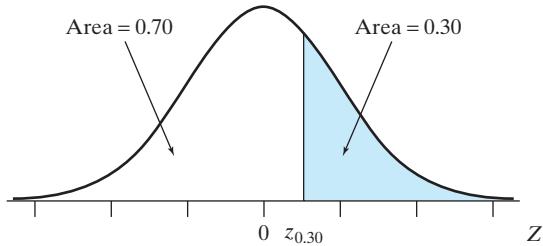
**Figure 4.3.11** Area under the normal curve above 1.96



**Figure 4.3.12** Area under the normal curve above  $\alpha$

We often need to determine a  $z_\alpha$  value when we want to determine a *percentile* of a normal distribution. The percentiles of a distribution divide the distribution into 100 equal parts, just as the quartiles divide it into 4 equal parts [from the Latin roots *centum* (“hundred”) and *quartus* (“fourth”)]. For example, suppose we want to find

the 70th percentile of a standard normal distribution. That means that we want to find the number  $z_{0.30}$  that divides the standard normal distribution into two parts: the bottom 70% and the top 30%. As Figure 4.3.13 illustrates, we need to look in Table 3 for an area of 0.7000. The closest value is an area of 0.6985, corresponding to a  $z$  value of 0.52. Thus,  $z_{0.30} = 0.52$ .

**Figure 4.3.13**

Determining the 70th percentile of a normal distribution

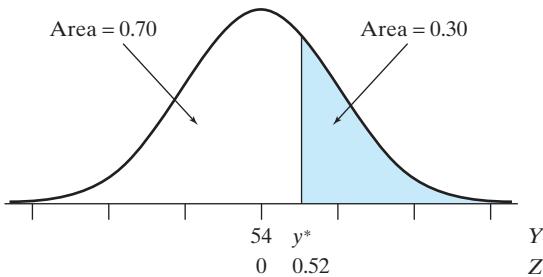
**Example  
4.3.2**
**Lengths of Fish**

- (a) Suppose we want to find the 70th percentile of the fish length distribution of Example 4.3.1. Let us denote the 70th percentile by  $y^*$ . By definition,  $y^*$  is the value such that 70% of the fish lengths are less than  $y^*$  and 30% are greater, as illustrated in Figure 4.3.14.

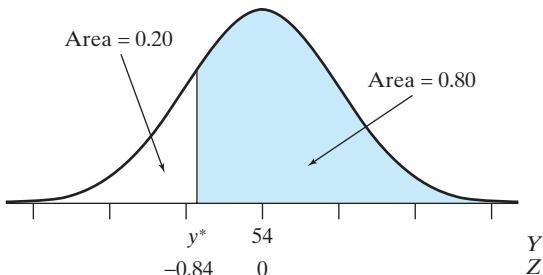
To find  $y^*$ , we use the value of  $z_{0.30} = 0.52$  that we just determined. Next we convert this  $z$  value to the  $Y$  scale. We know that if we were given the value of  $y^*$ , we could convert it to a standard normal ( $z$  scale) and the result would be 0.52. Thus, from the standardization formula we obtain the equation

$$0.52 = \frac{y^* - 54}{45}$$

which can be solved to give  $y^* = 54 + 0.52 \times 4.5 = 56.3$ . The 70th percentile of the fish length distribution is 56.3 mm.

**Figure 4.3.14** Determining the 70th percentile of a normal distribution, Example 4.3.2(a)

- (b) Suppose we want to find the 20th percentile of the fish length distribution of Example 4.3.1. Let us denote the 20th percentile by  $y^*$ . By definition,  $y^*$  is the value such that 20% of the fish lengths are less than  $y^*$  and 80% are greater, as illustrated in Figure 4.3.15.

**Figure 4.3.15** Determining the 20th percentile of a normal distribution, Example 4.3.2(b)

To find  $y^*$  we first determine the value of  $z_{0.80}$ , which is the 20th percentile in the  $Z$  scale. As Figure 4.3.15 illustrates, we need to look in Table 3 for an area of .2000. The closest value is an area of .2005, corresponding to  $z = -0.84$ . The next step is to convert this  $z$  value to the  $Y$  scale. From the standardization formula, we obtain the equation

$$-0.84 = \frac{y^* - 54}{45}$$

which can be solved to give  $y^* = 54 - 0.84 \times 4.5 = 50.2$ . The 20th percentile of the fish length distribution is 50.2 mm. ■

**Problem-Solving Tip** In solving problems that require the use of Table 3, a sketch of the distribution (as in Figures 4.3.7–4.3.10 and 4.3.14–4.3.15) is a very handy aid to straight thinking.

While Table 3 is handy for carrying out the sorts of computations discussed previously, computer software may also be used to find normal probabilities directly without the need for any standardization.

## Exercises 4.3.1–4.3.16

**4.3.1** Suppose a certain population of observations is normally distributed. What percentage of the observations in the population

- (a) are within  $\pm 1.5$  standard deviations of the mean?
- (b) are more than 2.5 standard deviations above the mean?
- (c) are more than 3.5 standard deviations away from (above or below) the mean?

### 4.3.2

- (a) The 90th percentile of a normal distribution is how many standard deviations above the mean?
- (b) The 10th percentile of a normal distribution is how many standard deviations below the mean?

**4.3.3** The brain weights of a certain population of adult Swedish males follow approximately a normal distribution with mean 1,400 gm and standard deviation 100 gm.<sup>6</sup> What percentage of the brain weights are

- (a) 1,500 gm or less?
- (b) between 1,325 and 1,500 gm?
- (c) 1,325 gm or more?
- (d) 1,475 gm or more?
- (e) between 1,475 and 1,600 gm?
- (f) between 1,200 and 1,325 gm?

**4.3.4** Let  $Y$  represent a brain weight randomly chosen from the population of Exercise 4.3.3. Find

- (a)  $\Pr\{Y \leq 1,325\}$
- (b)  $\Pr\{1,475 \leq Y \leq 1,600\}$

**4.3.5** In an agricultural experiment, a large uniform field was planted with a single variety of wheat. The field was divided into many plots (each plot being  $7 \times 100$  ft) and the yield (lb) of grain was measured for each plot. These plot yields followed approximately a normal distribution with mean 88 lb and standard deviation 7 lb.<sup>7</sup> What percentage of the plot yields were

- (a) 80 lb or more? (b) 90 lb or more?
- (c) 75 lb or less? (d) between 75 and 90 lb?
- (e) between 90 and 100 lb? (f) between 75 and 80 lb?

**4.3.6** Refer to Exercise 4.3.5. Let  $Y$  represent the yield of a plot chosen at random from the field. Find

- (a)  $\Pr\{Y > 90\}$  (b)  $\Pr\{75 < Y < 90\}$

**4.3.7** Consider a standard normal distribution,  $Z$ . Find

- (a)  $z_{0.10}$  (b)  $z_{0.25}$  (c)  $z_{0.05}$  (d)  $z_{0.01}$

**4.3.8** For the wheat-yield distribution of Exercise 4.3.5, find

- (a) the 65th percentile (b) the 35th percentile

**4.3.9** The serum cholesterol levels of 12- to 14-year-olds follow a normal distribution with mean 162 mg/dl and standard deviation 28 mg/dl. What percentage of 12 to 14-year-olds have serum cholesterol values

- (a) 171 or more? (b) 143 or less?
- (c) 194 or less? (d) 105 or more?
- (e) between 166 and 194? (f) between 105 and 138?
- (g) between 138 and 166?

**4.3.10** Refer to Exercise 4.3.9. Suppose a 13-year-old is chosen at random and let  $Y$  be the person's serum cholesterol value. Find

- (a)  $\Pr\{Y \geq 166\}$       (b)  $\Pr\{166 < Y < 194\}$

**4.3.11** For the serum cholesterol distribution of Exercise 4.3.9, find

- (a) the 80th percentile      (b) the 20th percentile

**4.3.12** When red blood cells are counted using a certain electronic counter, the standard deviation of repeated counts of the same blood specimen is about 0.8% of the true value, and the distribution of repeated counts is approximately normal.<sup>8</sup> For example, this means that if the true value is 5,000,000 cells/mm<sup>3</sup>, then the SD is 40,000.

- (a) If the true value of the red blood count for a certain specimen is 5,000,000 cells/mm<sup>3</sup>, what is the probability that the counter would give a reading between 4,900,000 and 5,100,000?  
 (b) If the true value of the red blood count for a certain specimen is  $\mu$ , what is the probability that the counter would give a reading between  $0.98\mu$  and  $1.02\mu$ ?  
 (c) A hospital lab performs counts of many specimens every day. For what percentage of these specimens does the reported blood count differ from the correct value by 2% or more?

**4.3.13** The amount of growth, in a 15-day period, for a population of sunflower plants was found to follow a normal distribution with mean 3.18 cm and standard deviation 0.53 cm.<sup>9</sup> What percentage of plants grow

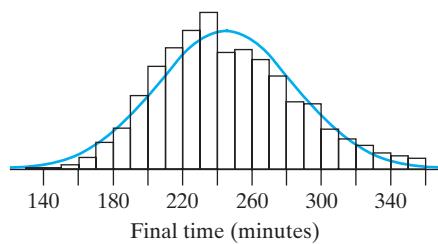
- (a) 4 cm or more?      (b) 3 cm or less?  
 (c) between 2.5 and 3.5 cm?

**4.3.14** Refer to Exercise 4.3.13. In what range do the middle 90% of all growth values lie?

**4.3.15** For the sunflower plant growth distribution of Exercise 4.3.13, what is the 25th percentile?

**4.3.16** Many cities sponsor marathons each year. The following histogram shows the distribution of times that it took for 10,002 runners to complete the Rome marathon in 2008, with a normal curve superimposed. The fastest runner completed the 26.3-mile course in 2 hours and 9 minutes, or 129 minutes. The average time was 245 minutes and the standard deviation was 40 minutes. Use the normal curve to answer the following questions.<sup>10</sup>

- (a) What percentage of times were greater than 200 minutes?  
 (b) What is the 60th percentile of the times?  
 (c) Notice that the normal curve approximation is fairly good except around the 240-minute mark. How can we explain this anomalous behavior of the distribution?



## 4.4 Assessing Normality

Many statistical procedures are based on having data from a normal population. In this section we consider ways to assess whether it is reasonable to use a normal curve model for a set of data and, if not, how we might proceed.

Recall from Section 4.3 that if the variable  $Y$  follows a normal distribution, then

- about 68% of the  $y$ 's are within  $\pm 1$  SD of the mean.
- about 95% of the  $y$ 's are within  $\pm 2$  SDs of the mean.
- about 99.7% of the  $y$ 's are within  $\pm 3$  SDs of the mean.

We can use these facts as a check of how closely a normal curve model fits a set of data.

**Example  
4.4.1**

**Serum Cholesterol** For the serum cholesterol data of Example 4.1.1, the sample mean is 162 and the sample SD is 28. The interval “mean  $\pm$  SD” is

$$(162 - 28, 162 + 28) \text{ or } (134, 190)$$

This interval contains 509 of the 727 observations, or 70.0% of the data. Likewise, the interval

$$(162 - 2 \times 28, 162 + 2 \times 28) \text{ is } (106, 218)$$

which contains 685, or 94.2%, of the 727 observations. Finally, the interval

$$(162 - 3 \times 28, 162 + 3 \times 28) \text{ is } (78, 246)$$

which contains 724, or 99.6%, of the 727 observations. The three observed percentages

70.0%, 94.2%, and 99.6%

agree quite well with the theoretical percentages of

68%, 95%, and 99.7%

This agreement supports the claim that serum cholesterol levels for 12- to 14-year-olds have a normal distribution. This reinforces the visual evidence of Figure 4.1.1. ■

**Example  
4.4.2**

**Moisture Content** Moisture content was measured in each of 83 freshwater fruit.<sup>11</sup> Figure 4.4.1 shows that this distribution is strongly skewed to the left. The sample mean of these data is 80.7 and the sample SD is 12.7. The interval

$$(80.7 - 12.7, 80.7 + 12.7)$$

contains 70, or 84.3%, of the 83 observations. The interval

$$(80.7 - 2 \times 12.7, 80.7 + 2 \times 12.7)$$

contains 78, or 94.0%, of the 83 observations. Finally, the interval

$$(80.7 - 3 \times 12.7, 80.7 + 3 \times 12.7)$$

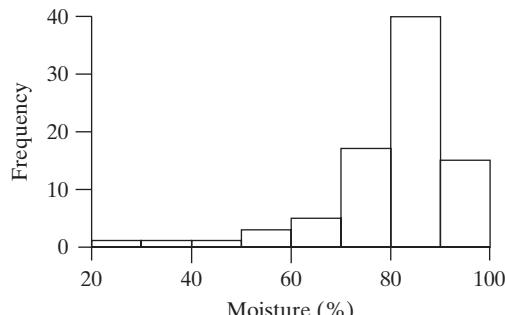
contains 80, or 96.4%, of the 83 observations. The three percentages

84.3%, 94.0%, and 96.4%

differ from the theoretical percentages of

68%, 95%, and 99.7%

because the distribution is far from being bell-shaped. This reinforces the visual evidence of Figure 4.4.1. ■



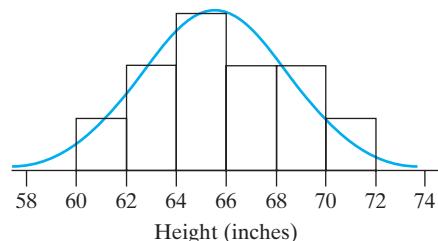
**Figure 4.4.1** Moisture content in freshwater fruit

## Normal Probability Plots

A **normal probability plot** is a special statistical graph that is used to assess normality. We present this statistical tool with an example using the heights (in inches) of a sample of 11 women, sorted from smallest to largest:

$$61, 62.5, 63, 64, 64.5, 65, 66.5, 67, 68, 68.5, 70.5$$

Based on these data, does it make sense to use a normal curve to model the distribution of women's heights? Figure 4.4.2 is a histogram of the data with a normal curve superimposed, using the sample mean of 65.5 and the sample standard deviation of 2.9 as the parameters of the normal curve. This histogram is fairly symmetric, but when we have a small sample, it can be hard to tell the shape of the population distribution by looking at a histogram.



**Figure 4.4.2** Histogram of the heights of 11 women

Because it is often difficult to visually examine a histogram and decide if it is bell-shaped or not, a visually simpler plot, the normal probability plot, was developed.\* A normal probability plot is a scatterplot that compares our observed data values to values we would expect to see if the population were normal. If the data come from a normal population, the points in this plot should follow a straight line, which is much easier to visually recognize than a bell shape of a jagged histogram. As many statistical procedures are based on the condition that the data came from a normal population, it is important to be able to assess normality.

## How Normal Probability Plots Work

In Examples 4.4.1 and 4.4.2 we compared the observed proportion of data that falls within 1, 2, and 3 SDs of the mean and then compared those values to the proportions we would expect to find if the data were from a normal population. It is natural to consider these intervals, but we could consider other intervals as well. For example, we would expect about 86.6% of normal data to fall within 1.5 SDs of the mean and 96.4% to within 2.1.<sup>†</sup> We could even consider one-sided intervals. For example, we would expect 84.1% of normal data values to be less than the mean plus 1 SD.

Rather than focus on comparing percentages, we could instead focus on comparing actual observed women's heights to heights we would expect to see if the data were from a normal population. For example, the shortest woman in our sample is 61 inches tall; that is, 1/11th (or 0.0909) of the sample is 61 inches or shorter. If heights of women really follow a normal distribution, with mean 65.5 and standard deviation 2.9, then we would expect the 9.09th percentile to be  $\mu + z_{(1-0.0909)}\sigma = 65.5 - 1.34 \times 2.9$  or 61.6 inches. This value is close to the observed

\*Though visually simple, the construction of these plots is complex and typically performed using statistical software.

<sup>†</sup>These values can be verified using the techniques of Section 4.3.

value of 61 inches. We could repeat this sort of calculation for each of the 11 observed data values. A normal probability plot provides a visual comparison of these values.

The first step in creating a normal probability plot, therefore, is to compute the sample percentiles. Example 4.4.3 presents this computation, which is typically performed by statistical software.

**Example  
4.4.3**

**Height of Eleven Women** Sorting the data from smallest to largest we observe that 1/11th (= 9.1%) of our sample is 61 inches or shorter, 2/11ths (= 18.2%) is 62.5 inches or shorter, ... 10/11ths (90.9%) is 68.5 inches or shorter and 11/11ths (100%) is 70.5 inches or shorter. Unfortunately, computing percentages in this simplistic way (i.e.,  $100 \times i/n$  where  $i$  is the sorted observation number) creates some implausible population estimates. For example, it seems unreasonable to believe that 100% of the *population* is 70.5 inches or shorter when, after all, we are observing only a small sample; a larger sample would likely observe some taller women. To correct for this, an alternative and more reasonable percentage for each data value is computed as  $100(i - \frac{1}{2})/n$  where  $i$  is the index of the data value in the sorted list.\* These adjusted percentiles are tabulated in Table 4.4.1. Note that these values actually do not depend on the data observed; they depend only on the number of data values in the sample. ■

**Table 4.4.1** Computing indices and percentiles for the heights of eleven women

$i$	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Percentile $100(i/11)$	9.09	18.18	27.27	36.36	45.45	54.55	63.64	72.73	81.82	90.91	100.00
Adjusted percentile $100(i - \frac{1}{2})/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45

Once we have the adjusted percentiles we find the corresponding  $z$  scores using Table 3 or a computer. Then, with these  $z$  scores we find the theoretical heights:  $\mu + z \times \sigma$  as in Example 4.4.4.

**Example  
4.4.4**

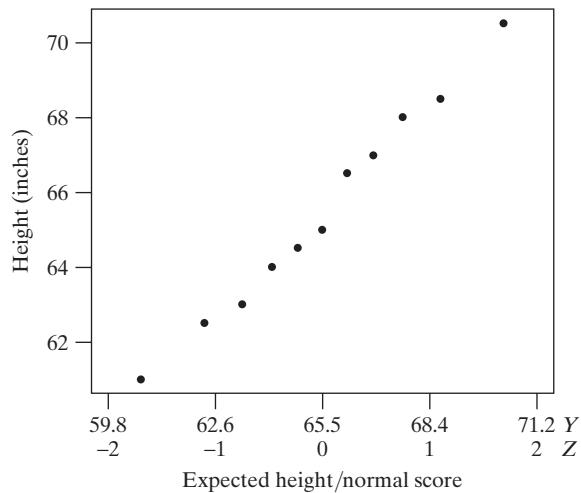
**Heights of Eleven Women** The shortest woman's adjusted percentile is 4.55%. The corresponding  $z$  score is  $z_{(1-0.0455)} = z_{0.9545} = -1.69$ . In this example, the sample mean and standard deviation are 65.5 and 2.9, respectively, so the expected height of the shortest woman in a sample of 11 women from a normal population is  $65.5 - 1.69 \times 2.9 = 60.6$  inches. The  $z$  scores and theoretical heights for this woman and the remaining 10 women appear in Table 4.4.2.

**Table 4.4.2** Computing theoretical  $z$  scores and heights for eleven women

$i$	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Adjusted percentile $100(i - \frac{1}{2})/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45
$z$	-1.69	-1.10	-0.75	-0.47	-0.23	0.00	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4

\*Different software packages may compute these proportions differently and may also modify the formula based on sample size. The preceding formula is used by the software package *R* when  $n > 10$ .

Next, by plotting the observed heights against the theoretical heights in a scatterplot as in Figure 4.4.3, we may visually compare the values. In this case our plot appears fairly linear, suggesting that the observed values generally agree with the theoretical values—that the normal model provides a reasonable approximation to the data. If the data do not agree with the normal model, then the plot will show strong nonlinear patterns such as curvature or S shapes.



**Figure 4.4.3** Normal probability plot of the heights of 11 women

Because of the one-to-one correspondence between the  $z$  scores and theoretical values, it is not common to put both sets of labels on the  $x$ -axis as in Figure 4.4.3. Traditionally only the  $z$  scores are displayed.\* ■

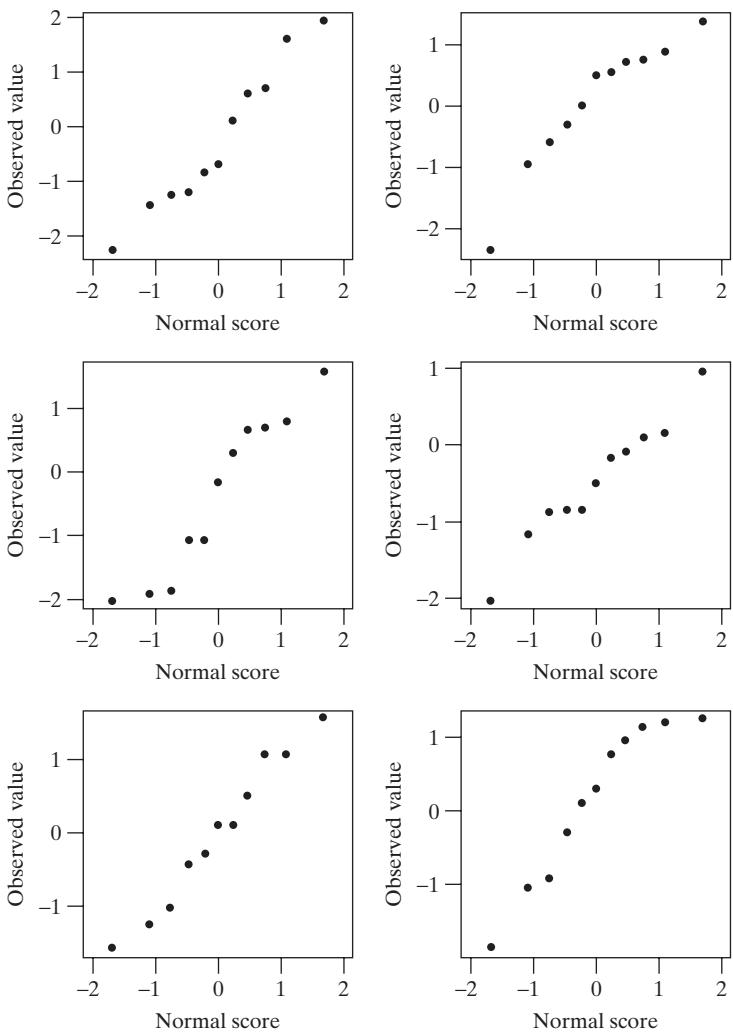
### Making Decisions about Normality

Of course, even when we sample from a perfectly normal distribution, we have to expect that there will be some variability between the sample we obtain and the theoretical normal scores. Figure 4.4.4 shows six normal probability plots based on samples taken from a  $N(0, 1)$  distribution. Notice that all six plots show a general linear pattern. It is true that there is a fair amount of “wiggle” in some of the plots, but the important feature of each of these plots is that we can draw a line that captures the trend in the bulk of the points, with little deviation away from this line, even at the extremes.

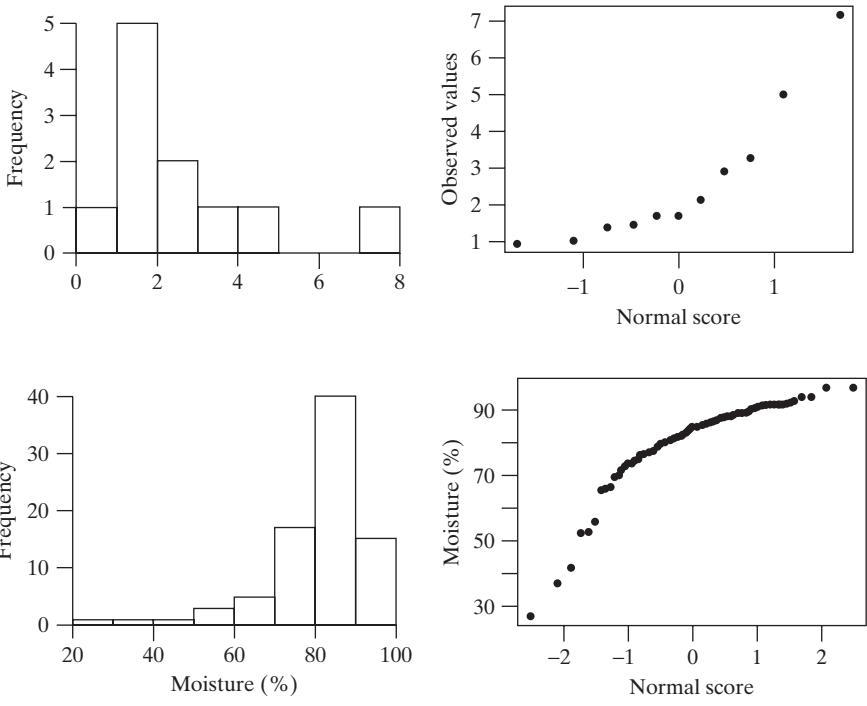
If the points in the normal probability plot do not fall more or less along a straight line, then there is an indication that the data are not from a normal population. For example, if the top of the plot bends up, that means the  $y$  values at the upper end of the distribution are too large for the distribution to be bell-shaped; that is, the distribution is skewed to the right or has large outliers, as in Figure 4.4.5.

If the bottom of the plot bends down, that means the  $y$  values at the lower end of the distribution are too small for the distribution to be bell-shaped; that is, the distribution is skewed to the left or has small outliers. Figure 4.4.6 shows the distribution of moisture content in the freshwater fruit from Example 4.4.2, which is strongly skewed to the left.

\*Some software programs create normal probability plots with the normal scores on the vertical axis and the observed data on the horizontal axis.

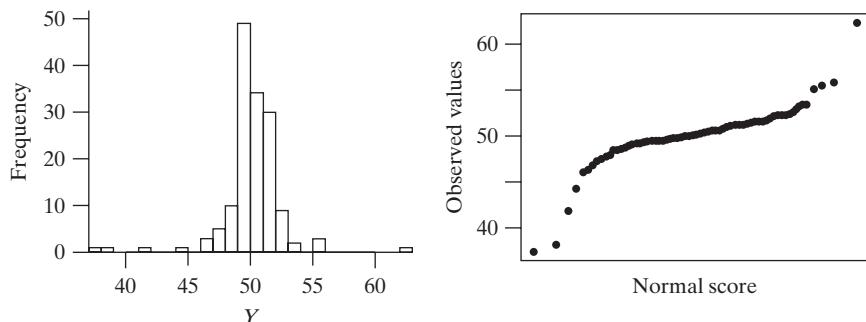


**Figure 4.4.4** Normal probability plots for normal data



**Figure 4.4.5** Histogram and normal probability plot of a distribution that is skewed to the right

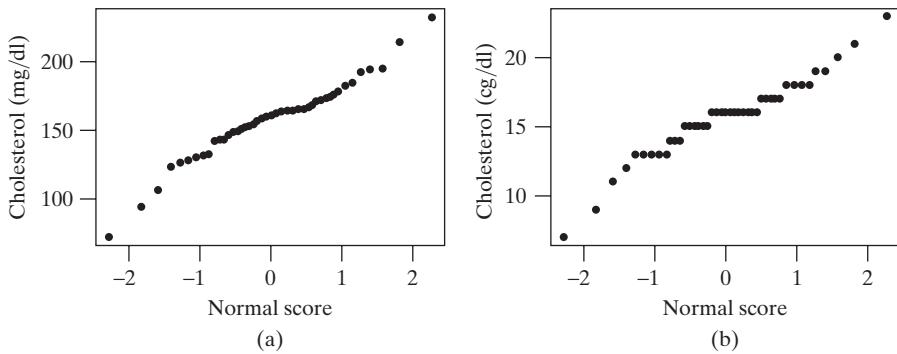
**Figure 4.4.6** Histogram and normal probability plot of a distribution that is skewed to the left



**Figure 4.4.7** Histogram and normal probability plot of a distribution that has long tails

If a distribution has a very long left-hand tail and a long right-hand tail, when compared to a normal curve, then the normal probability plot will have something of an S shape. Figure 4.4.7 shows such a distribution.

Sometimes the same value shows up repeatedly in a sample, due to rounding in the measurement process. This leads to *granularity* in the normal probability plot, as in Figure 4.4.8, but this does not stop us from inferring that the underlying distribution is normal.



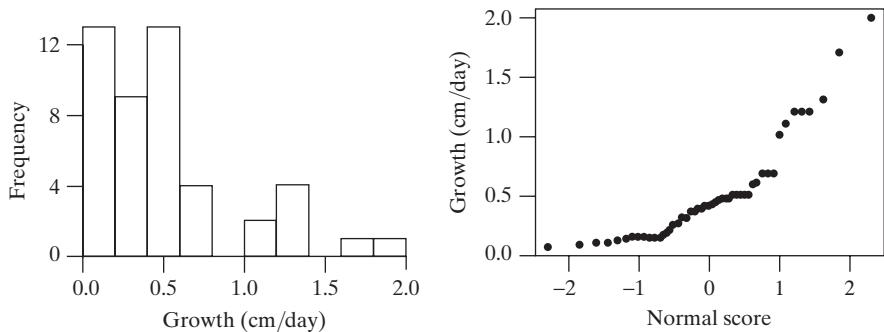
**Figure 4.4.8** Normal probability plots of cholesterol values of fifty 12- to 14-year-olds measured to (a) the nearest mg/dl and (b) the nearest cg/dl

## Transformations for Nonnormal Data

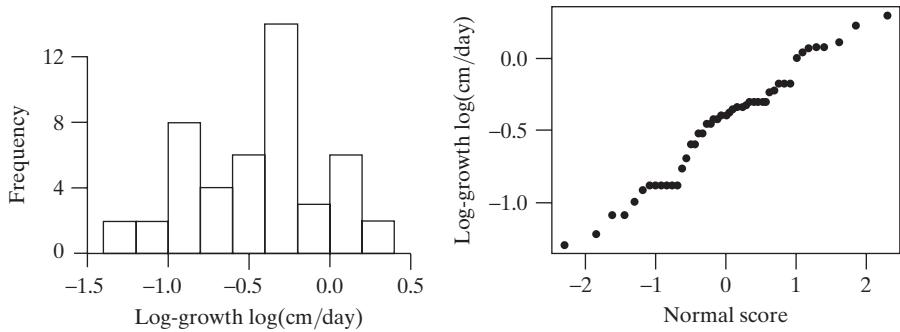
A normal probability plot can help us assess whether or not the data came from a normal distribution. Sometimes a histogram or normal probability plot shows that our data are nonnormal, but a transformation of the data gives us a symmetric, bell-shaped curve. In such a situation, we may wish to transform the data and continue our analysis in the new (transformed) scale.

### Example 4.4.5

**Lentil Growth** The histogram and normal probability plot in Figure 4.4.9 show the distribution of the growth rate, in cm per day, for a sample of 47 lentil plants.<sup>12</sup> This distribution is skewed to the right. If we take the logarithm of each observation, we



**Figure 4.4.9** Histogram and normal probability plot of growth rates of 47 lentil plants



**Figure 4.4.10** Histogram and normal probability plot of the logarithms of the growth rates of 47 lentil plants

get a distribution that is much more nearly symmetric. The plots in Figure 4.4.10 show that in log scale the growth rate distribution is approximately normal. (In Figure 4.4.10 the base 10 logarithm,  $\log_{10}$ , is used, but we could use any base, such as the natural log,  $\log_e = \ln$ , and the effect on the shape of the distribution would be the same.) ■

In general, if the distribution is skewed to the right then one of the following transformations should be considered:  $\sqrt{Y}$ ,  $\log Y$ ,  $1/\sqrt{Y}$ ,  $1/Y$ . These transformations will pull in the long right-hand tail and push out the short left-hand tail, making the distribution more nearly symmetric. Each of these is more drastic than the one before. Thus, a square root transformation will change a mildly skewed distribution into a symmetric distribution, but a log transformation may be needed if the distribution is more heavily skewed, and so on. For example, we saw in Example 2.7.6 how a square root transformation pulls in a long right-hand tail and how a log transformation pulls in the right-hand tail even more. If the distribution of a variable  $Y$  is skewed to the left, then raising  $Y$  to a power greater than 1 can be helpful.

### An Objective Measure of Abnormality: The Shapiro–Wilk Test (optional)

While normal probability plots are better than histograms to visually assess departures of normality, our visual perception is still subjective. The data appearing in the probability plots of Figure 4.4.4 come from a normal population, but to untrained eyes (and even to some trained ones) a few of the plots might be interpreted as being nonnormal. The **Shapiro–Wilk test** is a statistical procedure that numerically assesses evidence for certain types of nonnormality in data. As with the normal probability plot, the mechanics of the procedure is complex, but fortunately many statistical software packages will perform this or similar tests of normality.\*

The output of a Shapiro–Wilk test is a  $P$ -value<sup>†</sup> and is interpreted as follows:

$P$ -value < 0.001	Very strong evidence for nonnormality
$P$ -value < 0.01	Strong evidence for nonnormality
$P$ -value < 0.05	Moderate evidence for nonnormality
$P$ -value < 0.10	Mild or weak evidence for nonnormality
$P$ -value $\geq 0.10$	No compelling evidence for nonnormality

\*The Ryan–Joiner, Anderson–Darling, and Kolmogorov–Smirnov tests are other tests of nonnormality commonly found in statistical software packages.

<sup>†</sup>As we shall see in much greater detail in Chapter 7, a  $P$ -value is not unique to testing for normality. In a test of all sorts of hypotheses, the weight of evidence for the hypothesis in question (in this case—the Shapiro–Wilk test—the hypothesis is that the data are nonnormal) can be reported using this term. Small  $P$ -values are interpreted as evidence for the hypothesis in question.

Example 4.4.6 illustrates the Shapiro–Wilk test for the lentil growth data of Example 4.4.5.

**Example  
4.4.6**

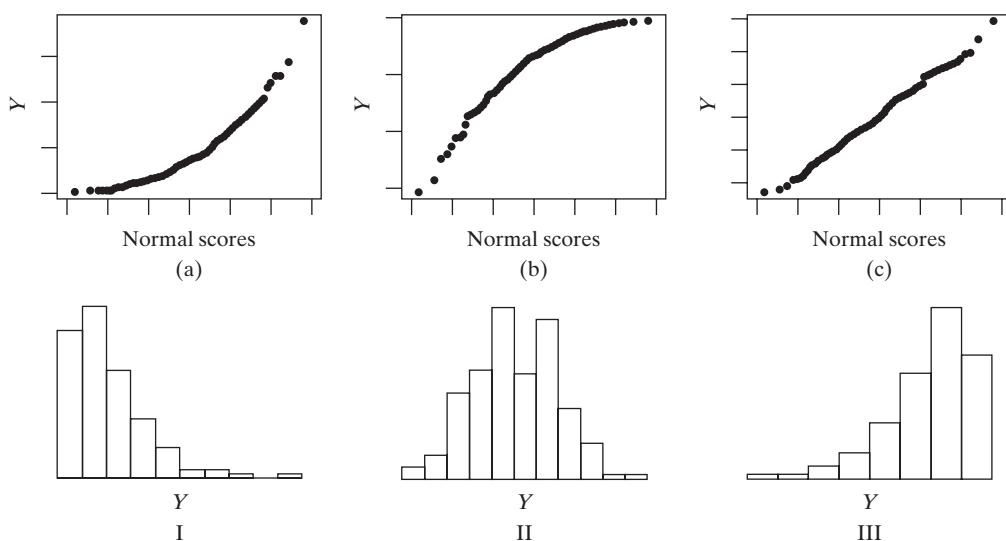
**Lentil Growth** For the untransformed lentil data in Figure 4.4.9, the  $P$ -value (reported from the statistical software package  $R$ ) for the Shapiro–Wilk test is 0.000006. Thus, there is very strong evidence that lentil growth does not follow a normal distribution. For the transformed data in Figure 4.4.10, however, the  $P$ -value for the Shapiro–Wilk test is 0.2090, indicating that there is no compelling evidence for nonnormality of the log-transformed growth data. ■

**Caution.** The use of this test procedure and  $P$ -value is somewhat like the use of the “check engine light” on a car. When the  $P$ -value is small, there is an indication of nonnormality. This is like your engine light coming on: You pull over and assess the situation. Likewise, as we shall see in future chapters, when we have nonnormal data, we will carefully have to assess how to proceed with our analyses. On the other hand, when the  $P$ -value is not small ( $\geq 0.10$ ) we don’t have evidence of nonnormality. This is similar to your engine light staying off: You continue to drive forward without worry, *but* this does not guarantee that your car is perfectly OK. Your car could break down at any time. Of course, if we were constantly worried about our car even when the check engine light were off, we would perpetually find ourselves paralyzed and pulled over at the side of the road. Analogously, when the  $P$ -value from the Shapiro–Wilk tests is not small (the light is off), this only means that there is no compelling evidence for nonnormality. It does not guarantee that the population is, in fact, normal.

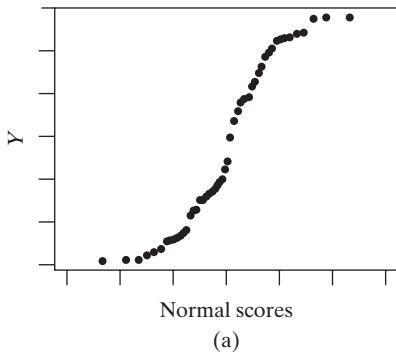
## Exercises 4.4.1–4.4.8

- 4.4.1** In Example 4.1.2 it was stated that shell thicknesses in a population of eggs follow a normal distribution with mean  $\mu = 0.38$  mm and standard deviation  $\sigma = 0.03$  mm. Use the 68%–95%–99.7% rule to determine intervals, centered at the mean, that include 68%, 95%, and 99.7% of the shell thicknesses in the distribution.

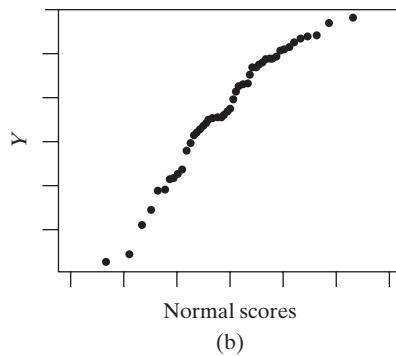
- 4.4.2** The following three normal probability plots, (a), (b), and (c), were generated from the distributions shown by histograms I, II, and III. Which normal probability plot goes with which histogram? How do you know?



**4.4.3** For each of the following normal probability plots, sketch the corresponding histogram of the data.



(a)



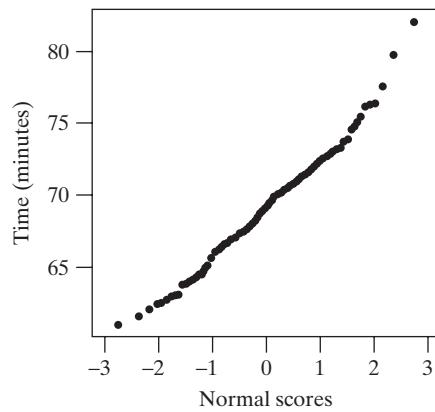
(b)

**4.4.4** The mean daily rainfall between January 1, 2007, through January 1, 2009, at Pismo Beach, California, was 0.02 inches with a standard deviation of 0.11 inches. Based on this information, do you think it is reasonable to believe that daily rainfall at Pismo Beach follows a normal distribution? Explain. (*Hint:* Think about the possible values for daily rainfall.)<sup>13</sup>

**4.4.5** The mean February 1 daily high temperature in Juneau, Alaska, between 1945 and 2005 was  $1.1^{\circ}\text{C}$  with a standard deviation of  $1.9^{\circ}\text{C}$ .<sup>14</sup>

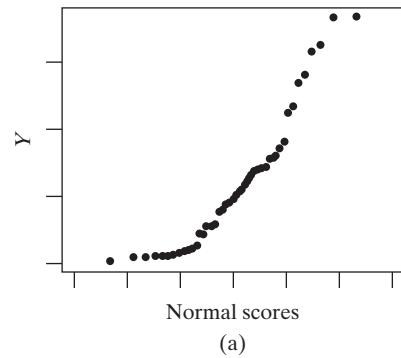
- (a) Based on this information, do you think it is reasonable to believe that the February 1 daily high temperatures in Juneau, Alaska, follow a normal distribution? Explain.
- (b) Does this information provide compelling evidence that the February 1 daily high temperatures in Juneau, Alaska, follow a normal distribution? Explain.

**4.4.6** The following normal probability plot was created from the times that it took 166 bicycle riders to complete the stage 11 time trial, from Grenoble to Chamrousse, France, in the 2001 Tour de France cycling race.

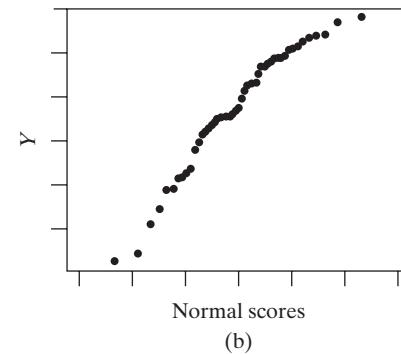


- (a) Consider the fastest riders. Are their times better than, worse than, or roughly equal to the times one would expect the fastest riders to have if the data came from a truly normal distribution?
- (b) Consider the slowest riders. Are their times better than, worse than, or roughly equal to the times one would expect the slowest riders to have if the data came from a truly normal distribution?

**4.4.7** The  $P$ -values for the Shapiro–Wilk test for the data appearing in probability plots (a) and (b) are 0.235 and 0.00015. Which  $P$ -value corresponds to which plot? What is the basis for your decision?



(a)



(b)

**4.4.8**

- (a) The  $P$ -value for the Shapiro–Wilk test of normality for the data in Exercise 4.4.3(b) is 0.039. Using this value to justify your answer, does it seem reasonable to believe that these data came from a normal population?
- (b) The  $P$ -value for the Shapiro–Wilk test of normality for the data in Exercise 4.4.2(c) is 0.770. Using this value to justify your answer, does it seem reasonable to believe that these data came from a normal population?
- (c) Does the  $P$ -value in part (b) prove that the data come from a normal population?

## 4.5 Perspective

The normal distribution is also called the Gaussian distribution, after the German mathematician K. F. Gauss. The term *normal*, with its connotations of “typical” or “usual,” can be seriously misleading. Consider, for instance, a medical context, where the primary meaning of “normal” is “not abnormal.” Thus, confusingly, the phrase “the normal population of serum cholesterol levels” may refer to cholesterol levels in ideally “healthy” people, or it may refer to a Gaussian distribution such as the one in Example 4.1.1. In fact, for many variables the distribution in the normal (nondiseased) population is decidedly not normal (i.e., not Gaussian).

The examples of this chapter have illustrated one use of the normal distribution—as an approximation to naturally occurring biological distributions. If a natural distribution is well approximated by a normal distribution, then the mean and standard deviation provide a complete description of the distribution: The mean is the center of the distribution: About 68% of the values are within 1 standard deviation of the mean, about 95% are within 2 standard deviations of the mean, and so on.

As noted in Section 2.6, the 68% and 95% benchmarks can roughly be applicable even to distributions that are rather skewed. (But if the distribution is skewed, then the 68% is not symmetrically divided on both sides of the mean, and similarly for the 95%.) However, the benchmarks do not apply to a distribution (even a symmetric one) for which one or both tails are long and thin (see Figures 2.2.13 and 2.2.16).

We will see in later chapters that many classical statistical methods are specifically designed for, and function best with, data that have been sampled from normal populations. We will further see that in many practical situations these methods also work very well for samples from nonnormal populations.

The normal distribution is of central importance in spite of the fact that many, perhaps most, naturally occurring biological distributions could be described better by a skewed curve than by a normal curve. A major use of the normal distribution is not to describe natural distributions, but rather to describe certain theoretical distributions, called sampling distributions, that are used in the statistical analysis of data. We will see in Chapter 5 that many sampling distributions are approximately normal even when the underlying data are not; it is this property that makes the normal distribution so important in the study of statistics.

## Supplementary Exercises 4.S.1–4.S.21

**4.S.1** The activity of a certain enzyme is measured by counting emissions from a radioactively labeled molecule. For a given tissue specimen, the counts in consecutive 10-second time periods may be regarded (approximately) as repeated independent observations from a normal distribution.<sup>15</sup> Suppose the mean 10-second count for a certain tissue specimen is 1,200 and

the standard deviation is 35. Let  $Y$  denote the count in a randomly chosen 10-second time period. Find

- $\Pr\{Y \geq 1,250\}$
- $\Pr\{Y \leq 1,175\}$
- $\Pr\{1,150 \leq Y \leq 1,250\}$
- $\Pr\{1,150 \leq Y \leq 1,175\}$

**4.S.2** The shell thicknesses of the eggs produced by a large flock of hens follow approximately a normal distribution with mean equal to 0.38 mm and standard deviation equal to 0.03 mm (as in Example 4.1.2). Find the 95th percentile of the thickness distribution.

**4.S.3** Refer to the eggshell thickness distribution of Exercise 4.S.2. Suppose an egg is defined as thin shelled if its shell is 0.32 mm thick or less.

- (a) What percentage of the eggs are thin shelled?
- (b) Suppose a large number of eggs from the flock are randomly packed into boxes of 12 eggs each. What percentage of the boxes will contain at least one thin-shelled egg? (*Hint:* First find the percentage of boxes that will contain no thin-shelled egg.)

**4.S.4** The heights of a certain population of corn plants follow a normal distribution with mean 145 cm and standard deviation 22 cm.<sup>16</sup> What percentage of the plant heights are

- (a) 100 cm or more?
- (b) 120 cm or less?
- (c) between 120 and 150 cm?
- (d) between 100 and 120 cm?
- (e) between 150 and 180 cm?
- (f) 180 cm or more?
- (g) 150 cm or less?

**4.S.5** Suppose four plants are to be chosen at random from the corn plant population of Exercise 4.S.4. Find the probability that none of the four plants will be more than 150 cm tall.

**4.S.6** Refer to the corn plant population of Exercise 4.S.4. Find the 90th percentile of the height distribution.

**4.S.7** For the corn plant population described in Exercise 4.S.4, find the quartiles and the interquartile range.

**4.S.8** Suppose a certain population of observations is normally distributed.

- (a) Find the value of  $z^*$  such that 95% of the observations in the population are between  $-z^*$  and  $+z^*$  on the  $Z$  scale.
- (b) Find the value of  $z^*$  such that 99% of the observations in the population are between  $-z^*$  and  $+z^*$  on the  $Z$  scale.

**4.S.9** In the nerve-cell activity of a certain individual fly, the time intervals between “spike” discharges follow approximately a normal distribution with mean 15.6 ms and standard deviation 0.4 ms (as in Example 4.1.3). Let  $Y$  denote a randomly selected interspike interval. Find

- (a)  $\Pr\{Y > 15\}$
- (b)  $\Pr\{Y > 16.5\}$
- (c)  $\Pr\{15 < Y < 16.5\}$
- (d)  $\Pr\{15 < Y < 15.5\}$

**4.S.10** For the distribution of interspike-time intervals described in Exercise 4.S.9, find the quartiles and the interquartile range.

**4.S.11** Among American women aged 20 to 29 years, 10% are less than 60.8 inches tall, 80% are between 60.8 and 67.6 inches tall, and 10% are more than 67.6 inches tall.<sup>17</sup> Assuming that the height distribution can adequately be approximated by a normal curve, find the mean and standard deviation of the distribution.

**4.S.12** The intelligence quotient (IQ) score, as measured by the Stanford-Binet IQ test, is normally distributed in a certain population of children. The mean IQ score is 100 points, and the standard deviation is 16 points.<sup>18</sup> What percentage of children in the population have IQ scores

- (a) 140 or more?
- (b) 80 or less?
- (c) between 80 and 120?
- (d) between 80 and 140?
- (e) between 120 and 140?

**4.S.13** Refer to the IQ distribution of Exercise 4.S.12. Let  $Y$  be the IQ score of a child chosen at random from the population. Find  $\Pr\{80 \leq Y \leq 140\}$ .

**4.S.14** Refer to the IQ distribution of Exercise 4.S.12. Suppose five children are to be chosen at random from the population. Find the probability that exactly one of them will have an IQ score of 80 or less and four will have scores higher than 80. (*Hint:* First find the probability that a randomly chosen child will have an IQ score of 80 or less.)

**4.S.15** A certain assay for serum alanine aminotransferase (ALT) is rather imprecise. The results of repeated assays of a single specimen follow a normal distribution with mean equal to the true ALT concentration for that specimen and standard deviation equal to 4 U/l (see Example 2.2.12). Suppose that a certain hospital lab measures many specimens every day, performing one assay for each specimen, and that specimens with ALT readings of 40 U/l or more are flagged as “unusually high.” If a patient’s true ALT concentration is 35 U/l, what is the probability that his specimen will be flagged as “unusually high”?

**4.S.16** Resting heart rate was measured for a group of subjects; the subjects then drank 6 ounces of coffee. Ten minutes later their heart rates were measured again. The change in heart rate followed a normal distribution, with a mean increase of 7.3 beats per minute and a standard deviation of 11.1.<sup>19</sup> Let  $Y$  denote the change in heart rate for a randomly selected person. Find

- (a)  $\Pr\{Y > 10\}$
- (b)  $\Pr\{Y > 20\}$
- (c)  $\Pr\{5 < Y < 15\}$

**4.S.17** Refer to the heart rate distribution of Exercise 4.S.16. The fact that the standard deviation is greater than the average and that the distribution is normal tells us

that some of the data values are negative, meaning that the person's heart rate went down, rather than up. Find the probability that a randomly chosen person's heart rate will go down. That is, find  $\Pr\{Y < 0\}$ .

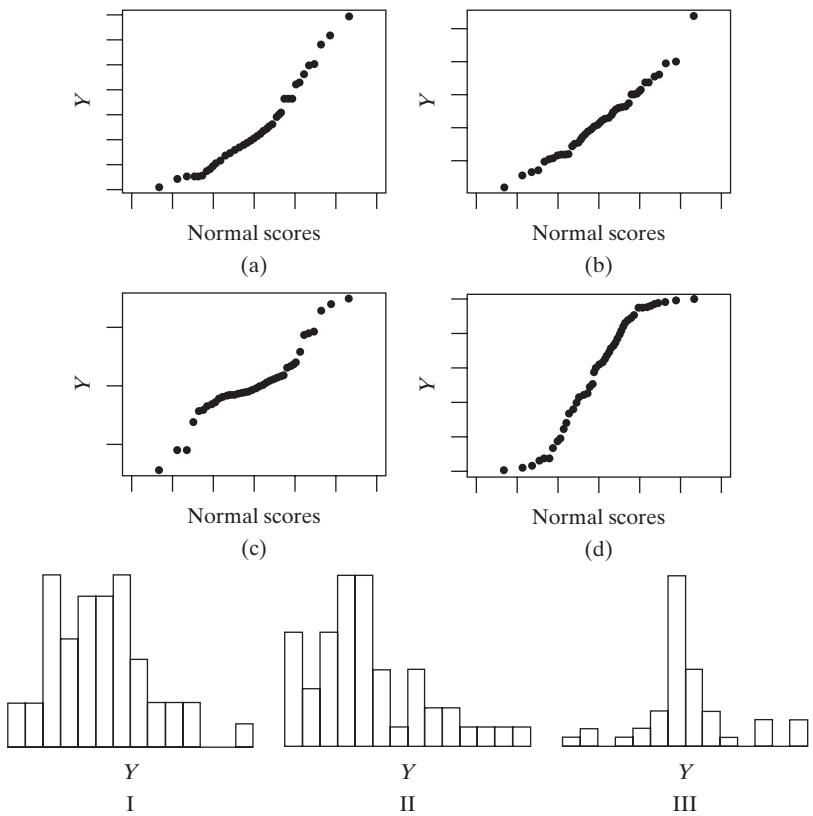
**4.S.18** Refer to the heart rate distribution of Exercise 4.S.16. Suppose we take a random sample of size 400 from this distribution. How many observations do we expect to obtain that fall between 0 and 15?

**4.S.19** Refer to the heart rate distribution of Exercise 4.S.16. If we use the  $1.5 \times \text{IQR}$  rule, from Chapter 2, to identify outliers, how large would an observation need to be in order to be labeled an outlier on the upper end?

**4.S.20** It is claimed that the heart rates of Exercise 4.S.16 follow a normal distribution. If this is true, which of the following Shapiro–Wilk's test  $P$ -values for a random sample of 15 subjects are consistent with this claim?

- (a)  $P$ -value = 0.0149
- (b)  $P$ -value = 0.1345
- (c)  $P$ -value = 0.0498
- (d)  $P$ -value = 0.0042

**4.S.21** The following four normal probability plots, (a), (b), (c), and (d), were generated from the distributions shown by histograms I, II, and III and another histogram that is not shown. Which normal probability plot goes with which histogram? How do you know? (There will be one normal probability plot that is not used.)



# SAMPLING DISTRIBUTIONS

## Objectives

In this chapter we will develop the idea of a sampling distribution, which is central to classical statistical inference. In particular, we will

- describe sampling distributions.
- show how the sample size is related to the accuracy of the sample mean.
- explore the Central Limit Theorem.
- demonstrate how the normal distribution can be used to approximate the binomial distribution.

### 5.1 Basic Ideas

An important goal of data analysis is to distinguish between features of the data that reflect real biological facts and features that may reflect only chance effects. As explained in Sections 1.3 and 2.8, the random sampling model provides a framework for making this distinction. The underlying reality is visualized as a population, the data are viewed as a random sample from the population, and chance effects are regarded as sampling error—that is, discrepancy between the sample and the population.

In this chapter we develop the theoretical background that will enable us to place specific limits on the degree of sampling error to be expected in a study. (Although in Chapter 1 we distinguished between an experimental study and an observational study, for the present discussion we will call any scientific investigation a *study*.) As in earlier chapters, we continue to confine the discussion to the simple context of a study with only one group (one sample).

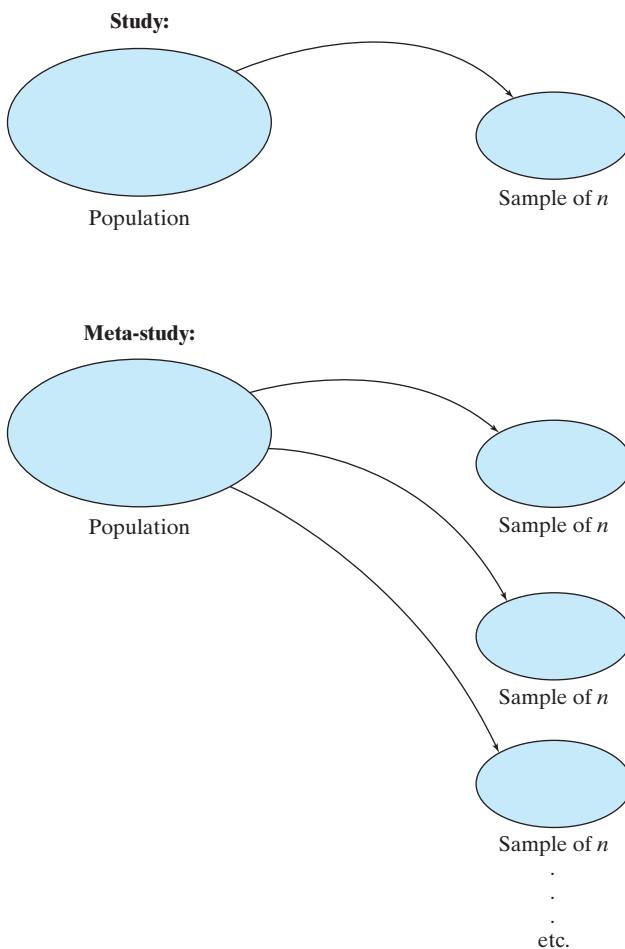
### Sampling Variability

The variability among random samples from the same population is called **sampling variability**. A probability distribution that characterizes some aspect of sampling variability is termed a **sampling distribution**. Usually a random sample will resemble the population from which it came. Of course, we have to expect a certain amount of discrepancy between the sample and the population. A sampling distribution tells us how close the resemblance between the sample and the population is likely to be.

In this chapter we will discuss several aspects of sampling variability and study an important sampling distribution. From this point forward, we will assume that the sample size is a negligibly small fraction of the population size. This assumption simplifies the theory because it guarantees that the process of drawing the sample does not change the population composition in any appreciable way.

## The Meta-Study

According to the random sampling model, we regard the data in a study as a random sample from a population. Generally we obtain only a single random sample, which comes from a very large population. However, to visualize sampling variability we must broaden our frame of reference to include not merely one sample, but all the possible samples that might be drawn from the population. This wider frame of reference we will call the **meta-study**. A meta-study consists of indefinitely many repetitions, or replications, of the same study.\* Thus, if the study consists of drawing a random sample of size  $n$  from some population, the corresponding meta-study involves drawing *repeated* random samples of size  $n$  from the same population. The process of repeated drawing is carried on indefinitely, with the members of each sample being replaced before the next sample is drawn. The study and the meta-study are schematically represented in Figure 5.1.1.



**Figure 5.1.1** Schematic representation of study and meta-study

---

\*The term *meta-study* is not a standard term. It is unrelated to the term *meta-analysis*, which denotes a particular type of statistical analysis.

The following two examples illustrate the notion of a meta-study.

---

**Example  
5.1.1**

**Rat Blood Pressure** A study consists of measuring the change in blood pressure in each of  $n = 10$  rats after administering a certain drug. The corresponding meta-study would consist of repeatedly choosing groups of  $n = 10$  rats from the same population and making blood pressure measurements under the same conditions. ■

---

**Example  
5.1.2**

**Bacterial Growth** A study consists of observing bacterial growth in  $n = 5$  petri dishes that have been treated identically. The corresponding meta-study would consist of repeatedly preparing groups of five petri dishes and observing them in the same way. ■

Note that a meta-study is a theoretical construct rather than an operation that is actually performed by a researcher.

The meta-study concept provides a link between sampling variability and probability. Recall from Chapter 3 that the probability of an event can be interpreted as the long-run relative frequency of occurrence of the event. Choosing a random sample is a chance operation; the meta-study consists of many repetitions of this chance operation, and so *probabilities concerning a random sample can be interpreted as relative frequencies in a meta-study*. Thus, the meta-study is a device for explicitly visualizing a sampling distribution: The sampling distribution describes the variability, for a chosen statistic, among the many random samples in a meta-study.

We consider a small (and artificial) example to illustrate the idea of a sampling distribution.

---

**Example  
5.1.3**

**Knee Replacement** Consider a population of women age 65 to 75 who are experiencing pain in their knees and are candidates for knee replacement surgery. A woman might have replacement surgery done on one knee at a cost of \$35,000, both knees at a cost of \$60,000 (a “double replacement,” which is less expensive than two single replacements), or neither knee. Consider the perspective of an insurance company regarding a sample of  $n = 3$  women it insures: What is the total cost for treating these three? The smallest the total could be is zero—if all three women skip surgery—while the largest possible cost would be \$180,000—if all three women have double replacements. To keep things relatively simple, suppose that one-fourth of women age 65 to 75 elect a double knee replacement, one-half elect a single knee replacement, and one-fourth choose not to have surgery.

The complete list of possible samples is given in Table 5.1.1, along with the sample total (in thousands of dollars) in each case and the probability of each case arising. For example, the probability that all three women skip surgery (“None, None, None”) is  $(1/4) \times (1/4) \times (1/4) = 1/64$  while the probability that the first two women skip surgery and the third has a single knee operation (“None, None, Single”) is  $(1/4) \times (1/4) \times (2/4) = 2/64$ . There are 10 possible values for the sample total: 0, 35, 60, 70, 95, 105, 120, 130, 155, and 180. The first and third columns of Table 5.1.2 give the sampling distribution of the sample total by combining the samples that yield the same total and summing their probabilities. For example, there are three ways for the total to be 70, each of which has probability 4/64; these sum to 12/64.

**Table 5.1.1** Total knee replacement costs for all possible samples of size  $n = 3$ 

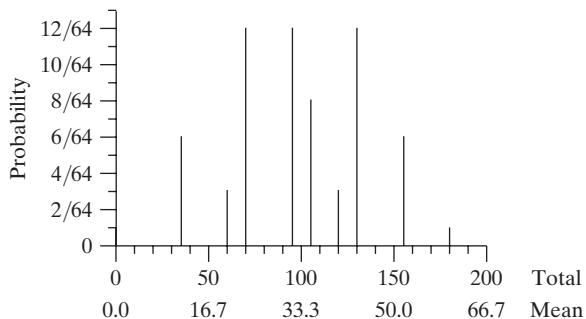
Sample	Costs (in units of \$1,000)	Sample total	Probability
None, None, None	0,0,0	0	1/64
None, None, Single	0,0,35	35	2/64
None, None, Double	0,0,60	60	1/64
None, Single, None	0,35,0	35	2/64
None, Single, Single	0,35,35	70	4/64
None, Single, Double	0,35,60	95	2/64
None, Double, None	0,60,0	60	1/64
None, Double, Single	0,60,35	95	2/64
None, Double, Double	0,60,60	120	1/64
Single, None, None	35,0,0	35	2/64
Single, None, Single	35,0,35	70	4/64
Single, None, Double	35,0,60	95	2/64
Single, Single, None	35,35,0	70	4/64
Single, Single, Single	35,35,35	105	8/64
Single, Single, Double	35,35,60	130	4/64
Single, Double, None	35,60,0	95	2/64
Single, Double, Single	35,60,35	130	4/64
Single, Double, Double	35,60,60	155	2/64
Double, None, None	60,0,0	60	1/64
Double, None, Single	60,0,35	95	2/64
Double, None, Double	60,0,60	120	1/64
Double, Single, None	60,35,0	95	2/64
Double, Single, Single	60,35,35	130	4/64
Double, Single, Double	60,35,60	155	2/64
Double, Double, None	60,60,0	120	1/64
Double, Double, Single	60,60,35	155	2/64
Double, Double, Double	60,60,60	180	1/64

The second column of Table 5.1.2 shows the sample mean (rounded to one decimal place) so that the last two columns of the table give the sampling distribution of the sample mean. These two distributions, shown graphically in Figure 5.1.2, are scaled versions of each other. An insurance company might speak in terms of total cost, but this is equivalent to looking at average cost. ■

## Relationship to Statistical Inference

Knowing a sampling distribution allows one to make probability statements about possible samples. For example, for the setting in Example 5.1.3 the insurance company might ask, What is the probability that the total knee replacement costs for a sample of three women will be less than \$110,000? We can answer this question by

Sample total	Sample mean	Probability
0	0.0	1/64
35	11.7	6/64
60	20.0	3/64
70	23.3	12/64
95	31.7	12/64
105	35.0	8/64
120	40.0	3/64
130	43.3	12/64
155	51.7	6/64
180	60.0	1/64



**Figure 5.1.2** Graph of the sampling distribution of total surgery costs for samples of size  $n = 3$

adding the probabilities of the first six outcomes listed in Table 5.1.2; the sum is 42/64. We will expand upon this idea as we formally develop ideas of statistical inference.

## Exercises 5.1.1–5.1.4

**5.1.1** Consider taking a random sample of size 3 from the knee replacement population of Example 5.1.3. What is the probability that the total cost for those in the sample will be greater than \$125,000?

**5.1.2** Consider taking a random sample of size 3 from the knee replacement population of Example 5.1.3. What is the probability that the total cost for those in the sample will be between \$80,000 and \$125,000?

**5.1.3** Consider taking a random sample of size 3 from the knee replacement population of Example 5.1.3. What is

the probability that the mean cost for those in the sample will be between \$40,000 and \$100,000?

**5.1.4** Consider a hypothetical population of dogs in which there are four possible weights, all of which are equally likely: 42, 48, 52, or 58 pounds. If a sample of size  $n = 2$  is drawn from this population, what is the sampling distribution of the total weight of the two dogs selected? That is, what are the possible values for the total and what are the probabilities associated with each of those values?

## 5.2 The Sample Mean

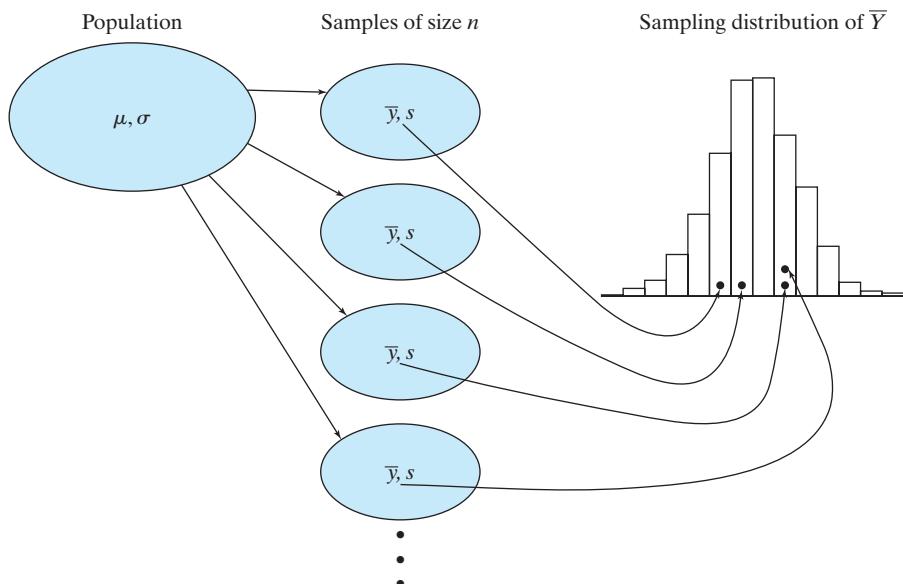
For a quantitative variable, the sample and the population can be described in various ways—by the mean, the median, the standard deviation, and so on. The natures (e.g., shape, center, spread) of the sampling distributions for these descriptive measures are not all the same. In this section we will focus primarily on the sampling distribution of the sample mean.

### The Sampling Distribution of $\bar{Y}$

The sample mean  $\bar{y}$  can be used, not only as a description of the data in the sample, but also as an estimate of the population mean  $\mu$ . It is natural to ask, “How close to  $\mu$  is  $\bar{y}$ ?” We cannot answer this question for the mean  $\bar{y}$  of a particular

sample, but we can answer it if we think in terms of the random sampling model and regard the sample mean as a random variable  $\bar{Y}$ . The question then becomes: “How close to  $\mu$  is  $\bar{Y}$  likely to be?” and the answer is provided by the **sampling distribution of  $\bar{Y}$** —that is, the probability distribution that describes sampling variability in  $\bar{Y}$ .

To visualize the sampling distribution of  $\bar{Y}$ , imagine the meta-study as follows: Random samples of size  $n$  are repeatedly drawn from a fixed population with mean  $\mu$  and standard deviation  $\sigma$ ; each sample has its own mean  $\bar{y}$ . The variation of the  $\bar{y}$ ’s among the samples is specified by the sampling distribution of  $\bar{Y}$ . This relationship is indicated schematically in Figure 5.2.1.



**Figure 5.2.1** Schematic representation of the sampling distribution of  $\bar{Y}$

When we think of  $\bar{Y}$  as a random variable, we need to be aware of two basic facts. The first of these is intuitive: On average, the sample mean equals the population mean. That is, the average of the sampling distribution of  $\bar{Y}$  is  $\mu$ . The second fact is not obvious: The standard deviation of  $\bar{Y}$  is equal to the standard deviation of  $Y$  divided by the square root of the sample size. That is, the standard deviation of  $\bar{Y}$  is  $\sigma/\sqrt{n}$ .

### Example 5.2.1

**Serum Cholesterol** The serum cholesterol levels of 12- to 14-year-olds follow a normal distribution with mean  $\mu = 162$  mg/dl and standard deviation  $\sigma = 28$  mg/dl.<sup>1</sup> If we take a random sample, then we expect the sample mean to be near 162, with the means of some samples being larger than 162 and the means of some samples being smaller than 162. As the preceding formula indicates, the amount of variability in the sample mean depends on the variability of cholesterol levels of the population,  $\sigma$ . If the population is very homogeneous (everyone has nearly the same cholesterol value so that  $\sigma$  is small), then samples and hence sample means would all be very similar and thus exhibit low variability. If the population is very heterogeneous ( $\sigma$  is large), then samples (and hence sample mean values) would vary more. While researchers have little control over the value of  $\sigma$ , we can control the sample size,  $n$ , and  $n$  affects the amount of variability in the sample mean. If we take a sample of

size  $n = 9$ , then the standard deviation of the sample mean is  $\frac{28}{\sqrt{9}} = \frac{28}{3} = 9.3$ . This means, loosely speaking, that the sample mean,  $\bar{Y}$ , will vary from one to sample to the next by about 9.3 mg/dl.\* If we took larger random samples of size  $n = 25$ , then the standard deviation of the sample mean would be smaller:  $\frac{28}{\sqrt{25}} = \frac{28}{5} = 5.6$ , which means that  $\bar{Y}$  would vary from one sample to the next by about 5.6. As the sample size goes up, the variability in the sample mean  $\bar{Y}$  goes down. ■

We now state as a theorem the basic facts about the sampling distribution of  $\bar{Y}$ . The theorem can be proved using the methods of mathematical statistics; we will state it without proof. The theorem describes the sampling distribution of  $\bar{Y}$  in terms of its mean (denoted by  $\mu_{\bar{Y}}$ ), its standard deviation (denoted by  $\sigma_{\bar{Y}}$ ), and its shape.\*\*

### Theorem 5.2.1: The Sampling Distribution of $\bar{Y}$

1. **Mean** The mean of the sampling distribution of  $\bar{Y}$  is equal to the population mean. In symbols,

$$\mu_{\bar{Y}} = \mu$$

2. **Standard deviation** The standard deviation of the sampling distribution of  $\bar{Y}$  is equal to the population standard deviation divided by the square root of the sample size. In symbols,

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

3. **Shape**

(a) If the population distribution of  $Y$  is normal, then the sampling distribution of  $\bar{Y}$  is normal, regardless of the sample size  $n$ .

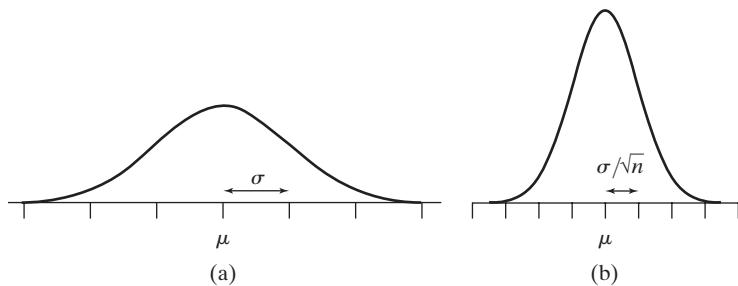
(b) *Central Limit Theorem* If  $n$  is large, then the sampling distribution of  $\bar{Y}$  is approximately normal, even if the population distribution of  $Y$  is not normal.

Parts 1 and 2 of Theorem 5.2.1 specify the relationship between the mean and standard deviation of the population being sampled, and the mean and standard deviation of the sampling distribution of  $\bar{Y}$ . Part 3(a) of the theorem states that, if the observed variable  $Y$  follows a normal distribution in the population being sampled, then the sampling distribution of  $\bar{Y}$  is also a normal distribution. These relationships are indicated in Figure 5.2.2.

---

\*Strictly speaking, the standard deviation measures deviation from the mean, not the difference between consecutive observations.

\*\*We are assuming here that the population is infinitely large or, equivalently, that we are sampling with replacement, so that we never exhaust the population. If we sample without replacement from a finite population then an adjustment is needed to get the right value for  $\sigma_{\bar{Y}}$ . Here  $\sigma_{\bar{Y}}$  is given by  $\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$ . The term  $\sqrt{\frac{N-n}{N-1}}$  is called the **finite population correction factor**. Note that if the sample size  $n$  is 10% of the population size  $N$ , then the correction factor is  $\sqrt{\frac{0.9N}{N-1}} \approx 0.95$ , so the adjustment is small. Thus, if  $n$  is small, in comparison to  $N$ , then the finite population correction factor is close to 1 and can be ignored.



**Figure 5.2.2** (a) The population distribution of a normally distributed variable  $Y$ ; (b) the sampling distribution of  $\bar{Y}$  in samples from the population of part (a)

The following example illustrates the meaning of parts 1, 2, and 3(a) of Theorem 5.2.1.

**Example 5.2.2**

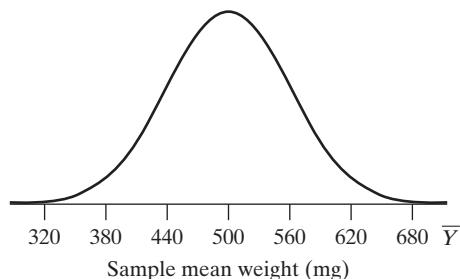
**Weights of Seeds** A large population of seeds of the princess bean *Phaseolus vulgaris* is to be sampled. The weights of the seeds in the population follow a normal distribution with mean  $\mu = 500$  mg and standard deviation  $\sigma = 120$  mg.<sup>2</sup> Suppose now that a random sample of four seeds is to be weighed, and let  $\bar{Y}$  represent the mean weight of the four seeds. Then, according to Theorem 5.2.1, the sampling distribution of  $\bar{Y}$  will be a normal distribution with mean and standard deviation as follows:

$$\mu_{\bar{Y}} = \mu = 500 \text{ mg}$$

and

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{120}{\sqrt{4}} = 60 \text{ mg}$$

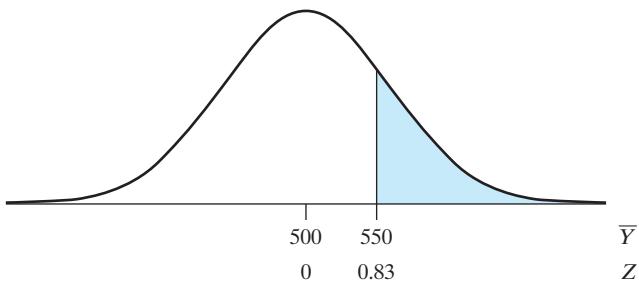
Thus, on average the sample mean will equal 500 mg, but the variability from one sample of size 4 to the next sample of size 4 is such that about two-thirds of the time  $\bar{Y}$  will be within 60 mg of 500 mg, that is, between  $500 - 60 = 440$  mg and  $500 + 60 = 560$  mg. Likewise, allowing for 2 standard deviations, we expect that  $\bar{Y}$  will be within 120 mg of 500 mg or between  $500 - 120 = 380$  mg and  $500 + 120 = 620$  mg about 95% of the time. The sampling distribution of  $\bar{Y}$  is shown in Figure 5.2.3; the ticks are 1 standard deviation apart. ■



**Figure 5.2.3** Sampling distribution of  $\bar{Y}$  for Example 5.2.2

The sampling distribution of  $\bar{Y}$  expresses the relative likelihood of the various possible values of  $\bar{Y}$ . For example, suppose we want to know the probability that the mean weight of the four seeds will be greater than 550 mg. This probability is shown as the shaded area in Figure 5.2.4. Notice that the value of  $\bar{y} = 550$  must be converted to the  $Z$  scale using the standard deviation  $\sigma_{\bar{Y}} = 60$ , not  $\sigma = 120$ .

$$z = \frac{\bar{y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \frac{550 - 500}{60} = 0.83$$



**Figure 5.2.4** Calculation of  $\Pr\{\bar{Y} > 550\}$  for Example 5.2.2

From Table 3,  $z = 0.83$  corresponds to an area of 0.7967. Thus,

$$\begin{aligned}\Pr\{\bar{Y} > 550\} &= \Pr\{Z > 0.83\} = 1 - 0.7967 \\ &= 0.2033 \approx 0.20\end{aligned}$$

This probability can be interpreted in terms of a meta-study as follows: If we were to choose many random samples of four seeds each from the population, then about 20% of the samples would have a mean weight exceeding 550 mg.

Part 3(b) of Theorem 5.2.1 is known as the **Central Limit Theorem**. The Central Limit Theorem states that, *no matter what distribution  $Y$  may have in the population*,\* if the sample size is large enough, then the sampling distribution of  $\bar{Y}$  will be approximately a normal distribution.

The Central Limit Theorem is of fundamental importance because it can be applied when (as often happens in practice) the form of the population distribution is not known. It is because of the Central Limit Theorem (and other similar theorems) that the normal distribution plays such a central role in statistics.

It is natural to ask how “large” a sample size is required by the Central Limit Theorem: How large must  $n$  be in order that the sampling distribution of  $\bar{Y}$  be well approximated by a normal curve? The answer is that the required  $n$  depends on the shape of the population distribution. If the shape is normal, any  $n$  will do. If the shape is moderately nonnormal, a moderate  $n$  is adequate. If the shape is highly nonnormal, then a rather large  $n$  will be required. (Some specific examples of this phenomenon are given in the optional Section 5.3.)

**Remark** We stated in Section 5.1 that the theory of this chapter is valid if the sample size is small compared to the population size. But the Central Limit Theorem is a statement about large samples. This may seem like a contradiction: How can a large sample be a small sample? In practice, there is no contradiction. In a typical biological application, the population size might be  $10^6$ ; a sample of size  $n = 100$  would be a small fraction of the population but would nevertheless be large enough for the Central Limit Theorem to be applicable (in most situations).

## Dependence on Sample Size

Consider the possibility of choosing random samples of various sizes from the same population. The sampling distribution of  $\bar{Y}$  will depend on the sample size  $n$  in two ways. First, its standard deviation is

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

---

\*Technically, the Central Limit Theorem requires that the distribution of  $Y$  have a standard deviation. In practice this condition is always met.

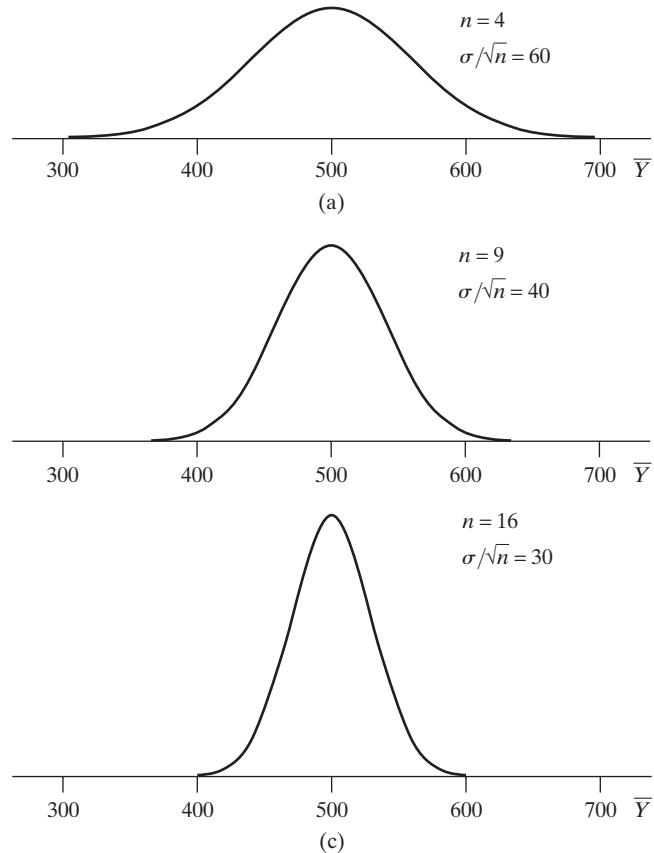
and this is inversely proportional to  $\sqrt{n}$ . Second, if the population distribution is not normal, then the *shape* of the sampling distribution of  $\bar{Y}$  depends on  $n$ , being more nearly normal for larger  $n$ . However, if the population distribution is normal, then the sampling distribution of  $\bar{Y}$  is always normal, and only the standard deviation depends on  $n$ .

The more important of the two effects of sample size is the first: Larger  $n$  gives a smaller value of  $\sigma_{\bar{Y}}$  and consequently a smaller expected sampling error if  $\bar{y}$  is used as an estimate of  $\mu$ . The following example illustrates this effect for sampling from a normal population.

**Example  
5.2.3**

**Weights of Seeds** Figure 5.2.5 shows the sampling distribution of  $\bar{Y}$  for samples of various sizes from the princess bean population of Example 5.2.2. Notice that for larger  $n$  the sampling distribution is more concentrated around the population mean  $\mu = 500$  mg. As a consequence, the probability that  $\bar{Y}$  is close to it is larger for larger  $n$ . For instance, consider the probability that  $\bar{Y}$  is within  $\pm 50$  mg of  $\mu$ , that is,  $\Pr\{450 \leq \bar{Y} \leq 550\}$ . Table 5.2.1 shows how this probability depends on  $n$ . ■

<b>Table 5.2.1</b>	
$n$	$\Pr\{450 \leq \bar{Y} \leq 550\}$
4	0.59
9	0.79
16	0.91
64	0.999



**Figure 5.2.5** Sampling distribution of  $\bar{Y}$  for various sample sizes  $n$

Example 5.2.3 illustrates how the closeness of  $\bar{Y}$  to  $\mu$  depends on sample size. The mean of a larger sample is not *necessarily* closer to it than the mean of a smaller sample, but it has a *greater probability* of being close. It is in this sense that a larger sample provides more information about the population mean than a smaller sample.

## Populations, Samples, and Sampling Distributions

In thinking about Theorem 5.2.1, it is important to distinguish clearly among three different distributions related to a quantitative variable  $Y$ : (1) the distribution of  $Y$  in the population; (2) the distribution of  $Y$  in a sample of data, and (3) the sampling distribution of  $\bar{Y}$ . The means and standard deviations of these distributions are summarized in Table 5.2.2.

<b>Table 5.2.2</b>		
Distribution	Mean	Standard deviation
$Y$ in population	$\mu$	$\sigma$
$Y$ in sample	$\bar{y}$	$s$
$\bar{Y}$ (in meta-study)	$\mu_{\bar{Y}} = \mu$	$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$

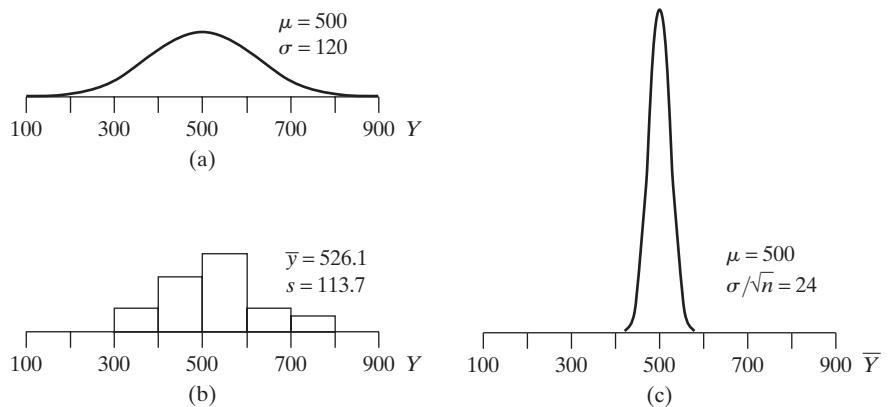
The following example illustrates the distinction among the three distributions.

### Example 5.2.4

**Weights of Seeds** For the princess bean population of Example 5.2.2, the population mean and standard deviation are  $\mu = 500$  mg and  $\sigma = 120$  mg; the population distribution of  $Y$  = weight is represented in Figure 5.2.6(a). Suppose we weigh a random sample of  $n = 25$  seeds from the population and obtain the data in Table 5.2.3.

For the data in Table 5.2.3, the sample mean is  $\bar{y} = 526.1$  mg and the sample standard deviation is  $s = 113.7$  mg. Figure 5.2.6(b) shows a histogram of the data; this histogram represents the distribution of  $Y$  in the sample. The sampling distribution of  $\bar{Y}$  is a theoretical distribution which relates, not to the particular sample shown in the histogram, but rather to the meta-study of repeated samples of size  $n = 25$ . The mean and standard deviation of the sampling distribution are

$$\mu_{\bar{Y}} = 500 \text{ mg and } \sigma_{\bar{Y}} = 120/\sqrt{25} = 24 \text{ mg}$$



**Figure 5.2.6** Three distributions related to  $Y$  = seed weight of princess beans:  
(a) population distribution of  $Y$ ; (b) distribution of 25 observations of  $Y$ ;  
(c) sampling distribution of  $\bar{Y}$  for  $n = 25$

**Table 5.2.3** Weights of twenty-five princess bean seeds

Weight (mg)						
343	755	431	480	516	469	694
659	441	562	597	502	612	549
348	469	545	728	416	536	581
433	583	570	334			

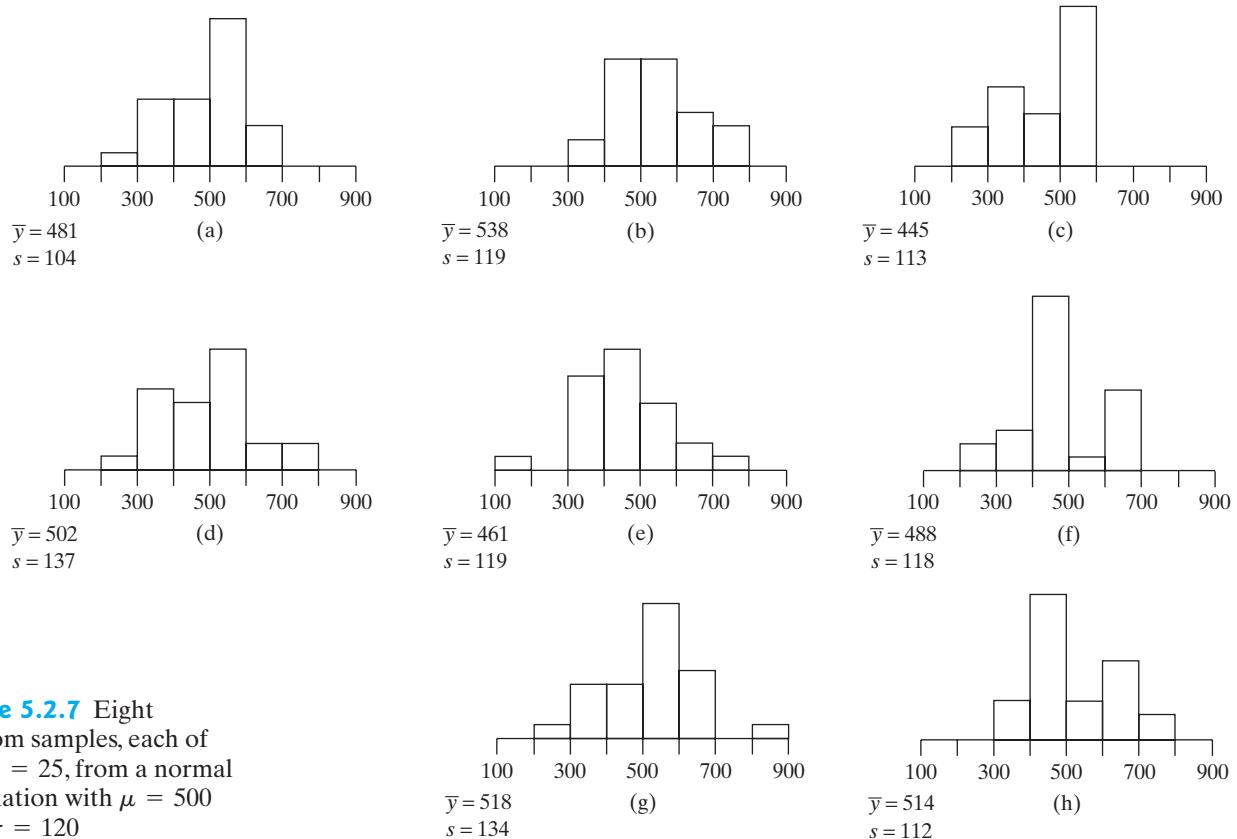
The sampling distribution is represented in Figure 5.2.6(c). Notice that the distributions in Figures 5.2.6(a) and (b) are more or less similar; in fact, the distribution in (b) is an estimate (based on the data in Table 5.2.3) of the distribution in (a). By contrast, the distribution in (c) is much narrower, because it represents a distribution of *means* rather than of individual observations.

## Other Aspects of Sampling Variability

The preceding discussion has focused on sampling variability in the sample mean,  $\bar{Y}$ . Two other important aspects of sampling variability are (1) sampling variability in the sample standard deviation,  $s$  and (2) sampling variability in the *shape* of the sample, as represented by the sample histogram. Rather than discuss these aspects formally, we illustrate them with the following example.

### Example 5.2.5

**Weights of Seeds** In Figure 5.2.6(b) we displayed a random sample of 25 observations from the princess bean population of Example 5.2.2; now we display in Figure 5.2.7 eight additional random samples from the same population. (All nine samples were actually simulated using a computer.) Notice that, even though the samples were drawn from a normal population [pictured in Figure 5.2.6(a)], there is very substantial variation in the forms of the histograms. Notice also that there is considerable variation in the sample standard deviations. Of course, if the sample size were larger (say,  $n = 100$  rather than  $n = 25$ ), there would be less sampling variation; the histograms would tend to resemble a normal curve more closely, and the standard deviations would tend to be closer to the population value ( $\sigma = 120$ ).



**Figure 5.2.7** Eight random samples, each of size  $n = 25$ , from a normal population with  $\mu = 500$  and  $\sigma = 120$

## Exercises 5.2.1–5.2.19

**5.2.1 (Sampling exercise)** Refer to Exercise 1.3.5. The collection of 100 ellipses shown there can be thought of as representing a natural population of the organism *C. ellipticus*. Use your judgment to choose a sample of 5 ellipses that you think should be reasonably representative of the population. (In order to best simulate the analogous judgment in a real-life setting, you should make your choice intuitively, without any detailed preliminary study of the population.) With a metric ruler, measure the length of each ellipse in your sample. Measure only the body, excluding any tail bristles; measurements to the nearest millimeter will be adequate. Compute the mean and standard deviation of the five lengths. To facilitate the pooling of results from the entire class, express the mean and standard deviation in millimeters, keeping two decimal places.

**5.2.2 (Sampling exercise)** Proceed as in Exercise 5.2.1, but use random sampling rather than “judgment” sampling. To do this, choose 10 random digits (from Table 1 or your calculator). Let the first 2 digits be the number of the first ellipse that goes into your sample, and so on. The 10 random digits will give you a random sample of five ellipses.

**5.2.3 (Sampling exercise)** Proceed as in Exercise 5.2.2, but choose a random sample of 20 ellipses.

**5.2.4** Refer to Exercise 5.2.2. The following scheme is proposed for choosing a sample of 5 ellipses from the population of 100 ellipses. (i) Choose a point at random in the ellipse “habitat” (that is, the figure); this could be done crudely by dropping a pencil point on the page, or much better by overlaying the page with graph paper and using random digits. (ii) If the chosen point is inside an ellipse, include that ellipse in the sample, otherwise start again at step (i). (iii) Continue until 5 ellipses have been selected. Explain why this scheme is not equivalent to random sampling. In what direction is the scheme biased—that is, would it tend to produce a  $\bar{y}$  that is too large, or a  $\bar{y}$  that is too small?

**5.2.5** The serum cholesterol levels of a population of 12- to 14-year-olds follow a normal distribution with mean 162 mg/dl and standard deviation 28 mg/dl (as in Example 4.1.1).

- What percentage of the 12- to 14-year-olds have serum cholesterol values between 152 and 172 mg/dl?
- Suppose we were to choose at random from the population a large number of groups of nine 12- to 14-year-olds each. In what percentage of the groups would the group mean cholesterol value be between 152 and 172 mg/dl?

- If  $\bar{Y}$  represents the mean cholesterol value of a random sample of nine 12- to 14-year-olds from the population, what is  $\Pr\{152 \leq \bar{Y} \leq 172\}$ ?

**5.2.6** An important indicator of lung function is forced expiratory volume (FEV), which is the volume of air that a person can expire in one second. Dr. Hernandez plans to measure FEV in a random sample of  $n$  young women from a certain population, and to use the sample mean  $\bar{y}$  as an estimate of the population mean. Let  $E$  be the event that Hernandez’s sample mean will be within  $\pm 100$  ml of the population mean. Assume that the population distribution is normal with mean 3,000 ml and standard deviation 400 ml.<sup>3</sup> Find  $\Pr\{E\}$  if

- $n = 15$
- $n = 60$
- How does  $\Pr\{E\}$  depend on the sample size? That is, as  $n$  increases, does  $\Pr\{E\}$  increase, decrease, or stay the same?

**5.2.7** Refer to Exercise 5.2.6. Assume that the population distribution of FEV is normal with standard deviation 400 ml.

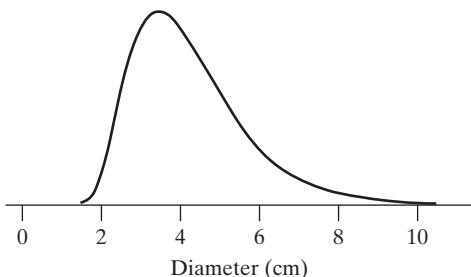
- Find  $\Pr\{E\}$  if  $n = 15$  and the population mean is 2,800 ml.
- Find  $\Pr\{E\}$  if  $n = 15$  and the population mean is 2,600 ml.
- How does  $\Pr\{E\}$  depend on the population mean?

**5.2.8** The heights of a certain population of corn plants follow a normal distribution with mean 145 cm and standard deviation 22 cm (as in Exercise 4.S.4).

- What percentage of the plants are between 135 and 155 cm tall?
- Suppose we were to choose at random from the population a large number of samples of 16 plants each. In what percentage of the samples would the sample mean height be between 135 and 155 cm?
- If  $\bar{Y}$  represents the mean height of a random sample of 16 plants from the population, what is  $\Pr\{135 \leq \bar{Y} \leq 155\}$ ?
- If  $\bar{Y}$  represents the mean height of a random sample of 36 plants from the population, what is  $\Pr\{135 \leq \bar{Y} \leq 155\}$ ?

**5.2.9** The basal diameter of a sea anemone is an indicator of its age. The density curve shown here represents the distribution of diameters in a certain large population of anemones; the population mean diameter is 4.2 cm, and the standard deviation is 1.4 cm.<sup>4</sup> Let  $\bar{Y}$  represent the

mean diameter of 25 anemones randomly chosen from the population.



- (a) Find the approximate value of  $\Pr\{4 \leq \bar{Y} \leq 5\}$ .  
 (b) Why is your answer to part (a) approximately correct even though the population distribution of diameters is clearly not normal? Would the same approach be equally valid for a sample of size 2 rather than 25? Why or why not?

**5.2.10** In a certain population of fish, the lengths of the individual fish follow approximately a normal distribution with mean 54.0 mm and standard deviation 4.5 mm. We saw in Example 4.3.1 that in this situation 65.68% of the fish are between 51 and 60 mm long. Suppose a random sample of four fish is chosen from the population. Find the probability that

- (a) all four fish are between 51 and 60 mm long.  
 (b) the mean length of the four fish is between 51 and 60 mm.

**5.2.11** In Exercise 5.2.10, the answer to part (b) was larger than the answer to part (a). Argue that this must necessarily be true, no matter what the population mean and standard deviation might be. [Hint: Can it happen that the event in part (a) occurs but the event in part (b) does not?]

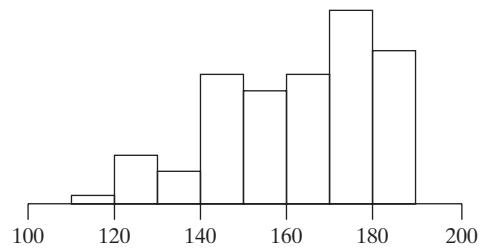
**5.2.12** Professor Smith conducted a class exercise in which students ran a computer program to generate random samples from a population that had a mean of 50 and a standard deviation of 9 mm. Each of Smith's students took a random sample of size  $n$  and calculated the sample mean. Smith found that about 68% of the students had sample means between 48.5 and 51.5 mm. What was  $n$ ? (Assume that  $n$  is large enough that the Central Limit Theorem is applicable.)

**5.2.13** A certain assay for serum alanine aminotransferase (ALT) is rather imprecise. The results of repeated assays of a single specimen follow a normal distribution with mean equal to the ALT concentration for that specimen and standard deviation equal to 4 U/l (as in Exercise 4.S.15). Suppose a hospital lab measures many specimens every day, and specimens with reported ALT values of 40 or more are flagged as "unusually high." If a patient's true ALT concentration is 35 U/l, find the probability that his specimen will be flagged as "unusually high".

- (a) if the reported value is the result of a single assay

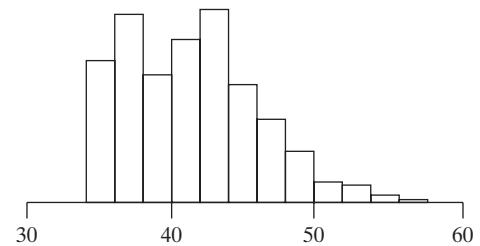
- (b) if the reported value is the mean of three independent assays of the same specimen.

**5.2.14** The mean of the distribution shown in the following histogram is 162 and the standard deviation is 18. Consider taking random samples of size  $n = 9$  from this distribution and calculating the sample mean,  $\bar{y}$ , for each sample.



- (a) What is the mean of the sampling distribution of  $\bar{Y}$ ?  
 (b) What is the standard deviation of the sampling distribution of  $\bar{Y}$ ?

**5.2.15** The mean of the distribution shown in the following histogram is 41.5 and the standard deviation is 4.7. Consider taking random samples of size  $n = 4$  from this distribution and calculating the sample mean,  $\bar{y}$ , for each sample.



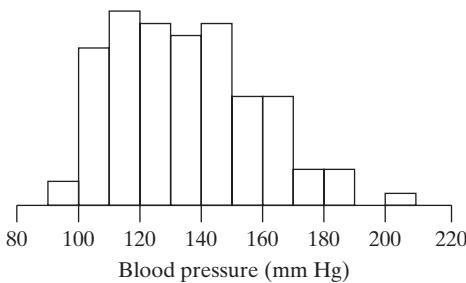
- (a) What is the mean of the sampling distribution of  $\bar{Y}$ ?  
 (b) What is the standard deviation of the sampling distribution of  $\bar{Y}$ ?

**5.2.16** Refer to the histogram in Exercise 5.2.15. Suppose that 100 random samples are taken from this population and the sample mean is calculated for each sample. If we were to make a histogram of the distribution of the sample means from 100 samples, what kind of shape would we expect the histogram to have?

- (a) if  $n = 2$  for each random sample?  
 (b) if  $n = 25$  for each random sample?

**5.2.17** Refer to the histogram in Exercise 5.2.15. Suppose that 100 random samples are taken from this population and the sample mean is calculated for each sample. If we were to make a histogram of the distribution of the sample means from 100 samples, what kind of shape would we expect the histogram to have if  $n = 1$  for each random sample? That is, what does the sampling distribution of the mean look like when the sample size is  $n = 1$ ?

**5.2.18** A medical researcher measured systolic blood pressure in 100 middle-aged men.<sup>5</sup> The results are displayed in the accompanying histogram; note that the distribution is rather skewed. According to the Central Limit Theorem, would we expect the distribution of blood pressure readings to be less skewed (and more bell shaped) if it were based on  $n = 400$  rather than  $n = 100$  men? Explain.



**5.2.19** The partial pressure of oxygen,  $\text{PaO}_2$ , is a measure of the amount of oxygen in the blood. Assume that the distribution of  $\text{PaO}_2$  levels among newborns has an average of 38 (mm Hg) and a standard deviation of 9.<sup>6</sup> If we take a sample of size  $n = 25$ ,

- what is the probability that the sample average will be greater than 36?
- what is the probability that the sample average will be greater than 41?

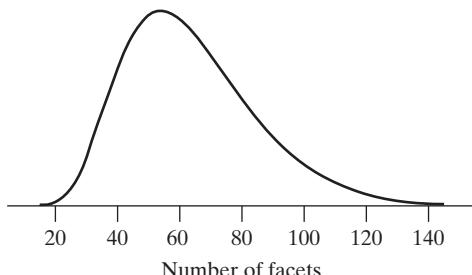
## 5.3 Illustration of the Central Limit Theorem (Optional)

The importance of the normal distribution in statistics is due largely to the Central Limit Theorem and related theorems. In this section we take a closer look at the Central Limit Theorem. According to the Central Limit Theorem, the sampling distribution of  $\bar{Y}$  is approximately normal if  $n$  is large. If we consider larger and larger samples from a fixed nonnormal population, then the sampling distribution of  $\bar{Y}$  will be more nearly normal for larger  $n$ . The following examples show the Central Limit Theorem at work for two nonnormal distributions: a moderately skewed distribution (Example 5.3.1) and a highly skewed distribution (Example 5.3.2).

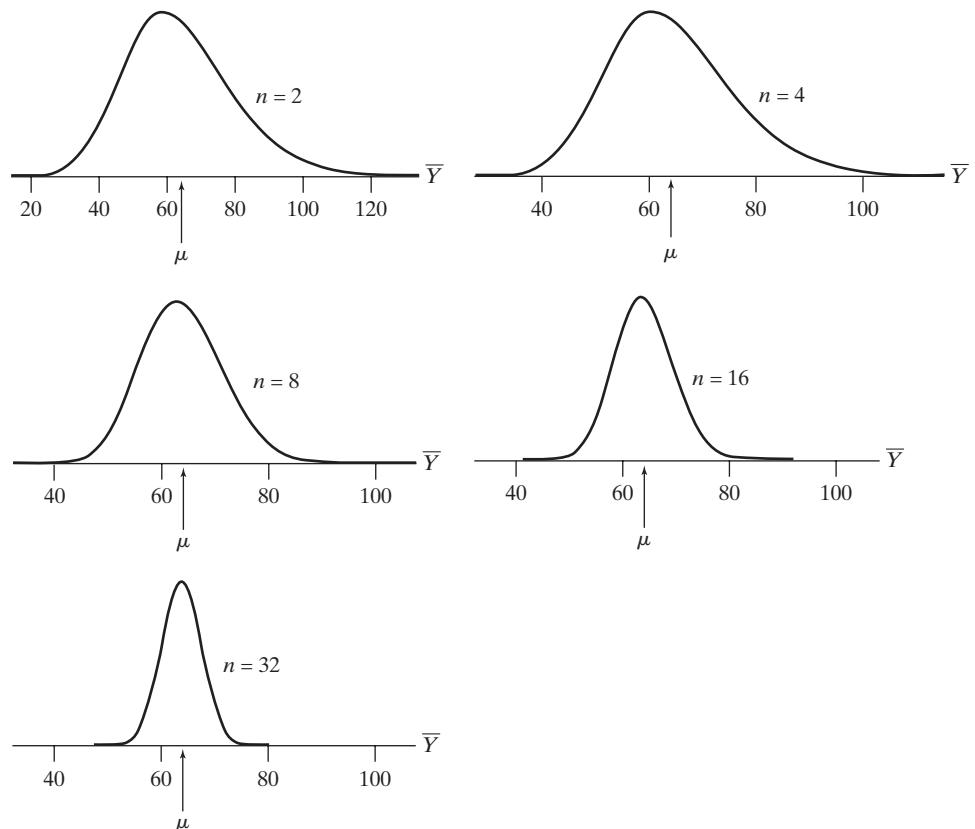
### Example 5.3.1

**Eye Facets** The number of facets in the eye of the fruitfly *Drosophila melanogaster* is of interest in genetic studies. The distribution of this variable in a certain *Drosophila* population can be approximated by the density function shown in Figure 5.3.1. The distribution is moderately skewed; the population mean and standard deviation are  $\mu = 64$  and  $\sigma = 22$ .<sup>7</sup>

Figure 5.3.2 shows the sampling distribution of  $\bar{Y}$  for samples of various sizes from the eye-facet population. In order to clearly show the shape of these distributions, we have plotted them to different scales; the horizontal scale is stretched more for larger  $n$ . Notice that the distributions are somewhat skewed to the right, but the skewness is diminished for larger  $n$ ; for  $n = 32$  the distribution looks very nearly normal. ■



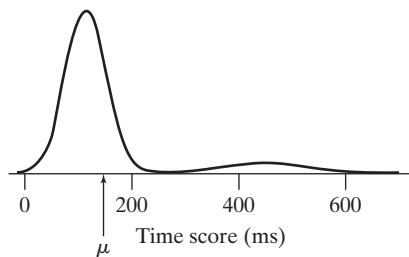
**Figure 5.3.1** Distribution of eye-facet number in a *Drosophila* population



**Figure 5.3.2** Sampling distributions of  $\bar{Y}$  for samples from the *Drosophila* eye-facet population

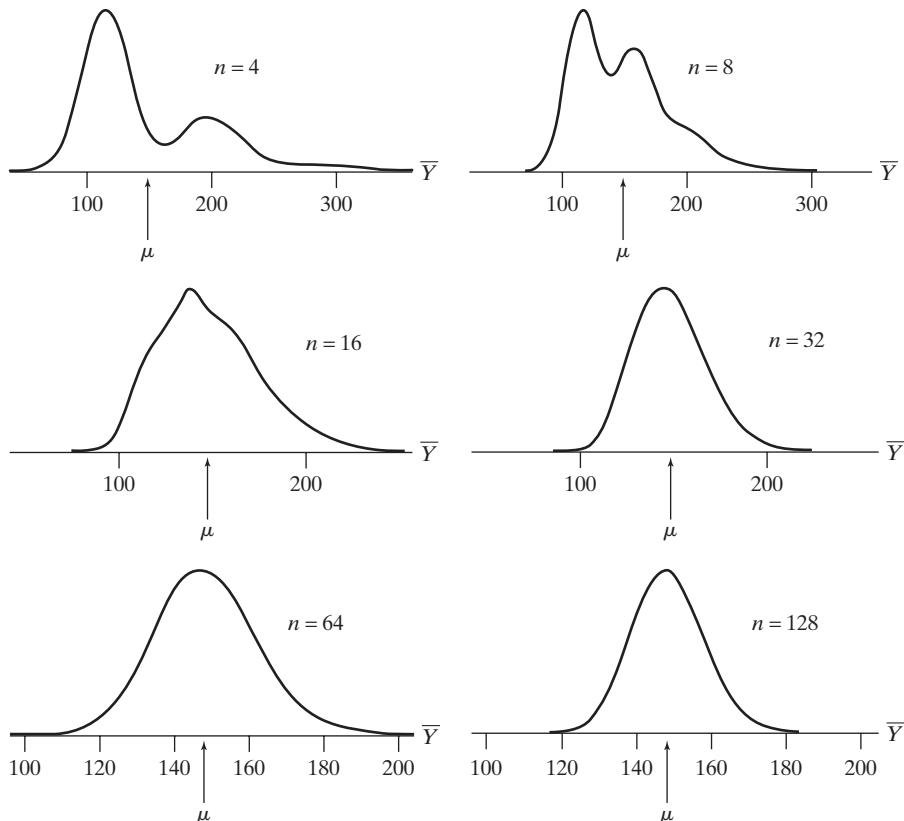
**Example 5.3.2**

**Reaction Time** A psychologist measured the time required for a person to reach up from a fixed position and operate a pushbutton with his or her forefinger. The distribution of time scores (in milliseconds) for a single person is represented by the density shown in Figure 5.3.3. About 10% of the time, the subject fumbled, or missed the button on the first thrust; the resulting delayed times appear as the second peak of the distribution.<sup>8</sup> The first peak is centered at 115 ms and the second at 450 ms; because of the two peaks, the overall distribution is violently skewed. The population mean and standard deviation are  $\mu = 148$  ms and  $\sigma = 105$  ms, respectively.



**Figure 5.3.3** Distribution of time scores in a button-pushing task

Figure 5.3.4 shows the sampling distribution of  $\bar{Y}$  for samples of various sizes from the time-score distribution. To show the shape clearly, the  $Y$  scale has been stretched more for larger  $n$ . Notice that for small  $n$  the distribution has several modes. As  $n$  increases, these modes are reduced to bumps and finally disappear, and the distribution becomes increasingly symmetric. ■



**Figure 5.3.4** Sampling distributions of  $\bar{Y}$  for samples from the time-score population

Examples 5.3.1 and 5.3.2 illustrate the fact, mentioned in Section 5.2, that the meaning of the requirement “ $n$  is large” in the Central Limit Theorem depends on the shape of the population distribution. Approximate normality of the sampling distribution of  $\bar{Y}$  will be achieved for a moderate  $n$  if the population distribution is only moderately nonnormal (as in Example 5.3.1), while a highly nonnormal population (as in Example 5.3.2) will require a larger  $n$ .

Note, however, that Example 5.3.2 indicates the remarkable strength of the Central Limit Theorem. The skewness of the time-score distribution is so extreme that one might be reluctant to consider the mean as a summary measure. Even in this “worst case,” you can see the effect of the Central Limit Theorem in the relative smoothness and symmetry of the sampling distribution for  $n = 64$ .

The Central Limit Theorem may seem rather like magic. To demystify it somewhat, we look at the time-score sampling distributions in more detail in the following example.

---

**Example 5.3.3**

**Reaction Time** Consider the sampling distributions of  $\bar{Y}$  displayed in Figure 5.3.4. Consider first the distribution for  $n = 4$ , which is the distribution of the mean of four button-pressing times. The high peak at the left of the distribution represents cases in which the subject did not fumble any of the 4 thrusts, so that all four times were about 115 ms; such an outcome would occur about 66% of the time [from the binomial distribution, because  $(0.9)^4 = 0.66$ ]. The next lower peak represents cases in which 3 thrusts took about 115 ms each, while one was fumbled and took about 450 ms. (Notice that the average of three 115's and one 450 is about 200, which is the center of the second peak.) Similarly, the third peak (which is barely visible)

represents cases in which the subject fumbled 2 of the 4 thrusts. The peaks representing 3 and 4 fumbles are too low to be visible in the plot.

Now consider the plot for  $n = 8$ . The first peak represents 8 good thrusts (no fumbles), the second represents 7 good thrusts and 1 fumble, the third represents 6 good thrusts and 2 fumbles, and so on. The fourth and later peaks are blended together. For  $n = 16$  it is more likely to see 15 good thrusts and 1 fumble than 16 good thrusts (as you can verify from the binomial distribution) and thus there is a bump, corresponding to 16 good thrusts, below the overall peak, which corresponds to 15 good thrusts; the bump to the right of the peak corresponds to 14 good thrusts and 2 fumbles. For  $n = 32$ , the most likely outcome is 3 fumbles and 29 good thrusts; this outcome gives a mean time of about

$$\frac{(3)(450) + (29)(115)}{32} \approx 146 \text{ ms}$$

which is the location of the central peak. For similar reasons, the distribution for larger  $n$  is centered at about 148 ms, which is the population mean. ■

### Exercises 5.3.1–5.3.3

**5.3.1** Refer to Example 5.3.3. In the sampling distribution of  $\bar{Y}$  for  $n = 4$  (Figure 5.3.4), approximately what is the area under

- (a) the first peak?
  - (b) the second peak?
- (Hint: Use the binomial distribution.)

**5.3.2** Refer to Example 5.3.3. Consider the sampling distribution of  $\bar{Y}$  for  $n = 2$  (which is not shown in Figure 5.3.4).

- (a) Make a rough sketch of the sampling distribution. How many peaks does it have? Show the location (on the  $Y$ -axis) of each peak.

(b) Find the approximate area under each peak. (Hint: Use the binomial distribution.)

**5.3.3** Refer to Example 5.3.3. Consider the sampling distribution of  $\bar{Y}$  for  $n = 1$  (which is not shown in Figure 5.3.4). Make a rough sketch of the sampling distribution. How many peaks does it have? Show the location (on the  $Y$ -axis) of each peak.

## 5.4 The Normal Approximation to the Binomial Distribution (Optional)

The Central Limit Theorem tells us that the sampling distribution of a mean becomes bell shaped as the sample size increases. Suppose we have a large dichotomous population for which we label the two types of outcomes as “1” (for “success”) and “0” (for “failure”). If we take a sample and calculate the average number of 1’s, then this sample average is just the sample proportion of 1’s—commonly labeled as  $\hat{P}$ —and is governed by the Central Limit Theorem. This means that if the sample size  $n$  is large, then the distribution of  $\hat{P}$  will be approximately normal.

Note that if we know the number of 1’s (i.e., the number of successes in  $n$  trials), then we know the proportion of 1’s and vice versa. Thus, the normal approximation to the binomial distribution can be expressed in two equivalent ways: in terms of the number of successes,  $Y$ , or in terms of the proportion of successes,  $\hat{P}$ . We state both forms in the following theorem. In this theorem,  $n$  represents the sample size (or, more generally, the number of independent trials) and  $p$  represents the population proportion (or, more generally, the probability of success in each independent trial).

**Theorem 5.4.1: Normal Approximation to Binomial Distribution**

- (a) If  $n$  is large, then the binomial distribution of the number of successes,  $Y$ , can be approximated by a normal distribution with

$$\text{Mean} = np$$

and

$$\text{Standard deviation} = \sqrt{np(1 - p)}$$

- (b) If  $n$  is large, then the sampling distribution of  $\hat{P}$  can be approximated by a normal distribution with

$$\text{Mean} = p$$

and

$$\text{Standard deviation} = \sqrt{\frac{p(1 - p)}{n}}$$

**Remarks**

1. Appendix 5.1 provides more detailed explanation of the relationship between the normal approximation to the binomial and the Central Limit Theorem.
2. As shown in Appendix 3.2, for a population of 0's and 1's, where the proportion of 1's is given by  $p$ , the standard deviation is  $\sigma = \sqrt{p(1 - p)}$ . Theorem 5.2.1 stated that the standard deviation of a mean is given by  $\frac{\sigma}{\sqrt{n}}$ . We think of  $\hat{P}$  in part (b) of Theorem 5.2.1 as a special kind of sample average, for the setting in which all of the data are 0's and 1's. Thus, Theorem 5.2.1 tells us that the standard deviation of  $\hat{P}$  should be  $\frac{\sqrt{p(1 - p)}}{\sqrt{n}}$ , or  $\sqrt{\frac{p(1 - p)}{n}}$ , which agrees with the result stated in Theorem 5.4.1(b).

The following example illustrates the use of Theorem 5.4.1.

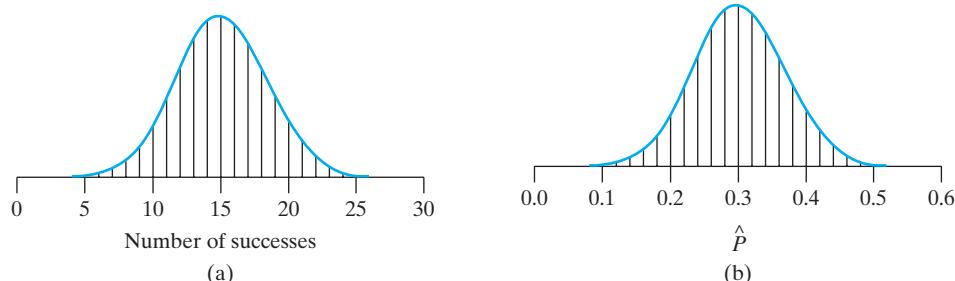
**Example  
5.4.1**

**Normal Approximation to Binomial** We consider a binomial distribution with  $n = 50$  and  $p = 0.3$ . Figure 5.4.1(a) shows this binomial distribution, using spikes to represent probabilities; superimposed is a normal curve with

$$\text{Mean} = np = (50)(0.3) = 15$$

and

$$\text{SD} = \sqrt{np(1 - p)} = \sqrt{(50)(0.3)(0.7)} = 3.24$$



**Figure 5.4.1** The normal approximation (blue curve) to the binomial distribution (black spikes) with  $n = 50$  and  $p = 0.3$

Note that the curve fits the distribution fairly well. Figure 5.4.1(b) shows the sampling distribution of  $\hat{P}$  for  $n = 50$  and  $p = 0.3$ ; superimposed is a normal curve with

$$\text{Mean} = p = 0.3$$

and

$$\text{SD} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.3)(0.7)}{50}} = 0.0648$$

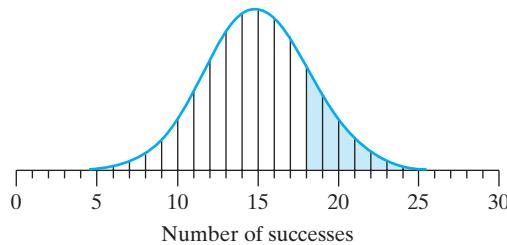
Note that Figure 5.4.1(b) is just a relabeled version of Figure 5.4.1(a).

To illustrate the use of the normal approximation, let us find the probability that 50 independent trials result in at least 18 successes. We could use the binomial formula to find the probability of exactly 18 successes in 50 trials and add this to the probability of exactly 19 successes, exactly 20 successes, and so on:

$$\begin{aligned}\Pr\{\text{at least 18 successes}\} &= {}_{50}C_{18}(0.3)^{18}(1 - 0.3)^{50-18} \\ &\quad + {}_{50}C_{19}(0.3)^{19}(1 - 0.3)^{50-19} + \dots \\ &= 0.0772 + 0.0558 + \dots = 0.2178\end{aligned}$$

This probability can be visualized as the area above and to the right of the “18” in Figure 5.4.2. The normal approximation to the probability is the corresponding area under the normal curve, which is shaded in Figure 5.4.2. The  $z$  value that corresponds to 18 is

$$z = \frac{18 - 15}{3.2404} = 0.93$$



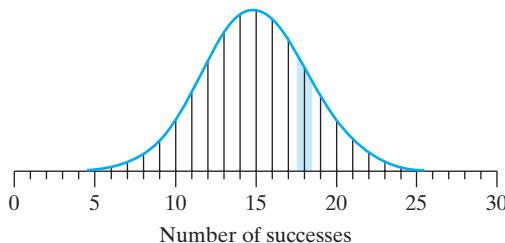
**Figure 5.4.2** Normal approximation to the probability of at least 18 successes

From Table 3, we find that the area is  $1 - 0.8238 = 0.1762$ , which is reasonably close to the exact value of 0.2178. This approximation can be improved by accounting for the fact that the binomial distribution is discrete and the normal distribution is continuous as we shall see below. ■

## The Continuity Correction

As we have seen in Chapter 4, because the normal distribution is continuous, probabilities are computed areas under the normal curve, rather than being the height of the normal curve at any particular value. Because of this, to compute  $\Pr\{Y = 18\}$ , the probability of 18 successes, we think of “18” as covering the space from 17.5 to 18.5 and thus we consider the area under the normal curve between 17.5 and 18.5; this is illustrated in Figure 5.4.3. Likewise, to get a more accurate approximation in Example 5.4.1, we can use 17.5 in place of 18 when finding the  $z$  value. Each of these is an example of a continuity correction.

**Figure 5.4.3** Normal approximation to the probability of exactly 18 successes



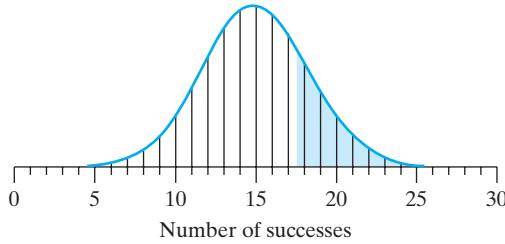
**Example 5.4.2**

Applying continuity correction within the normal approximation, the probability of at least 18 successes in 50 trials, when  $p = 0.3$ , is approximated by finding

$$z = \frac{17.5 - 15}{3.2404} = 0.77$$

From Table 3, we find that the area above 0.77 is  $1 - 0.7794 = 0.2206$ , which agrees quite well with the exact value of 0.2178. This area is displayed in Figure 5.4.4. ■

**Figure 5.4.4** Improved normal approximation to the probability of at least 18 successes



**Example 5.4.3**

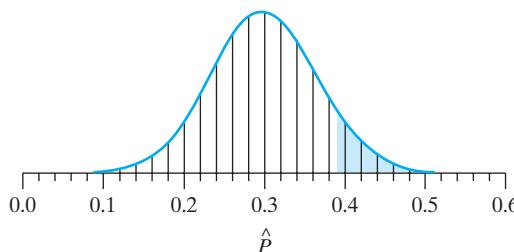
To illustrate part (b) of Theorem 5.4.1, we again assume that  $n = 50$  and  $p = 0.3$ . Consider finding the probability that at least 40% of the 50 trials in a binomial experiment with  $p = 0.3$  result in successes. That is, we wish to find  $\Pr\{\hat{P} \geq 0.40\}$ . The normal approximation to this probability is the shaded area in Figure 5.4.5. Using continuity correction, the boundary of the area is  $\hat{p} = 19.5/50 = 0.39$ , which corresponds on the Z scale to

$$z = \frac{0.39 - 0.30}{0.0648} = 1.39$$

The resulting approximation (from Table 3) is then

$$\Pr\{\hat{P} \geq 0.40\} \approx 1 - 0.9177 = 0.0823$$

**Figure 5.4.5** Normal approximation to  $\Pr\{\hat{P} \geq 0.40\}$



which agrees very well with the exact value of 0.0848 (found by using the binomial formula). ■

**Remark** Any problem involving the normal approximation to the binomial can be solved in two ways: in terms of  $Y$ , using part (a) of Theorem 5.4.1, or in terms of  $\hat{P}$ , using part (b) of the theorem. Although it is natural to state questions in terms of proportions (e.g., “What is  $\Pr\{\hat{P} > 0.70\}?$ ”), it is often easier to solve problems in terms of the binomial count  $Y$  (e.g., “What is  $\Pr\{Y > 35\}?$ ”), particularly when using continuity correction. The following example illustrates the approach of converting a question about a sample proportion into a question about the number of successes for a binomial random variable.

**Example  
5.4.4**

Consider a binomial distribution with  $n = 50$  and  $p = 0.3$ . The sample proportion of successes, out of the 50 trials, is  $\hat{P}$ . Figure 5.4.1(b) shows the sampling distribution of  $\hat{P}$  with a normal curve superimposed.

Suppose we wish to find the probability that  $0.24 \leq \hat{P} \leq 0.36$ . Since  $\hat{P} = Y/50$ , this is the probability that  $0.24 \leq Y/50 \leq 0.36$ , which is the same as the probability that  $12 \leq Y \leq 18$ . That is,  $\Pr\{0.24 \leq \hat{P} \leq 0.36\} = \Pr\{12 \leq Y \leq 18\}$ .

We know that  $Y$  has a binomial distribution with mean  $= np = (50)(0.3) = 15$  and  $SD = \sqrt{np(1 - p)} = \sqrt{(50)(0.3)(0.7)} = 3.24$ . Using continuity correction, we would find the  $Z$  scale values of

$$z = \frac{11.5 - 15}{3.24} = -1.08$$

and

$$z = \frac{18.5 - 15}{3.24} = 1.08$$

Then, using Table 3, we have  $\Pr\{0.24 \leq \hat{P} \leq 0.36\} = \Pr\{12 \leq Y \leq 18\} \approx 0.8599 - 0.1401 = 0.7198$ . ■

## How Large Must $n$ Be?

Theorem 5.4.1 states that the binomial distribution can be approximated by a normal distribution if  $n$  is “large.” It is helpful to know how large  $n$  must be in order for the approximation to be adequate. The required  $n$  depends on the value of  $p$ . If  $p = 0.5$ , then the binomial distribution is symmetric and the normal approximation is quite good even for  $n$  as small as 10. However, if  $p = 0.1$ , the binomial distribution for  $n = 10$  is quite skewed, and is poorly fitted by a normal curve; for larger  $n$  the skewness is diminished and the normal approximation is better. A simple rule of thumb is the following:

The normal approximation to the binomial distribution is fairly good if both  $np$  and  $n(1 - p)$  are at least equal to 5.

For example, if  $n = 50$  and  $p = 0.3$ , as in Example 5.4.4, then  $np = 15$  and  $n(1 - p) = 35$ ; since  $15 \geq 5$  and  $35 \geq 5$ , the rule of thumb indicates that the normal approximation is fairly good.

## Exercises 5.4.1–5.4.13

**5.4.1** A fair coin is to be tossed 20 times. Find the probability that 10 of the tosses will fall heads and 10 will fall tails,

- (a) using the binomial distribution formula.
- (b) using the normal approximation with the continuity correction.

**5.4.2** In the United States, 44% of the population has type O blood. Suppose a random sample of 12 persons is taken. Find the probability that 6 of the persons will have type O blood (and 6 will not)

- (a) using the binomial distribution formula.
- (b) using the normal approximation.

**5.4.3** Refer to Exercise 5.4.2. Find the probability that at most 6 of the persons will have type O blood by using the normal approximation

- (a) without the continuity correction.
- (b) with the continuity correction.

**5.4.4** An epidemiologist is planning a study on the prevalence of oral contraceptive use in a certain population.<sup>9</sup> She plans to choose a random sample of  $n$  women and to use the sample proportion of oral contraceptive users ( $\hat{P}$ ) as an estimate of the population proportion ( $p$ ). Suppose that in fact  $p = 0.12$ . Use the normal approximation (with the continuity correction) to determine the probability that  $\hat{P}$  will be within  $\pm 0.03$  of  $p$  if

- (a)  $n = 100$ .
- (b)  $n = 200$ .

[Hint: If you find using part (b) of Theorem 5.4.1 to be difficult here, try using part (a) of the theorem instead.]

**5.4.5** In a study of how people make probability judgments, college students (with no background in probability or statistics) were asked the following question.<sup>10</sup> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of one year, each hospital recorded the days on which at least 60% of the babies born were boys. Which hospital do you think recorded more such days?

- The larger hospital
- The smaller hospital
- About the same (i.e., within 5% of each other)

- (a) Imagine that you are a participant in the study. Which answer would you choose, based on intuition alone?
- (b) Determine the correct answer by using the normal approximation (without the continuity correction) to calculate the appropriate probabilities.

**5.4.6** Consider random sampling from a dichotomous population with  $p = 0.3$ , and let  $E$  be the event that  $\hat{P}$  is

within  $\pm 0.05$  of  $p$ . Use the normal approximation (without the continuity correction) to calculate  $\Pr\{E\}$  for a sample of size  $n = 400$ .

**5.4.7** Refer to Exercise 5.4.6. Calculate  $\Pr\{E\}$  for  $n = 40$  (rather than 400) without the continuity correction.

**5.4.8** Refer to Exercise 5.4.6. Calculate  $\Pr\{E\}$  for  $n = 40$  (rather than 400) with the continuity correction.

**5.4.9** A certain cross between sweet-pea plants will produce progeny that are either purple flowered or white flowered;<sup>11</sup> the probability of a purple-flowered plant is  $p = \frac{9}{16}$ . Suppose  $n$  progeny are to be examined, and let  $\hat{P}$  be the sample proportion of purple-flowered plants. It might happen, by chance, that  $\hat{P}$  would be closer to  $\frac{1}{2}$  than to  $\frac{9}{16}$ . Find the probability that this misleading event would occur if

- (a)  $n = 1$ .
- (b)  $n = 64$ .
- (c)  $n = 320$ .

(Use the normal approximation without the continuity correction.)

**5.4.10** Cytomegalovirus (CMV) is a (generally benign) virus that infects one-half of young adults.<sup>12</sup> If a random sample of 10 young adults is taken, find the probability that between 30% and 40% (inclusive) of those sampled will have CMV,

- (a) using the binomial distribution formula.
- (b) using the normal approximation with the continuity correction.

**5.4.11** In a certain population of mussels (*Mytilus edulis*), 80% of the individuals are infected with an intestinal parasite.<sup>13</sup> A marine biologist plans to examine 100 randomly chosen mussels from the population. Find the probability that 85% or more of the sampled mussels will be infected, using the normal approximation without the continuity correction.

**5.4.12** Refer to Exercise 5.4.11. Find the probability that 85% or more of the sampled mussels will be infected, using the normal approximation with the continuity correction.

**5.4.13** Refer to Exercise 5.4.11. Suppose that the biologist takes a random sample of size 50. Find the probability that fewer than 35 of the sampled mussels will be infected, using the normal approximation

- (a) without the continuity correction.
- (b) with the continuity correction.

## 5.5 Perspective

In this chapter we have presented the concept of a sampling distribution and have focused on the sampling distribution of  $\bar{Y}$ . Of course, there are many other important sampling distributions, such as the sampling distribution of the sample standard deviation and the sampling distribution of the sample median.

Let us take another look at the random sampling model in the light of Chapter 5. As we have seen, a *random sample* is not necessarily a *representative sample*.<sup>\*</sup> But using sampling distributions, one can specify the degree of representativeness to be expected in a random sample. For instance, it is intuitively plausible that a larger sample is likely to be more representative than a smaller sample from the same population. In Sections 5.1 and 5.2 we saw how a sampling distribution can make this vague intuition precise by specifying the probability that a specified degree of representativeness will be achieved by a random sample. Thus, sampling distributions provide what has been called “certainty about uncertainty.”<sup>14</sup>

In Chapter 6 we will see for the first time how the theory of sampling distributions can be put to practical use in the analysis of data. We will find that, although the calculations of Chapter 5 seem to require the knowledge of unknowable quantities (such as  $\mu$  and  $\sigma$ ), when analyzing data one can nevertheless estimate the probable magnitude of sampling error using only information contained in the sample itself.

In addition to their application to data analysis, sampling distributions provide a basis for comparing the relative merits of different methods of analysis. For example, consider sampling from a normal population with mean  $\mu$ . Of course, the sample mean  $\bar{Y}$  is an estimator of  $\mu$ . But since a normal distribution is symmetric, it is also the population median, so the sample *median* is also an estimator of  $\mu$ . How, then, can we decide which estimator is better? This question can be answered in terms of sampling distributions, as follows: Statisticians have determined that, if the population is normal, the sample median is inferior to the sample mean in the sense that its sampling distribution, while centered at  $\mu$ , has a standard deviation larger than  $\frac{\sigma}{\sqrt{n}}$ .

Consequently, the sample median is less efficient (as an estimator of  $\mu$ ) than the sample mean; for a given sample size  $n$ , the sample median provides less information about  $\mu$  than does the sample mean. (If the population is not normal, however, the sample median can be much more efficient than the mean.)

---

<sup>\*</sup>It is true, however, that sometimes the investigator can force the sample to be representative with respect to some variable (not the one under study) whose population distribution is known; for example, a stratified random sample as discussed in Section 1.3. The methods of analysis given in this book, however, are only appropriate for *simple* random samples and cannot be applied without suitable modification.

## Supplementary Exercises 5.S.1–5.S.12

(Note: Exercises preceded by an asterisk refer to optional sections.)

**5.S.1** In an agricultural experiment, a large field of wheat was divided into many plots (each plot being  $7 \times 100$  ft) and the yield of grain was measured for each plot. These plot yields followed approximately a normal distribution with mean 88 lb and standard deviation 7 lb (as in Exercise 4.3.5). Let  $\bar{Y}$  represent the mean yield of five plots chosen at random from the field. Find  $\Pr\{\bar{Y} > 90\}$ .

**5.S.2** Consider taking a random sample of size 14 from the population of students at a certain college and measuring the diastolic blood pressure each of the 14 students. In the context of this setting, explain what is meant by the sampling distribution of the sample mean.

**5.S.3** Refer to the setting of Exercise 5.S.2. Suppose that the population mean is 70 mmHg and the population standard deviation is 10 mmHg. If the sample size is 14, what is the standard deviation of the sampling distribution of the sample mean?

**5.S.4** The heights of men in a certain population follow a normal distribution with mean 69.7 inches and standard deviation 2.8 inches.<sup>15</sup>

- (a) If a man is chosen at random from the population, find the probability that he will be more than 72 inches tall.
- (b) If two men are chosen at random from the population, find the probability that (i) both of them will be more than 72 inches tall; (ii) their mean height will be more than 72 inches.

**5.S.5** Suppose a botanist grows many individually potted eggplants, all treated identically and arranged in groups of four pots on the greenhouse bench. After 30 days of growth, she measures the total leaf area  $Y$  of each plant. Assume that the population distribution of  $Y$  is approximately normal with mean =  $800 \text{ cm}^2$  and SD =  $90 \text{ cm}^2$ .<sup>16</sup>

- (a) What percentage of the plants in the population will have leaf area between  $750 \text{ cm}^2$  and  $850 \text{ cm}^2$ ?
- (b) Suppose each group of four plants can be regarded as a random sample from the population. What percentage of the groups will have a group mean leaf area between  $750 \text{ cm}^2$  and  $850 \text{ cm}^2$ ?

**5.S.6** Refer to Exercise 5.S.5. In a real greenhouse, what factors might tend to invalidate the assumption that each group of plants can be regarded as a random sample from the same population?

**\*5.S.7** Consider taking a random sample of size 25 from a population in which 42% of the people have type A blood. What is the probability that the sample proportion with type A blood will be greater than 0.44? Use the normal approximation to the binomial with continuity correction.

**5.S.8** The activity of a certain enzyme is measured by counting emissions from a radioactively labeled molecule. For a given tissue specimen, the counts in consecutive 10-second time periods may be regarded (approximately) as repeated independent observations from a normal distribution (as in Exercise 4.S.1). Suppose the mean 10-second count for a certain tissue specimen is 1,200 and the standard deviation is 35. For that specimen, let  $Y$  represent a 10-second count and let  $\bar{Y}$  represent the mean of six 10-second counts. Both  $Y$  and  $\bar{Y}$  are unbiased—they each have an average of 1,200—but that doesn’t imply that they are equally good. Find  $\Pr\{1,175 \leq Y \leq 1,225\}$  and  $\Pr\{1,175 \leq \bar{Y} \leq 1,225\}$ , and

compare the two. Does the comparison indicate that counting for one minute and dividing by 6 would tend to give a more precise result than merely counting for a single 10-second time period? How?

**5.S.9** In a certain lab population of mice, the weights at 20 days of age follow approximately a normal distribution with mean weight =  $8.3 \text{ gm}$  and standard deviation =  $1.7 \text{ gm}$ .<sup>17</sup> Suppose many litters of 10 mice each are to be weighed. If each litter can be regarded as a random sample from the population, what percentage of the litters will have a total weight of  $90 \text{ gm}$  or more? (*Hint:* How is the total weight of a litter related to the mean weight of its members?)

**5.S.10** Refer to Exercise 5.S.9. In reality, what factors would tend to invalidate the assumption that each litter can be regarded as a random sample from the same population?

**5.S.11** Consider taking a random sample of size 25 from a population of plants, measuring the weight of each plant, and adding the weights to get a sample total. In the context of this setting, explain what is meant by the sampling distribution of the sample total.

**5.S.12** The skull breadths of a certain population of rodents follow a normal distribution with a standard deviation of 10 mm. Let  $\bar{Y}$  be the mean skull breadth of a random sample of 64 individuals from this population, and let  $\mu$  be the population mean skull breadth.

- (a) Suppose  $\mu = 50 \text{ mm}$ . Find  $\Pr\{\bar{Y} \text{ is within } \pm 2 \text{ mm of } \mu\}$ .
- (b) Suppose  $\mu = 100 \text{ mm}$ . Find  $\Pr\{\bar{Y} \text{ is within } \pm 2 \text{ mm of } \mu\}$ .
- (c) Suppose  $\mu$  is unknown. Can you find  $\Pr\{\bar{Y} \text{ is within } \pm 2 \text{ mm of } \mu\}$ ? If so, do it. If not, explain why not.

# CONFIDENCE INTERVALS

## Objectives

In this chapter we will begin a formal study of statistical inference. We will

- introduce the concept of the standard error to quantify the degree of uncertainty in an estimated quantity and compare it with the standard deviation.
- demonstrate the construction and interpretation of confidence intervals for means.
- provide a method to determine the sample size that is needed to achieve a desired level of accuracy.
- consider the conditions under which the use of a confidence interval is valid.
- introduce the standard error of a difference in sample means.
- demonstrate the construction and interpretation of confidence intervals for differences between means.

## 6.1 Statistical Estimation

In this chapter we undertake our first substantial adventure into statistical inference. Recall that statistical inference is based on the random sampling model: We view our data as a random sample from some population, and we use the information in the sample to infer facts about the population. Statistical estimation is a form of statistical inference in which we use the data to (1) determine an estimate of some feature of the population and (2) assess the precision of the estimate. Let us consider an example.

**Example  
6.1.1**

**Butterfly Wings** As part of a larger study of body composition, researchers captured 14 male Monarch butterflies at Oceano Dunes State Park in California and measured wing area (in  $\text{cm}^2$ ). The data are given in Table 6.1.1.<sup>1</sup>

**Table 6.1.1** Wing areas of male Monarch butterflies

Wing area ( $\text{cm}^2$ )				
33.9	33.0	30.6	36.6	36.5
34.0	36.1	32.0	28.0	32.0
32.2	32.2	32.3	30.0	

For these data, the mean and standard deviation are

$$\bar{y} = 32.8143 \approx 32.81 \text{ cm}^2 \quad \text{and} \quad s = 2.4757 \approx 2.48 \text{ cm}^2$$

Suppose we regard the 14 observations as a random sample from a population; the population could be described by (among other things) its mean,  $\mu$ , and its standard deviation,  $\sigma$ . We might define  $\mu$  and  $\sigma$  verbally as follows:

$\mu$  = the (population) mean wing area of male Monarch butterflies in the Oceano Dunes region

$\sigma$  = the (population) SD of wing area of male Monarch butterflies in the Oceano Dunes region

It is natural to estimate  $\mu$  by the sample mean and  $\sigma$  by the sample standard deviation. Thus, from the data on the 14 butterflies,

32.81 is an estimate of  $\mu$ .

2.48 is an estimate of  $\sigma$ .

We know that these estimates are subject to sampling error. Note that we are not speaking merely of measurement error; no matter how accurately each individual butterfly was measured, the sample information is imperfect due to the fact that only 14 butterflies were measured, rather than the entire population of butterflies. ■

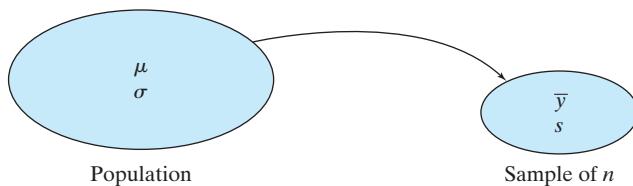
In general, for a sample of observations on a quantitative variable  $Y$ , the sample mean and SD are estimates of the population mean and SD:

$\bar{y}$  is an estimate of  $\mu$ .

$s$  is an estimate of  $\sigma$ .

The notation for these means and SDs is summarized schematically in Figure 6.1.1.

**Figure 6.1.1** Notation for means and SDs of sample and population



Our goal is to estimate  $\mu$ . We will see how to assess the reliability or precision of this estimate, and how to plan a study large enough to attain a desired precision.

## 6.2 Standard Error of the Mean

It is intuitively reasonable that the sample mean  $\bar{y}$  should be an estimate of  $\mu$ . It is not so obvious how to determine the reliability of the estimate. As an estimate of  $\mu$ , the sample mean  $\bar{y}$  is imprecise to the extent that it is affected by sampling error. In Section 5.3 we saw that the magnitude of the sampling error—that is, the amount of discrepancy between  $\bar{y}$  and  $\mu$ —is described (in a probability sense) by the sampling distribution of  $\bar{Y}$ . The standard deviation of the sampling distribution of  $\bar{Y}$  is

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

Since  $s$  is an estimate of  $\sigma$ , a natural estimate of  $\frac{\sigma}{\sqrt{n}}$  would be  $\frac{s}{\sqrt{n}}$ ; this quantity is called the **standard error of the mean**. We will denote it as  $\text{SE}_{\bar{Y}}$  or sometimes simply SE.\*

**Definition** The **standard error of the mean** is defined as

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

The following example illustrates the definition.

**Example  
6.2.1**

**Butterfly Wings** For the Monarch butterfly data of Example 6.1.1, we have  $n = 14$ ,  $\bar{y} = 32.8143 \approx 32.81 \text{ cm}^2$  and  $s = 2.4757 \approx 2.48 \text{ cm}^2$ . The standard error of the mean is

$$\begin{aligned}\text{SE}_{\bar{Y}} &= \frac{s}{\sqrt{n}} \\ &= \frac{2.4757}{\sqrt{14}} = 0.6617 \text{ cm}^2, \text{ which we will round to } 0.66 \text{ cm}^2\end{aligned}$$
■

As we have seen, the SE is an estimate of  $\sigma_{\bar{Y}}$ . On a more practical level, the SE can be interpreted in terms of the expected sampling error: Roughly speaking, the difference between  $\bar{y}$  and  $\mu$  is rarely more than a few standard errors. Indeed, we expect  $\bar{y}$  to be within about one standard error of  $\mu$  quite often. Thus, the standard error is a measure of the reliability or precision of  $\bar{y}$  as an estimate of  $\mu$ ; the smaller the SE, the more precise the estimate. Notice how the SE incorporates the two factors that affect reliability: (1) the inherent variability of the observations (expressed through  $s$ ), and (2) the sample size ( $n$ ).

## Standard Error versus Standard Deviation

The terms “standard error” and “standard deviation” are sometimes confused. It is extremely important to distinguish between standard error (SE) and standard deviation ( $s$ , or SD). These two quantities describe entirely different aspects of the data. The SD describes the dispersion of the data, while the SE describes the unreliability (due to sampling error) in the *mean* of the sample as an estimate of the *mean* of the population. Let us consider a concrete example.

**Example  
6.2.2**

**Lamb Birthweights** A geneticist weighed 28 female lambs at birth. The lambs were all born in April, were all the same breed (Rambouillet), and were all single births (no

\*Some statisticians prefer to reserve the term “standard error” for  $\sigma/\sqrt{n}$  and to call  $s/\sqrt{n}$  the “estimated standard error.”

†Rounding Summary Statistics

For reporting the mean, standard deviation, and standard error of the mean, the following procedure is recommended:

1. Round the SE to two significant digits.
2. Round  $\bar{y}$  and  $s$  to match the SE with respect to the decimal position of the last significant digit. (The concept of significant digits is reviewed in Appendix 6.1.) For example, if the SE is rounded to the nearest hundredth, then  $\bar{y}$  and  $s$  are also rounded to the nearest hundredth.

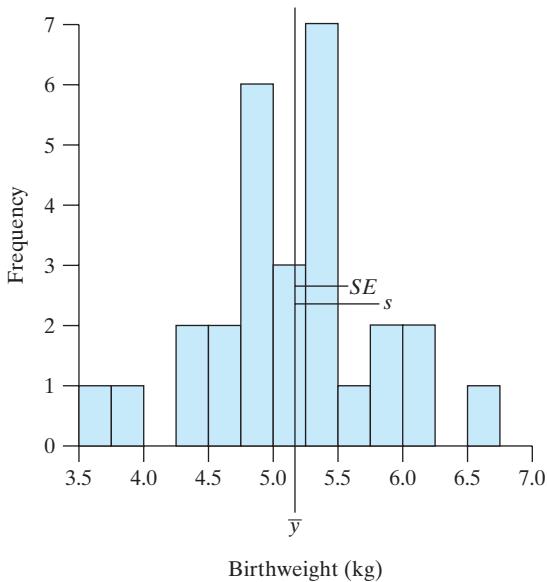
twins). The diet and other environmental conditions were the same for all the parents. The birthweights are shown in Table 6.2.1.<sup>2</sup>

**Table 6.2.1** Birthweights of twenty-eight Rambouillet lambs

Birthweight (kg)							
4.3	5.2	6.2	6.7	5.3	4.9	4.7	
5.5	5.3	4.0	4.9	5.2	4.9	5.3	
5.4	5.5	3.6	5.8	5.6	5.0	5.2	
5.8	6.1	4.9	4.5	4.8	5.4	4.7	

For these data, the mean is  $\bar{y} = 5.17\text{ kg}$ , the standard deviation is  $s = 0.65\text{ kg}$ , and the standard error is  $SE = 0.12\text{ kg}$ . The SD,  $s$ , describes the variability of birthweights among the lambs in the sample, while the SE indicates the variability associated with the sample mean ( $5.17\text{ kg}$ ), viewed as an estimate of the population mean birthweight. This distinction is emphasized in Figure 6.2.1, which shows a histogram of the lamb birthweight data; the SD is indicated as a deviation from  $\bar{y}$ , while the SE is indicated as variability associated with  $\bar{y}$  itself.

**Figure 6.2.1** Birthweights of twenty-eight lambs

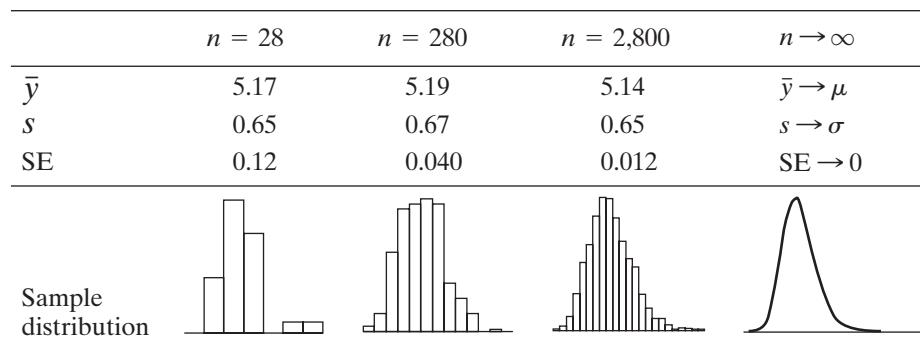


Another way to highlight the contrast between the SE and the SD is to consider samples of various sizes. As the sample size increases, the sample mean and SD tend to approach more closely the population mean and SD; indeed, the distribution of the data tends to approach the population distribution. The standard error, by contrast, tends to decrease as  $n$  increases; when  $n$  is very large, the SE is very small and so the sample mean is a very precise estimate of the population mean. The following example illustrates this effect.

**Example  
6.2.3**

**Lamb Birthweights** Suppose we regard the birthweight data of Example 6.2.2 as a sample of size  $n = 28$  from a population, and consider what would happen if we were to choose larger samples from the same population—that is, if we were to

**Figure 6.2.2** Samples of various sizes from the lamb birthweight population



measure the birthweights of additional female Rambouillet lambs born under the specified conditions. Figure 6.2.2 shows the kind of results we might expect; the values given are fictitious but realistic. For very large  $n$ ,  $\bar{y}$  and  $s$  would be very close to  $\mu$  and  $\sigma$ , where

$\mu$  = Mean birthweight of female Rambouillet lambs born under the conditions described

and

$\sigma$  = Standard deviation of birthweights of female Rambouillet lambs born under the conditions described.



## Graphical Presentation of the SE and the SD

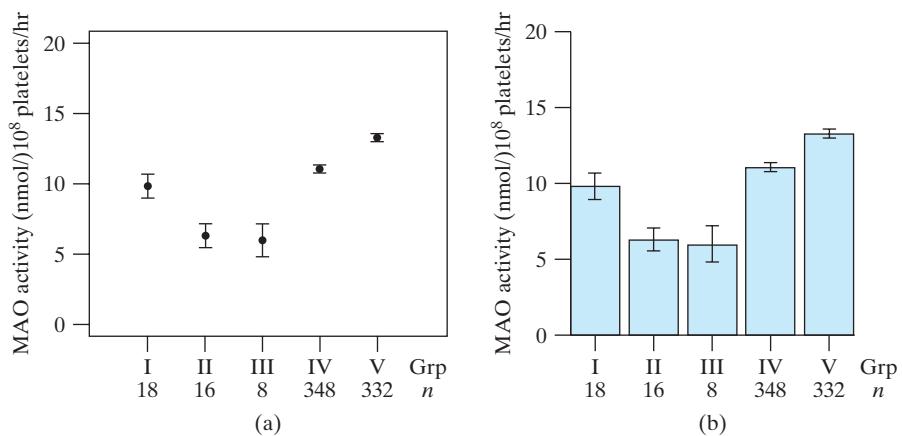
The clarity and impact of a scientific report can be greatly enhanced by well-designed displays of the data. Data can be displayed graphically or in a table. We briefly discuss some of the options.

Let us first consider graphical presentation of data. Here is an example.

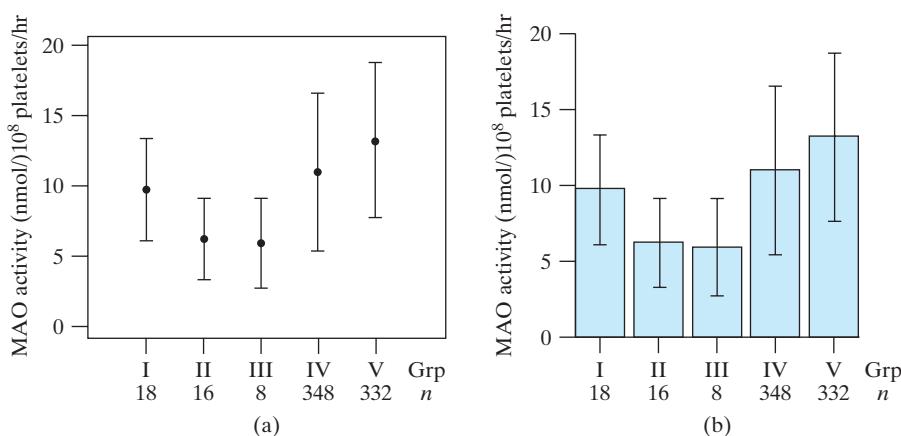
### Example 6.2.4

**MAO and Schizophrenia** The enzyme monoamine oxidase (MAO) is of interest in the study of human behavior. Figures 6.2.3 and 6.2.4 display measurements of MAO activity in the blood platelets in five groups of people: Groups I, II, and III are three

**Figure 6.2.3** MAO data displayed as  $\bar{y} \pm SE$  using (a) an interval plot and (b) a bargraph with standard error bars



**Figure 6.2.4** MAO data displayed as  $\bar{y} \pm SD$  using (a) an interval plot and (b) a bargraph with standard deviation bars



diagnostic categories of schizophrenic patients (see Example 1.1.4), and groups IV and V are healthy male and female controls.<sup>3</sup> The MAO activity values are expressed as nmol benzylaldehyde product per  $10^8$  platelets per hour. In both Figures 6.2.3 and 6.2.4, the dots (a) or bars (b) represent the group means; the vertical lines represent  $\pm SE$  in Figure 6.2.3 and  $\pm SD$  in Figure 6.2.4.

Figures 6.2.3 and 6.2.4 convey very different information. Figure 6.2.3 conveys (1) the mean MAO value in each group, and (2) the reliability of each group mean, viewed as an estimate of its respective population mean. Figure 6.2.4 conveys (1) the mean MAO value in each group, and (2) the variability of MAO within each group. For instance, group V shows greater variability of MAO than group I (Figure 6.2.4) but has a much smaller standard error (Figure 6.2.3) because it is a much larger group.

Figure 6.2.3 invites the viewer to compare the means and gives some indication of the reliability of the comparisons. (But a full discussion of comparison of two or more means must wait until Chapter 7 and later chapters.) Figure 6.2.4 invites the viewer to compare the means and also to compare the standard deviations. Furthermore, Figure 6.2.4 gives the viewer some information about the extent of overlap of the MAO values in the various groups. For instance, consider groups IV and V; whereas they appear quite “separate” in Figure 6.2.3, we can easily see from Figure 6.2.4 that there is considerable overlap of individual MAO values in the two groups. ■

While we have displayed the MAO data using four individual plots in Figures 6.2.3 and 6.2.4, we typically would choose only one of these to publish in a report. Choosing between the interval plots and bargraphs is a matter of personal preference and style. And, as previously mentioned, choosing whether the interval bars represent the SD or SE will depend on whether we wish to emphasize a comparison of the means (SE), or more simply a summary of the variability in our observed data (SD).\*

In some scientific reports, data are summarized in tables rather than graphically. Table 6.2.2 shows a tabular summary for the MAO data of Example 6.2.4. As with the preceding graphs, when formally presenting results, one typically displays either the SD or SE, but not both.

\*To present a slightly simpler graphic, often only the “upper” error bars (SE or SD) on bargraphs are displayed.

**Table 6.2.2** MAO activity in five groups of people

MAO activity ( $\text{nmol}/10^8 \text{ platelets}/\text{hr}$ )				
Group	n	Mean	SE	SD
I	18	9.81	0.85	3.62
II	16	6.28	0.72	2.88
III	8	5.97	1.13	3.19
IV	348	11.04	0.30	5.59
V	332	13.29	0.30	5.50

## Exercises 6.2.1–6.2.7

**6.2.1** A pharmacologist measured the concentration of dopamine in the brains of several rats. The mean concentration was 1,269 ng/gm and the standard deviation was 145 ng/gm.<sup>4</sup> What was the standard error of the mean if

- (a) 8 rats were measured?
- (b) 30 rats were measured?

**6.2.2** An agronomist measured the heights of  $n$  corn plants.<sup>5</sup> The mean height was 220 cm and the standard deviation was 15 cm. Calculate the standard error of the mean if

- (a)  $n = 25$
- (b)  $n = 100$

**6.2.3** In evaluating a forage crop, it is important to measure the concentration of various constituents in the plant tissue. In a study of the reliability of such measurements, a batch of alfalfa was dried, ground, and passed through a fine screen. Five small (0.3 gm) aliquots of the alfalfa were then analyzed for their content of insoluble ash.<sup>6</sup> The results (gm/kg) were as follows:

10.0      8.9      9.1      11.7      7.9

For these data, calculate the mean, the standard deviation, and the standard error of the mean.

**6.2.4** A zoologist measured tail length in 86 individuals, all in the one-year age group, of the deer mouse *Peromyscus*. The mean length was 60.43 mm and the standard deviation was 3.06 mm. The table presents a frequency distribution of the data.<sup>7</sup>

TAIL LENGTH (mm)	NUMBER OF MICE
[52, 54)	1
[54, 56)	3
[56, 58)	11
[58, 60)	18
[60, 62)	21
[62, 64)	20
[64, 66)	9
[66, 68)	2
[68, 70)	1
Total	86

- (a) Calculate the standard error of the mean.

- (b) Construct a histogram of the data and indicate the intervals  $\bar{y} \pm SD$  and  $\bar{y} \pm SE$  on your histogram. (See Figure 6.2.1.)

**6.2.5** Refer to the mouse data of Exercise 6.2.4. Suppose the zoologist were to measure 500 additional animals from the same population. Based on the data in Exercise 6.2.4

- (a) What would you predict would be the standard deviation of the 500 new measurements?
- (b) What would you predict would be the standard error of the mean for the 500 new measurements?

**6.2.6** In a report of a pharmacological study, the experimental animals were described as follows.<sup>8</sup> “Rats weighing  $150 \pm 10 \text{ gm}$  were injected ...” with a certain chemical, and then certain measurements were made on the rats. If the author intends to convey the degree of homogeneity of the group of experimental animals, then should the 10 gm be the SD or the SE? Explain.

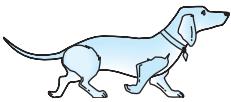
**6.2.7** For each of the following, decide whether the description fits the SD or the SE.

- (a) This quantity is a measure of the accuracy of the sample mean as an estimate of the population mean.
- (b) This quantity tends to stay the same as the sample size goes up.
- (c) This quantity tends to go down as the sample size goes up.

## 6.3 Confidence Interval for $\mu$

In Section 6.2 we said that the standard error of the mean (the SE) measures how far  $\bar{y}$  is likely to be from the population mean  $\mu$ . In this section we make that idea precise.

### Confidence Interval for $\mu$ : Basic Idea



**Figure 6.3.1** Invisible man walking his dog

Figure 6.3.1 is a drawing of an invisible man walking his dog. The dog, which is visible, is on an invisible spring-loaded leash. The tension on the spring is such that the dog is within 1 SE of the man about two-thirds of the time. The dog is within 2 standard errors of the man 95% of the time. Only 5% of the time is the dog more than 2 SEs from the man—unless the leash breaks, in which case the dog could be anywhere. We can see the dog, but we would like to know where the man is. Since the man and the dog are usually within 2 SEs of each other, we can take the interval “dog  $\pm 2 \times \text{SE}$ ” as an interval that typically would include the man. Indeed, we could say that we are 95% confident that the man is in this interval.

This is the basic idea of a confidence interval. We would like to know the value of the population mean  $\mu$ —which corresponds to the man—but we cannot see it directly. What we *can* see is the sample mean  $\bar{y}$ —which corresponds to the dog. We use what we can see,  $\bar{y}$ , together with the standard error, which we can calculate from the data, as a way of constructing an interval that we hope will include what we cannot see, the population mean  $\mu$ . We call the interval “position of the dog  $\pm 2 \times \text{SE}$ ” a 95% confidence interval for the position of the man. [This all depends on having a model that is correct: We said that if the leash breaks, then knowing where the dog is doesn’t tell us much about where the man is. Likewise, if our statistical model is wrong (for example, if we have a biased sample), then knowing  $\bar{y}$  doesn’t tell us much about  $\mu$ !]]

### Confidence Interval for $\mu$ : Mathematics

In the invisible man analogy,\* we said that the dog is within 1 SE of the man about two-thirds of the time and within 2 SEs of the man 95% of the time. This is based on the idea of the sampling distribution of  $\bar{Y}$  when we have a random sample from a normal distribution. If  $Z$  is a standard normal random variable, then the probability that  $Z$  is between  $\pm 2$  is about 95%. More precisely,  $\Pr\{-1.96 < Z < 1.96\} = 0.95$ .

From Chapter 5 we know that if  $Y$  has a normal distribution, then  $\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$  has a standard normal ( $Z$ ) distribution, so

$$\Pr\left\{-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right\} = 0.95 \quad (6.3.1)$$

Thus,

$$\Pr\{-1.96 \times \sigma/\sqrt{n} < \bar{Y} - \mu < 1.96 \times \sigma/\sqrt{n}\} = 0.95$$

and

$$\Pr\{-\bar{Y} - 1.96 \times \sigma/\sqrt{n} < -\mu < -\bar{Y} + 1.96 \times \sigma/\sqrt{n}\} = 0.95$$

so

$$\Pr\{\bar{Y} - 1.96 \times \sigma/\sqrt{n} < \mu < \bar{Y} + 1.96 \times \sigma/\sqrt{n}\} = 0.95$$

---

\*Credit for this analogy is due to Geoff Jowett.

That is, the interval

$$\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}} \quad (6.3.2)$$

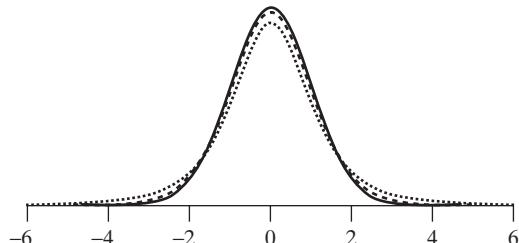
will contain  $\mu$  for 95% of all samples.

The interval (6.3.2) cannot be used for data analysis because it contains a quantity—namely,  $\sigma$ —that cannot be determined from the data. If we replace  $\sigma$  by its estimate—namely,  $s$ —then we can calculate an interval from the data, but what happens to the 95% interpretation? Fortunately, it turns out that there is an escape from this dilemma. The escape was discovered by a British scientist named W. S. Gosset, who was employed by the Guinness Brewery. He published his findings in 1908 under the pseudonym “Student,” and the method has borne his name ever since.<sup>9</sup> “Student” discovered that, if the data come from a normal population and if we replace  $\sigma$  in the interval (6.3.2) by the sample SD,  $s$ , then the 95% interpretation can be preserved if the multiplier of  $\frac{\sigma}{\sqrt{n}}$  (that is, 1.96) is replaced by a suitable quantity; the new quantity is denoted  $t_{0.025}$  and is related to a distribution known as Student’s  $t$  distribution.

## Student’s $t$ Distribution

The **Student’s  $t$  distributions** are theoretical continuous distributions that are used for many purposes in statistics, including the construction of confidence intervals. The exact shape of a Student’s  $t$  distribution depends on a quantity called “degrees of freedom,” abbreviated “df.” Figure 6.3.2 shows the density curves of two Student’s  $t$  distributions with  $df = 3$  and  $df = 10$ , and also a normal curve. A  $t$  curve is symmetric and bell shaped like the normal curve but has a larger standard deviation. As the df increase, the  $t$  curves approach the normal curve; thus, the normal curve can be regarded as a  $t$  curve with infinite df ( $df = \infty$ ).

**Figure 6.3.2** Two Student’s  $t$  curves (dotted,  $df = 3$  and dashed,  $df = 10$ ) and a normal curve ( $df = \infty$ )

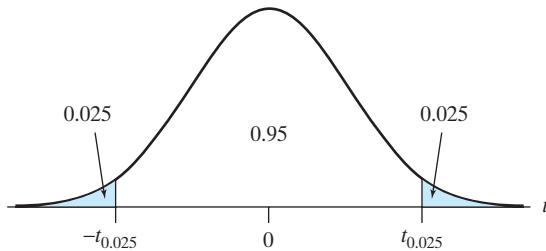


The quantity  $t_{0.025}$  is called the “two-tailed 5% critical value” of Student’s  $t$  distribution and is defined to be the value such that the interval between  $-t_{0.025}$  and  $+t_{0.025}$  contains 95% of the area under the curve, as shown in Figure 6.3.3.\* That is, the combined area in the two tails—below  $-t_{0.025}$  and above  $+t_{0.025}$ —is 5%. The total shaded area in Figure 6.3.3 is equal to 0.05; note that the shaded area consists of two “pieces” of area 0.025 each.

Critical values of Student’s  $t$  distribution are tabulated in Table 4. The values of  $t_{0.025}$  are shown in the column headed “Upper Tail Probability 0.025.” If you glance down this column, you will see that the values of  $t_{0.025}$  decrease as the df increase; for  $df = \infty$  (that is, for the normal distribution) the value is  $t_{0.025} = 1.960$ . You can confirm from Table 3 that the interval  $\pm 1.96$  (on the  $Z$  scale) contains 95% of the area under a normal curve.

\*In some statistics textbooks, you may find other notations, such as  $t_{0.05}$  or  $t_{0.975}$ , rather than  $t_{0.025}$ .

**Figure 6.3.3** Definition of the critical value  $t_{0.025}$



Other columns of Table 4 show other critical values, which are defined analogously; for instance, the interval  $\pm t_{0.05}$  contains 90% of the area under a Student's  $t$  curve.

### Confidence Interval for $\mu$ : Method

We describe Student's method for constructing a confidence interval for  $\mu$ , based on a random sample from a normal population. First, suppose we have chosen a confidence level equal to 95% (i.e., we wish to be 95% confident). To construct a 95% confidence interval for  $\mu$ , we compute the lower and upper limits of the interval as

$$\bar{y} - t_{0.025} \text{SE}_{\bar{Y}} \quad \text{and} \quad \bar{y} + t_{0.025} \text{SE}_{\bar{Y}}$$

that is,

$$\bar{y} \pm t_{0.025} \frac{s}{\sqrt{n}}$$

where the critical value  $t_{0.025}$  is determined from Student's  $t$  distribution with

$$\text{df} = n - 1$$

The following example illustrates the construction of a confidence interval.

#### Example 6.3.1

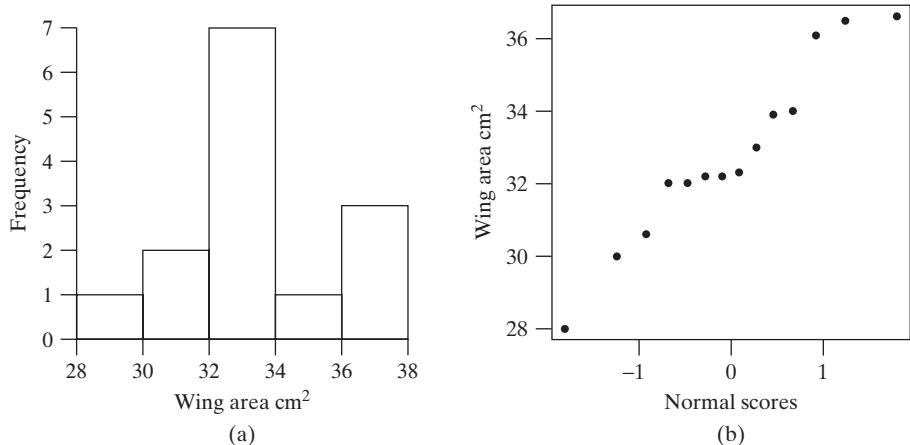
**Butterfly Wings** For the Monarch butterfly data of Example 6.1.1, we have  $n = 14$ ,  $\bar{y} = 32.8143 \text{ cm}^2$ , and  $s = 2.4757 \text{ cm}^2$ . Figure 6.3.4 shows a histogram and a normal probability plot of the data; these support the belief that the data came from a normal population. We have 14 observations, so the value of  $\text{df}$  is

$$\text{df} = n - 1 = 14 - 1 = 13$$

From Table 4 we find

$$t_{0.025} = 2.160$$

**Figure 6.3.4** Histogram (a) and normal probability plot (b) of butterfly wings data



The 95% confidence interval for  $\mu$  is

$$32.8143 \pm 2.160 \frac{2.4757}{\sqrt{14}}$$

$$32.8143 \pm 2.160(0.6617)$$

$$32.8143 \pm 1.4293$$

or, approximately,

$$32.81 \pm 1.43$$

The confidence interval may be left in this form. Alternatively, the endpoints of the interval may be explicitly calculated as

$$32.81 - 1.43 = 31.38 \quad \text{and} \quad 32.81 + 1.43 = 34.24$$

and the interval may be written compactly as

$$(31.4, 34.2)$$

or in a more complete form as the following “confidence statement”:

$$31.4 \text{ cm}^2 < \mu < 34.2 \text{ cm}^2$$

The confidence statement asserts that the population mean wing area of male Monarch butterflies in the Oceano Dunes region of California is between  $31.4 \text{ cm}^2$  and  $34.2 \text{ cm}^2$  with 95% confidence. ■

The interpretation of the “95% confidence” will be discussed after the next example.

Confidence coefficients other than 95% are used analogously. For instance, a 90% confidence interval for  $\mu$  is constructed using  $t_{0.05}$  instead of  $t_{0.025}$  as follows:

$$\bar{y} \pm t_{0.05} \frac{s}{\sqrt{n}}$$

The following is an example.

---

**Example  
6.3.2**

**Butterfly Wings** From Table 4, we find that  $t_{0.05} = 1.771$  with  $df = 13$ . Thus, the 90% confidence interval for  $\mu$  from the butterfly wings data is

$$32.8143 \pm 1.771 \frac{2.4757}{\sqrt{14}}$$

$$32.8143 \pm 1.1718$$

or

$$31.6 < \mu < 34.0$$

As you see, the choice of a confidence level is somewhat arbitrary. For the butterfly wings data, the 95% confidence interval is

$$32.81 \pm 1.43$$

and the 90% confidence interval is

$$32.81 \pm 1.17$$

Thus, the 90% confidence interval is narrower than the 95% confidence interval. If we want to be 95% confident that our interval contains  $\mu$ , then we need a wider interval than we would need if we wanted to be only 90% confident: The higher the confidence level, the wider the confidence interval (for a fixed sample size; but note that as  $n$  increases the intervals get smaller).

**Remark** The quantity  $(n - 1)$  is referred to as “degrees of freedom” because the deviations  $(y_i - \bar{y})$  must sum to zero, and so only  $(n - 1)$  of them are “free” to vary. A sample of size  $n$  provides only  $(n - 1)$  independent pieces of information about variability, that is, about  $\sigma$ . This is particularly clear if we consider the case  $n = 1$ ; a sample of size 1 provides some information about  $\mu$ , but no information about  $\sigma$ , and so no information about sampling error. It makes sense, then, that when  $n = 1$ , we cannot use Student’s  $t$  method to calculate a confidence interval: the sample standard deviation does not exist (see Example 2.6.5) and there is no critical value with  $df = 0$ . A sample of size 1 is sometimes called an “anecdote”; for instance, an individual medical case history is an anecdote. Of course, a case history can contribute greatly to medical knowledge, but it does not (in itself) provide a basis for judging how closely the individual case resembles the population at large.

## Confidence Intervals and Randomness

In what sense can we be “confident” in a confidence interval? To answer this question, let us assume that we are dealing with a random sample from a normal population. Consider, for instance, a 95% confidence interval. One way to interpret the confidence level (95%) is to refer to the meta-study of repeated samples from the same population. If a 95% confidence interval for  $\mu$  is constructed for each sample, then 95% of the confidence intervals will contain  $\mu$ . Of course, the observed data in an experiment comprise only *one* of the possible samples; we can hope “confidently” that this sample is one of the lucky 95%, but we will never know.

The following example provides a more concrete visualization of the meta-study interpretation of a confidence level.

### Example 6.3.3

**Eggshell Thickness** In a certain large population of chicken eggs (described in Example 4.1.3), the distribution of eggshell thickness is normal with mean  $\mu = 0.38\text{ mm}$  and standard deviation  $\sigma = 0.03\text{ mm}$ . Figure 6.3.5 shows some typical samples from this population; plotted on the right are the associated 95% confidence intervals. The sample sizes are  $n = 5$  and  $n = 20$ . Notice that the second confidence interval with  $n = 5$  does not contain  $\mu$ . In the totality of potential confidence intervals, the percentage that would contain  $\mu$  is 95% for either sample size; as Figure 6.3.5 shows, the larger samples tend to produce narrower confidence intervals. ■

A confidence level can be interpreted as a probability, but caution is required. If we consider 95% confidence intervals, for instance, then the following statement is correct:

$$\Pr\{\text{the next sample will give us a confidence interval that contains } \mu\} = 0.95$$

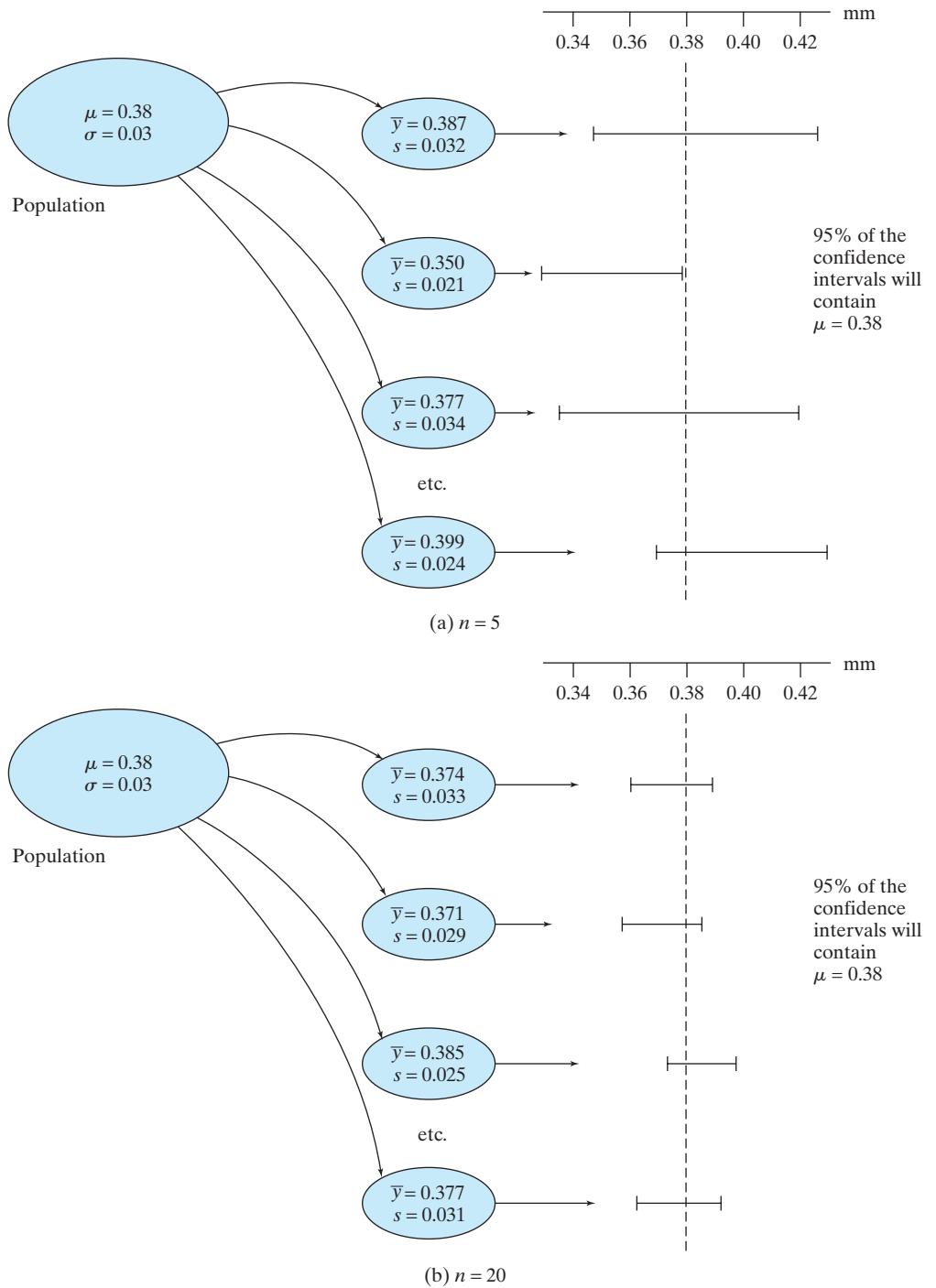
However, one should realize that it is *the confidence interval* that is the random item in this statement, and it is not correct to replace this item with its value from the data. Thus, for instance, we found in Example 6.3.1 that the 95% confidence interval for the mean butterfly wings is

$$31.4\text{ cm}^2 < \mu < 34.2\text{ cm}^2 \quad (6.3.3)$$

Nevertheless, it is *not* correct to say that

$$\Pr\{31.4\text{ cm}^2 < \mu < 34.2\text{ cm}^2\} = 0.95$$

because this statement has no chance element; either  $\mu$  is between 20.6 and 22.1 or it is not. If  $\mu = 32$ , then  $\Pr\{31.4\text{ cm}^2 < \mu < 34.2\text{ cm}^2\} = \Pr\{31.4\text{ cm}^2 < 32 < 34.2\text{ cm}^2\} = 1$  (not 0.95). The following analogy may help to clarify this point.

**Figure 6.3.5** Confidence intervals for mean eggshell thickness

Suppose we let  $Y$  represent the number of spots showing when a balanced die is tossed; then

$$\Pr\{Y = 2\} = \frac{1}{6}$$

On the other hand, if we now toss the die and observe 5 spots, it is obviously *not* correct to substitute this “datum” in the probability statement to conclude that

$$\Pr\{S = 2\} = \frac{1}{6}^*$$

As the preceding discussion indicates, the confidence level (for instance, 95%) is a property of the *method* rather than of a particular interval. An individual statement—such as (6.3.3)—is either true or false, but in the long run, if the researcher constructs 95% confidence intervals in various experiments, each time producing a statement such as (6.3.3), then 95% of the statements will be true.

## Interpretation of a Confidence Interval

### Example 6.3.4

**Bone Mineral Density** Low bone mineral density often leads to hip fractures in the elderly. In an experiment to assess the effectiveness of hormone replacement therapy, researchers gave conjugated equine estrogen (CEE) to a sample of 94 women between the ages of 45 and 64.<sup>10</sup> After taking the medication for 36 months, the bone mineral density was measured for each of the 94 women. The average density was  $0.878 \text{ g/cm}^2$ , with a standard deviation of  $0.126 \text{ g/cm}^2$ .

The standard error of the mean is thus  $\frac{0.126}{\sqrt{94}} = 0.013$ . It is not clear that the

distribution of bone mineral density is a normal distribution, but as we will see in Section 6.5, when the sample size is large, the condition of normality is not crucial. There were 94 observations, so there are 93 degrees of freedom. To find the *t* multiplier for a 95% confidence interval, we will use 100 degrees of freedom (since Table 4 doesn’t list 93 degrees of freedom); the *t* multiplier is  $t_{0.025} = 1.984$ . A 95% confidence interval for  $\mu$  is

$$0.878 \pm 1.984(0.013)$$

or, approximately,

$$0.878 \pm 0.026$$

or

$$(0.852, 0.904)^{\dagger}$$

Thus, we are 95% confident that the mean hip bone mineral density of all women age 45 to 64 who take CEE for 36 months is between  $0.852 \text{ g/cm}^2$  and  $0.904 \text{ g/cm}^2$ . ■

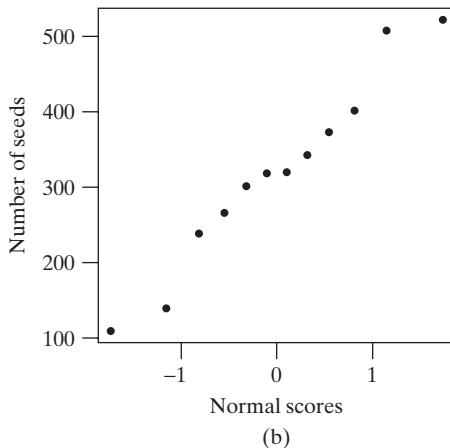
### Example 6.3.5

**Seeds per Fruit** The number of seeds per fruit for the freshwater plant *Vallisneria Americana* varies considerably from one fruit to another. A researcher took a random sample of 12 fruit and found that the average number of seeds was 320, with a standard deviation of 125.<sup>11</sup> The researcher expected the number of seeds to follow, at least approximately, a normal distribution. A normal probability plot of the data is shown in Figure 6.3.6. This supports the use of a normal distribution model for these data.

\*Even if the die rolls under a chair and we can’t immediately see that the top face of the die has 5 spots, it would be wrong (given our definition of probability) to say “The probability that the top of the die is showing 2 spots is 1/6.”

<sup>†</sup>If we use a computer to calculate the confidence interval, we get  $(0.8522, 0.9038)$ ; there is very little difference between the *t* multipliers for 100 versus 93 degrees of freedom.

**Figure 6.3.6** Normal probability plot of seeds per fruit for *Vallisneria Americana*



The standard error of the mean is  $\frac{125}{\sqrt{12}} = 36$ . There are 11 degrees of freedom.

The  $t$  multiplier for a 90% confidence interval is  $t_{0.05} = 1.796$ . A 90% confidence interval for  $\mu$  is

$$320 \pm 1.796(36)$$

or, approximately,

$$320 \pm 65$$

or

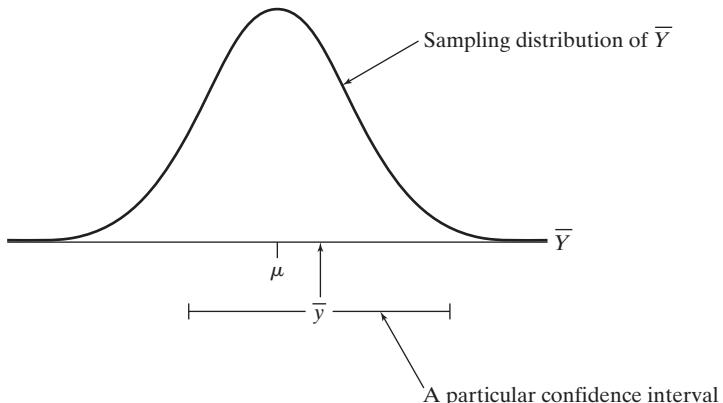
$$(255, 385)$$

Thus, we are 90% confident that the (population) mean number of seeds per fruit for *Vallisneria Americana* is between 255 and 385. ■

### Relationship to Sampling Distribution of $\bar{Y}$

At this point it may be helpful to look back and see how a confidence interval for  $\mu$  is related to the sampling distribution of  $\bar{Y}$ . Recall from Section 5.3 that the mean of the sampling distribution is  $\mu$  and its standard deviation is  $\frac{\sigma}{\sqrt{n}}$ . Figure 6.3.7 shows a particular sample mean ( $\bar{y}$ ) and its associated 95% confidence interval for  $\mu$ , superimposed on the sampling distribution of  $\bar{Y}$ . Notice that the particular confidence interval does contain  $\mu$ ; this will happen for 95% of samples.

**Figure 6.3.7** Relationship between a particular confidence interval for  $\mu$  and the sampling distribution of  $\bar{Y}$



## One-Sided Confidence Intervals

Most confidence intervals are of the form “estimate  $\pm$  margin of error”; these are known as two-sided intervals. However, it is possible to construct a one-sided confidence interval, which is appropriate when only a lower bound, or only an upper bound, is of interest. The following two examples illustrate 90% and 95% one-sided confidence intervals.

### Example 6.3.6

**Seeds per Fruit—One-Sided, 90%** Consider the seed data from Example 6.3.5, which are used to estimate the number of seeds per fruit for *Vallisneria Americana*. It might be that we want a lower bound on  $\mu$ , the population mean, but we are not concerned with how large  $\mu$  might be. Whereas a two-sided 90% confidence interval is based on capturing the middle 90% of a  $t$  distribution and thus uses the  $t$  multipliers of  $\pm t_{0.05}$ , a one-sided 90% (lower) confidence interval uses the fact that  $\Pr(-t_{0.10} < t < \infty) = 0.90$ . Thus, the lower limit of the confidence interval is  $\bar{y} - t_{0.10}SE_{\bar{Y}}$  and the upper limit of the interval is infinity. In this case, with 11 degrees of freedom the  $t$  multiplier is  $t_{11,0.10} = 1.363$  and we get

$$320 - 1.363(36) = 320 - 49 = 271$$

as the lower limit. The resulting interval is  $(271, \infty)$ . Thus, we are 90% confident that the (population) mean number of seeds per fruit for *Vallisneria Americana* is at least 271. ■

### Example 6.3.7

**Seeds per Fruit—One-Sided, 95%** A one-sided 95% confidence interval is constructed in the same manner as a one-sided 90% confidence interval, but with a different  $t$  multiplier. For the *Vallisneria Americana* seeds data we have  $t_{11,0.05} = 1.796$  and we get

$$320 - 1.796(36) = 320 - 65 = 255$$

as the lower limit. The resulting interval is  $(255, \infty)$ . Thus, we are 95% confident that the (population) mean number of seeds per fruit for *Vallisneria Americana* is at least 255. ■

## Exercises 6.3.1–6.3.20

**6.3.1 (Sampling exercise)** Refer to Exercise 5.3.1. Use your sample of five ellipse lengths to construct an 80% confidence interval for  $\mu$ , using the formula  $\bar{y} \pm (1.533)s/\sqrt{n}$ .

**6.3.2 (Sampling exercise)** Refer to Exercise 5.3.3. Use your sample of 20 ellipse lengths to construct an 80% confidence interval for  $\mu$  using the formula  $\bar{y} \pm (1.328)s/\sqrt{n}$ .

**6.3.3** As part of a study of the development of the thymus gland, researchers weighed the glands of five chick embryos after 14 days of incubation. The thymus weights (mg) were as follows:<sup>12</sup>

29.6    21.5    28.0    34.6    44.9

For these data, the mean is 31.7 and the standard deviation is 8.7.

- (a) Calculate the standard error of the mean.
- (b) Construct a 90% confidence interval for the population mean.

**6.3.4** Consider the data from Exercise 6.3.3.

- (a) Construct a 95% confidence interval for the population mean.
- (b) Interpret the confidence interval you found in part (a). That is, explain what the numbers in the interval mean. (See Examples 6.3.4 and 6.3.5.)

**6.3.5** Six healthy three-year-old female Suffolk sheep were injected with the antibiotic Gentamicin, at a dosage of 10 mg/kg body weight. Their blood serum concentrations ( $\mu\text{g}/\text{ml}$ ) of Gentamicin 1.5 hours after injection were as follows:<sup>13</sup>

33    26    34    31    23    25

For these data, the mean is 28.7 and the standard deviation is 4.6.

- (a) Construct a 95% confidence interval for the population mean.
- (b) Define in words the population mean that you estimated in part (a). (See Example 6.1.1.)

- (c) The interval constructed in part (a) nearly contains all of the observations; will this typically be true for a 95% confidence interval? Explain.

**6.3.6** A zoologist measured tail length in 86 individuals, all in the one-year age group, of the deer mouse *Peromyscus*. The mean length was 60.43 mm and the standard deviation was 3.06 mm. A 95% confidence interval for the mean is (59.77, 61.09).

- (a) True or false (and say why): We are 95% confident that the average tail length of the 86 individuals in the sample is between 59.77 mm and 61.09 mm.  
 (b) True or false (and say why): We are 95% confident that the average tail length of all the individuals in the population is between 59.77 mm and 61.09 mm.

**6.3.7** Refer to Exercise 6.3.6.

- (a) Without doing any computations, would an 80% confidence interval for the data in Exercise 6.3.6 be wider, narrower, or about the same? Explain.  
 (b) Without doing any computations, if 500 mice were sampled rather than 86, would the 95% confidence interval listed in Exercise 6.3.6 be wider, narrower, or about the same? Explain.

**6.3.8** Researchers measured the bone mineral density of the spines of 94 women who had taken the drug CEE. (See Example 6.3.4, which dealt with hip bone mineral density.) The mean was  $1.016 \text{ g/cm}^2$  and the standard deviation was  $0.155 \text{ g/cm}^2$ . A 95% confidence interval for the mean is (0.984, 1.048).

- (a) True or false (and say why): 95% of the sampled bone mineral density measurements are between 0.984 and 1.048.  
 (b) True or false (and say why): 95% of the population bone mineral density measurements are between 0.984 and 1.048.

**6.3.9** There was a control group in the study described in Example 6.3.4. The 124 women in the control group were given a placebo, rather than an active medication. At the end of the study they had an average bone mineral density of  $0.840 \text{ g/cm}^2$ . Shown are three confidence intervals: One is a 90% confidence interval, one is an 85% confidence interval, and the other is an 80% confidence interval. Without doing any calculations, match the intervals with the confidence levels and explain how you determined which interval goes with which level.

Confidence levels:

90%    85%    80%

Intervals (in scrambled order):

(0.826, 0.854)    (0.824, 0.856)    (0.822, 0.858)

**6.3.10** Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress. A researcher conducted a study to investigate whether a

program of regular exercise might affect the resting (unstressed) concentration of HBE in the blood. He measured blood HBE levels, in January and again in May, from 10 participants in a physical fitness program. The results were as shown in the table.<sup>14</sup>

- (a) Construct a 95% confidence interval for the population mean difference in HBE levels between January and May. (*Hint:* You need to use only the values in the right-hand column.)

PARTICIPANT	HBE LEVEL (pg/ml)		
	JANUARY	MAY	DIFFERENCE
1	42	22	20
2	47	29	18
3	37	9	28
4	9	9	0
5	33	26	7
6	70	36	34
7	54	38	16
8	27	32	-5
9	41	33	8
10	18	14	4
Mean	37.8	24.8	13.0
SD	17.6	10.9	12.4

- (b) Interpret the confidence interval from part (a). That is, explain what the interval tells you about HBE levels. (See Examples 6.3.4 and 6.3.5.)  
 (c) Using your interval to support your answer, is there evidence that HBE levels are lower in May than January? (*Hint:* Does your interval include the value zero?)

**6.3.11** Consider the data from Exercise 6.3.10. If the sample size is small, as it is in this case, then in order for a confidence interval based on Student's *t* distribution to be valid, the data must come from a normally distributed population. Is it reasonable to think that difference in HBE level is normally distributed? How do you know?

**6.3.12** Invertase is an enzyme that may aid in spore germination of the fungus *Colletotrichum graminicola*. A botanist incubated specimens of the fungal tissue in petri dishes and then assayed the tissue for invertase activity. The specific activity values for nine petri dishes incubated at 90% relative humidity for 24 hours are summarized as follows:<sup>15</sup>

Mean = 5,111 units    SD = 818 units

- (a) Assume that the data are a random sample from a normal population. Construct a 95% confidence interval for the mean invertase activity under these experimental conditions.

- (b) Interpret the confidence interval you found in part (a). That is, explain what the numbers in the interval mean. (See Examples 6.3.4 and 6.3.5.)
- (c) If you had the raw data, how could you check the condition that the data are from a normal population?

**6.3.13** As part of a study of the treatment of anemia in cattle, researchers measured the concentration of selenium in the blood of 36 cows who had been given a dietary supplement of selenium (2 mg/day) for one year. The cows were all the same breed (*Santa Gertrudis*) and had borne their first calf during the year. The mean selenium concentration was 6.21  $\mu\text{g}/\text{dl}$  and the standard deviation was 1.84  $\mu\text{g}/\text{dl}$ .<sup>16</sup> Construct a 95% confidence interval for the population mean.

**6.3.14** In a study of larval development in the tufted apple budmoth (*Platynota idaealis*), an entomologist measured the head widths of 50 larvae. All 50 larvae had been reared under identical conditions and had moulted six times. The mean head width was 1.20 mm and the standard deviation was 0.14 mm. Construct a 90% confidence interval for the population mean.<sup>17</sup>

**6.3.15** In a study of the effect of aluminum intake on the mental development of infants, a group of 92 infants who had been born prematurely were given a special aluminum-depleted intravenous-feeding solution.<sup>18</sup> At age 18 months the neurologic development of the infants was measured using the Bayley Mental Development Index. (The Bayley Mental Development Index is similar to an IQ score, with 100 being the average in the general population.) A 95% confidence interval for the mean is (93.8, 102.1).

- (a) Interpret this interval. That is, what does the interval tell us about neurologic development in the

population of prematurely born infants who receive intravenous-feeding solutions?

- (b) Does this interval indicate that the mean IQ of the sampled population is below the general population average of 100?

**6.3.16** A group of 101 patients with end-stage renal disease were given the drug epoetin.<sup>19</sup> The mean hemoglobin level of the patients was 10.3 (g/dl), with an SD of 0.9. Construct a 95% confidence interval for the population mean.

**6.3.17** In Table 4 we find that  $t_{0.025} = 1.960$  when  $\text{df} = \infty$ . Show how this value can be verified using Table 3.

**6.3.18** Use Table 3 to find the value of  $t_{0.0025}$  when  $\text{df} = \infty$ . (Do not attempt to interpolate in Table 4.)

**6.3.19** Data are often summarized in this format:  $\bar{y} \pm \text{SE}$ . Suppose this interval is interpreted as a confidence interval. If the sample size is large, what would be the confidence level of such an interval? That is, what is the chance that an interval computed as

$$\bar{y} \pm (1.00)\text{SE}$$

will actually contain the population mean? [Hint: Recall that the confidence level of the interval  $\bar{y} \pm (1.96)\text{SE}$  is 95%.]

#### 6.3.20 (Continuation of Exercise 6.3.19)

- (a) If the sample size is small but the population distribution is normal, is the confidence level of the interval  $\bar{y} \pm \text{SE}$  larger or smaller than the answer to Exercise 6.3.19? Explain.
- (b) How is the answer to Exercise 6.3.19 affected if the population distribution of  $Y$  is not approximately normal?

## 6.4 Planning a Study to Estimate $\mu$

Before collecting data for a research study, it is wise to consider in advance whether the estimates generated from the data will be sufficiently precise. It can be painful indeed to discover after a long and expensive study that the standard errors are so large that the primary questions addressed by the study cannot be answered.

The precision with which a population mean can be estimated is determined by two factors: (1) the population variability of the observed variable  $Y$ , and (2) the sample size.

In some situations the variability of  $Y$  cannot, and perhaps should not, be reduced. For example, a wildlife ecologist may wish to conduct a field study of a natural population of fish; the heterogeneity of the population is not controllable and in fact is a proper subject of investigation. As another example, in a medical investigation, in addition to knowing the average response to a treatment, it may also be important to know how much the response varies from one patient to another, and so it may not be appropriate to use an overly homogeneous group of patients.

On the other hand, it is often appropriate, especially in comparative studies, to reduce the variability of  $Y$  by holding *extraneous* conditions as constant as possible. For example, physiological measurements may be taken at a fixed time of day; tissue may be held at a controlled temperature; all animals used in an experiment may be the same age.

Suppose, then, that plans have been made to reduce the variability of  $Y$  as much as possible, or desirable. What sample size will be sufficient to achieve a desired degree of precision in estimation of the population mean? If we use the standard error as our measure of precision, then this question can be approached in a straightforward manner. Recall that the SE is defined as

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

In order to decide on a value of  $n$ , one must (1) specify what value of the SE is considered desirable to achieve and (2) have available a preliminary guess of the SD, either from a pilot study or other previous experience, or from the scientific literature. The required sample size is then determined from the following equation:

$$\text{Desired SE} = \frac{\text{Guessed SD}}{\sqrt{n}}$$

The following example illustrates the use of this equation.

---

**Example  
6.4.1**

**Butterfly Wings** The butterfly wing data of Example 6.1.1 yielded the following summary statistics:

$$\bar{y} = 32.81 \text{ cm}^2$$

$$s = 2.48 \text{ cm}^2$$

$$\text{SE} = 0.66 \text{ cm}^2$$

Suppose the researcher is now planning a new study of butterflies and has decided that it would be desirable that the SE be no more than  $0.4 \text{ cm}^2$ . As a preliminary guess of the SD, she will use the value from the old study, namely  $2.48 \text{ cm}^2$ . Thus, the desired  $n$  must satisfy the following relation:

$$\text{SE} = \frac{2.48}{\sqrt{n}} \leq 0.4$$

This equation is easily solved to give  $n \geq 38.4$ . Since one cannot have 38.4 butterflies, the new study should include at least 39 butterflies. ■

You may wonder how a researcher would arrive at a value such as  $0.4 \text{ cm}^2$  for the desired SE. Such a value is determined by considering how much error one is willing to tolerate in the estimate of  $\mu$ . For example, suppose the researcher in Example 6.4.1 has decided that she would like to be able to estimate the population mean,  $\mu$ , to within  $\pm 0.8$  with 95% confidence. That is, she would like her 95% confidence interval for  $\mu$  to be  $\bar{y} \pm 0.8$ . The “ $\pm$  part” of the confidence interval, which is sometimes called the **margin of error for 95% confidence**, is  $t_{0.025} \times \text{SE}$ . The precise value of  $t_{0.025}$  depends on the degrees of freedom, but typically  $t_{0.025}$  is approximately 2. Thus, the researcher wants  $2 \times \text{SE}$  to be no more than 0.8. This means that the SE should be no more than  $0.4 \text{ cm}^2$ .

In comparative studies, the primary consideration is usually the size of anticipated treatment effects. For instance, if one is planning to compare two experimental

groups or distinct populations, the anticipated SE for each population or experimental group should be substantially smaller than (preferably less than one-fourth of) the anticipated difference between the two group means.\* Thus, the butterfly researcher of Example 6.4.1 might arrive at the value  $0.4 \text{ cm}^2$  if she were planning to compare male and female Monarch butterflies and she expected the wing areas for the sexes to differ (on the average) by about  $1.6 \text{ cm}^2$ . She would then plan to capture 39 male and 39 female butterflies.

To see how the required  $n$  depends on the specified precision, suppose the butterfly researcher specified the desired SE to be  $0.2 \text{ cm}^2$  rather than  $0.4 \text{ cm}^2$ . Then the relation would be

$$\text{SE} = \frac{2.48}{\sqrt{n}} \leq 0.2$$

which yields  $n = 153.76$ , so that she would plan to capture 154 butterflies of each sex. Thus, to double the precision (by cutting the SE in half) requires not twice as many but four times as many observations. This phenomenon of “diminishing returns” is due to the square root in the SE formula.

## Exercises 6.4.1–6.4.5

**6.4.1** An experiment is being planned to compare the effects of several diets on the weight gain of beef cattle, measured over a 140-day test period.<sup>20</sup> In order to have enough precision to compare the diets, it is desired that the standard error of the mean for each diet should not exceed 5 kg.

- (a) If the population standard deviation of weight gain is guessed to be about 20 kg on any of the diets, how many cattle should be put on each diet in order to achieve a sufficiently small standard error?
- (b) If the guess of the standard deviation is doubled, to 40 kg, does the required number of cattle double? Explain.

**6.4.2** A medical researcher proposes to estimate the mean serum cholesterol level of a certain population of middle-aged men, based on a random sample of the population. He asks a statistician for advice. The ensuing discussion reveals that the researcher wants to estimate the population mean to within  $\pm 6 \text{ mg/dl}$  or less, with 95% confidence. Thus, the standard error of the mean should be 3 mg/dl or less. Also, the researcher believes that the standard deviation of serum cholesterol in the population is probably about 40 mg/dl.<sup>21</sup> How large a sample does the researcher need to take?

**6.4.3** A plant physiologist is planning to measure the stem lengths of soybean plants after two weeks of growth when using a new fertilizer. Previous experiments suggest

that the standard deviation of stem length is around 1.2 cm.<sup>22</sup> Using this as a guess of  $\sigma$ , determine how many soybean plants the researcher should have if she wants the standard error of the group mean to be no more than 0.2 cm.

**6.4.4** Suppose you are planning an experiment to test the effects of various diets on the weight gain of young turkeys. The observed variable will be  $Y$  = weight gain in three weeks (measured over a period starting one week after hatching and ending three weeks later). Previous experiments suggest that the standard deviation of  $Y$  under a standard diet is approximately 80 g.<sup>23</sup> Using this as a guess of  $\sigma$ , determine how many turkeys you should have in a treatment group, if you want the standard error of the group mean to be no more than

- (a) 20 g
- (b) 15 g

**6.4.5** A researcher is planning to compare the effects of two different types of lights on the growth of bean plants. She expects that the means of the two groups will differ by about 1 inch and that in each group the standard deviation of plant growth will be around 1.5 inches. Consider the guideline that the anticipated SE for each experimental group should be no more than one-fourth of the anticipated difference between the two group means. How large should the sample be (for each group) in order to meet this guideline?

---

\*This is a rough guideline for obtaining adequate sensitivity to discriminate between treatments. Such sensitivity, technically called *power*, is discussed in Chapter 7.

## 6.5 Conditions for Validity of Estimation Methods

For any sample of quantitative data, one can use the methods of this chapter to compute the mean, its standard error, and various confidence intervals; indeed, computers can make this rather easy to carry out. However, the *interpretations* that we have given for these descriptions of the data are valid only under certain conditions.

### Conditions for Validity of the SE Formula

First, the very notion of regarding the sample mean as an estimate of a population mean requires that the data be viewed “as if” they had been generated by random sampling from some population. To the extent that this is not possible, any inference beyond the actual data is questionable. The following example illustrates the difficulty.

#### Example 6.5.1

**Marijuana and Intelligence** Ten people who used marijuana heavily were found to be quite intelligent; their mean IQ was 128.4, whereas the mean IQ for the general population is known to be 100. The 10 people belonged to a religious group that uses marijuana for ritual purposes. Since their decision to join the group might very well be related to their intelligence, it is not clear that the 10 can be regarded (with respect to IQ) as a random sample from any particular population, and therefore there is no apparent basis for thinking of the sample mean (128.4) as an estimate of the mean IQ of a particular population (such as, for instance, all heavy marijuana users). An inference about the *effect* of marijuana on IQ would be even more implausible, especially because data were not available on the IQs of the 10 people *before* they began marijuana use.<sup>24</sup> ■

Second, the use of the standard error formula  $SE = s/\sqrt{n}$  requires two further conditions:

1. The population size must be large compared to the sample size. This requirement is rarely a problem in the life sciences; the sample can be as much as 5% of the population without seriously invalidating the SE formula.\*
2. The observations must be independent of each other. This requirement means that the  $n$  observations actually give  $n$  independent pieces of information about the population.

Data often fail to meet the independence requirement if the experiment or sampling regime has a **hierarchical structure**, in which observational units are “nested” within sampling units, as illustrated by the following example.

#### Example 6.5.2

**Canine Anatomy** The coccygeus muscle is a bilateral muscle in the pelvic region of the dog. As part of an anatomical study, the left side and the right side of the coccygeus muscle were weighed for each of 21 female dogs. There were thus

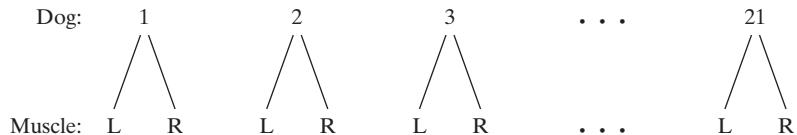
---

\*If the sample size,  $n$ , is a substantial fraction of the population size,  $N$ , then the “finite population correction factor” should be applied. This factor is  $\sqrt{\frac{N-n}{N-1}}$ . The standard error of the mean then becomes

$$\frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}.$$

$2 \times 21 = 42$  observations, but only 21 units chosen from the population of interest (female dogs). Because of the symmetry of the coccygeus, the information contained in the right and left sides is largely redundant, so that the data contain not 42, but only 21, independent pieces of information about the coccygeus muscle of female dogs. It would therefore be incorrect to apply the SE formula as if the data comprised a sample of size  $n = 42$ . The hierarchical nature of the data set is indicated in Figure 6.5.1.<sup>25</sup>

**Figure 6.5.1** Hierarchical data structure of Example 6.5.2



Hierarchical data structures are rather common in the life sciences. For instance, observations may be made on 90 nerve cells that come from only three different cats; on 80 kernels of corn that come from only four ears; on 60 young mice who come from only 10 litters. A particularly clear example of nonindependent observations is replicated measurements on the same individual; for instance, if a physician makes triplicate blood pressure measurements on each of 10 patients, she clearly does not have 30 independent observations. In some situations a correct treatment of hierarchical data is obvious; for instance, the triplicate blood pressure measurements could be averaged to give a single value for each patient. In other situations, however, lack of independence can be more subtle. For instance, suppose 60 young mice from 10 litters are included in an experiment to compare two diets. Then the choice of a correct analysis depends on the *design* of the experiment—on such aspects as whether the diets are fed to the young mice themselves or to the mothers, and how the animals are allocated to the two diets.

Sometimes variation arises at several different hierarchical levels in an experiment, and it can be a challenge to sort these out, and particularly, to correctly identify the quantity  $n$ . Example 6.5.3 illustrates this issue.

---

**Example  
6.5.3**

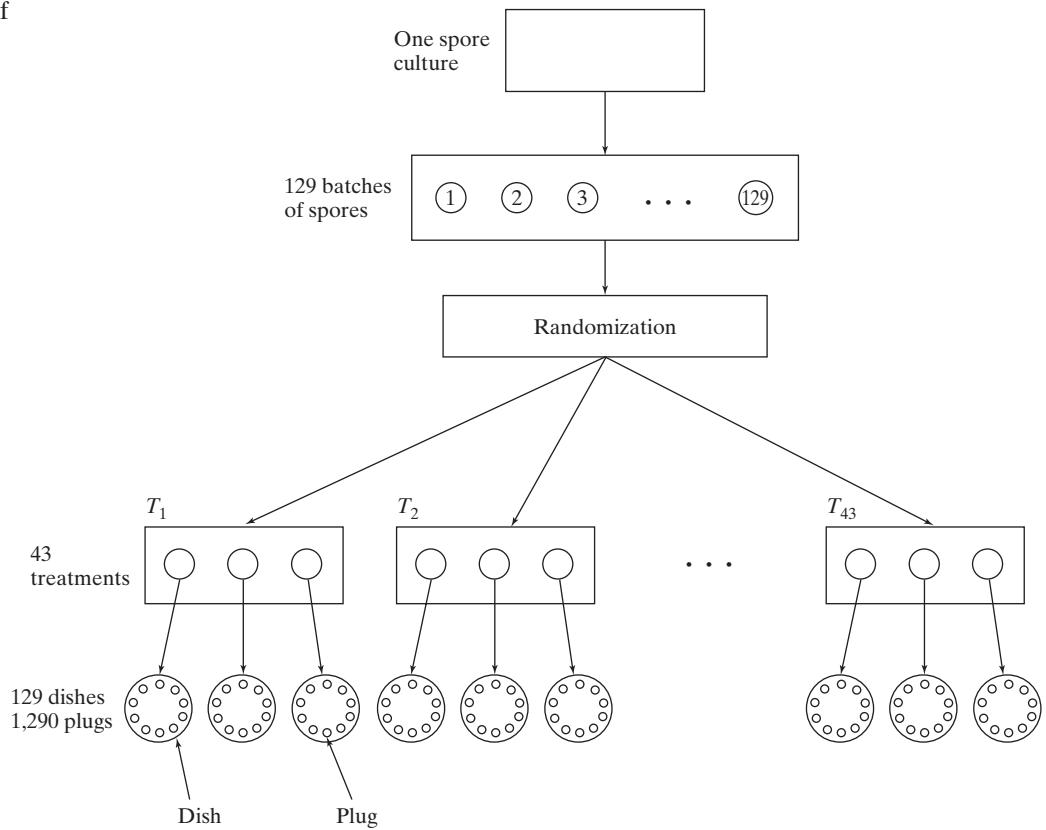
**Germination of Spores** In a study of the fungus that causes the anthracnose disease of corn, interest focused on the survival of the fungal spores.<sup>26</sup> Batches of spores, all prepared from a single culture of the fungus, were stored in chambers under various environmental conditions and then assayed for their ability to germinate, as follows. Each batch of spores was suspended in water and then plated on agar in a petri dish. Ten “plugs” of 3-mm diameter were cut from each petri dish and were incubated at 25 °C for 12 hours. Each plug was then examined with a microscope for germinated and ungerminated spores. The environmental conditions of storage (the “treatments”) included the following:

- $T_1$ : Storage at 70% relative humidity for one week
- $T_2$ : Storage at 60% relative humidity for one week
- $T_3$ : Storage at 60% relative humidity for two weeks
- and so on.

All together there were 43 treatments.

The design of the experiment is indicated schematically in Figure 6.5.2. There were 129 batches of spores, which were randomly allocated to the 43 treatments, three batches to each treatment. Each batch of spores resulted in one petri dish, and each petri dish resulted in 10 plugs.

**Figure 6.5.2** Design of spore germination experiment



To get a feeling for the issues raised by this design, let us look at some of the raw data. Table 6.5.1 shows the percentage of the spores that had germinated for each plug assayed for treatment 1.

Table 6.5.1 shows that there is considerable variability both *within* each petri dish and *between* the dishes. The variability within the dishes reflects local variation in the percent germination, perhaps due largely to differences among the spores themselves (some of the spores were more mature than others). The variability

**Table 6.5.1** Percentage germination under treatment 1

	Dish I	Dish II	Dish III
49	66	49	
58	84	60	
48	83	54	
69	69	72	
45	72	57	
43	85	70	
60	59	65	
44	60	68	
44	75	66	
68	68	60	
Mean	52.8	72.1	62.1
SD	10.1	9.5	7.4

between dishes is even larger, because it includes not only local variation, but also larger-scale variation such as the variability among the original batches of spores, and temperature and relative humidity variations within the storage chambers.

Now consider the problem of comparing treatment 1 to the other treatments. Would it be legitimate to take the point of view that we have 30 observations for each treatment? To focus this question, let us consider the matter of calculating the standard error for the mean of treatment 1. The mean and SD of all 30 observations are

$$\text{Mean} = 62.33$$

$$\text{SD} = 11.88$$

Is it legitimate to calculate the SE of the mean as

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{11.88}{\sqrt{30}} = 2.2$$

As you may suspect, **this is not legitimate**. There is a hierarchical structure in the data, and so we cannot apply the SE formula so naively. An acceptable way to calculate the SE is to consider the mean for each dish as an observation; thus, we obtain the following:<sup>\*</sup>

Observations: 52.8, 72.1, 62.1

$$n = 3$$

$$\text{Mean} = 62.33$$

$$\text{SD} = 9.65$$

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{9.65}{\sqrt{3}} = 5.6$$

Notice that the incorrect analysis gave the same mean (62.33) as this analysis, but an inappropriately small SE (2.2 rather than 5.6). If we were comparing several treatments, the same pattern would tend to hold; the incorrect analysis would tend to produce SEs that were (individually and pooled) too small, which might cause us to “overinterpret” the data, in the sense of suggesting there is significant evidence of treatment differences where none exists.

We should emphasize that, even though the correct analysis requires combining the measurements on the 10 plugs in a dish into a single observation for that dish, the experimenter was not wasting effort by measuring 10 plugs per dish instead of, say, only one plug per dish. The mean of 10 plugs is a much better estimate of the average for the entire dish than is a measurement on one plug; the improved precision for measuring 10 plugs is reflected in a smaller between-dish SD. For instance, for treatment 1 the SD was 9.65; if fewer plugs per dish had been measured, this SD would probably have been larger. ■

The pitfall illustrated by Example 6.5.3 has trapped many an unwary researcher. When hierarchical structures result from repeated measurements on the same individual organism (as in Example 6.5.2), they are relatively easy to recognize. But the hierarchical structure in Example 6.5.3 has a different origin; it is due to the fact that the unit of observation is an individual plug, but individual plugs are not randomly allocated to the treatment groups. Rather, the unit that is randomly allocated to treatment is a batch of spores, which later is plated in a petri dish, which then gives

---

<sup>\*</sup>An alternative way to aggregate the data from the 10 plugs in a dish would be to combine the raw counts of germinated and ungerminated spores for the whole dish and express these as an overall percent germination.

rise to 10 plugs. In the language of experimental design, *plugs* are **nested** within petri dishes. *Whenever observational units are nested within the units that are randomly allocated to treatments, a hierarchical structure may potentially exist in the data.* Note that the difficulty is only “potential”; in some cases a nonhierarchical analysis may be acceptable. For instance, if experience had shown that the differences between petri dishes were negligible, then we might ignore the hierarchical structure in analyzing the data. The decision can be a difficult one and may require expert statistical advice.

The issue of hierarchical data structures has important implications for the design of an experiment as well as its analysis. The sample size ( $n$ ) must be appropriately identified in order to determine whether the experiment includes enough replication. As a simple example, suppose it is proposed to do a spore germination experiment such as that of Example 6.5.3, but with only *one* dish per treatment, rather than three. To see the flaw in this proposal, suppose that the proposed experiment is to include three treatments, with one dish per treatment. With this design, would we then be able to distinguish treatment differences from inherent differences between the dishes? No. The intertreatment differences and the interdish differences would be mutually entangled, or confounded. You can easily visualize this situation if you look at the data in Table 6.5.1 and pretend that those data came from the proposed experiment; that is, pretend that dishes I, II, and III had received different treatments, and that we had no other data. It would be difficult to extract meaningful information about intertreatment differences unless we knew for *certain* that interdish variation was negligible.

We saw in Section 6.4 how to use a preliminary estimate of the SD to determine the sample size ( $n$ ) required to attain a desired degree of precision, as expressed by the SE. These ideas carry over to experiments involving hierarchical data structures. For example, suppose a botanist is planning a spore germination experiment such as that of Example 6.5.3. If she has already decided to use 10 plugs per dish, the remaining problem would be to decide on the number of dishes per treatment. This question could be approached as in Section 6.4, considering the dish as the experimental unit, and using a preliminary estimate of the SD between dishes (which was 9.65 in Example 6.5.3). If, however, she wants to choose optimal values for *both* the number of plugs per dish *and* the number of dishes per treatment, she may wish to consult a statistician.

## Conditions for Validity of a Confidence Interval for $\mu$

A confidence interval for  $\mu$  provides a definite quantitative interpretation for  $SE_{\bar{Y}}$ . Note that the data must be a random sample from the population of interest. If there is bias in the sampling process, then the sampling distribution concepts on which the confidence interval method is based do not hold: Knowing the mean of a biased sample does not provide information about the population mean  $\mu$ . The validity of Student’s  $t$  method for constructing confidence intervals also depends on the form of the population distribution of the observed variable  $Y$ . If  $Y$  follows a normal distribution in the population, then Student’s  $t$  method is exactly **valid**—that is to say, the probability that the confidence interval will contain  $\mu$  is actually equal to the confidence level (for example, 95%). By the same token, this interpretation is approximately valid if the population distribution is approximately normal. Even if the population distribution is not normal, the Student’s  $t$  confidence interval is approximately valid *if* the sample size is large. This fact can often be used to justify the use of the confidence interval even in situations where the population distribution cannot be assumed to be approximately normal.

From a practical point of view, the important question is: How large must the sample be in order for the confidence interval to be approximately valid? Not surprisingly, the answer to this question depends on the *degree* of nonnormality of the population distribution: If the population is only moderately nonnormal, then  $n$  need not be very large. Table 6.5.2 shows the actual probability that a Student's  $t$  confidence interval will contain  $\mu$  for samples from three different populations.<sup>27</sup> The forms of the population distributions are shown in Figure 6.5.3.

**Table 6.5.2** Actual probability that confidence intervals will contain the population mean

(a) 95% confidence interval

	Sample size						
	2	4	8	16	32	64	Very large
Population 1	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Population 2	0.94	0.93	0.94	0.94	0.95	0.95	0.95
Population 3	0.87	0.53	0.57	0.80	0.88	0.92	0.95

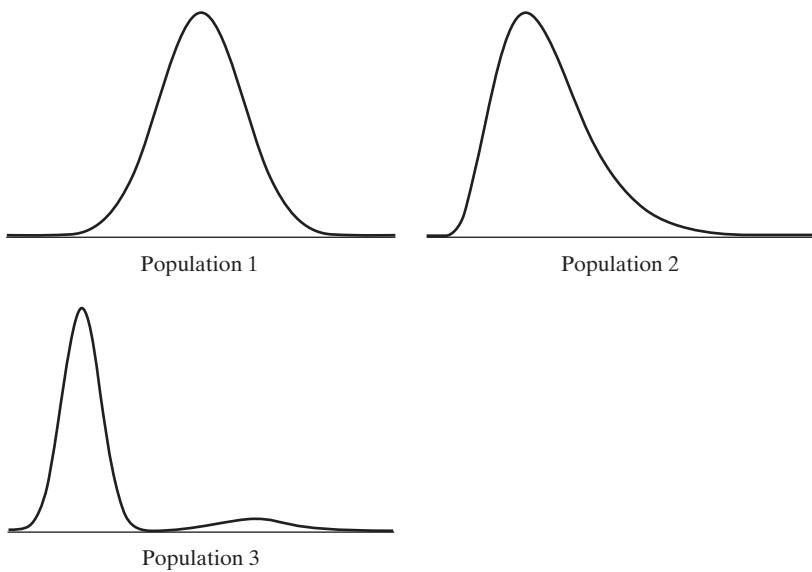
(b) 99% confidence interval

	Sample size						
	2	4	8	16	32	64	Very large
Population 1	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Population 2	0.99	0.98	0.98	0.98	0.99	0.99	0.99
Population 3	0.97	0.82	0.60	0.81	0.93	0.96	0.99

Population 1 is a normal population, population 2 is moderately skewed, and population 3 is an extremely skewed, "L-shaped" distribution. (Populations 2 and 3 were discussed in optional Section 5.3.)

For population 1, Table 6.5.2 shows that the confidence interval method is exactly valid for all sample sizes, even  $n = 2$ . For population 2, the method is approximately valid even for fairly small samples. For population 3 the approximation

**Figure 6.5.3** Three population distributions: (1) normal, (2) slightly skewed right, (3) heavily skewed right



is very poor for small samples and is only fair for samples as large as  $n = 64$ . In a sense, population 3 is a “worst case”; it could be argued that the mean is not a meaningful measure for population 3, because of its bizarre shape.

## Summary of Conditions

In summary, Student’s  $t$  method of constructing a confidence interval for  $\mu$  is appropriate if the following conditions hold.

1. **Conditions on the design of the study**
  - (a) It must be reasonable to regard the data as a random sample from a large population.
  - (b) The observations in the sample must be independent of each other.
2. **Conditions on the form of the population distribution**
  - (a) If  $n$  is small, the population distribution must be approximately normal.
  - (b) If  $n$  is large, the population distribution need not be approximately normal.

The requirement that the data are a random sample is the most important condition.

The required “largeness” in condition 2(b) depends (as shown in Example 6.5.3) on the degree of nonnormality of the population. In many practical situations, moderate sample sizes (say,  $n = 20$  to 30) are large enough.

## Verification of Conditions

In practice, the preceding “conditions” are often “assumptions” rather than known facts. However, it is always important to check whether the conditions are reasonable in a given case.

To determine whether the random sampling model is applicable to a particular study, the design of the study should be scrutinized, with particular attention to possible biases in the choice of experimental material and to possible nonindependence of the observations due to hierarchical data structures.

As to whether the population distribution is approximately normal, information on this point may be available from previous experience with similar data. If the only source of information is the data at hand, then normality can be roughly checked by making a histogram and normal probability plot of the data. Unfortunately, for a small or moderate sample size, this check is fairly crude; for instance, if you look back at Figure 5.2.7, you will see that even samples of size 25 from a normal population often do not appear particularly normal.\* Of course, if the sample is large, then the sample histogram gives us good information about the population shape; however, if  $n$  is large, the requirement of normality is less important anyway.

In any case, a crude check is better than none, and *every* data analysis should begin with inspection of a graph of the data, with special attention to any observations that lie very far from the center of the distribution.

Sometimes a histogram or normal probability plot of the data indicate that the data did not come from a normal population. If the sample size is small, then

---

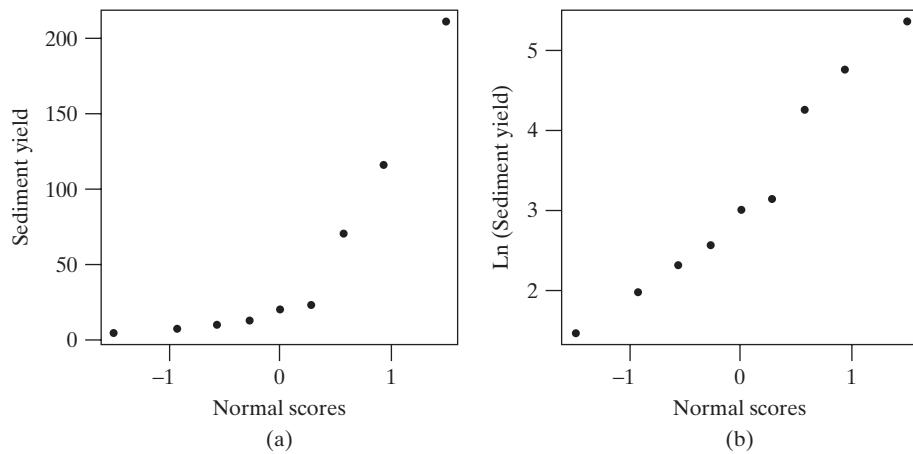
\*We could aid our graphical assessment of normality by using a more objective method such as the Shapiro-Wilk test of Section 4.4.

Student's  $t$  method will not give valid results. However, it may be possible to transform the data to achieve approximate normality and then analyze the data in the transformed scale.

**Example  
6.5.3**

**Sediment Yield** Sediment yield, which is a measure of the amount of suspended sediment in water, is a measure of water quality for a river. The distribution of sediment yield often has a skewed distribution. However, taking the logarithm of each observation can produce a distribution that follows a normal curve quite well. Figure 6.5.4 shows normal probability plots of sediment yields of water samples from the Black River in northern Ohio for  $n = 9$  days (a) in mg/l and (b) in log scale (i.e.,  $\ln(\text{mg/l})$ ).<sup>28</sup>

**Figure 6.5.4** Normal probability plots of sediment yields of water samples from the Black River for nine days (a) in mg/l and (b) after taking the natural logarithm of each observation\*



The natural logarithms of the sediment yields have an average of  $\bar{y} = 3.21$  and a standard deviation of  $s = 1.33$ . Thus, the standard error of the mean is  $\frac{1.33}{\sqrt{9}} = 0.44$ . The  $t$  multiplier for a 95% confidence interval is  $t_{8,0.025} = 2.306$ . A 95% confidence interval for  $\mu$  is

$$3.21 \pm 2.306(0.44)$$

or, approximately,

$$3.21 \pm 1.01$$

or

$$(2.20, 4.22)$$

Thus, we are 95% confident that the mean natural logarithm of sediment yield for the Black River is between 2.20 and 4.22.<sup>†</sup>

\*The Shapiro-Wilk test of normality (from Section 4.4) for the raw data yields a  $P$ -value of 0.0039 providing strong evidence of abnormality for the untransformed data. In contrast, for the natural-log transformed data, the Shapiro-Wilk  $P$ -value is 0.6551, showing no significant evidence for abnormality. Note that we could also have taken the base 10 log to normalize the data.

<sup>†</sup>Note that we have constructed a confidence interval for the population average logarithm of sediment yield. Because the logarithm transformation is not linear, the mean of the logarithms is not the logarithm of the mean, so applying the inverse transformation to the endpoints of the confidence interval will not convert it properly into a confidence interval for the population mean in the original scale of mg/l. However, we can get an approximate confidence interval by taking  $\exp(2.2 + 1.33^2/2)$  and  $\exp(4.22 + 1.33^2/2)$ . [This is based on the fact that the mean of a log normal distribution (which is bell shaped after taking logarithms) is  $\exp(\mu + \sigma^2/2)$ .]

## Exercises 6.5.1–6.5.8

**6.5.1** SGOT is an enzyme that shows elevated activity when the heart muscle is damaged. In a study of 31 patients who underwent heart surgery, serum levels of SGOT were measured 18 hours after surgery.<sup>29</sup> The mean was 49.3 U/l and the standard deviation was 68.3 U/l. If we regard the 31 observations as a sample from a population, what feature of the data would cause one to doubt that the population distribution is normal?

**6.5.2** A dendritic tree is a branched structure that emanates from the body of a nerve cell. In a study of brain development, researchers examined brain tissue from seven adult guinea pigs. The investigators randomly selected nerve cells from a certain region of the brain and counted the number of dendritic branch segments emanating from each selected cell. A total of 36 cells was selected, and the resulting counts were as follows.<sup>30</sup>

38	42	25	35	35	33	48	53	17
24	26	26	47	28	24	35	38	26
38	29	49	26	41	26	35	38	44
25	45	28	31	46	32	39	59	53

The mean of these counts is 35.67 and the standard deviation is 9.99.

Suppose we want to construct a 95% confidence interval for the population mean. We could calculate the standard error as

$$SE_{\bar{Y}} = \frac{9.99}{\sqrt{36}} = 1.67$$

and obtain the confidence interval as

$$35.67 \pm (2.042)(1.67)$$

or

$$32.3 < \mu < 39.1$$

- (a) On what grounds might the above analysis be criticized? (*Hint:* Are the observations independent?)
- (b) Using the classes [15, 20], [20, 25], and so on, construct a histogram of the data. Does the shape of the distribution support the criticism you made in part (a)? If so, explain how.

**6.5.3** In an experiment to study the regulation of insulin secretion, blood samples were obtained from seven dogs before and after electrical stimulation of the vagus nerve. The following values show, for each animal, the increase (after minus before) in the immunoreactive insulin concentration ( $\mu\text{U/ml}$ ) in pancreatic venous plasma.<sup>31</sup>

30	100	60	30	130	1,060	30
----	-----	----	----	-----	-------	----

For these data, Student's  $t$  method yields the following 95% confidence interval for the population mean:

$$-145 < \mu < 556$$

Is Student's  $t$  method appropriate in this case? Why or why not?

**6.5.4** In a study of parasite–host relationships, 242 larvae of the moth *Ephestia* were exposed to parasitization by the Ichneumon fly. The following table shows the number of Ichneumon eggs found in each of the *Ephestia* larva.<sup>32</sup>

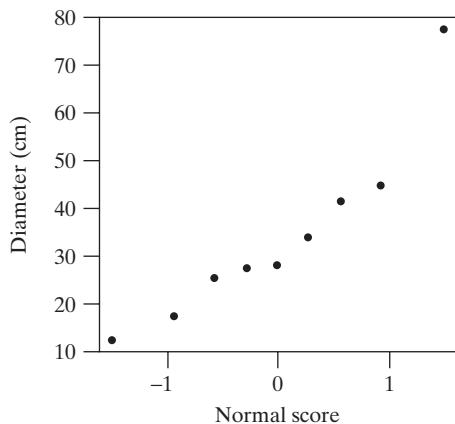
NUMBER OF EGGS (Y)	NUMBER OF LARVAE
0	21
1	77
2	52
3	41
4	23
5	13
6	9
7	1
8	2
9	0
10	2
11	0
12	0
13	0
14	0
15	1
Total	242

For these data,  $\bar{y} = 2.368$  and  $s = 1.950$ . Student's  $t$  method yields the following 95% confidence interval for  $\mu$ , the population mean number of eggs per larva:

$$2.12 < \mu < 2.61$$

- (a) Does it appear reasonable to assume that the population distribution of  $Y$  is approximately normal? Explain.
- (b) In view of your answer to part (a), on what grounds can you defend the application of Student's  $t$  method to these data?

**6.5.5** The following normal probability plot shows the distribution of the diameters, in cm, of each of nine American Sycamore trees.<sup>33</sup>



The normal probability plot is not linear, which suggests that a transformation of the data is needed before a confidence interval can be constructed using Student's  $t$  method. The raw data are

12.4 44.8 28.2 77.6 34 17.5 41.5 25.5 27.5

- (a) Take the square root of each observation and construct a 90% confidence interval for the mean.
- (b) Interpret the confidence interval from part (a). That is, explain what the interval tells you about the square root of the diameters of these trees.

**6.5.6** Four treatments were compared for their effect on the growth of spinach cells in cell culture flasks. The

experimenter randomly allocated two flasks to each treatment. After a certain time on treatment, he randomly drew three aliquots (1 cc each) from each flask and measured the cell density in each aliquot; thus, he had six cell density measurements for each treatment. In calculating the standard error of a treatment mean, the experimenter calculated the standard deviation of the six measurements and divided by  $\sqrt{6}$ . On what grounds might an objection be raised to this method of calculating the SE?

**6.5.7** In an experiment on soybean varieties, individually potted soybean plants were grown in a greenhouse, with 10 plants of each variety used in the experiment. From the harvest of each plant, five seeds were chosen at random and individually analyzed for their percentage of oil. This gave a total of 50 measurements for each variety. To calculate the standard error of the mean for a variety, the experimenter calculated the standard deviation of the 50 observations and divided by  $\sqrt{50}$ . Why would this calculation be of doubtful validity?

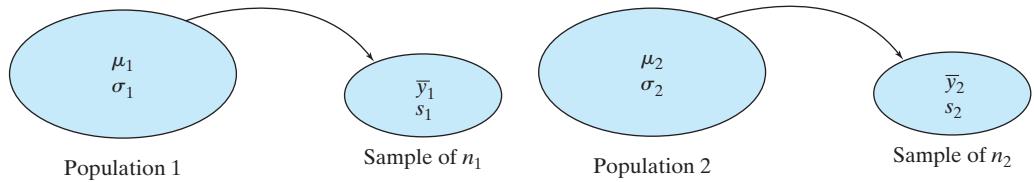
**6.5.8** In a plant mitigation project, an entire local (endangered) population of 255 Congdon's tarplants was transplanted to a new location.<sup>34</sup> One year after transplant, 30 of the 255 plants were randomly selected and the diameter at the root caudix junction (the top of the root just beneath the surface of the soil) was measured. If the population of plants under consideration consists of only the local 255 plants, explain why it would be improper to use Student's  $t$  method of constructing a confidence interval for  $\mu$ , the population mean root caudix junction diameter.

## 6.6 Comparing Two Means

In previous sections we have considered the analysis of a single sample of quantitative data. In practice, however, much scientific research involves the comparison of two or more samples from different populations. When the observed variable is quantitative, the comparison of two samples can include several aspects, notably (1) comparison of means, (2) comparison of standard deviations, and (3) comparison of shapes. In this section, and indeed throughout this book, the primary emphasis will be on comparison of means and on other comparisons related to shift. We will begin by discussing the confidence interval approach to comparing means, which is a natural extension of the material in Section 6.3; in Chapter 7 we will consider an approach known as hypothesis testing.

### Notation

Figure 6.6.1 presents our notation for comparison of two samples. The notation is exactly parallel to our earlier notation, but now a subscript (1 or 2) is used to differentiate between the two samples. The two "populations" can be naturally

**Figure 6.6.1** Notation for comparison of two samples

occurring populations (as in Example 6.1.1) or they can be conceptual populations defined by certain experimental conditions (as in Example 6.3.4). In either case, the data in each sample are viewed as a random sample from the corresponding population.

We begin by describing, in the next section, some simple computations that are used for both confidence intervals and hypothesis testing.

### Standard Error of $(\bar{Y}_1 - \bar{Y}_2)$

In this section we introduce a fundamental quantity for comparing two samples: the standard error of the difference between two sample means.

#### Basic Ideas

We saw in Chapter 6 that the precision of a sample mean  $\bar{Y}$  can be expressed by its standard error, which is equal to

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

To compare two sample means, it is natural to consider the difference between them:

$$\bar{Y}_1 - \bar{Y}_2$$

which is an estimate of the quantity  $(\mu_1 - \mu_2)$ . To characterize the sampling error of estimation, we need to be concerned with the standard error of the difference  $(\bar{Y}_1 - \bar{Y}_2)$ . We illustrate this idea with an example.

---

**Example  
6.6.1**

**Vital Capacity** Vital capacity is a measure of the amount of air that someone can exhale after taking a deep breath. One might expect that musicians who play brass instruments would have greater vital capacities, on average, than would other persons of the same age, sex, and height. In one study the vital capacities of eight brass players were compared to the vital capacities of seven control subjects; Table 6.6.1 shows the data.<sup>35</sup>

The difference between the sample means is

$$\bar{y}_1 - \bar{y}_2 = 4.83 - 4.74 = 0.09$$

We know that both  $\bar{y}_1$  and  $\bar{y}_2$  are subject to sampling error, and consequently the difference (0.09) is subject to sampling error. The standard error of  $\bar{Y}_1 - \bar{Y}_2$  tells us how much precision to attach to this difference between  $\bar{Y}_1$  and  $\bar{Y}_2$ .

<b>Table 6.6.1</b> Vital capacity (liters)		
	Brass player	Control
	4.7	4.2
	4.6	4.7
	4.3	5.1
	4.5	4.7
	5.5	5.0
	4.9	
	5.3	
<i>n</i>	7	5
$\bar{y}$	4.83	4.74
<i>s</i>	0.435	0.351

### Definition

The **standard error of  $\bar{Y}_1 - \bar{Y}_2$**  is defined as

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The following alternative form of the formula shows how the SE of the difference is related to the individual SEs of the means:

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\text{SE}_1^2 + \text{SE}_2^2}$$

where

$$\text{SE}_1 = \text{SE}_{\bar{Y}_1} = \frac{s_1}{\sqrt{n_1}}$$

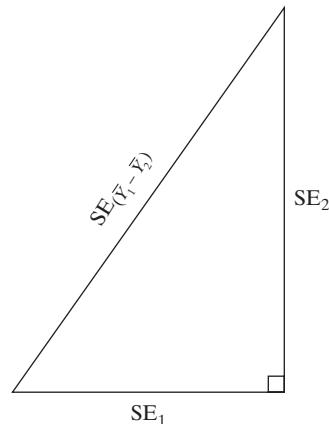
$$\text{SE}_2 = \text{SE}_{\bar{Y}_2} = \frac{s_2}{\sqrt{n_2}}$$

Notice that this version of the formula shows that “SEs add like Pythagorus.” When we have two independent samples, we take the SE of each mean, square them, add them, and then take the square root of the sum. Figure 6.6.2 illustrates this idea.

It may seem odd that in calculating the SE of a difference we *add* rather than subtract within the formula  $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\text{SE}_1^2 + \text{SE}_2^2}$ . However, as was discussed in Section 3.5, the variability of the difference depends on the variability of each part. Whether we add  $\bar{Y}_2$  to  $\bar{Y}_1$  or subtract  $\bar{Y}_2$  from  $\bar{Y}_1$ , the “noise” associated with  $\bar{Y}_2$  (i.e.,  $\text{SE}_2$ ) adds to the overall uncertainty. The greater the variability in  $\bar{Y}_2$ , the greater the variability in  $\bar{Y}_1 - \bar{Y}_2$ . The formula  $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\text{SE}_1^2 + \text{SE}_2^2}$  accounts for this variability.

We illustrate the formulas in the following example.

**Figure 6.6.2** SE for a difference



**Example 6.6.2**

**Vital Capacity** For the vital capacity data, preliminary computations yield the results in Table 6.6.2.

The SE of  $(\bar{Y}_1 - \bar{Y}_2)$  is

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{0.1892}{7} + \frac{0.1232}{5}} = 0.227 \approx 0.23$$

Note that

$$0.227 = \sqrt{(0.164)^2 + (0.157)^2}$$

Notice that the SE of the difference is greater than either of the individual SEs but less than their sum. ■

<b>Table 6.6.2</b>		
	Brass player	Control
$s^2$	0.1892	0.1232
$n$	7	5
SE	0.164	0.157

**Example 6.6.3**

**Tonsillectomy** An experiment was conducted to compare conventional surgery to a newer procedure called Coblation-assisted intracapsular tonsillectomy for children who needed to have their tonsils removed. A key measurement taken during the study was the pain score that each child reported, on a scale of 0–10, four days after surgery. Table 6.6.3 gives the means and standard deviations of pain scores for the two groups.<sup>36</sup>

<b>Table 6.6.3</b> Pain score		
Type of surgery		
	Conventional	Coblation
Mean	4.3	1.9
SD	2.8	1.8
$n$	49	52

The data in Table 6.6.3 show that the standard deviation of pain scores in 49 children given conventional surgery was 2.8. Thus, the SE for the conventional mean is  $\frac{2.8}{\sqrt{49}} = 0.40$ . For the 52 children in the coblation group, the SD was 1.8, which gives an SE of  $\frac{1.8}{\sqrt{52}} = 0.2496$ . The SE for the difference in the two means is  $\sqrt{0.40^2 + 0.25^2} = 0.4717 \approx 0.47$ . ■

## The Pooled Standard Error (Optional)

The preceding standard error is known as the “unpooled” standard error. Many statistics software packages allow the user to specify use of what is known as the “pooled” standard error, which we will discuss briefly.

Recall that the square of the standard deviation,  $s$ , is the sample variance,  $s^2$ , defined as

$$s^2 = \frac{\sum(\bar{y}_i - \bar{y})^2}{n - 1}$$

The pooled variance is a weighted average of  $s_1^2$ , the variance of the first sample, and  $s_2^2$ , the variance of the second sample, with weights equal to the degrees of freedom from each sample,  $n_i - 1$ :

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}.$$

The pooled standard error is defined as

$$\text{SE}_{\text{pooled}} = \sqrt{s_{\text{pooled}}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

We illustrate with an example.

---

**Example  
6.6.4**

**Vital Capacity** For the vital capacity data we found that  $s_1^2 = 0.1892$  and  $s_2^2 = 0.1232$ . The pooled variance is

$$s_{\text{pooled}}^2 = \frac{(7 - 1)0.1892 + (5 - 1)0.1232}{(7 + 5 - 2)} = 0.1628$$

and the pooled SE is

$$\text{SE}_{\text{pooled}} = \sqrt{0.1628 \left( \frac{1}{7} + \frac{1}{5} \right)} = 0.236.$$

Recall from Example 6.6.2 that the unpooled SE for the same data was 0.227. ■

If the sample sizes are equal ( $n_1 = n_2$ ) or if the sample standard deviations are equal ( $s_1 = s_2$ ), then the unpooled and the pooled method will give the same answer for  $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$ . The two answers will not differ substantially unless both the sample sizes and the sample SDs are quite discrepant.

To show the analogy between the two SE formulas, we can write them as follows:

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and

$$\text{SE}_{\text{pooled}} = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}}$$

In the pooled method, the separate variances— $s_1^2$  and  $s_2^2$ —are replaced by the single variance  $s_{\text{pooled}}^2$ , which is calculated from both samples.

Both the unpooled and the pooled SE have the same purpose—to estimate the standard deviation of the sampling distribution of  $(\bar{Y}_1 - \bar{Y}_2)$ . In fact, it can be shown that the standard deviation is

$$\sigma_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note the resemblance between this formula and the formula for  $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$ .

In analyzing data when the sample sizes are unequal ( $n_1 \neq n_2$ ), one needs to decide whether to use the pooled or unpooled method for calculating the standard error. The choice depends on whether one is willing to assume that the population SDs ( $\sigma_1$  and  $\sigma_2$ ) are equal. It can be shown that if  $\sigma_1 = \sigma_2$ , then the pooled method should be used, because in this case  $s_{\text{pooled}}$  is the best estimate of the population SD. However, in this case the unpooled method will typically give an SE that is quite similar to that given by the pooled method. If  $\sigma_1 \neq \sigma_2$ , then the unpooled method should be used, because in this case  $s_{\text{pooled}}$  is not an estimate of either  $\sigma_1$  or  $\sigma_2$ , so that pooling would accomplish nothing. Because the two methods substantially agree when  $\sigma_1 = \sigma_2$  and the pooled method is not valid when  $\sigma_1 \neq \sigma_2$ , most statisticians prefer the unpooled method. There is little to be gained by pooling when pooling is appropriate and there is much to be lost when pooling is not appropriate. Many software packages use the unpooled method by default; the user must specify use of the pooled method if she or he wishes to pool the variances.

## Exercises 6.6.1–6.6.9

**6.6.1** Data from two samples gave the following results:

	SAMPLE 1	SAMPLE 2
$n$	6	12
$\bar{y}$	40	50
$s$	4.3	5.7

**6.6.2** Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$  for the following data:

	SAMPLE 1	SAMPLE 2
$n$	10	10
$\bar{y}$	125	217
$s$	44.2	28.7

Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$ .

**6.6.3** Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$  for the following data:

	SAMPLE 1	SAMPLE 2
$n$	5	7
$\bar{y}$	44	47
$s$	6.5	8.4

**6.6.4** Consider the data from Exercise 6.6.3. Suppose the sample sizes were doubled, but the means and SDs stayed the same, as follows. Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$ .

	SAMPLE 1	SAMPLE 2
$n$	10	14
$\bar{y}$	44	47
$s$	6.5	8.4

**6.6.5** Data from two samples gave the following results:

	SAMPLE 1	SAMPLE 2
$\bar{y}$	96.2	87.3
SE	3.7	4.6

Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$ .

**6.6.6** Data from two samples gave the following results:

	SAMPLE 1	SAMPLE 2
$n$	22	21
$\bar{y}$	1.7	2.4
SE	0.5	0.7

Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$ .

**6.6.7** Example 6.6.3 reports measurements of pain for children who have had their tonsils removed. Another variable measured in that experiment was the number of doses of Tylenol taken by the children in the two groups. Those data are

TYPE OF SURGERY	
CONVENTIONAL	COBLATION
$n$	49
$\bar{y}$	3.0
SD	2.4
	52
	2.3
	2.0

Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$ .

**6.6.8** Two varieties of lettuce were grown for 16 days in a controlled environment. The following table shows the total dry weight (in grams) of the leaves of nine plants of the variety “Salad Bowl” and six plants of the variety “Bibb.”<sup>37</sup>

SALAD BOWL	BIBB
3.06	1.31
2.78	1.17
2.87	1.72
3.52	1.20
3.81	1.55
3.60	1.53
3.30	
2.77	
3.62	
$\bar{y}$	3.259
$s$	.400
	.220

Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$  for these data.

**6.6.9** Some soap manufacturers sell special “antibacterial” soaps. However, one might expect ordinary soap to also kill bacteria. To investigate this, a researcher prepared a solution from ordinary, nonantibiotic soap and a control solution of sterile water. The two solutions were placed onto petri dishes and *E. coli* bacteria were added. The dishes were incubated for 24 hours and the number of bacteria colonies on each dish were counted.<sup>38</sup> The data are given in the following table.

CONTROL (GROUP 1)	SOAP (GROUP 2)
30	76
36	27
66	16
21	30
63	26
38	46
35	6
45	
$n$	8
$\bar{y}$	41.8
$s$	15.6
SE	5.5
	8.6

Compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$  for these data.

## 6.7 Confidence Interval for $(\mu_1 - \mu_2)$

One way to compare two sample means is to construct a confidence interval for the difference in the population means—that is, a confidence interval for the quantity  $(\mu_1 - \mu_2)$ . Recall from Chapter 6 that a 95% confidence interval for the mean  $\mu$  of a single population that is normally distributed is constructed as

$$\bar{y} \pm t_{0.025} \text{SE}_{\bar{Y}}$$

Analogously, a 95% confidence interval for  $(\mu_1 - \mu_2)$  is constructed as

$$(\bar{y}_1 - \bar{y}_2) \pm t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$$

The critical value  $t_{0.025}$  is determined from Student's  $t$  distribution using degrees of freedom\* given as

$$df = \frac{(\text{SE}_1^2 + \text{SE}_2^2)^2}{\text{SE}_1^4/(n_1 - 1) + \text{SE}_2^4/(n_2 - 1)} \quad (6.7.1)$$

where  $\text{SE}_1 = s_1/\sqrt{n_1}$  and  $\text{SE}_2 = s_2/\sqrt{n_2}$ .

Of course, calculating the degrees of freedom from formula (6.7.1) is complicated and time consuming. Most computer software uses formula (6.7.1), as do some graphing calculators. A simpler method to obtain the approximate degrees of freedom is to use the smaller of  $(n_1 - 1)$  and  $(n_2 - 1)$ . This option gives a confidence interval that is somewhat conservative, in the sense that the true confidence level is a bit larger than 95% when  $t_{0.025}$  is used. A third approach is to approximate the degrees of freedom as  $n_1 + n_2 - 2$ . This approach is somewhat liberal, in the sense that the true confidence level is a bit smaller than 95% when  $t_{0.025}$  is used.

Intervals with other confidence coefficients are constructed analogously; for example, for a 90% confidence interval one would use  $t_{0.05}$  instead of  $t_{0.025}$ .

The following example illustrates the construction of a confidence interval for  $(\mu_1 - \mu_2)$ .

### Example 6.7.1

**Fast Plants** The Wisconsin Fast Plant, *Brassica campestris*, has a very rapid growth cycle that makes it particularly well suited for the study of factors that affect plant growth. In one such study, seven plants were treated with the substance Ancymidol (ancy) and were compared to eight control plants that were given ordinary water. Heights of all of the plants were measured, in cm, after 14 days of growth.<sup>39</sup> The data are given in Table 6.7.1.

Parallel dotplots and normal probability plots (Figure 6.7.1) show that both sample distributions are reasonably symmetric and bell shaped. Moreover, we would expect that a distribution of plant heights might well be normally distributed, since height distributions often follow a normal curve. The dotplots show that the ancy distribution is shifted down a bit from the control distribution; the difference in sample means is  $15.9 - 11.0 = 4.9$ . The SE for the difference in sample means is

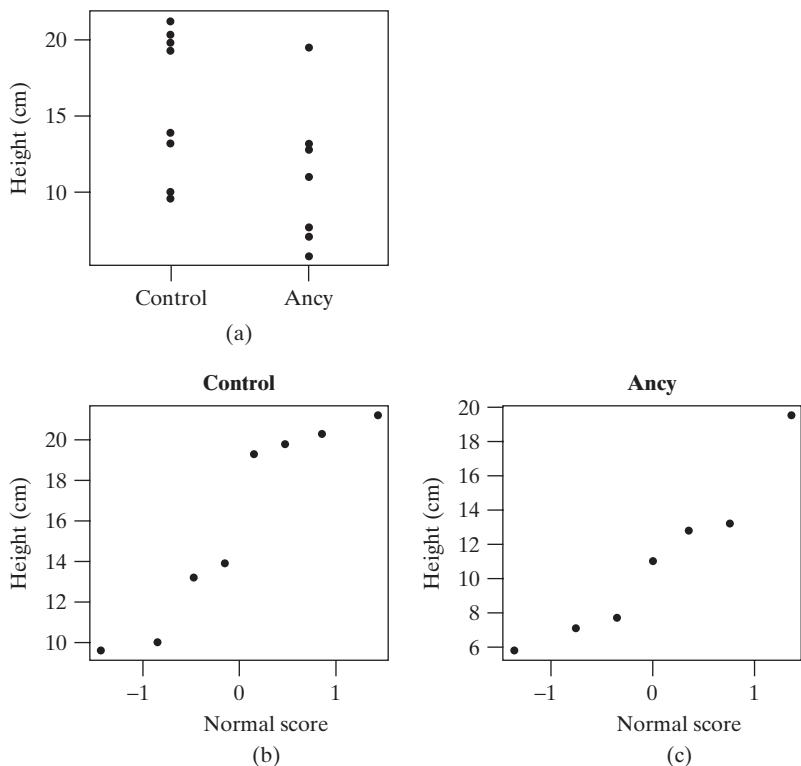
$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{4.8^2}{8} + \frac{4.7^2}{7}} = 2.46$$

---

\*Strictly speaking, the distribution needed to construct a confidence interval here depends on the unknown population standard deviations  $\sigma_1$  and  $\sigma_2$  and is not a Student's  $t$  distribution. However, Student's  $t$  distribution with degrees of freedom given by formula (6.7.1) is a very good approximation. This is sometimes known as Welch's method or Satterthwaite's method.

**Table 6.7.1** Fourteen-day height of control and of ancy plants (cm)

	Control (Group 1)	Ancy (Group 2)
	10.0	13.2
	13.2	19.5
	19.8	11.0
	19.3	5.8
	21.2	12.8
	13.9	7.1
	20.3	7.7
	9.6	
<i>n</i>	8	7
$\bar{y}$	15.9	11.0
<i>s</i>	4.8	4.7
SE	1.7	1.8

**Figure 6.7.1** Parallel dotplots (a) and normal probability plots of heights of fast plants receiving Control (b) and Ancy (c)

Using Formula (6.7.1), we find the degrees of freedom to be 12.8:

$$df = \frac{(1.7^2 + 1.8^2)^2}{1.7^4/7 + 1.8^4/6} = 12.8$$

Using a computer, we can find that for a 95% confidence interval the *t* multiplier for 12.8 degrees of freedom is  $t_{12.8, 0.025} = 2.164$ . (Without a computer, we could round down the degrees of freedom to 12, in which case the *t* multiplier is 2.179.

This change from 12.8 to 12 degrees of freedom has little effect on the final answer.) The confidence interval formula gives

$$(15.9 - 11.0) \pm (2.164)(2.46)$$

or

$$4.9 \pm 5.32$$

The 95% confidence interval for  $(\mu_1 - \mu_2)$  is

$$(-0.42, 10.22)$$

Rounding off, we have

$$(-0.4, 10.2)$$

Thus, we are 95% confident that the population average 14-day height of fast plants when water is used ( $\mu_1$ ) is between 0.4 cm lower and 10.2 cm higher than the average 14-day height of fast plants when water is used ( $\mu_2$ ). ■

---

**Example  
6.7.2**

**Fast Plants** We said that a conservative method of constructing a confidence interval for a difference in means is to use the smaller of  $n_1 - 1$  and  $n_2 - 1$ . For the data given in Example 6.7.1, this method would use 6 degrees of freedom and a  $t$  multiplier of 2.447. In this case, the 95% confidence interval for  $(\mu_1 - \mu_2)$  is

$$(15.9 - 11.0) \pm (2.447)(2.46)$$

or

$$4.9 \pm 6.02$$

The 95% confidence interval for  $(\mu_1 - \mu_2)$  is

$$(-1.1, 10.9)$$

This interval is a bit conservative in the sense that the interval is wider than the interval found in Example 6.7.1. ■

---

**Example  
6.7.3**

**Thorax Weight** Biologists have theorized that male Monarch butterflies have, on average, a larger thorax than do females. A sample of seven male and eight female Monarchs yielded the data in Table 6.7.2, which are displayed in Figure 6.7.2. (These data come from another part of the study described in Example 6.1.1.)

For the data in Table 6.7.2, the SE for  $(\bar{Y}_1 - \bar{Y}_2)$  is

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{8.4^2}{7} + \frac{7.5^2}{8}} = 4.14$$

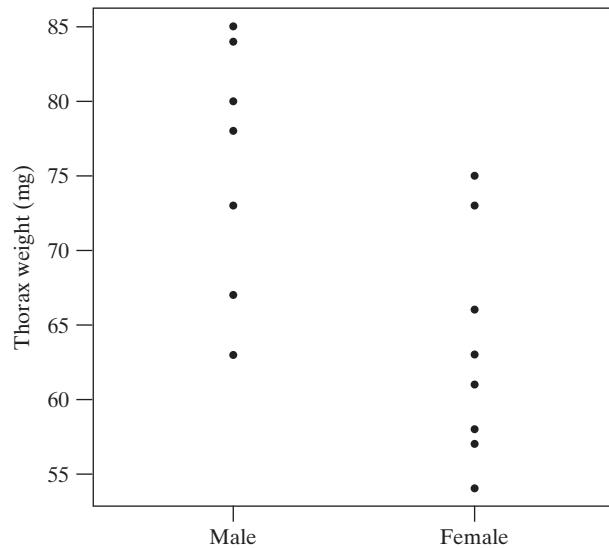
Formula (6.7.1) gives degrees of freedom

$$\text{df} = \frac{(3.2^2 + 2.7^2)^2}{\frac{3.2^4}{6} + \frac{2.7^4}{7}} = 12.3$$

For a 95% confidence interval the  $t$  multiplier is  $t_{12.3, 0.025} = 2.173$ . (We could round the degrees of freedom to 12, in which case the  $t$  multiplier is 2.179. This change

<b>Table 6.7.2</b> Thorax weight (mg)	
Male	Female
67	73
73	54
85	61
84	63
78	66
63	57
80	75
	58
<i>n</i>	7
$\bar{y}$	75.7
<i>s</i>	8.4
SE	3.2
	63.4
	7.5
	2.7

**Figure 6.7.2** Parallel dotplots of thorax weights



from 12.3 to 12 degrees of freedom has only a small effect on the final answer.) The confidence interval formula gives

$$(75.7 - 63.4) \pm (2.173)(4.14)$$

or

$$12.3 \pm 9.0$$

and the 95% confidence interval for  $(\mu_1 - \mu_2)$  is

$$(3.3, 21.3)$$

According to the confidence interval, we can be 95% confident that the population mean thorax weight for male Monarch butterflies ( $\mu_1$ ) is larger than that for females ( $\mu_2$ ) by an amount that might be as small as 3.3 mg or as large as 21.3 mg.

Likewise, for a 90% confidence interval the  $t$  multiplier is  $t_{12.3, 0.05} = 1.779$ . The confidence interval formula gives

$$(75.7 - 63.4) \pm (1.779)(4.14)$$

or

$$12.3 \pm 7.4$$

and the 90% confidence interval for  $(\mu_1 - \mu_2)$  is

$$(4.9, 19.7)$$

According to the confidence interval, we can be 90% confident that the population mean thorax weight for male Monarch butterflies ( $\mu_1$ ) is larger than that for females ( $\mu_2$ ) by an amount that might be as small as 4.9 mg or as large as 19.7 mg. ■

**Conditions for Validity** In Section 6.5 we stated the conditions that make a confidence interval for a mean valid: We require that the data can be thought of as (1) a random sample from (2) a normal population. Likewise, when comparing two means, we require two independent, random samples from normal populations. If the sample sizes are large, then the condition of normality is not crucial (due to the Central Limit Theorem).

## Exercises 6.7.1–6.7.14

**6.7.1** In Table 6.6.3, data were presented from an experiment that compared two types of surgery. The average pain score of the 49 children given conventional tonsillectomies was 4.3, with an SD of 2.8. For the 52 children in the Coblation group the average was 1.9 with an SD of 1.8. Use these data to construct a 95% confidence interval for the difference in population average pain scores. [Note: Formula (6.7.1) yields 81.1 degrees of freedom for these data.]

**6.7.2** Ferulic acid is a compound that may play a role in disease resistance in corn. A botanist measured the concentration of soluble ferulic acid in corn seedlings grown in the dark or in a light/dark photoperiod. The results (nmol acid per gm tissue) were as shown in the table.<sup>40</sup>

	DARK	PHOTOPERIOD
$n$	4	4
$\bar{y}$	92	115
$s$	13	13

- (a) Construct a 95% confidence interval for the difference in ferulic acid concentration under the two lighting conditions. (Assume that the two populations from which the data came are normally distributed.) [Note: Formula (6.7.1) yields 6 degrees of freedom for these data.]
- (b) Repeat part (a) for a 90% level of confidence.

**6.7.3 (Continuation of 6.7.2)** Using your work from Exercise 6.7.2(a), fill in the blank: “We are 95% confident

that the difference in population means is at least \_\_\_\_\_ nmol/g.”

**6.7.4** A study was conducted to determine whether relaxation training, aided by biofeedback and meditation, could help in reducing high blood pressure. Subjects were randomly allocated to a biofeedback group or a control group. The biofeedback group received training for eight weeks. The table reports the reduction in systolic blood pressure (mm Hg) after eight weeks.<sup>41</sup> [Note: Formula (6.7.1) yields 190 degrees of freedom for these data.]

- (a) Construct a 95% confidence interval for the difference in mean response.
- (b) Interpret the confidence interval from part (a) in the context of this setting.

	BIOFEEDBACK	CONTROL
$n$	99	93
$\bar{y}$	13.8	4.0
SE	1.34	1.30

**6.7.5** Consider the data in Exercise 6.7.4. Suppose we are worried that the blood pressure data do not come from normal distributions. Does this mean that the confidence interval found in Exercise 6.7.3 is not valid? Why or why not?

**6.7.6** Prothrombin time is a measure of the clotting ability of blood. For 10 rats treated with an antibiotic and 10 control rats, the prothrombin times (in seconds) were reported as follows:<sup>42</sup>

	ANTIBIOTIC	CONTROL
$n$	10	10
$\bar{y}$	25	23
$s$	10	8

- (a) Construct a 90% confidence interval for the difference in population means. (Assume that the two populations from which the data came are normally distributed.) [Note: Formula (6.7.1) yields 17.2 degrees of freedom for these data.]
- (b) Why is it important that we assume that the two populations are normally distributed in part (a)?
- (c) Interpret the confidence interval from part (a) in the context of this setting.

**6.7.7** The accompanying table summarizes the sucrose consumption (mg in 30 minutes) of black blowflies injected with Pargyline or saline (control).<sup>43</sup>

	CONTROL	PARGYLINE
$n$	900	905
$\bar{y}$	14.9	46.5
$s$	5.4	11.7

- (a) Construct a 95% confidence interval for the difference in population means. [Note: Formula (6.7.1) yields 1,274 degrees of freedom for these data.]
- (b) Repeat part (a) using a 99% level of confidence.

**6.7.8** In a field study of mating behavior in the Mormon cricket (*Anabrus simplex*), a biologist noted that some females mated successfully while others were rejected by the males before coupling was complete. The question arose whether some aspect of body size might play a role in mating success. The accompanying table summarizes measurements of head width (mm) in the two groups of females.<sup>44</sup>

- (a) Construct a 95% confidence interval for the difference in population means. [Note: Formula (6.7.1) yields 35.7 degrees of freedom for these data.]
- (b) Interpret the confidence interval from part (a) in the context of this setting.
- (c) Using your interval computed in (a) to support your answer, is there strong evidence that the population mean head width is indeed larger for successful maters than unsuccessful maters?

	SUCCESSFUL	UNSUCCESSFUL
$n$	22	17
$\bar{y}$	8.498	8.440
$s$	0.283	0.262

**6.7.9** In an experiment to assess the effect of diet on blood pressure, 154 adults were placed on a diet rich in fruits and vegetables. A second group of 154 adults was placed on a standard diet. The blood pressures of the 308 subjects were recorded at the start of the study. Eight weeks later, the blood pressures of the subjects were measured again and the change in blood pressure was recorded for each person. Subjects on the fruits-and-vegetables diet had an average drop in systolic blood pressure of 2.8 mm Hg more than did subjects on the standard diet. A 97.5% confidence interval for the difference between the two population means is (0.9, 4.7).<sup>45</sup> Interpret this confidence interval. That is, explain what the numbers in the interval mean. (See Examples 6.7.1 and 6.7.3.)

**6.7.10** Consider the experiment described in Exercise 6.7.9. For the same subjects, the change in diastolic blood pressure was 1.1 mm Hg greater, on average, for the subjects on the fruits-and-vegetables diet than for subjects on the standard diet. A 97.5% confidence interval for the difference between the two population means is (-0.3, 2.4). Interpret this confidence interval. That is, explain what the numbers in the interval mean. (See Examples 6.7.1 and 6.7.3.)

**6.7.11** Researchers were interested in the short-term effect that caffeine has on heart rate. They enlisted a group of volunteers and measured each person's resting heart rate. Then they had each subject drink 6 ounces of coffee. Nine of the subjects were given coffee containing caffeine and 11 were given decaffeinated coffee. After 10 minutes each person's heart rate was measured again. The data in the table show the change in heart rate; a positive number means that heart rate went up and a negative number means that heart rate went down.<sup>46</sup>

	CAFFEINE	DECAF
$n$	9	11
$\bar{y}$	7.3	5.9
$s$	11.1	11.2
SE	3.7	3.4
	18	-10

- (a) Use these data to construct a 90% confidence interval for the difference in mean affect that caffeinated coffee has on heart rate, in comparison to decaffeinated coffee. [Note: Formula (6.7.1) yields 17.3 degrees of freedom for these data.]
- (b) Using the interval computed in part (a) to justify your answer, is it reasonable to believe that caffeine may not affect heart rates?
- (c) Using the interval computed in part (a) to justify your answer, is it reasonable to believe that caffeine may affect heart rates? If so, by how much?
- (d) Are your answers to (b) and (c) contradictory? Explain.

**6.7.12** Consider the data from Exercise 6.7.11. Given that there are only a small number of observations in each group, the confidence interval calculated in Exercise 6.7.11 is only valid if the underlying populations are normally distributed. Is the normality condition reasonable here? Support your answer with appropriate graphs.

**6.7.13** A researcher investigated the effect of green light, in comparison to red light, on the growth rate of bean plants. The following table shows data on the heights of plants (in inches) from the soil to the first branching stem, two weeks after germination.<sup>47</sup> Use these data to construct a 95% confidence interval for the difference in mean affect that red light has on bean plant growth, in comparison to green light. [Note: Formula (6.7.1) yields 38 degrees of freedom for these data.]

**6.7.14** The distributions of the data from Exercise 6.7.13 are somewhat skewed, particularly the red group. Does this mean that the confidence interval calculated in Exercise 6.7.13 is not valid? Why or why not?

	RED	GREEN
	8.4	8.6
	8.4	5.9
	10.0	4.6
	8.8	9.1
	7.1	9.8
	9.4	10.1
	8.8	6.0
	4.3	10.4
	9.0	10.8
	8.4	9.6
	7.1	10.5
	9.6	9.0
	9.3	8.6
	8.6	10.5
	6.1	9.9
	8.4	11.1
	10.4	5.5
		8.2
		8.3
		10.0
		8.7
		9.8
		9.5
		11.0
		8.0
<i>n</i>	17	25
$\bar{y}$	8.36	8.94
<i>s</i>	1.50	1.78
SE	0.36	0.36

## 6.8 Perspective and Summary

In this section we place Chapter 6 in perspective by relating it to other chapters and also to other methods for analyzing a single sample of data. We also present a condensed summary of the methods of Chapter 6.

### Sampling Distributions and Data Analysis

The theory of the sampling distribution of  $\bar{Y}$  (Section 5.3) seemed to require knowledge of quantities— $\mu$  and  $\sigma$ —that in practice are unknown. In Chapter 6, however, we have seen how to make an inference about  $\mu$  and  $(\mu_1 - \mu_2)$ , including an assessment of the precision of that inference, using only information provided by the sample. Thus, the theory of sampling distributions has led to a practical method of analyzing data.

In later chapters we will study more complex methods of data analysis. Each method is derived from an appropriate sampling distribution; in most cases, however, we will not study the sampling distribution in detail.

## Choice of Confidence Level

In illustrating the confidence interval methods, we have often chosen a confidence level equal to 95%. However, it should be remembered that the confidence level is arbitrary. It is true that in practice the 95% level is the confidence level that is most widely used; however, there is nothing wrong with an 80% confidence interval, for example.

## Characteristics of Other Measures

This chapter has primarily discussed estimation of a population mean,  $\mu$ , and the difference of two population means ( $\mu_1 - \mu_2$ ). In some situations, one may wish to estimate other parameters of a population such as a population proportion (which we shall address in Chapter 9). The methods in this chapter can be extended to even more complex situations; for example, in evaluating a measurement technique, interest may focus on the repeatability of the technique, as indicated by the standard deviation of repeated determinations. As another example, in defining the limits of health, a medical researcher might want to estimate the 95th percentile of serum cholesterol levels in a certain population. Just as the precision of the mean can be indicated by a standard error or a confidence interval, statistical techniques are also available to specify the precision of estimation of parameters such as the population standard deviation or 95th percentile.

## Summary of Estimation Methods

For convenient reference, we summarize in the box the confidence interval methods presented in this chapter.

### Standard Error of the Mean

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

### Confidence Interval for $\mu$

$$95\% \text{ confidence interval: } \bar{y} \pm t_{0.025} \text{SE}_{\bar{Y}}$$

Critical value  $t_{0.025}$  from Student's  $t$  distribution with  $\text{df} = n - 1$ .

Intervals with other confidence levels (such as 90%, 99%, etc.) are constructed analogously (using  $t_{0.05}$ ,  $t_{0.005}$ , etc.).

The confidence interval formula is valid if (1) the data can be regarded as a random sample from a large population, (2) the observations are independent, and (3) the population is normal. If  $n$  is large then condition (3) is less important.

### Standard Error of $\bar{Y}_1 - \bar{Y}_2$

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\text{SE}_1^2 + \text{SE}_2^2}$$

### Confidence Interval for $\mu_1 - \mu_2$

95% confidence interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$$

Critical value  $t_{0.025}$  from Student's  $t$  distribution with

$$\text{df} = \frac{(\text{SE}_1^2 + \text{SE}_2^2)^2}{\text{SE}_1^4/(n_1 - 1) + \text{SE}_2^4/(n_2 - 1)}$$

where  $\text{SE}_1 = s_1/\sqrt{n_1}$  and  $\text{SE}_2 = s_2/\sqrt{n_2}$ .

Confidence intervals with other confidence levels (90%, 99%, etc.) are constructed analogously (using  $t_{0.05}$ ,  $t_{0.005}$ , etc.).

The confidence interval formula is valid if (1) the data can be regarded as coming from two independently chosen random samples, (2) the observations are independent within each sample, and (3) each of the populations is normally distributed. If  $n$  is large, condition (3) is less important.

## Supplementary Exercises 6.S.1–6.S.20

**6.S.1** To study the conversion of nitrite to nitrate in the blood, researchers injected four rabbits with a solution of radioactively labeled nitrite molecules. Ten minutes after injection, they measured for each rabbit the percentage of the nitrite that had been converted to nitrate. The results were as follows:<sup>48</sup>

51.1      55.4      48.0      49.5

- (a) For these data, calculate the mean, the standard deviation, and the standard error of the mean.
- (b) Construct a 95% confidence interval for the population mean percentage.
- (c) Without doing any calculations, would a 99% confidence interval be wider, narrower, or the same width as the confidence interval you found in part (b)? Why?

**6.S.2** The diameter of the stem of a wheat plant is an important trait because of its relationship to breakage of the stem, which interferes with harvesting the crop. An agronomist measured stem diameter in eight plants of the Tetrastichon cultivar of soft red winter wheat. All observations were made three weeks after flowering of the plant. The stem diameters (mm) were as follows:<sup>49</sup>

2.3    2.6    2.4    2.2    2.3    2.5    1.9    2.0

The mean of these data is 2.275 and the standard deviation is 0.238.

- (a) Calculate the standard error of the mean.
- (b) Construct a 95% confidence interval for the population mean.
- (c) Define in words the population mean that you estimated in part (b). (See Example 6.1.1.)

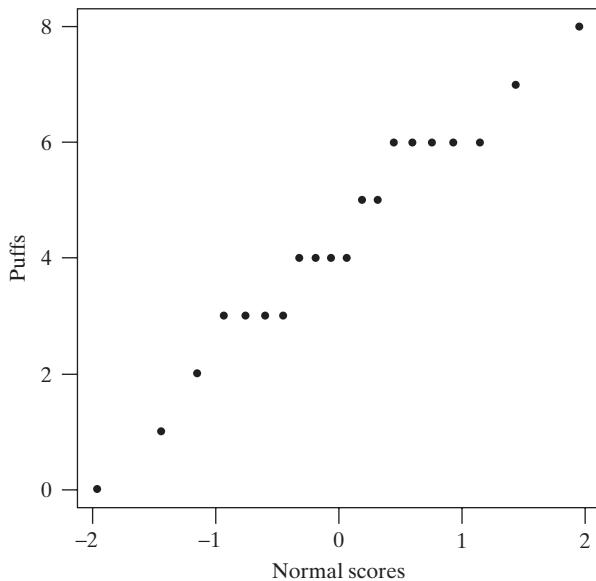
**6.S.3** Refer to Exercise 6.S.2.

- (a) What conditions are needed for the confidence interval to be valid?
- (b) Are these conditions met? How do you know?
- (c) Which of these conditions is most important?

**6.S.4** Refer to Exercise 6.S.2. Suppose that the data on the eight plants are regarded as a pilot study, and that the agronomist now wishes to design a new study for which he wants the standard error of the mean to be only 0.03 mm. How many plants should be measured in the new study?

**6.S.5** A sample of 20 fruitfly (*Drosophila melanogaster*) larva were incubated at 37 °C for 30 minutes. It is theorized that such exposure to heat causes polytene chromosomes located in the salivary glands of the fly to unwind, creating puffs on the chromosome arm that are visible under a microscope. The following normal probability

plot supports the use of a normal curve to model the distribution of puffs.<sup>50</sup>



The average number of puffs for the 20 observations was 4.30, with a standard deviation of 2.03.

- (a) Construct a 95% confidence interval for  $\mu$ .
- (b) In the context of this problem, describe what  $\mu$  represents. That is, the confidence interval from part (a) is a confidence interval for what quantity?
- (c) The normal probability plot shows the dots lining up on horizontal bands. Is this sort of behavior surprising for this type of data? Explain.

**6.S.6** Over a period of about nine months, 1,353 women reported the timing of each of their menstrual cycles. For the first cycle reported by each woman, the mean cycle time was 28.86 days, and the standard deviation of the 1,353 times was 4.24 days.<sup>51</sup>

- (a) Construct a 99% confidence interval for the population mean cycle time.
- (b) Because environmental rhythms can influence biological rhythms, one might hypothesize that the population mean menstrual cycle time is 29.5 days, the length of the lunar month. Is the confidence interval of part (a) consistent with this hypothesis?

**6.S.7** Refer to the menstrual cycle data of Exercise 6.S.6.

- (a) Over the entire time period of the study, the women reported a total of 12,247 cycles. When all of these cycles are included, the mean cycle time is 28.22 days. Explain why one would expect that this mean would be smaller than the value 28.86 given in Exercise 6.5.6. (*Hint:* If each woman reported for a fixed time

period, which women contributed more cycles to the total of 12,247 observations?)

- (b) Instead of using only the first reported cycle as in Exercise 6.5.6, one could use the first four cycles for each woman, thus obtaining  $1,353 \times 4 = 5,412$  observations. One could then calculate the mean and standard deviation of the 5,412 observations and divide the SD by  $\sqrt{5412}$  to obtain the SE; this would yield a much smaller value than the SE found in Exercise 6.51. Why would this approach not be valid?

**6.S.8** For the 28 lamb birthweights of Example 6.2.2, the mean is 5.1679 kg, the SD is 0.6544 kg, and the SE is 0.1237 kg.

- (a) Construct a 95% confidence interval for the population mean.
- (b) Construct a 99% confidence interval for the population mean.
- (c) Interpret the confidence interval you found in part (a). That is, explain what the numbers in the interval mean. (*Hint:* See Examples 6.3.4 and 6.3.5.)
- (d) Often researchers will summarize their data in reports and articles by writing  $\bar{y} \pm \text{SD}$  ( $5.17 \pm 0.65$ ) or  $\bar{y} \pm \text{SE}$  ( $5.17 \pm 0.12$ ). If the researcher of this study is planning to compare the mean birthweight of these Rambouillet lambs to another breed, Booroolas, which style of presentation should she use?

**6.S.9** Refer to Exercise 6.S.8.

- (a) What conditions are required for the validity of the confidence intervals?
- (b) Which of the conditions of part (a) can be checked (roughly) from the histogram of Figure 6.2.1?
- (c) Twin births were excluded from the lamb birthweight data. If twin births had been included, would the confidence intervals be valid? Why or why not?

**6.S.10** Researchers measured the number of tree species in each of 69 vegetational plots in the Lama Forest of Benin, West Africa.<sup>52</sup> The number of species ranged from a low of 1 to a high of 12. The sample mean was 6.8 and the sample SD was 2.4, which results in a 95% confidence interval of (6.2, 7.4). However, the number of tree species in a plot takes on only integer values. Does this mean that the confidence interval should be (7, 7)? Or does it mean that we should round off the endpoints of the confidence interval and report it as (6, 7)? Or should the confidence interval really be (6.2, 7.4)? Explain.

**6.S.11** As part of a study of natural variation in blood chemistry, serum potassium concentrations were measured in 84 healthy women. The mean concentration was 4.36 mEq/l, and the standard deviation was 0.42 mEq/l.

The table presents a frequency distribution of the data.<sup>53</sup>

SERUM POTASSIUM (mEq/l)	NUMBER OF WOMEN
[3.1, 3.4)	1
[3.4, 3.7)	2
[3.7, 4.0)	7
[4.0, 4.3)	22
[4.3, 4.6)	28
[4.6, 4.9)	16
[4.9, 5.2)	4
[5.2, 5.5)	3
[5.5, 5.8)	1
Total	84

- (a) Calculate the standard error of the mean.
- (b) Construct a histogram of the data and indicate the intervals  $\bar{y} \pm SD$  and  $\bar{y} \pm SE$  on the histogram. (See Figure 6.2.1.)
- (c) Construct a 95% confidence interval for the population mean.
- (d) Interpret the confidence interval you found in part (c). That is, explain what the numbers in the interval mean. (*Hint:* See Examples 6.3.4 and 6.3.5.)

**6.S.12** Refer to Exercise 6.S.11. In medical diagnosis, physicians often use “reference limits” for judging blood chemistry values; these are the limits within which we would expect to find 95% of healthy people. Would a 95% confidence interval for the mean be a reasonable choice of “reference limits” for serum potassium in women? Why or why not?

**6.S.13** Refer to Exercise 6.S.11. Suppose a similar study is to be conducted next year, to include serum potassium measurements on 200 healthy women. Based on the data in Exercise 6.S.11, what would you predict would be

- (a) the SD of the new measurements?
- (b) the SE of the new measurements?

**6.S.14** An agronomist selected six wheat plants at random from a plot, and then, for each plant, selected 12 seeds from the main portion of the wheat head; by weighing, drying, and reweighing, she determined the percent moisture in each batch of seeds. The results were as follows:<sup>54</sup>

62.7    63.6    60.9    63.0    62.7    63.7

- (a) Calculate the mean, the standard deviation, and the standard error of the mean.
- (b) Construct a 90% confidence interval for the population mean.

**6.S.15** As part of the National Health and Nutrition Examination Survey (NHANES), hemoglobin levels were checked for a sample of 1139 men age 70 and over.<sup>55</sup> The sample mean was 145.3 g/l and the standard deviation was 12.87 g/l.

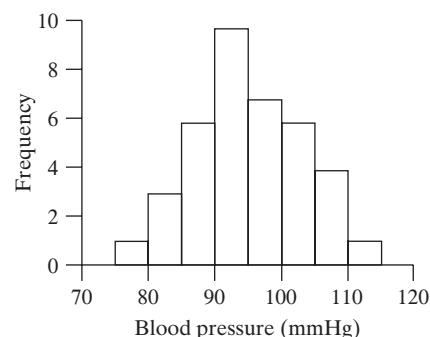
- (a) Use these data to construct a 95% confidence interval for  $\mu$ .
- (b) Does the confidence interval from part (a) give limits in which we expect 95% of the sample data to lie? Why or why not?
- (c) Does the confidence interval from part (a) give limits in which we expect 95% of the population to lie? Why or why not?

**6.S.16** The following data are 16 weeks of weekly fecal coliform counts (MPN/100 ml) at Dairy Creek in San Luis Obispo County, California.<sup>56</sup>

203	215	240	236	217	296	301	190
197	203	210	215	270	290	310	287

- (a) Counts above 225 MPN/100ml are considered unsafe. What type of one-sided interval (upper- or lower-bound) would be appropriate to assess the safety of this creek? Explain your reasoning.
- (b) Using 95% confidence, construct the interval chosen in part (a).
- (c) Based on your interval in part (b), what conclusions can you make regarding the safety of the water?

**6.S.17** The blood pressure (average of systolic and diastolic measurements) of each of 38 persons were measured.<sup>57</sup> The average was 94.5 (mm Hg). A histogram of the data is shown.



Which of the following is an approximate 95% confidence interval for the population mean blood pressure? Explain.

- (i)  $94.5 \pm 16$
- (ii)  $94.5 \pm 8$
- (iii)  $94.5 \pm 2.6$
- (iv)  $94.5 \pm 1.3$

**6.S.18** Suppose you wished to estimate the mean blood pressure of students at your school to within 2 mmHg with 95% confidence.

- (a) Using the data displayed in Exercise 6.S.17 as pilot data for your study, determine the (approximate) sample size necessary to achieve your goals. (*Hint:* You will need to use the graph to make some visual estimates).
- (b) Suppose your school is a small private college that only has 500 students. Would the interval based on your sample size be valid? Explain. Do you think it would be too wide or too narrow?

**6.S.19** It is known that alcohol consumption during pregnancy can harm the fetus. To study this phenomenon, 10 pregnant mice will receive a low dose of alcohol. When each mouse gives birth, the birthweight of each pup will be measured. Suppose the mice give birth to a total of 85 pups, so the experimenter has 85 observations of

$Y$  = birthweight. To calculate the standard error of the mean of these 85 observations, the experimenter could calculate the standard deviation of the 85 observations and divide by  $\sqrt{85}$ . On what grounds might an objection be raised to this method of calculating the SE?

**6.S.20** Is the nutrition information on commercially produced food accurate? In one study, researchers sampled 13 packages of a certain frozen reduced-calorie chicken entrée with a reported calorie content of 252 calories per package. The mean calorie count of the sampled entrées was 306 with a sample standard deviation of 51 calories.<sup>58</sup>

- (a) Compute a 95% confidence interval for the population mean calorie content of the frozen entrée.
- (b) Based on this interval computed in part (a), what do you think about the reported calorie content for this entrée?
- (c) Manufacturers are punished if they provide *less* food than advertised. How does this fact relate to your results in (a) and (b)?

# COMPARISON OF TWO INDEPENDENT SAMPLES

## Objectives

In this chapter we continue our study of comparisons of two independent samples by introducing hypothesis testing. We will

- explore how randomization can be used to form the basis of a statistical inference.
- demonstrate how to conduct a two-sample  $t$  test to compare sample means and explain how this test relates to the confidence interval for the difference of two means.
- discuss the interpretation of  $P$ -values.
- take a closer look at how confounding and spurious association can limit the utility of a study.
- compare causal versus associative inferences and their relationships to experiments and observational studies.
- discuss the concepts of significance level, effect size, Type I and II errors, and power.
- distinguish between directional and nondirectional tests and examine how the  $P$ -values of these tests compare.
- consider the conditions under which the use of a  $t$  test is valid.
- show how to compare distributions using the Wilcoxon-Mann-Whitney test.

## 7.1 Hypothesis Testing: The Randomization Test

Consider taking a sample from a population and then randomly dividing the sample into two parts. We would expect the two parts of the sample to look similar, but not exactly alike. Now suppose that we have samples from two populations. If the two samples look quite similar to each other, we might infer that the two populations are identical; if the samples look quite different, we would infer that the populations differ. The question is, “How different do two samples have to be in order for us to infer that the populations that generated them are actually different?”

One way to approach this question is to compare the two sample means and to see how much they differ in comparison to the amount of difference we would expect to see due to chance.\* The randomization test gives us a way to measure the variability in the difference of two sample means.

### Example 7.1.1

**Flexibility** A researcher studied the flexibility of each of seven women, four of whom were in an aerobics class and three of whom were dancers. One measure she recorded was the “trunk flexion”—how far forward each of the women could

\*One could compare the two sample medians rather than the means. We compare means so that we have a process similar to the  $t$  test, which is introduced in the next section and is based on means.

<b>Table 7.1.1</b>	
Aerobics	Dance
38	48
45	59
58	61
64	
mean 51.25	56.00

stretch while seated on the floor.\* The measures (in centimeters) are shown in Table 7.1.1.<sup>1</sup>

Do the data provide evidence that the flexibility is associated with being a dancer?

If being a dancer has no effect on flexibility, then one could argue that the seven data points in the study came from a common population: Some women have greater trunk flexion than others, but this has nothing to do with being a dancer.

Another way of saying this is

**Claim:** The seven trunk flexion measures came from a single population; the labels “aerobics” and “dance” are arbitrary and have nothing to do with flexibility (as measured by trunk flexion). ■

If the claim stated in Example 7.1.1 is true, then any rearrangement of the seven observations into two groups, with four “aerobics” and three “dance” women, is as likely as any other rearrangement. Indeed, we could imagine writing the seven observations onto seven cards, shuffling the cards, and then drawing four of them to be the observations for the “aerobics” group, with the other three being the observations for the “dance” group.

---

**Example  
7.1.2**

**Flexibility** There are 35 possible ways to divide the trunk flexion measures of the seven observations into two groups, of sizes 4 and 3. Table 7.1.2 lists each of the 35 possibilities, along with the difference in sample means for each. (We report the means to three decimal places, since we will be using these values in future calculations.) The two samples obtained in the study are listed first, followed by the other 34 ways that the samples might have turned out.

**Table 7.1.2**

Sample 1 (“aerobics”)	Sample 2 (“dance”)	Mean of sample 1	Mean of sample 2	Difference in means
38 45 58 64	48 59 61	51.25	56.00	<b>-4.75</b>
38 45 58 48	64 59 61	47.25	61.33	<b>-14.08</b>
38 45 58 59	64 48 61	50.00	57.67	<b>-7.67</b>
38 45 58 61	64 48 59	50.50	57.00	<b>-6.50</b>
38 45 64 48	58 59 61	48.75	59.33	<b>-10.58</b>
38 45 64 59	58 48 61	51.50	55.67	<b>-4.17</b>
38 45 64 61	58 48 59	52.00	55.00	<b>-3.00</b>
38 45 48 59	58 64 61	47.50	61.00	<b>-13.50</b>
38 45 48 61	58 64 59	48.00	60.33	<b>-12.33</b>
38 45 59 61	58 64 48	50.75	56.67	<b>-5.92</b>
38 58 64 48	45 59 61	52.00	55.00	-3.00
38 58 64 59	45 48 61	54.75	51.33	3.42
38 58 64 61	45 48 59	55.25	50.67	4.58
38 58 48 59	45 64 61	50.75	56.67	<b>-5.92</b>
38 58 48 61	45 64 59	51.25	56.00	<b>-4.75</b>

(Continues on next page)

---

\*These data are part of a larger study—we are working with a subset of the full study in order to simplify matters.

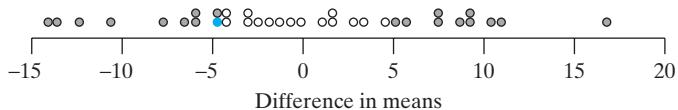
**Table 7.1.2 (Continued)**

Sample 1 ("aerobics")	Sample 2 ("dance")	Mean of sample 1	Mean of sample 2	Difference in means
38 58 59 61	45 64 48	54.00	52.33	1.67
38 64 48 59	45 58 61	52.25	54.67	-2.42
38 64 48 61	45 58 59	52.75	54.00	-1.25
38 64 59 61	45 58 48	55.50	50.33	<b>5.17</b>
38 48 59 61	45 58 64	51.50	55.67	-4.17
45 58 64 48	38 59 61	53.75	52.67	1.08
45 58 64 59	38 48 61	56.50	49.00	<b>7.50</b>
45 58 64 61	38 48 59	57.00	48.33	<b>8.67</b>
45 58 48 59	38 64 61	52.50	54.33	-1.83
45 58 48 61	38 64 59	53.00	53.67	-0.67
45 58 59 61	38 64 48	55.75	50.00	<b>5.75</b>
45 64 48 59	38 58 61	54.00	52.33	1.67
45 64 48 61	38 58 59	54.50	51.67	2.83
45 64 59 61	38 58 48	57.25	48.00	<b>9.25</b>
45 48 59 61	38 58 64	53.25	53.33	-0.08
58 64 48 59	38 45 61	57.25	48.00	<b>9.25</b>
58 64 48 61	38 45 59	57.75	47.33	<b>10.42</b>
58 64 59 61	38 45 48	60.50	43.67	<b>16.83</b>
58 48 59 61	38 45 64	56.50	49.00	<b>7.50</b>
64 48 59 61	38 45 58	58.00	47.00	<b>11.00</b>

Figure 7.1.1 gives a visual display of these 35 possible values. The observed result of -4.75, which is highlighted, falls not far from the middle of the distribution.

Suppose that the labels “aerobics” and “dance” are, in fact, arbitrary and have nothing to do with trunk flexion. Then each of the 35 outcomes listed in Table 7.1.2, and shown in Figure 7.1.1, is equally likely. This means that the differences, shown in the last column of the table, are equally likely. Of the 35 differences, 20 of them are at least as large in magnitude as the -4.75 obtained in the study; these are shown in bold type in the table and filled in black or gray in the figure. Thus, if the claim is true (that the labels “aerobics” and “dance” are arbitrary), there is a 20/35 chance of obtaining a difference in sample means as large, in magnitude, as the difference that was observed.

The fraction 20/35 is approximately equal to 0.57, which is rather large. Thus, the observed data are consistent with the claim that the labels “aerobics” and “dance” are arbitrary and have nothing to do with flexibility. If the claim is true, we would expect to see a difference in sample means of 4.75 or more over half of the time, just due to chance alone. Therefore, this data provides little evidence that flexibility is associated with dancing. ■



**Figure 7.1.1** Distribution of “Difference in means” values, with the observed result of -4.75 colored blue, and values with observed results as or more extreme (in magnitude) than 4.75 colored gray

The process shown in Example 7.1.2 is called the **randomization test**.<sup>\*</sup> In a randomization test one randomly divides the observed data into groups in order to see how likely it is that the observed difference is to arise due to chance alone.

**Note:** In Section 7.2 we will introduce a procedure known as the *t* test, which often provides a good approximation to the randomization test. The value of 20/35 (0.57) computed in Example 7.1.2 is called a *P*-value. (We have seen this term used earlier for the decision making in the context of the Shapiro–Wilk test for normality in Section 4.4. The general use of this term, and others, will be explained more fully in Section 7.2.) For the data in Example 7.1.1 the *t* test yields a *P*-value of 0.54. We can think of the 0.54 *P*-value from the *t* test as an approximation to the 0.57 *P*-value found with the randomization test.

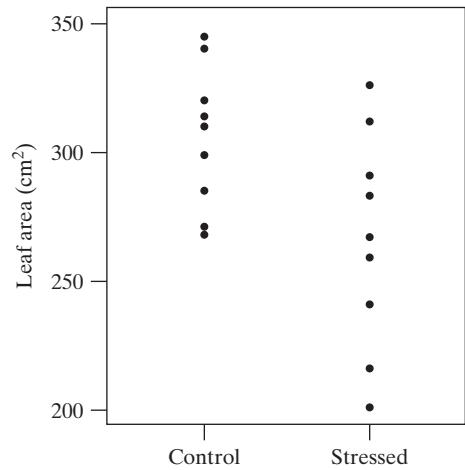
## Larger Samples

When we are dealing with small samples, such as in Example 7.1.1, we can list all of the possible outcomes from randomly assigning observations to groups. The following example shows how to handle large samples, where no such listing is possible.

### Example 7.1.3

**Leaf Area** A plant physiologist investigated the effect of mechanical stress on the growth of soybean plants. Individually potted seedlings were divided into two groups. Those in the first group were stressed by shaking for 20 minutes twice daily, while those in the second group (the control group) were not shaken. After 16 days of growth the plants were harvested and total leaf area ( $\text{cm}^2$ ) was measured for each plant. The data are given in Table 7.1.3 and are graphed in Figure 7.1.2.<sup>2</sup>

<b>Table 7.1.3</b>	
Control	Stressed
314	283
320	312
310	291
340	259
299	216
268	201
345	267
271	326
285	241
mean 305.8	266.2



**Figure 7.1.2** Parallel dotplots of leaf areas

The mean for the stressed plants is lower than for the control plants and Figure 7.1.2 provides some visual evidence of a difference between the two groups. On the other hand, the dotplots overlap quite a bit. Perhaps stressing the seedlings by shaking them has no actual effect on leaf area and the difference observed in this experiment ( $305.8 - 266.2 = 39.6$ ) was simply due to chance. That is, it might be

\*Many people would call this a permutation test, since it involves listing all possible permutations of the data.

that the “control” and “stressed” conditions have nothing to do with leaf area. If this is the case, then we can think of the 18 seedlings as having come from one population, with the division into “control” and “stressed” groups being arbitrary.

In Example 7.1.2 we could list all of the possible ways that the two groups could have been formed. However, in the current example there are 48,620 possible ways to select 9 of the 18 seedlings as the control group (and the other 9 as the stressed group). Thus, it is not feasible to create a table similar to Table 7.1.2 and list all the possibilities. What we can do, however, is to randomly sample from the 48,620 possibilities. One way to do this would be to (1) write the 18 observations on each of 18 cards; (2) shuffle the cards; (3) randomly deal out 9 of them as the control group, with the other 9 being the stress group; (4) calculate the difference in sample means; (5) record whether the magnitude of the difference in sample means is at least 39.6; (6) repeat steps (1)–(5) many times.

Consider the fraction of times that the magnitude of the difference in sample means is at least as large as the value of 39.6 obtained in the experiment. This is a measure of the evidence against the claim that “Stressing the seedlings by shaking them has no actual effect on leaf area.”

Rather than use 18 cards, we could use a computer simulation to accomplish the same thing. In one simulation with 1,000 trials there were only 36 trials that gave a difference in sample means as large in magnitude as 39.6.\* This indicates that the observed difference of 39.6 is unlikely to arise by chance—the chance is only 3.6%—so we have evidence that stressing the plants has an effect. Indeed, it appears that shaking the seedlings led to a reduction in average leaf area. ■

**Note:** The  $t$  test procedure (to be introduced in Section 7.2) yields a  $P$ -value of 0.033, which is a good approximation to the 0.036  $P$ -value given by the randomization test.

## Exercises 7.1.1–7.1.3

---

**7.1.1** Suppose we have samples of five men and of five women and have conducted a randomization test to compare the sexes on the variable  $Y = \text{pulse}$ . Further, suppose we have found that in 120 out of the 252 possible outcomes under randomization the difference in means is at least as large as the difference in the two observed sample means. Does the randomization test provide evidence that the sexes differ with regard to pulse? Justify your answer using the randomization results.

**7.1.2** In an investigation of the possible influence of dietary chromium on diabetic symptoms, some rats were fed a low-chromium diet and others were fed a normal diet. One response variable was activity of the liver enzyme GITH, which was measured using a radioactively labeled molecule. The accompanying table shows the

results, expressed as thousands of counts per minute per gram of liver.<sup>3</sup> The sample means are 49.17 for the low-chromium diet and 51.90 for the normal diet; thus the difference in sample means is  $-2.73$ . There are 10 possible randomizations of the five observations into two groups, of sizes three and two.

- Create a list of these 10 randomizations (one of which is the original assignment of observations to the two groups) and for each case calculate the low-chromium diet mean minus the normal diet mean.
- How many of the 10 randomizations yield a difference in sample means as far from zero as  $-2.73$ , the difference in sample means for our observed samples?

---

\*In this instance, we could also use a computer to consider the difference in means for each of the 48,620 possibilities and note how many of these yield differences larger than 39.6 in magnitude. However, as samples grow larger, listing all possibilities can be computationally expensive (even with fast computers) and only marginally more accurate than conducting simulations as we have described.

- (c) Is there evidence that dietary chromium affects GITH liver enzyme activity? Justify your answer using the randomization results.

LOW-CHROMIUM DIET	NORMAL DIET
42.3	53.1
51.5	50.7
53.7	

**7.1.3** The following table shows the number of bacteria colonies present in each of several petri dishes, after *E. coli* bacteria were added to the dishes and they were incubated for 24 hours. The “soap” dishes contained a solution prepared from ordinary soap; the “control” dishes contained a solution of sterile water. (These data are a subset of the larger data set seen in Exercise 6.6.9.) The sample means are 44 for the control group and 39.7 for the soap group; thus the difference in sample means is

4.3, with the control mean being larger, as would be expected if the soap were effective. There are 20 possible randomizations of the six observations into two groups, each of size three.

- (a) Create a list of these 20 randomizations (one of which is the original assignment of observations to the two groups) and for each case calculate the control mean minus the soap mean.
- (b) How many of the 20 randomizations produce a difference in means at least as large as 4.3?
- (c) Is there evidence that the soap inhibits *E. coli* growth? Justify your answer using the randomization results.

CONTROL	SOAP
30	76
36	27
66	16

## 7.2 Hypothesis Testing: The *t* Test

In Chapter 6 we saw that two means can be compared by using a confidence interval for the difference ( $\mu_1 - \mu_2$ ). Now we will explore another approach to the comparison of means: the procedure known as *hypothesis testing*. The general idea is to formulate as a hypothesis the statement that  $\mu_1$  and  $\mu_2$  differ and then to see whether the data provide sufficient evidence in support of that hypothesis.

### The Null and Alternative Hypotheses

The hypothesis that  $\mu_1$  and  $\mu_2$  are *not* equal is called an **alternative hypothesis** (or a research hypothesis) and is abbreviated  $H_A$ . It can be written as

$$H_A: \mu_1 \neq \mu_2$$

Its antithesis is the **null hypothesis**,

$$H_0: \mu_1 = \mu_2$$

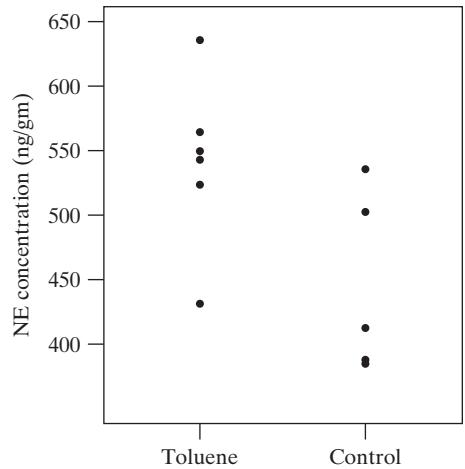
which asserts that  $\mu_1$  and  $\mu_2$  are equal. A researcher would usually express these hypotheses more informally, as in the following example.

#### Example 7.2.1

**Toluene and the Brain** Abuse of substances containing toluene (for example, glue) can produce various neurological symptoms. In an investigation of the mechanism of these toxic effects, researchers measured the concentrations of various chemicals in the brains of rats that had been exposed to a toluene-laden atmosphere, and also in unexposed control rats. The concentrations of the brain chemical norepinephrine (NE) in the medulla region of the brain, for six toluene-exposed rats and five control rats, are given in Table 7.2.1 and displayed in Figure 7.2.1.<sup>4</sup>

The observed mean NE in the toluene group ( $\bar{y}_1 = 540.8$  ng/gm) is substantially higher than the mean in the control group ( $\bar{y}_2 = 444.2$  ng/gm). One might ask whether this observed difference indicates a real biological phenomenon—the effect of toluene—or whether the truth might be that toluene has no effect and that

<b>Table 7.2.1</b> NE concentration (ng/gm)	
Toluene (Group 1)	Control (Group 2)
543	535
523	385
431	502
635	412
564	387
549	
<i>n</i>	6
$\bar{y}$	540.8
<i>s</i>	66.1
SE	27
	5
	444.2
	69.6
	31



**Figure 7.2.1** Parallel dotplots of NE concentration

the observed difference between  $\bar{y}_1$  and  $\bar{y}_2$  reflects only chance variation. Corresponding hypotheses, informally stated, would be

$H_0^*$ : Toluene has no effect on NE concentration in rat medulla.

$H_A^*$ : Toluene has some effect on NE concentration in rat medulla. ■

We denote the informal statements by different symbols ( $H_0^*$  and  $H_A^*$  rather than  $H_0$  and  $H_A$ ) because they make different assertions. In Example 7.2.1 the informal alternative hypothesis makes a very strong claim—not only that there is a difference, but that the difference is *caused* by toluene.\*

A statistical **test of hypothesis** is a procedure for assessing the strength of evidence present in the data in support of  $H_A$ . The data are considered to demonstrate evidence for  $H_A$  if any discrepancies from  $H_0$  (the opposite of  $H_A$ ) could not be readily attributed to chance (that is, to sampling error).

## The *t* Statistic

We consider the problem of testing the null hypothesis

$$H_0: \mu_1 = \mu_2$$

against the alternative hypothesis

$$H_A: \mu_1 \neq \mu_2$$

Note that the null hypothesis says that the two population means are equal, which is the same as saying that the difference between them is zero:

$$H_0: \mu_1 = \mu_2 \leftrightarrow H_0: \mu_1 - \mu_2 = 0$$

The alternative hypothesis asserts that the difference is not zero:

$$H_A: \mu_1 \neq \mu_2 \leftrightarrow H_A: \mu_1 - \mu_2 \neq 0$$

The ***t* test** is a standard method of choosing between these two hypotheses. To carry out the *t* test, the first step is to compute the **test statistic**, which for a *t* test is defined as

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\text{SE}(\bar{Y}_1 - \bar{Y}_2)}$$

---

\*Of course, our statements of  $H_0^*$  and  $H_A^*$  are abbreviated. Complete statements would include all relevant conditions of the experiment—adult male rats, toluene 1,000 ppm atmosphere for 8 hours, and so on. Our use of abbreviated statements should not cause any confusion.

Note that we subtract zero from  $\bar{y}_1 - \bar{y}_2$  because  $H_0$  states that  $\mu_1 - \mu_2$  equals zero; writing “ $(\bar{y}_1 - \bar{y}_2) - 0$ ” reminds us of what we are testing. The subscript “*s*” on  $t_s$  serves as a reminder that this value is calculated from the data (“*s*” for “sample”). The quantity  $t_s$  is the test statistic for the *t* test; that is,  $t_s$  provides the data summary that is the basis for the test procedure. Notice the structure of  $t_s$ : It is a measure of how far the difference between the sample means ( $\bar{y}$ 's) is from the difference we would expect to see if  $H_0$  were true (zero difference), expressed in relation to the SE of the difference—the amount of variation we expect to see in differences of means from random samples. We illustrate with an example.

**Example 7.2.2**

**Toluene and the Brain** For the brain NE data of Example 7.2.1, the SE for  $(\bar{Y}_1 - \bar{Y}_2)$  is

$$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{66.1^2}{6} + \frac{69.6^2}{5}} = 41.195$$

and the value of  $t_s$  is

$$t_s = \frac{(540.8 - 444.2) - 0}{41.195} = 2.34$$

The *t* statistic shows that the difference between  $\bar{y}_1$  and  $\bar{y}_2$  is about 2.3 SEs from zero, the difference we'd expect to see if toluene had no effect on NE. ■

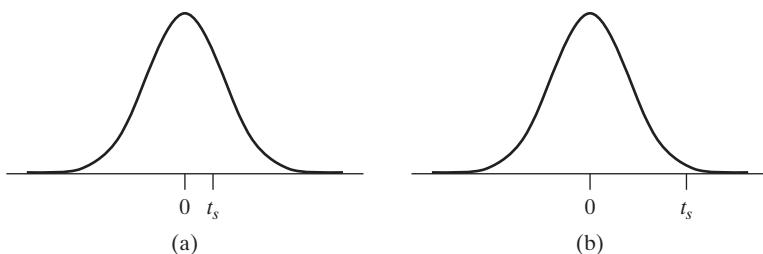
How shall we judge whether our data are sufficient evidence for  $H_A$ ? A complete lack of evidence (*perfect* agreement with  $H_0$ ) would be expressed by sample means that were identical and a resulting *t* statistic equal to zero ( $t_s = 0$ ). But, even if the null hypothesis  $H_0$  were true, we would not expect  $t_s$  to be exactly zero; we expect the sample means to differ from one another because of sampling variability (measured via  $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$ ). Fortunately, we know what to expect regarding this sampling variability; in fact, the chance difference in the  $\bar{Y}$ 's is not likely to exceed a couple of standard errors when the null hypothesis is true. To put this more precisely, it can be shown mathematically that

If  $H_0$  is true, then the sampling distribution of  $t_s$  is well approximated by a Student's *t* distribution with degrees of freedom given by formula (6.7.1).\*

The preceding statement is true if certain conditions are met. Briefly: We require independent random samples from normally distributed populations. These conditions will be considered in detail in Section 7.9.

The essence of the *t* test procedure is to identify where the observed value  $t_s$  falls in the Student's *t* distribution, as indicated in Figure 7.2.2. If  $t_s$  is near the center, as in Figure 7.2.2(a), then the data are regarded as compatible with  $H_0$  because the observed difference between  $(\bar{Y}_1 - \bar{Y}_2)$  and the null difference of zero can readily be attributed to chance variation caused by sampling error. ( $H_0$  predicts that the sample means will be equal, since  $H_0$  says that the population means are equal.)

**Figure 7.2.2** Essence of the *t* test. (a) Data compatible with  $H_0$  (and thus a lack of significant evidence for  $H_A$ ); (b) data incompatible with  $H_0$  (and thus significant evidence for  $H_A$ ).



\*As we stated in Section 6.8, a conservative approximation to formula (6.7.1) is to use degrees of freedom given by the smaller of  $n_1 - 1$  and  $n_2 - 1$ .

If, on the other hand,  $t_s$  falls in the far tail of the  $t$  distribution, as in Figure 7.2.2(b), then the data are regarded as evidence for  $H_A$ , because the observed deviation cannot be readily explained as being due to chance variation. To put this another way, if  $H_0$  is true, then it is unlikely that  $t_s$  would fall in the far tails of the  $t$  distribution.

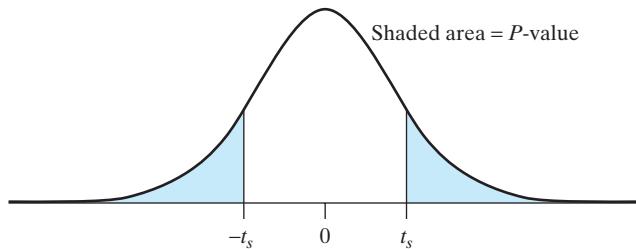
### The $P$ -Value

To judge whether an observed value  $t_s$  is “far” in the tail of the  $t$  distribution, we need a quantitative yardstick for locating  $t_s$  within the distribution. This yardstick is provided by the  $P$ -value, which can be defined (in the present context) as follows:

The  **$P$ -value** of the test is the area under Student’s  $t$  curve in the double tails beyond  $-t_s$  and  $+t_s$ .

Thus, the  $P$ -value, which is sometimes abbreviated as simply “ $P$ ,” is the shaded area in Figure 7.2.3. Note that we have defined the  $P$ -value as the total area in *both* tails; this is sometimes called the “two-tailed”  $P$ -value.

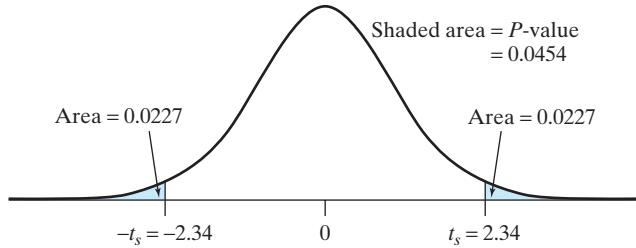
**Figure 7.2.3** The two-tailed  $P$ -value for the  $t$  test



#### Example 7.2.3

**Toluene and the Brain** For the brain NE data of Example 7.2.1, the value of  $t_s$  is 2.34. We can ask, “If  $H_0$  were true so that one would expect  $\bar{Y}_1 - \bar{Y}_2 = 0$ , on average, what is the probability that  $\bar{Y}_1 - \bar{Y}_2$  would differ from zero by as many as 2.34 SEs?” The  $P$ -value answers this question. Formula (6.7.1) yields 8.47 degrees of freedom for these data. Thus, the  $P$ -value is the area under the  $t$  curve (with 8.47 degrees of freedom) beyond  $\pm 2.34$ . This area, which was found using a computer, is shown in Figure 7.2.4 to be 0.0454. ■

**Figure 7.2.4** The two-tailed  $P$ -value for the toluene data



**Definition** The  **$P$ -value** for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme as the value of the test statistic that was actually obtained.

From the definition of  $P$ -value, it follows that the  **$P$ -value is a measure of compatibility between the data and  $H_0$**  and thus measures the **evidence for  $H_A$** : A large  $P$ -value (close to 1) indicates a value of  $t_s$  near the center of the  $t$  distribution (lack of evidence for  $H_A$ ), whereas a small  $P$ -value (close to 0) indicates a value of  $t_s$  in the far tails of the  $t$  distribution (evidence for  $H_A$ ).

## Drawing Conclusions from a *t* Test

The *P*-value is a measure of the evidence in the data for  $H_A$ , but where does one draw the line in determining how much evidence is sufficient? Most people would agree that  $P$ -value = 0.0001 indicates very strong evidence, and that  $P$ -value = 0.80 indicates a lack of evidence, but what about intermediate values? For example, should  $P$ -value = 0.10 be regarded as sufficient evidence for  $H_A$ ? The answer is not intuitively obvious.

In much scientific research, it is not necessary to draw a sharp line. However, in many situations a *decision* must be reached. For example, the Food and Drug Administration (FDA) must decide whether the data submitted by a pharmaceutical manufacturer are sufficient to justify approval of a medication. As another example, a fertilizer manufacturer must decide whether the evidence favoring a new fertilizer is sufficient to justify the expense of further research.

Making a decision requires drawing a definite line between sufficient and insufficient evidence. The threshold value, on the *P*-value scale, is called the **significance level** of the test and is denoted by the Greek letter  $\alpha$  (alpha). The value of  $\alpha$  is chosen by whoever is making the decision. Common choices are  $\alpha = 0.10, 0.05$ , and  $0.01$ . *If the P-value of the data is less than or equal to  $\alpha$ , the data are judged to provide statistically significant evidence in favor of  $H_A$ ; we also may say that  $H_0$  is rejected.* If the *P*-value of the data is greater than  $\alpha$ , we say that the data provide insufficient evidence to claim that  $H_A$  is true, and thus  $H_0$  is **not rejected**.

The following example illustrates the use of the *t* test to make a decision.

### Example 7.2.4

**Toluene and the Brain** For the brain NE experiment of Example 7.2.1, the data are summarized in Table 7.2.2. Suppose we choose to make a decision at the 5% significance level,  $\alpha = 0.05$ . In Example 7.2.3 we found that the *P*-value of these data is 0.0454. This means that one of two things happened: Either (1)  $H_0$  is true and we got a strange set of data just by chance or (2)  $H_0$  is false. If  $H_0$  is true, the kind of discrepancy we observed between  $\bar{y}_1$  and  $\bar{y}_2$  would happen only about 4.5% of the time. Because the *P*-value, 0.0454, is less than 0.05, we reject  $H_0$  and conclude that the data provide statistically significant evidence in favor of  $H_A$ . The strength of the evidence is expressed by the statement that the *P*-value is 0.0454.

<b>Table 7.2.2</b> NE concentration (ng/gm)		
	Toluene	Control
$n$	6	5
$\bar{y}$	540.8	444.2
$s$	66.1	69.6

**Conclusion:** The data provide sufficient evidence at the 0.05 level of significance (*P*-value = 0.0454) that toluene increases NE concentration.\* ■

The next example illustrates a *t* test in which there is a lack of sufficient evidence at the 0.05 level of significance for  $H_A$ .

### Example 7.2.5

**Fast Plants** In Example 6.7.1 we saw that the mean height of fast plants was smaller when a nancy was used than when water (the control) was used. Table 7.2.3 summarizes

\*Because the alternative hypothesis was  $H_A: \mu_1 \neq \mu_2$ , some authors would say, "We conclude that toluene affects NE concentration," rather than saying that toluene increases NE concentration.

**Table 7.2.3** Fourteen-day height of control and of ancy plants

	Control	Ancy
$n$	8	7
$\bar{y}$	15.9	11.0
$s$	4.8	4.7

the data. The difference between the sample means is  $15.9 - 11.0 = 4.9$ . The SE for the difference is

$$\text{SE}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{4.8^2}{8} + \frac{4.7^2}{7}} = 2.46$$

Suppose we choose to use  $\alpha = 0.05$  in testing

$$H_0: \mu_1 = \mu_2 \text{ (i.e., } \mu_1 - \mu_2 = 0\text{)}$$

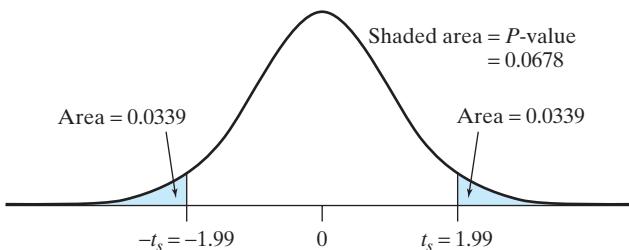
against the alternative hypothesis

$$H_A: \mu_1 \neq \mu_2 \text{ (i.e., } \mu_1 - \mu_2 \neq 0\text{)}$$

The value of the test statistic is

$$t_s = \frac{(15.9 - 11.0) - 0}{2.46} = 1.99$$

Formula (6.7.1) gives 12.8 degrees of freedom for the  $t$  distribution. The  $P$ -value for the test is the probability of getting a  $t$  statistic that is at least as far away from zero as 1.99. Figure 7.2.5 shows that this probability is 0.0678. (This 4-digit  $P$ -value was found using a computer.) Because the  $P$ -value is greater than  $\alpha$ , we have insufficient evidence for  $H_A$ ; thus, we do not reject  $H_0$ . That is, these data do not provide sufficient evidence to conclude that  $\mu_1$  and  $\mu_2$  differ; the difference we observed between  $\bar{y}_1$  and  $\bar{y}_2$  could easily have happened by chance.

**Figure 7.2.5** The two-sided  $P$ -value for the ancy data

**Conclusion:** The data do *not* provide sufficient evidence ( $P$ -value = 0.0678) at the 0.05 level of significance to conclude that ancy and water differ in their effects on fast plant growth (under the conditions of the experiment that was conducted). ■

Note carefully the phrasing of the conclusion in Example 7.2.5. We do *not* say that there is evidence *for* the null hypothesis, but only that there is insufficient evidence *against* it. When we do not reject  $H_0$ , this indicates a lack of evidence that  $H_0$  is false, which is *not* the same thing as evidence that  $H_0$  is true. The astronomer Carl Sagan (in another context) summed up this principle of evidence in this succinct statement:<sup>5</sup>

Absence of evidence is not evidence of absence.

In other words, nonrejection of  $H_0$  is *not* the same as *acceptance* of  $H_0$ . (To avoid confusion, it may be best not to use the phrase “accept  $H_0$ ” at all.)

Nonrejection of  $H_0$  indicates that the data are compatible with  $H_0$ , but the data may *also* be quite compatible with  $H_A$ . For instance, in Example 7.2.5 we found that the observed difference between the sample means could be due to sampling variation, but this finding does not rule out the possibility that the observed difference is actually due to a real effect caused by ancy. (Methods for such ruling out of possible alternatives will be discussed in Section 7.7 and optional Section 7.8.)

In testing a hypothesis, the researcher starts out with the assumption that  $H_0$  is true and then asks whether the data contradict that assumption. This logic can make sense even if the researcher regards the null hypothesis as implausible. For instance, in Example 7.2.5 it could be argued that there is almost certainly *some* difference (perhaps very small) between using ancy and not using ancy. The fact that we did not reject  $H_0$  does not mean that we accept  $H_0$ .

## Using Tables versus Using Technology

In analyzing data, how do we determine the *P*-value of a test? Statistical computer software, and some calculators, will provide exact *P*-values. If such technology is not available, then we can use formula (6.7.1) to find the degrees of freedom but round down to make the value an integer. A conservative alternative to using formula (6.7.1) is to use the smaller of  $n_1 - 1$  and  $n_2 - 1$  as the degrees of freedom for the test. A liberal approach is to use  $n_1 + n_2 - 2$  as the degrees of freedom. (Formula (6.7.1) will always give degrees of freedom between the conservative value of the smaller of  $n_1 - 1$  and  $n_2 - 1$  and the liberal value of  $n_1 + n_2 - 2$ .) We can rely on the limited information in Table 4 to *bracket* the *P*-value, rather than to determine it exactly. The *P*-value found using the conservative approach will be somewhat larger than the exact *P*-value; the *P*-value found using the liberal approach will be somewhat smaller than the exact *P*-value. The following example illustrates the bracketing process.

### Example 7.2.6

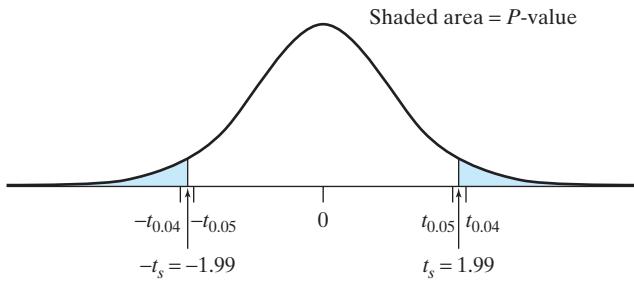
**Fast Plants** For the fast plant growth data, the value of the *t* statistic (as determined in Example 7.2.5) is  $t_s = 1.99$ . The smaller of  $n_1 - 1$  and  $n_2 - 1$  is  $7 - 1 = 6$ , so the conservative degrees of freedom are 6. The liberal degrees of freedom are  $8 + 7 - 2 = 13$ . Here is a copy of part of Table 4, with key numbers highlighted.

Upper Tail Probability			
df	.05	.04	.03
6	<b>1.943</b>	<b>2.104</b>	2.313
7	1.895	2.046	2.241
8	1.860	2.004	2.189
9	1.833	1.973	2.150
10	1.812	1.948	2.120
11	1.796	1.928	2.096
12	1.782	1.912	2.076
13	1.771	<b>1.899</b>	<b>2.060</b>

We begin with the conservative degrees of freedom, 6. From the preceding table (or from Table 4) we find  $t_{6,0.05} = 1.943$  and  $t_{6,0.04} = 2.104$ . The corresponding conservative *P*-value, based on a *t* distribution with 6 degrees of freedom, is shaded in

Figure 7.2.6. Because  $t_s$  is between the 0.04 and 0.05 critical values, the upper tail area must be between 0.04 and 0.05; thus, the conservative  $P$ -value must be between 0.08 and 0.10.

**Figure 7.2.6** Conservative  $P$ -value for Example 7.2.6



The liberal degrees of freedom are  $8 + 7 - 2 = 13$ . From the preceding table (or from Table 4) we find  $t_{13, 0.04} = 1.899$  and  $t_{13, 0.03} = 2.060$ . Because  $t_s$  is between these 0.03 and 0.04 critical values, the upper tail area must be between 0.06 and 0.08; thus, the liberal  $P$ -value must be between 0.06 and 0.08.

Putting these two together, we have

$$0.06 < P\text{-value} < 0.10 \quad \blacksquare$$

If the observed  $t_s$  is not within the boundaries of Table 4, then the  $P$ -value is bracketed on only one side. For example, if  $t_s$  is greater than  $t_{0.0005}$ , then the two-sided  $P$ -value is bracketed as

$$P\text{-value} < 0.001$$

## Reporting the Results of a $t$ Test

In reporting the results of a  $t$  test, a researcher may choose to make a definite decision (to claim there is significant evidence for  $H_A$  or not significant evidence to support  $H_A$ ) at a specified significance level  $\alpha$ , or the researcher may choose simply to describe the results in phrases such as “There is very strong evidence that . . .” or “The evidence suggests that . . .” or “There is virtually no evidence that . . .”. In writing a report for publication, it is very desirable to state the  $P$ -value so that the reader can make a decision on his or her own.

The term *significant* is often used in reporting results. For instance, an observed difference is said to be “statistically significant at the 5% level” if it is large enough to justify significant evidence for  $H_A$  at  $\alpha = 0.05$ . In Example 7.2.4 we saw that the observed difference between the two sample means in the toluene data is statistically significant at the 5% level, since the  $P$ -value is 0.0454, which is less than 0.05. In contrast, the fast plant data of Example 7.2.5 do not show a statistically significant difference at the 5% level, since the  $P$ -value for the fast plant data is 0.0678. However, the difference in sample means in the fast plant data is statistically significant at the  $\alpha = 0.10$  level, since the  $P$ -value is less than 0.10. When  $\alpha$  is not specified, it is usually understood to be 0.05; we should emphasize, however, that  $\alpha$  is an arbitrarily chosen value and there is nothing “official” about 0.05. Unfortunately, the term “significant” is easily misunderstood and should be used with care; we will return to this point in Section 7.7.

**Note:** In this section we have considered tests of the form  $H_0: \mu_1 = \mu_2$  (i.e.,  $\mu_1 - \mu_2 = 0$ ) versus  $H_A: \mu_1 \neq \mu_2$  (i.e.,  $\mu_1 - \mu_2 \neq 0$ ); this is the most common pair of hypotheses. However, it may be that we wish to test that  $\mu_1$  is greater than  $\mu_2$ .

by some specific, nonzero amount, say  $c$ . To test  $H_0: \mu_1 - \mu_2 = c$  versus  $H_A: \mu_1 - \mu_2 \neq c$  we use the *t* test with test statistic given by

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - c}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

From this point on, the test proceeds as before (i.e., as for the case when  $c = 0$ ).

## Exercises 7.2.1–7.2.17

[Note: Answers to hypothesis testing questions should include a statement of the conclusion in the context of the setting. (See Examples 7.2.4 and 7.2.5.)]

**7.2.1** For each of the following data sets, use Table 4 to bracket the two-tailed *P*-value of the data as analyzed by the *t* test.

(a)	SAMPLE 1	SAMPLE 2
$n$	4	3
$\bar{y}$	735	854
$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 38$ with $\text{df} = 4$		

(b)	SAMPLE 1	SAMPLE 2
$n$	7	7
$\bar{y}$	5.3	5.0
$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 0.24$ with $\text{df} = 12$		

(c)	SAMPLE 1	SAMPLE 2
$n$	15	20
$\bar{y}$	36	30
$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 1.3$ with $\text{df} = 30$		

**7.2.2** For each of the following data sets, use Table 4 to bracket the two-tailed *P*-value of the data as analyzed by the *t* test.

(a)	SAMPLE 1	SAMPLE 2
$n$	8	5
$\bar{y}$	100.2	106.8
$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 5.7$ with $\text{df} = 10$		

(b)	SAMPLE 1	SAMPLE 2
$n$	8	8
$\bar{y}$	49.8	44.3
$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 1.9$ with $\text{df} = 13$		

(c)

	SAMPLE 1	SAMPLE 2
$n$	10	15
$\bar{y}$	3.58	3.00
$\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = 0.12$ with $\text{df} = 19$		

**7.2.3** For each of the following situations, suppose  $H_0: \mu_1 = \mu_2$  is being tested against  $H_A: \mu_1 \neq \mu_2$ . State whether or not there is significant evidence for  $H_A$ .

- (a)  $P$ -value = 0.085,  $\alpha = 0.10$ .
- (b)  $P$ -value = 0.065,  $\alpha = 0.05$ .
- (c)  $t_s = 3.75$  with 19 degrees of freedom,  $\alpha = 0.01$ .
- (d)  $t_s = 1.85$  with 12 degrees of freedom,  $\alpha = 0.05$ .

**7.2.4** For each of the following situations, suppose  $H_0: \mu_1 = \mu_2$  is being tested against  $H_A: \mu_1 \neq \mu_2$ . State whether or not there is significant evidence for  $H_A$ .

- (a)  $P$ -value = 0.046,  $\alpha = 0.02$ .
- (b)  $P$ -value = 0.033,  $\alpha = 0.05$ .
- (c)  $t_s = 2.26$  with 5 degrees of freedom,  $\alpha = 0.10$ .
- (d)  $t_s = 1.94$  with 16 degrees of freedom,  $\alpha = 0.05$ .

**7.2.5** In a study of the nutritional requirements of cattle, researchers measured the weight gains of cows during a 78-day period. For two breeds of cows, Hereford (HH) and Brown Swiss/Hereford (SH), the results are summarized in the following table.<sup>6</sup> [Note: Formula (6.7.1) yields 71.9 df.]

	HH	SH
$n$	33	51
$\bar{y}$	18.3	13.9
$s$	17.8	19.1

Use a *t* test to compare the means. Use  $\alpha = 0.10$ .

**7.2.6** Backfat thickness is a variable used in evaluating the meat quality of pigs. An animal scientist measured backfat thickness (cm) in pigs raised on two different diets, with the results given in the table.<sup>7</sup>

	DIET 1	DIET 2
$\bar{y}$	3.49	3.05
$s$	0.40	0.40

Consider using the  $t$  test to compare the diets. Bracket the  $P$ -value, assuming that the number of pigs on each diet was

- (a) 5
- (b) 10
- (c) 15

Use  $n_1 + n_2 - 2$  as the approximate degrees of freedom.

**7.2.7** Heart disease patients often experience spasms of the coronary arteries. Because biological amines may play a role in these spasms, a research team measured amine levels in coronary arteries that were obtained postmortem from patients who had died of heart disease and also from a control group of patients who had died from other causes. The accompanying table summarizes the concentration of the amine serotonin.<sup>8</sup>

SEROTONIN (NG/GM)	
HEART DISEASE	CONTROLS
$n$	8
$\bar{y}$	3,840
SE	850

- (a) For these data, the SE of  $(\bar{Y}_1 - \bar{Y}_2)$  is 1,064 and  $df = 14.3$  (which can be rounded to 14). Use a  $t$  test to compare the means at the 5% significance level.  
 (b) Verify the value of  $SE_{(\bar{Y}_1 - \bar{Y}_2)}$  given in part (a).

**7.2.8** In a study of the periodical cicada (*Magicicada septendecim*), researchers measured the hind tibia lengths of the shed skins of 110 individuals. Results for males and females are shown in the accompanying table.<sup>9</sup>

GROUP	TIBIA LENGTH (μm)		
	$n$	MEAN	SD
Males	60	78.42	2.87
Females	50	80.44	3.52

- (a) Use a  $t$  test to investigate the association of tibia length on gender in this species. Use the 5% significance level. [Note: Formula (6.7.1) yields 94.3 df.]  
 (b) Given the preceding data, if you were told the tibia length of an individual of this species, could you make a fairly confident prediction of its sex? Why or why not?  
 (c) Repeat the  $t$  test of part (a), assuming that the means and standard deviations were as given in the table, but that they were based on only one-tenth as many individuals (6 males and 5 females). [Note: Formula (6.7.1) yields 7.8 df.]

**7.2.9** Myocardial blood flow (MBF) was measured for two groups of subjects after five minutes of bicycle exercise. The normoxia (“normal oxygen”) group was provided normal air to breathe whereas the hypoxia group was provided with a gas mixture with reduced oxygen, to simulate high altitude. The results (ml/min/g) are shown in the table.<sup>10</sup> [Note: Formula (6.7.1) yields 12.2 df.]

	NORMOXIA	HYPOXIA
	3.45	6.37
	3.09	5.69
	3.09	5.58
	2.65	5.27
	2.49	5.11
	2.33	4.88
	2.28	4.68
	2.24	3.50
	2.17	
	1.34	
	$n$	8
	$\bar{y}$	2.51
	$s$	0.60

Use a  $t$  test to investigate the effect of hypoxia on MBF. Use  $\alpha = 0.05$ .

**7.2.10** In a study of the development of the thymus gland, researchers weighed the glands of 10 chick embryos. Five of the embryos had been incubated 14 days and 5 had been incubated 15 days. The thymus weights were as shown in the table.<sup>11</sup> [Note: Formula (6.7.1) yields 7.7 df.]

THYMUS WEIGHT (MG)	
14 DAYS	15 DAYS
29.6	32.7
21.5	40.3
28.0	23.7
34.6	25.2
44.9	24.2
$n$	5
$\bar{y}$	31.72
$s$	8.73

- (a) Use a  $t$  test to compare the means at  $\alpha = 0.10$ .  
 (b) Note that the chicks that were incubated longer had a smaller mean thymus weight. Is this “backward” result surprising, or could it easily be attributed to chance? Explain.

**7.2.11** As part of an experiment on root metabolism, a plant physiologist grew birch tree seedlings in the greenhouse. He flooded four seedlings with water for one day and kept four others as controls. He then harvested the seedlings and analyzed the roots for ATP content. The results (nmol ATP per mg tissue) are shown in the table.<sup>12</sup> [Note: Formula (6.7.1) yields 5.6 df.]

FLOODED	CONTROL
1.45	1.70
1.19	2.04
1.05	1.49
1.07	1.91
<i>n</i>	4
$\bar{y}$	1.190
<i>s</i>	0.184
	0.241

Use a *t* test to investigate the effect of flooding. Use  $\alpha = 0.05$ .

**7.2.12** After surgery a patient's blood volume is often depleted. In one study, the total circulating volume of blood plasma was measured for each patient immediately after surgery. After infusion of a "plasma expander" into the bloodstream, the plasma volume was measured again and the increase in plasma volume (ml) was calculated. Two of the plasma expanders used were albumin (25 patients) and polygelatin (14 patients). The accompanying table reports the increase in plasma volume.<sup>13</sup> [Note: Formula (6.7.1) yields 33.6 df.]

Use a *t* test to compare the mean increase in plasma volume under the two treatments. Let  $\alpha = 0.01$ .

	ALBUMIN	POLYGELATIN
<i>n</i>	25	14
mean increase	490	240
SE	60	30

**7.2.13** Nutritional researchers conducted an investigation of two high-fiber diets intended to reduce serum cholesterol level. Twenty men with high serum cholesterol were randomly allocated to receive an "oat" diet or a "bean" diet for 21 days. The table summarizes the fall (before minus after) in serum cholesterol levels.<sup>14</sup> Use a *t* test to compare the diets at the 5% significance level. [Note: Formula (6.7.1) yields 17.9 df.]

DIET	FALL IN CHOLESTEROL (MG/DL)		
	<i>n</i>	MEAN	SD
Oat	10	53.6	31.1
Bean	10	55.5	29.4

**7.2.14** Suppose we have conducted a *t* test, with  $\alpha = 0.05$ , and the *P*-value is 0.03. For each of the following statements, say whether the statement is true or false and explain why.

- (a) We reject  $H_0$  with  $\alpha = 0.05$ .
- (b) We have significant evidence for  $H_A$  with  $\alpha = 0.05$ .
- (c) We would reject  $H_0$  if  $\alpha$  were 0.10.
- (d) We do not have significant evidence for  $H_A$  with  $\alpha = 0.10$ .
- (e) If  $H_0$  is true, the probability of getting a test statistic at least as extreme as the value of the  $t_s$  that was actually obtained is 3%.
- (f) There is a 3% probability that  $H_0$  is true.

**7.2.15** Suppose we have conducted a *t* test, with  $\alpha = 0.10$ , and the *P*-value is 0.07. For each of the following statements, say whether the statement is true or false and explain why.

- (a) We reject  $H_0$  with  $\alpha = 0.10$ .
- (b) We have significant evidence for  $H_A$  with  $\alpha = 0.10$ .
- (c) We would reject  $H_0$  if  $\alpha$  were 0.05.
- (d) We do not have significant evidence for  $H_A$  with  $\alpha = 0.05$ .
- (e) The probability that  $\bar{Y}_1$  is greater than  $\bar{Y}_2$  is 0.07.

**7.2.16** The following table shows the number of bacteria colonies present in each of several petri dishes, after *E. coli* bacteria were added to the dishes and they were incubated for 24 hours. The "soap" dishes contained a solution prepared from ordinary soap; the "control" dishes contained a solution of sterile water. (These data were seen in Exercise 6.6.9.)

	CONTROL	SOAP
<i>n</i>	30	76
	36	27
	66	16
	21	30
	63	26
	38	46
	35	6
	45	
<i>n</i>	8	7
$\bar{y}$	41.8	32.4
<i>s</i>	15.6	22.8
SE	5.5	8.6

Use a *t* test to investigate whether soap affects the number of bacteria colonies that form. Use  $\alpha = 0.10$ . [Note: Formula (6.7.1) yields 10.4 degrees of freedom for these data.]

**7.2.17** Researchers studied the effect of a houseplant fertilizer on radish sprout growth. They randomly selected some radish seeds to serve as controls, while others were planted in aluminum planters to which fertilizer sticks were added. Other conditions were held constant between the two groups. The following table shows data on the heights of plants (in cm) two weeks after germination.<sup>15</sup>

Use a  $t$  test to investigate whether the fertilizer has an effect on average radish sprout growth. Use  $\alpha = 0.05$ . [Note: Formula (6.7.1) yields 53.5 degrees of freedom for these data.]

CONTROL		FERTILIZED	
3.4	1.6	2.8	1.9
4.4	2.9	1.9	2.7
3.5	2.3	3.6	2.3
2.9	2.8	1.2	1.8
2.7	2.5	2.4	2.7
2.6	2.3	2.2	2.6
3.7	1.6	3.6	1.3
2.7	1.6	1.2	3.0
2.3	3.0	0.9	1.4
2.0	2.3	1.5	1.2
1.8	3.2	2.4	2.6
2.3	2.0	1.7	1.8
2.4	2.6	1.4	1.7
2.5	2.4	1.8	1.5
$n$	28	28	
$\bar{y}$	2.58	2.04	
$s$	0.65	0.72	

## 7.3 Further Discussion of the $t$ Test

In this section we discuss more fully the method and interpretation of the  $t$  test.

### Relationship between Test and Confidence Interval

There is a close connection between the confidence interval approach and the hypothesis testing approach to the comparison of  $\mu_1$  and  $\mu_2$ . Consider, for example, a 95% confidence interval for  $(\mu_1 - \mu_2)$  and its relationship to the  $t$  test at the 5% significance level. The  $t$  test and the confidence interval use the same three quantities— $(\bar{Y}_1 - \bar{Y}_2)$ ,  $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$  and  $t_{0.025}$ —but manipulate them in different ways.

In the  $t$  test, when  $\alpha = 0.05$ , we have significant evidence for  $H_A$  (and so we reject  $H_0$ ) if the  $P$ -value is less than or equal to 0.05. This happens if and only if the test statistic,  $t_s$ , is in the tail of the  $t$  distribution, at or beyond  $\pm t_{0.025}$ . If the magnitude of  $t_s$  (symbolized as  $|t_s|$ ) is greater than or equal to  $t_{0.025}$ , then the  $P$ -value is less than or equal to 0.05 and we have significant evidence for  $H_A$ ; if  $|t_s|$  is less than  $t_{0.025}$ , then the  $P$ -value is greater than 0.05 and we do *not* have significant evidence for  $H_A$ . Figure 7.3.1 shows this relationship.

Thus, we lack significant evidence for  $H_A$ :  $\mu_1 - \mu_2 \neq 0$  if and only if  $|t_s| < t_{0.025}$ . That is, we lack significant evidence for  $H_A$  when

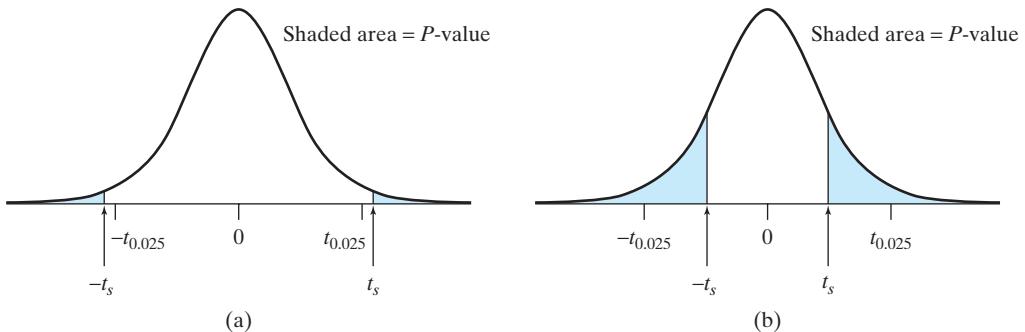
$$\frac{|\bar{y}_1 - \bar{y}_2|}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}} < t_{0.025}$$

This is equivalent to

$$|\bar{y}_1 - \bar{y}_2| < t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$$

or

$$-t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} < (\bar{y}_1 - \bar{y}_2) < t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$$



**Figure 7.3.1** Possible outcomes of the  $t$  test at  $\alpha = 0.05$ . (a) If  $|t_s| \geq t_{0.025}$  then  $P\text{-value} \leq 0.05$  and there is significant evidence for  $H_A$  (so  $H_0$  is rejected). (b) If  $|t_s| < t_{0.025}$ , then  $P\text{-value} > 0.05$  and there is a lack of significant evidence for  $H_A$ .

which is equivalent to

$$-(\bar{y}_1 - \bar{y}_2) - t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} < 0 < -(\bar{y}_1 - \bar{y}_2) + t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$$

or

$$(\bar{y}_1 - \bar{y}_2) + t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} > 0 > (\bar{y}_1 - \bar{y}_2) - t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$$

or

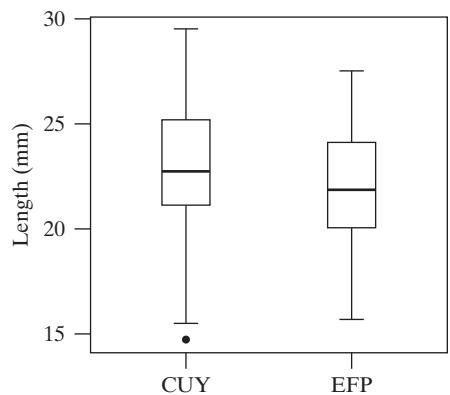
$$(\bar{y}_1 - \bar{y}_2) - t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} < 0 < (\bar{y}_1 - \bar{y}_2) + t_{0.025} \text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$$

Thus, we have shown that we lack significant evidence for  $H_A: \mu_1 - \mu_2 \neq 0$  if and only if the confidence interval for  $(\mu_1 - \mu_2)$  includes zero. Conversely, if the 95% confidence interval for  $(\mu_1 - \mu_2)$  does not cover zero, then we have significant evidence for  $H_A: \mu_1 - \mu_2 \neq 0$  when  $\alpha = 0.05$ . (The same relationship holds between the 90% confidence interval and the test at  $\alpha = 0.10$ , and so on.) We illustrate with an example.

### Example 7.3.1

**Crawfish Lengths** Biologists took samples of the crawfish species *Orconectes sanborii* from two rivers in central Ohio, the Upper Cuyahoga River (CUY) and East Fork of Pine Creek (EFP), and measured the length (mm) of each crawfish captured.<sup>16</sup> Table 7.3.1 shows the summary statistics; Figure 7.3.2 shows parallel boxplots of the data. The EFP sample distribution is shifted down from the CUY distribution; both distributions are reasonably symmetric.

<b>Table 7.3.1</b> Crawfish data: length (mm)		
	CUY	EFP
$n$	30	30
$\bar{y}$	22.91	21.97
$s$	3.78	2.90



**Figure 7.3.2** Boxplots of the crawfish data

For these data the two SEs are  $3.78/\sqrt{30} = 0.69$  and  $2.90/\sqrt{30} = 0.53$  for CUY and EFP, respectively. The degrees of freedom are

$$\text{df} = \frac{(0.69^2 + 0.53^2)^2}{0.69^4/30 + 0.53^4/30} = 56.3$$

The quantities needed for a  $t$  test with  $\alpha = 0.05$  are

$$\bar{y}_1 - \bar{y}_2 = 22.91 - 21.97 = 0.94$$

and

$$SE_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{0.69^2 + 0.53^2} = 0.87$$

The test statistic is

$$t_s = \frac{(22.91 - 21.97) - 0}{0.87} = \frac{0.94}{0.87} = 1.08$$

The  $P$ -value for this test (found using a computer) is 0.2850, which is greater than 0.05, so we do not reject  $H_0$ . (A quick look at Table 4, using  $df = 50$ , shows that the  $P$ -value is between 0.20 and 0.40.)

If we construct a 95% confidence interval for  $(\mu_1 - \mu_2)$  we get

$$0.94 \pm 2.006 \times 0.87$$

or  $(-2.68, 0.81)$ .\*

The confidence interval includes zero, which is consistent with not having significant evidence for  $H_A: \mu_1 - \mu_2 \neq 0$  in the  $t$  test. Note that this equivalence between the test and the confidence interval makes common sense; according to the confidence interval,  $\mu_1$  may be as much as 2.68 less, or as much as 0.81 more, than  $\mu_2$ ; it is natural, then, to say that we are uncertain as to whether  $\mu_1$  is greater than (or less than, or equal to)  $\mu_2$ . ■

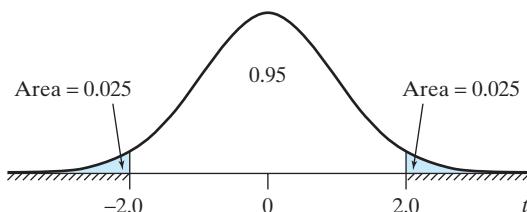
In the context of the Student's  $t$  method, the confidence interval approach and hypothesis testing approach are different ways of using the same basic information. The confidence interval has the advantage that it indicates the magnitude of the difference between  $\mu_1$  and  $\mu_2$ . The testing approach has the advantage that the  $P$ -value describes on a continuous scale the strength of the evidence that  $\mu_1$  and  $\mu_2$  are really different. In Section 7.7 we will explore further the use of a confidence interval to supplement the interpretation of a  $t$  test. In later chapters we will encounter other hypothesis tests that cannot so readily be supplemented by a confidence interval.

### Interpretation of $\alpha$

In analyzing data or making a decision based on data, you will often need to choose a significance level  $\alpha$ . How do you know whether to choose  $\alpha = 0.05$  or  $\alpha = 0.01$  or some other value? To make this judgment, it is helpful to have an *operational* interpretation of  $\alpha$ . We now give such an interpretation.

Recall from Section 7.2 that the sampling distribution of  $t_s$ , if  $H_0$  is true, is a Student's  $t$  distribution. Let us assume for definiteness that  $df = 60$  and that  $\alpha$  is chosen equal to 0.05. The critical value (from Table 4) is  $t_{0.025} = 2.00$ . Figure 7.3.3

**Figure 7.3.3** A  $t$  test at  $\alpha = 0.05$ . There is significant evidence for  $H_A$  if  $t_s$  falls in the hatched region

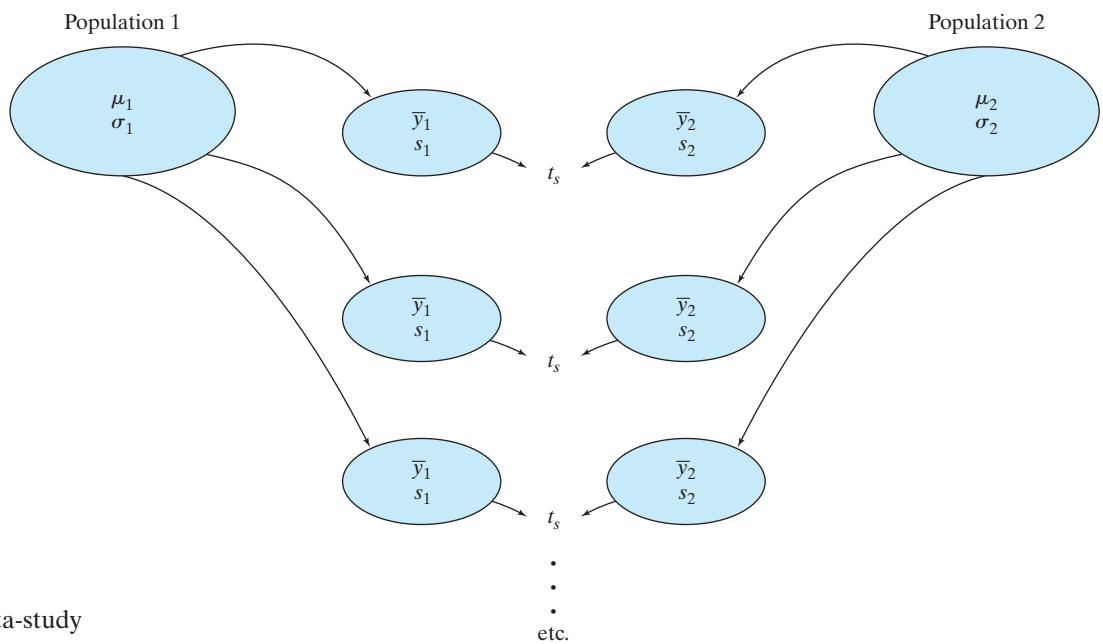


\*The value of  $t_{0.025} = 2.006$  is based on 56.3 degrees of freedom. If we were to use 50 degrees of freedom (i.e., if we had to rely on Table 4, rather than a computer) the  $t$  multiplier would be 2.009. This makes almost no difference in the resulting confidence interval.

shows the Student's *t* distribution and the values  $\pm 2.000$ . The total shaded area in the figure is 0.05; it is split into two equal parts of area 0.025 each. We can think of Figure 7.3.3 as a formal guide for deciding whether the evidence is strong enough to significantly support  $H_A$ : If the observed value of  $t_s$  falls in the hatched regions of the  $t_s$  axis, then there is significant evidence for  $H_A$ . But the chance of this happening is 5%, if  $H_0$  is true. Thus, we can say that

$$\Pr\{\text{data provide significant evidence for } H_A\} = 0.05 \text{ if } H_0 \text{ is true}$$

This probability has meaning in the context of a meta-study (depicted in Figure 7.3.4) in which we repeatedly sample from two populations and calculate a value of  $t_s$ . It is important to realize that the probability refers to a situation in which  $H_0$  is true. In order to concretely picture such a situation, you are invited to suspend disbelief for a moment and come on an imaginary trip in Example 7.3.2.



**Figure 7.3.4** Meta-study for the *t* test

### Example 7.3.2

**Music and Marigolds\*** Imagine that the scientific community has developed great interest in the influence of music on the growth of marigolds. One school of investigation centers on whether music written by Bach or Mozart produces taller plants. Plants are randomly allocated to listen to Bach (treatment 1) or Mozart (treatment 2) and, after a suitable period of listening, data are collected on plant height. The null hypothesis is

$$H_0: \text{Marigolds respond equally well to Bach or Mozart.}$$

or

$$H_0: \mu_1 = \mu_2$$

where

$$\mu_1 = \text{Mean height of marigolds if exposed to Bach}$$

$$\mu_2 = \text{Mean height of marigolds if exposed to Mozart}$$

---

\*This example is intentionally fanciful.

Assume for the sake of argument that  $H_0$  is in fact true. Imagine now that many investigators perform the Bach versus Mozart experiment, and that each experiment results in data with 60 degrees of freedom. Suppose each investigator analyzes his or her data with a  $t$  test at  $\alpha = 0.05$ . What conclusions will the investigators reach? In the meta-study of Figure 7.3.4, suppose each pair of samples represents a different investigator. Since we are assuming that  $\mu_1$  and  $\mu_2$  are actually equal, the values of  $t_s$  will deviate from 0 only because of chance sampling error. If all the investigators were to get together and make a frequency distribution of their  $t_s$  values, that distribution would follow a Student's  $t$  curve with 60 degrees of freedom. The investigators would make their decisions as indicated by Figure 7.3.3, so we would expect them to have the following experiences:

- 95% of them would (correctly) not find significant evidence for  $H_A$ ;
- 2.5% of them would find significant evidence for  $H_A$  and conclude (incorrectly) that the plants prefer Bach.
- 2.5% of them would find significant evidence for  $H_A$  and conclude (incorrectly) that the plants prefer Mozart.

Thus, a total of 5% of the investigators would find significant evidence for the alternative hypothesis. ■

Example 7.3.2 provides an image for interpreting  $\alpha$ . Of course, in analyzing data, we are not dealing with a meta-study but rather with a single experiment. When we perform a  $t$  test at the 5% significance level, we are playing the role of one of the investigators in Example 7.3.2, and the others are imaginary. If we find significant evidence for  $H_A$ , there are two possibilities:

1.  $H_A$  is in fact true; or
2.  $H_0$  is in fact true, but we are one of the unlucky 5% who obtained data that provided significant evidence for  $H_A$  anyway. In this case, we can think of the significant evidence for  $H_A$  as “setting off a false alarm.”

We feel “confident” in claiming our evidence for  $H_A$  is significant because the second possibility is unlikely (assuming that we regard 5% as a small percentage). Of course, we never know (unless someone replicates the experiment) whether or not we are one of the unlucky 5%.

**Significance Level versus  $P$ -Value** Students sometimes find it hard to distinguish between significance level ( $\alpha$ ) and  $P$ -value.\* For the  $t$  test, both  $\alpha$  and the  $P$ -value are tail areas under Student's  $t$  curve. But  $\alpha$  is an arbitrary prespecified value; it can be (and should be) chosen before looking at the data. By contrast, the  $P$ -value is determined from the data; indeed, giving the  $P$ -value is a way of describing the data. You may find it helpful at this point to compare Figure 7.2.3 with Figure 7.3.3. The shaded area represents  $P$ -value in the former and  $\alpha$  in the latter figure.

## Type I and Type II Errors

We have seen that  $\alpha$  can be interpreted as a probability:

$$\alpha = \Pr\{\text{finding significant evidence for } H_A\} \text{ if } H_0 \text{ is true}$$

---

\*Unfortunately, the term “significance level” is not used consistently by all people who write about statistics. A few authors use the terms “significance level” or “significance probability” where we have used “ $P$ -value.”

Claiming that data provide evidence that significantly supports  $H_A$  when  $H_0$  is true is called a **Type I error**. In choosing  $\alpha$ , we are choosing our level of protection against Type I error. Many researchers regard 5% as an acceptably small risk. If we do not regard 5% as small enough, we might choose to use a more conservative value of  $\alpha$  such as  $\alpha = 0.01$ ; in this case the percentage of true null hypotheses that we reject would be not 5% but 1%.

In practice, the choice of  $\alpha$  may depend on the context of the particular experiment. For example, a regulatory agency might demand more exacting proof of efficacy for a toxic drug than for a relatively innocuous one. Also, a person's choice of  $\alpha$  may be influenced by his or her prior opinion about the phenomenon under study. For instance, suppose an agronomist is skeptical of claims for a certain soil treatment; in evaluating a new study of the treatment, he might express his skepticism by choosing a very conservative significance level (say,  $\alpha = 0.001$ ), thus indicating that it would take a lot of evidence to convince him that the treatment is effective. For this reason, written reports of an investigation should include a  $P$ -value, so that each reader is free to choose his or her own value of  $\alpha$  in evaluating the reported results.

If  $H_A$  is true, but we do not observe sufficient evidence to support  $H_A$ , then we have made a **Type II error**. Table 7.3.2 displays the situations in which Type I and Type II errors can occur. For example, if we find significant evidence for  $H_A$ , then we eliminate the possibility of a Type II error, but by rejecting  $H_0$  we may have made a Type I error.

**Table 7.3.2** Possible outcomes of testing  $H_0$

		True situation	
		$H_0$ true	$H_A$ true
OUR DECISION	Lack of significant evidence for $H_A$	Correct	Type II error
	Significant evidence for $H_A$	Type I error	Correct

The consequences of Type I and Type II errors can be very different. The following two examples show some of the variety of these consequences.

### Example 7.3.3

**Marijuana and the Pituitary** Cannabinoids, which are substances contained in marijuana, can be transmitted from mother to young through the placenta and through the milk. Suppose we conduct the following experiment on pregnant mice: We give one group of mice a dose of cannabinoids and keep another group as controls. We then evaluate the function of the pituitary gland in the offspring. The hypotheses would be

$$H_0: \text{Cannabinoids do not affect pituitary of offspring.}$$

$$H_A: \text{Cannabinoids do affect pituitary of offspring.}$$

If in fact cannabinoids do not affect the pituitary of the offspring, but we conclude that our data provide significant evidence for  $H_A$ , we would be making a Type I error; the consequence might be unnecessary alarm if the conclusion were made public. On the other hand, if cannabinoids do affect the pituitary of the offspring, but our  $t$  test results in a lack of significant evidence for  $H_A$ , this would be a Type II error; one consequence might be unjustifiable complacency on the part of marijuana-smoking mothers. ■

**Example  
7.3.4**

**Immunotherapy** Chemotherapy is standard treatment for a certain cancer. Suppose we conduct a clinical trial to study the efficacy of supplementing the chemotherapy with immunotherapy (stimulation of the immune system). Patients are given either chemotherapy or chemotherapy plus immunotherapy. The hypotheses would be

$H_0$ : Immunotherapy is not effective in enhancing survival.

$H_A$ : Immunotherapy does affect survival.

If immunotherapy is actually not effective, but we conclude that our data provide significant evidence for  $H_A$  and thus conclude that immunotherapy is effective, then we have made a Type I error. The consequence, if this conclusion is acted on by the medical community, might be the widespread use of unpleasant, dangerous, and worthless immunotherapy. If, on the other hand, immunotherapy is actually effective, but our data do not enable us to detect that fact (perhaps because our sample sizes are too small), then we have made a Type II error, with consequences quite different from those of a Type I error: The standard treatment will continue to be used until someone provides convincing evidence that supplementary immunotherapy is effective. If we still “believe” in immunotherapy, we can conduct another trial (perhaps with larger samples) to try again to establish its effectiveness. ■

As the foregoing examples illustrate, the consequences of a Type I error are usually quite different from those of a Type II error. The likelihoods of the two types of error may be very different, also. The significance level  $\alpha$  is the probability of obtaining significant evidence for  $H_A$  if  $H_0$  is true. Because  $\alpha$  is chosen at will, the hypothesis testing procedure “protects” you against Type I error by giving you control over the risk of such an error. This control is independent of the sample size and other factors. The chance of a Type II error, by contrast, depends on many factors, and may be large or small. In particular, an experiment with small sample sizes often has a high risk of Type II error.

We are now in a position to reexamine Carl Sagan’s aphorism that “Absence of evidence is not evidence of absence.” Because the risk of Type I error is controlled and that of Type II error is not, our state of knowledge is much stronger after rejection of a null hypothesis than after nonrejection. For example, suppose we are testing whether a certain soil additive is effective in increasing the yield of field corn. If we find significant evidence for  $H_A$  and claim the additive is effective, then either (1) we are right; or (2) we have made a Type I error. Since the risk of a Type I error is controlled, we can be relatively confident of our conclusion that the additive is effective (although not necessarily very effective). Suppose, on the other hand, that the data are such that there is a lack of evidence for the additive’s effectiveness—we do not have evidence for  $H_A$ . Then either (1) we are right (that is,  $H_0$  is true), or (2) we have made a Type II error. Since the risk of a Type II error may be quite high, we cannot say confidently that the additive is ineffective. In order to justify a claim that the additive is ineffective, we would need to supplement our test of hypothesis with further analysis, such as a confidence interval or an analysis of the chance of Type II error. We will consider this in more detail in Sections 7.6 and 7.7.

## Power

As we have seen, Type II error is an important concept. The probability of making a Type II error is denoted by  $\beta$ :

$$\beta = \Pr\{\text{lack of significant evidence for } H_A\} \text{ if } H_A \text{ is true}$$

The chance of not making a Type II error when  $H_A$  is true—that is, the chance of having significant evidence for  $H_A$  when  $H_A$  is true—is called the **power** of a statistical test:

$$\text{Power} = 1 - \beta = \Pr\{\text{significant evidence for } H_A\} \text{ if } H_A \text{ is true}$$

Thus, the power of a  $t$  test is a measure of the sensitivity of the test, or the ability of the test procedure to detect a difference between  $\mu_1$  and  $\mu_2$  when such a difference really *does* exist. In this way the power is analogous to the resolving power of a microscope.

The power of a statistical test depends on many factors in an investigation, including the sample sizes, the inherent variability of the observations, and the magnitude of the difference between  $\mu_1$  and  $\mu_2$ . All other things being equal, using larger samples gives more information and thereby increases power. In addition, we will see that some statistical tests can be more powerful than others, and that some study designs can be more powerful than others.

The planning of a scientific investigation should always take power into consideration. No one wants to emerge from lengthy and perhaps expensive labor in the lab or the field, only to discover upon analyzing the data that the sample sizes were insufficient or the experimental material too variable, so that experimental effects that were considered important were not detected. Two techniques are available to aid the researcher in planning for adequate sample sizes. One technique is to decide how small each standard error ought to be and choose  $n$  using an analysis such as that of Section 6.4. A second technique is a quantitative analysis of the power of the statistical test. Such an analysis for the  $t$  test is discussed in Section 7.7.

## Exercises 7.3.1–7.3.8

---

**7.3.1 (Sampling exercise)** Refer to the collection of 100 ellipses shown with Exercise 3.1.1, which can be thought of as representing a natural population of the organism *C. ellipticus*. Use random digits (from Table 1 or your calculator) to choose two random samples of five ellipses each. Use a metric ruler to measure the body length of each ellipse; measurements to the nearest millimeter will be adequate.

- (a) Compare the means of your two samples, using a  $t$  test at  $\alpha = 0.05$ .
- (b) Did the analysis of part (a) lead you to a Type I error, a Type II error, or no error?

**7.3.2 (Sampling exercise)** Simulate choosing random samples from two different populations, as follows. First, proceed as in Exercise 7.3.1 to choose two random samples of five ellipses each and measure their lengths. Then add 6 mm to *each* measurement in one of the samples.

- (a) Compare the means of your two samples, using a  $t$  test at  $\alpha = 0.05$ .
- (b) Did the analysis of part (a) lead you to a Type I error, a Type II error, or no error?

**7.3.3 (Sampling exercise)** Prepare simulated data as follows. First, proceed as in Exercise 7.3.1 to choose two random samples of five ellipses each and measure their lengths. Then, toss a coin. If the coin falls heads, add

6 mm to *each* measurement in one of the samples. If the coin falls tails, do not modify either sample.

- (a) Prepare two copies of the simulated data. On the Student Copy, show the data only; on the Instructor Copy, indicate also which sample (if any) was modified.
- (b) Give your Instructor Copy to the instructor and trade your Student Copy with another student when you are told to do so.
- (c) After you have received another student's paper, compare the means of his or her two samples using a two-tailed  $t$  test at  $\alpha = 0.05$ . If you reject  $H_0$ , decide which sample was modified.

**7.3.4** Suppose a new drug is being considered for approval by the Food and Drug Administration. The null hypothesis is that the drug is not effective. If the FDA approves the drug, what type of error, Type I or Type II, could not possibly have been made?

**7.3.5** In Example 7.3.1, the null hypothesis was not rejected. What type of error, Type I or Type II, might have been made in that  $t$  test?

**7.3.6** Suppose that a 95% confidence interval for  $(\mu_1 - \mu_2)$  is calculated to be (1.4, 6.7). If we test  $H_0: \mu_1 - \mu_2 = 0$  versus  $H_A: \mu_1 - \mu_2 \neq 0$  using  $\alpha = 0.05$ , will we reject  $H_0$ ? Why or why not?

**7.3.7** Suppose that a 95% confidence interval for  $(\mu_1 - \mu_2)$  is calculated to be  $(-7.4, -2.3)$ . If we test  $H_0: \mu_1 = \mu_2$  versus  $H_A: \mu_1 \neq \mu_2$  using  $\alpha = 0.10$ , will we reject  $H_0$ ? Why or why not?

**7.3.8** A dairy researcher has developed a new technique for culturing cheese that is purported to age cheese in substantially less time than traditional methods without affecting other properties of the cheese. Retrofitting cheese manufacturing plants with this new technology will initially cost millions of dollars, but if it indeed reduces aging time—even marginally—it will lead to higher company profits in the long run. If, on the other hand, the new method is no better than the old, the retro-

fit would be a financial mistake. Before making the decision to retrofit, an experiment will be performed to compare culture times of the new and old methods.

- In plain English, what are the null and alternative hypotheses for this experiment?
- In the context of the problem, what would be the consequence of a Type I error?
- In the context of the problem, what would be the consequence of a Type II error?
- In your opinion, which type of error would be more serious? Justify your answer. (It is possible to argue both sides.)

## 7.4 Association and Causation

When we are comparing two populations we often focus on the nature of the relationship between a **response variable**,  $Y$ —a variable that measures an outcome of interest—and an **explanatory variable**  $X$ —a variable used to explain or predict an outcome. As we will explore next, with data collected from an **experiment** we can assess whether or not there is evidence that  $X$  *affects* the mean value of  $Y$ . That is, we can ask, Do changes in  $X$  *cause* changes in  $Y$ ? (For example, does toluene affect the mean amount of norepinephrine in the brain?) With **observational studies** our conclusions are more limited—we are not able to make causal claims, but rather only conclusions regarding association between  $X$  and  $Y$ . For example, we can ask, Are changes in  $X$  associated with changes in the mean value of  $Y$ ? Or, Is there evidence that the mean values of  $Y$  differ for two populations? (For example, do crawfish captured from two different locations have different mean lengths?)

Thus, our ability to investigate such questions depends on how the data were collected: experimentally or with an observational study. Below are examples of each type of study as they pertain to comparing the means of two samples, followed by a more formal discussion of these study types.

### Example 7.4.1

**Hematocrit in Males and Females** Hematocrit level is a measure of the concentration of red cells in blood. Table 7.4.1 gives the sample means and standard deviations of hematocrit values for two samples of 17-year-old American youths—489 males and 469 females.<sup>17</sup>

<b>Table 7.4.1</b> Hematocrit (percent)		
	Males	Females
Mean	45.8	40.6
SD	2.8	2.9

### Example 7.4.2

**Pargyline and Sucrose Consumption** A study was conducted to determine the effect of the psychoactive drug Pargyline on feeding behavior in the black blowfly *Phormia regina*. The response variable was the amount of sucrose (sugar) solution a fly would drink in 30 minutes. The experimenters used two separate groups of flies: a group injected with Pargyline (905 flies) and a control group injected with saline (900 flies). Comparing the responses of the two groups provides an indirect assessment of the effect of Pargyline. (One might propose that a more *direct* way to determine

the effect of the drug would be to measure each fly twice—on one occasion after injecting Pargyline and on another occasion after injecting saline. However, this direct method is not practical because the measurement procedure disturbs the fly so much that each fly can be measured only once.) Table 7.4.2 shows the means and standard deviations for the two groups.<sup>18</sup> ■

<b>Table 7.4.2</b> Sucrose consumption (mg)		
	Control	Pargyline
Mean	14.9	46.5
SD	5.4	11.7

Examples 7.4.1 and 7.4.2 both involve two-sample comparisons, but notice that the two studies differ in a fundamental way. In Example 7.4.1 the samples come from populations that occur naturally; the investigator is merely an observer:

- Population 1: Hematocrit values of 17-year-old U.S. males
- Population 2: Hematocrit values of 17-year-old U.S. females

By contrast, the two populations in Example 7.4.2 do not actually exist but rather are defined in terms of specific experimental conditions; in a sense, the populations are created by experimental intervention:

- Population 1: Sucrose consumptions of blowflies when injected with saline
- Population 2: Sucrose consumptions of blowflies when injected with Pargyline

These two types of two-sample comparisons—the observational and the experimental—are both widely used in research. The formal methods of analysis are often the same for the two types, but the interpretation of the results is often somewhat different. For instance, in Example 7.4.2 it might be reasonable to say that Pargyline *causes* the increase in sucrose consumption, while no such notion applies in Example 7.4.1.

## Observational versus Experimental Studies

A major consideration in interpreting the results of a biological study is whether the study was observational or experimental. In an **experiment**, the researcher intervenes in or manipulates the experimental conditions.\* In an **observational study**, the researcher merely observes an existing situation, as in the following example.

---

### Example 7.4.3

**Cigarette Smoking** In studies of the effects of smoking cigarettes, both experimental and observational approaches have been used. Effects in animals can be studied experimentally, because animals (for instance, dogs) can be allocated to treatment groups and the groups can be given various doses of cigarette smoke. Effects in humans are usually studied observationally. In one study, for example, pregnant women were questioned about their smoking habits, dietary habits, and so on.<sup>19</sup> When the babies were born, their physical and mental development was followed.

---

\*The conditions being manipulated must be those defining the populations being compared. For example, if five men and five women are given the same drug and then the sexes are compared, the comparison of men to women is observational, not experimental.

One striking finding related to the babies' birthweights: The smokers tended to have smaller babies than the nonsmokers. The difference was not attributable to chance (the  $P$ -value was less than  $10^{-5}$ ). Nevertheless, it was far from clear that the difference was *caused* by smoking, because the women who smoked differed from the nonsmokers in many other aspects of their lifestyle besides smoking—for instance, they had very different dietary habits.

As Example 7.4.3 illustrates, it can be difficult to determine the exact nature of a cause–effect relationship in an observational study. In an experiment, on the other hand, a cause–effect relationship may be easy to see, based on the way in which the researcher manipulated the experimental conditions. To help fix the ideas, consider studying cholesterol level. Suppose a group of patients with high cholesterol levels enrolls in a clinical trial—that is, in a medical experiment—in which some of the patients are randomly chosen to receive a new drug and others are given a standard drug that has shown only modest effects in the past. If a two-sample  $t$  test shows that the mean cholesterol level decreased more for those on the new drug than for those on the standard drug, then the researcher can conclude that the new drug *caused* the superior outcome and is better than the standard drug.

Now consider a two-sample  $t$  test to compare average cholesterol level in a random sample of 50-year-olds to average cholesterol level in a random sample of 25-year-olds. Suppose a two-sample  $t$  test gives a small  $P$ -value, with the 50-year-olds having higher cholesterol than the 25-year-olds. We could be fairly confident that cholesterol level tends to increase with age. However, it would be *possible* that some other explanation were at work. For example, maybe diets have changed over time and the 25-year-olds are eating foods that the 50-year-olds don't eat, causing the 25-year-olds to have low cholesterol; perhaps if the 25-year-olds keep the same diet until they are 50, they will still have low cholesterol at age 50.

As a third example, consider comparing a random sample of home owners to a random sample of renters. Suppose a two-sample  $t$  test shows a significantly higher mean cholesterol level among the home owners than among the renters. We should not conclude that buying a home causes one's cholesterol level to rise. Rather, we should consider that people who own homes tend to be older than are renters. It might very well be the case that age is the causal factor, which explains why the home owners have higher cholesterol than do the renters.

All three of these cases might involve a two-sample  $t$  test and the rejection of  $H_0$ . Indeed, we might get the same  $P$ -value in each test. However, the conclusions we can draw from the three situations are quite different. The scope of the inference we can draw depends on the way in which the data are collected. Experiments allow us to infer cause–effect relationships that can only be guessed at in observational studies. Sometimes an observational study will leave us feeling reasonably confident that we understand the causal mechanism at work; however, we will see that drawing such conclusions is fraught with danger. For this reason, researchers interested in drawing causal conclusions should make great efforts to conduct controlled experiments rather than observational studies.

## More on Observational Studies

The difficulties in interpreting observational studies arise from two primary sources:

- Nonrandom selection from populations
- Uncontrolled extraneous variables

The following example illustrates both of these.

**Example  
7.4.4**

**Race and Brain Size** In the nineteenth century, much effort was expended in the attempt to show “scientifically” that certain human races were inferior to others. A leading researcher on this subject was the American physician S. G. Morton, who won widespread admiration for his studies of human brain size. Throughout his life, Morton collected human skulls from various sources, and he carefully measured the cranial capacities of hundreds of these skulls. His data appeared to suggest that (as he suspected) the “inferior” races had smaller cranial capacities. Table 7.4.3 gives a summary of Morton’s data comparing Caucasian skulls to those of Native Americans.<sup>20</sup> According to a *t* test, the difference between these two samples is “statistically significant” (*P*-value < 0.001). But is it *meaningful*?

<b>Table 7.4.3</b> Cranial capacity (in <sup>3</sup> )		
	Caucasian	Native American
Mean	87	82
SD	8	10
<i>n</i>	52	144

In the first place, the notion that cranial capacity is a measure of intelligence is no longer taken seriously. Leaving that question aside, one can still ask whether it is true that the mean cranial capacity of Native Americans is less than that of Caucasians. Such an inference beyond the actual data requires that the data be viewed as random samples from their respective populations. Of course, in actuality, Morton’s data are not random samples but “samples of convenience,” because Morton measured those skulls that he happened to obtain. But might the data be viewed “as if” they were generated by random sampling? One way to approach this question is to look for sources of bias. In 1977, the noted biologist Stephen Jay Gould reexamined Morton’s data with this goal in mind, and indeed Gould found several sources of bias. For instance, the 144 Native American skulls represent many different groups of Native Americans; as it happens, 25% of the skulls (that is, 36 of them) were from Inca Peruvians, who were a small-boned people with small skulls, while relatively few were from large-skulled tribes such as the Iroquois. Clearly a comparison between Native Americans and Caucasians is meaningless unless somehow adjusted for such imbalances. When Gould made such an adjustment, he found that the difference between Native Americans and Caucasians vanished. ■

Even though the story of Morton’s skulls is more than 100 years old, it can still serve to alert us to the pitfalls of inference. Morton was a conscientious researcher and took great care to make accurate measurements; Gould’s reexamination did not reveal any suggestion of conscious fraud on Morton’s part. Morton may have overlooked the biases in his data because they were *invisible* biases; that is, they related to aspects of the selection process rather than aspects of the measurements themselves.

When we look at a set of observational data, we can sometimes become so hypnotized by its apparent *solidity* and *objectivity* that we forget to ask how the observational units—the persons or things that were observed—were selected. The question should always be asked. If the selection was haphazard rather than truly random, the results can be severely distorted.

## Confounding

Many observational studies are aimed at discovering some kind of causal relationship. Such discovery can be very difficult because of extraneous variables that enter in an uncontrolled (and perhaps unknown) way. The investigator must be guided by the maxim:

Association is not causation.

For instance, it is known that some populations whose diets are high in fiber enjoy a reduced incidence of colon cancer. But this observation does not in itself show that it is the high-fiber diet, rather than some other factor, that provides the protection against colon cancer.

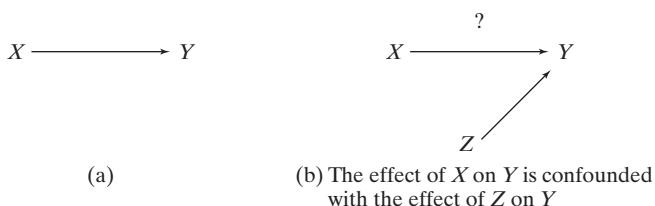
The following example shows how uncontrolled extraneous variables can cloud an observational study, and what kinds of steps can be taken to clarify the picture.

### Example 7.4.5

**Smoking and Birthweight** In a large observational study of pregnant women, it was found that the women who smoked cigarettes tended to have smaller babies than the nonsmokers.<sup>19</sup> (This study was mentioned in Example 7.4.3.) It is plausible that smoking could cause a reduction in birthweight, for instance, by interfering with the flow of oxygen and nutrients across the placenta. But of course plausibility is not proof. In fact, the investigators found that the smokers differed from the nonsmokers with respect to many other variables. For instance, the smokers drank more whiskey than the nonsmokers. Alcohol consumption might plausibly be linked to a deficit in growth. ■

In Example 7.4.5 three variables are presented; let us refer to these as  $X$  = smoking,  $Y$  = birthweight, and  $Z$  = alcohol consumption. There is an association between  $X$  and  $Y$ , but is there a *causal* link between them? Or is there a causal link between  $Z$  and  $Y$ ? Figure 7.4.1 gives a schematic representation of the situation. Changes in  $X$  are associated with changes in  $Y$ . However, changes in  $Z$  are also associated with changes in  $Y$ . We say that the effect that  $X$  has on  $Y$  is **confounded** with the effect that  $Z$  has on  $Y$ . In the context of Example 7.4.5, we say that the effect that smoking has on birthweight is confounded with the effect that alcohol consumption has on birthweight. In observational studies, confounding of effects is a common problem.

**Figure 7.4.1** Schematic representation of causation (a) and of confounding (b)



### Example 7.4.6

**Smoking and Birthweight** The study presented in Example 7.4.5 uncovered many confounding variables. For example, the smokers drank more coffee than the nonsmokers. In addition—and this is especially puzzling—it was found that the smokers began to menstruate at younger ages than the nonsmokers. This phenomenon (early onset of menstruation) could not possibly have been *caused* by smoking, because it occurred (in almost all instances) *before* the woman began to smoke. One interpretation that has been proposed is that the two populations—women who choose to smoke and those who do not—are different in some biological way; thus, it has been suggested that the reduced birthweight is due “to the *smoker*, not the *smoking*.<sup>21</sup>

A number of more recent studies have attempted to shed some light on the relationship between maternal smoking and infant development. Researchers in one study observed, in addition to smoking habits, about 50 extraneous variables, including the mother's age, weight, height, blood type, upper arm circumference, religion, education, income, and so on.<sup>22</sup> After applying complex statistical methods of adjustment, they concluded that birthweight varies with smoking even when these extraneous factors are held constant. This says that there quite likely is a link between  $X = \text{smoking}$  and  $Y = \text{birthweight}$  as shown in Figure 7.4.1, although several other variables also affect birthweight. The point is that the presence of confounding doesn't mean that a link does not exist between  $X$  and  $Y$ , only that it is tangled up with other effects, so that we have to be cautious when interpreting the findings of an observational study.

In another study of pregnant women, researchers measured various quantities related to the functioning of the placenta.<sup>23</sup> They found that, compared to non-smokers, women who smoked had more abnormalities of the placenta, and that their infants had very much higher blood levels of cotinine, a substance derived from nicotine. They also found evidence that, in the women who smoked, the circulation of blood in the placenta was notably improved by abstaining from smoking for three hours.

A third study used a matched design to try to isolate the effect of smoking behavior. The investigators identified 159 women who had smoked during one pregnancy but quit smoking before the next pregnancy.<sup>24</sup> These women were individually matched with 159 women who smoked during two consecutive pregnancies; pairs were matched with respect to the birthweight of the first child, amount of smoking during the first pregnancy, and several other factors. Thus, the members of a pair were believed to have identical "reproductive potential." The researchers then considered the birthweight of the second child; they found that the women who had quit smoking gave birth to infants who weighed more than the infants of their matched controls who continued to smoke. Of course, we cannot rule out the possibility that the women who quit smoking also quit other harmful habits, such as drinking too much alcohol, and that the increased birthweight was not really caused by giving up smoking. ■

Example 7.4.6 shows that observational studies can provide information about causality but must be interpreted cautiously. Researchers generally agree that a causal interpretation of an observed association requires extra support—for instance, that the association be observed consistently in observational studies conducted under various conditions and taking various extraneous factors into account, and also, ideally, that the causal link be supported by experimental evidence. We do not mean to say that an observed association *cannot* be causally interpreted, but only that such interpretation requires particular caution.

## Spurious Association

---

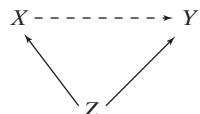
### Example 7.4.7

**Ultrasound** It is quite common for a physician to use ultrasound examination of the fetus of a pregnant woman. However, when ultrasound technology was first used, there were concerns that the procedure might be harmful to the baby. An early study seemed to bear this out: On average, babies exposed to ultrasound in the womb were lighter at birth than were babies not exposed to ultrasound.<sup>25</sup> Later, a study was done in which some women were randomly chosen to have ultrasounds and others were not given ultrasounds. This study found no difference in birthweight between the two groups.<sup>26</sup> It seems that the reason a difference appeared in the first

study was that ultrasound was being used mostly for women who were experiencing problem pregnancies. The complications with the pregnancy were leading to low birthweight, not the use of ultrasound.

Figure 7.4.2 gives a schematic representation of the situation in Example 7.4.7. Changes in  $X$  (having an ultrasound examination) are associated with changes in  $Y$  (lower birthweight). However,  $X$  and  $Y$  are both dependent on a third variable  $Z$  (whether or not there are problems with the pregnancy), which is the variable that is driving the relationship. Changes in  $X$  and changes in  $Y$  are a common response to the third variable  $Z$ . We say that the association between  $X$  and  $Y$  is **spurious**: When we control for the “lurking variable”  $Z$ , the link between  $X$  and  $Y$  disappears. In the case of Example 7.4.7, it was not having an ultrasound that influenced birthweight; what mattered was whether or not there were problems with the pregnancy.

**Figure 7.4.2** Schematic representation of spurious association



The association between  $X$  and  $Y$  is spurious; controlling for the lurking variable  $Z$  eliminates the  $X$ - $Y$  link.

## More on Experiments

An experiment is a study in which the researcher intervenes and imposes treatment conditions. The following is a simple example.

### Example 7.4.8

**Headache Pain** Suppose a researcher gives ibuprofen to some people who have headaches and aspirin to others and then measures how long it takes for each person’s headache to disappear. In this case, there are two treatments: ibuprofen and aspirin. By assigning people to treatment groups—ibuprofen and aspirin—the researcher is conducting an experiment.

When we are discussing an experiment, we refer to the units to which the treatments are assigned as **experimental units**. In an agricultural experiment, an experimental unit might be a plot of land. In general, an experimental unit is the smallest unit to which a treatment is applied in an experiment. Thus, in Example 7.4.8 the experimental units are individual people, since treatment is assigned on a person-by-person basis.

If treatments are assigned at random, for example, by tossing a coin and letting heads mean the person gets ibuprofen, while tails means the person gets aspirin, then the experiment is a *randomized* experiment. Sometimes an experiment is conducted in which one group is given a treatment and a second—the control group—is given nothing. For example, one could investigate the effectiveness of ibuprofen in treating headache pain by giving it to some people, while giving no painkiller to others. In contrast, the experiment in which some people are given ibuprofen and others are given aspirin is said to have an “active” control—the aspirin group.

## Randomization Distributions

In Section 5.2 we developed the concept of a sampling distribution for the sample mean,  $\bar{Y}$ , by considering how  $\bar{Y}$  varies from one random sample to another. Strictly

speaking, this provides the foundation for inference when analyzing an observational study, but not when the data arise from an experiment—in which treatments are assigned to experimental units, rather than a random sample being taken from a population. However, the concepts of Section 5.2 can be extended in a natural way to develop the **randomization distribution** of  $\bar{Y}$ , which is the distribution that  $\bar{Y}$  takes on under all possible random assignments within an experiment. Randomization distributions then form the foundation for inference for experiments.

## Only Statistical?

The term “statistical” is sometimes used—or, rather, misused—as an epithet. For instance, some people say that the evidence linking dietary cholesterol and heart disease is “only statistical.” What they really mean is “only observational.” Statistical evidence can be very strong indeed, if it flows from a randomized experiment rather than an observational study. As we have seen in the preceding examples, statistical evidence from an observational study must be interpreted with great care, because of potential distortions caused by extraneous variables.

### Exercises 7.4.1–7.4.9

**7.4.1** In 2005, 5.3% of the deaths in the United States were caused by chronic lower respiratory diseases (e.g., asthma and emphysema). In Arizona, 6.2% of deaths were due to chronic lower respiratory diseases.<sup>27</sup> Does this mean that living in Arizona exacerbates respiratory problems? If not, how can we explain the Arizona rate being above the national rate?

**7.4.2** It has been hypothesized that silicone breast implants cause illness. In one study it was found that women with implants were more likely to smoke, to be heavy drinkers, to use hair dye, and to have had an abortion than were women in a comparison group who did not have implants.<sup>28</sup> Use the language of statistics to explain why this study casts doubt on the claim that implants cause illness.

**7.4.3** Consider the setting of Exercise 7.4.2.

- What is the explanatory variable?
- What is the response variable?
- What are the observational units?

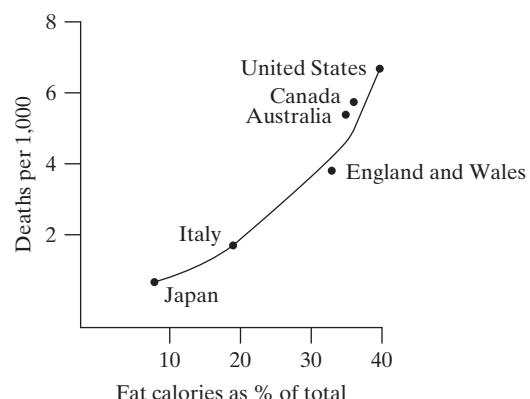
**7.4.4** In a study of 1,040 subjects, researchers found that the prevalence of coronary heart disease increased as the number of cups of coffee consumed per day increased.<sup>29</sup>

- What is the explanatory variable?
- What is the response variable?
- What are the observational units?

**7.4.5** For an early study of the relationship between diet and heart disease, the investigator obtained data on heart disease mortality in various countries and on national

average dietary compositions in the same countries. The accompanying graph shows, for six countries, the 1948–1949 death rate from degenerative heart disease (among men aged 55–59 years) plotted against the amount of fat in the diet.<sup>30</sup>

In what ways might this graph be misleading? Which extraneous variables might be relevant here? Discuss.



**7.4.6** Shortly before Valentine’s Day in 1999, a newspaper article was printed with the headline “Marriage makes for healthier, longer life, studies show.” The headline was based on studies that showed that married persons live longer and have lower rates of cancer, heart disease, and stroke than do those who never marry.<sup>31</sup> Use the language of statistics to discuss the headline. Use a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 to support your explanation of the situation.

**7.4.7** In June 2009, the *New York Times* published an article entitled “Alcohol’s Good for You? Some Scientists Doubt It.” The author wrote, “Study after study suggests that alcohol in moderation may promote heart health and even ward off diabetes and dementia. The evidence is so plentiful that some experts consider moderate drinking—about one drink a day for women, about two for men—a central component of a healthy lifestyle.” Later in the article, the author wrote, “For some scientists, the question will not go away. No study, these critics say, has ever proved a causal relationship between moderate drinking and lower risk of death.” Explain using the language of statistics and a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 why the critics say no study has ever proved a causal relationship.

**7.4.8** In a study of the relationship between birthweight and race, birth records of babies born in Illinois were examined. The researchers found that the percentage of low birthweight babies among babies born to U.S.-born white women was much lower than the percentage of low birthweight babies among babies born to U.S.-born black women. This suggests that race plays an important role in determining the chance that a baby will have a low birthweight. However, the percentage of low birthweight babies among babies born to African-born black women was roughly equal to the percentage among babies born to U.S.-born white women.<sup>32</sup> Use the language of statistics to discuss what these data say about the relationships between low birthweight, race, and mother’s birthplace. Use a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 to support your explanation.

**7.4.9** Does the release of a Harry Potter book lead children to spend more time reading and thus reduce the number of accidents they have? Doctors in England compared the number of emergency room visits due to

musculoskeletal injuries to children aged 7 to 15 during two types of weekends: (1) following the release dates of two books in the Harry Potter series and (2) during 24 “control” weekends, for one hospital. The following table shows the data, with the “Harry Potter weekends” in italics.<sup>33</sup>

WEEKEND	INJURIES	WEEKEND	INJURIES
6/7/03	63	7/10/04	57
6/14/03	77	7/17/04	66
<i>6/21/03</i>	<i>36*</i>	<i>7/24/04</i>	62
6/28/03	63	6/4/05	51
7/5/03	75	6/11/05	83
7/12/03	71	6/18/05	60
7/19/03	60	6/25/05	66
7/26/03	52	7/2/05	74
6/5/04	78	7/9/05	75
<i>6/12/04</i>	<i>84</i>	<i>7/16/05</i>	<i>37*</i>
6/19/04	70	7/23/05	46
6/26/04	75	7/30/05	68
<i>7/3/04</i>	<i>81</i>	<i>8/6/05</i>	60

- (a) Given the nature of the data, can we make an inference about the release of Harry Potter books *causing* a change in accidents? Why or why not?
- (b) The average for the Harry Potter weekends is 36.5, with a standard deviation of 0.7. The corresponding numbers for the other (control) weekends are 67.4 and 10.4. Use a *t* test to investigate the claim that the small number of injuries during Harry Potter weekends is consistent with chance variation. Use  $\alpha = 0.01$ . [Note: Formula (6.7.1) yields 23.9 degrees of freedom for these data.]

## 7.5 One-Tailed *t* Tests

The *t* test described in the preceding sections is called a **two-tailed *t* test** or a **two-sided *t* test** because the null hypothesis is rejected if  $t_s$  falls in either tail of the Student’s *t* distribution and the *P*-value of the data is a two-tailed area under Student’s *t* curve. A two-tailed *t* test is used to test the null hypothesis

$$H_0: \mu_1 = \mu_2$$

against the alternative hypothesis

$$H_A: \mu_1 \neq \mu_2$$

This alternative  $H_A$  is called a **nondirectional alternative**.

## Directional Alternative Hypotheses

In some studies it is apparent from the beginning—before the data are collected—that there is only one reasonable direction of deviation from  $H_0$ . In such situations it is appropriate to formulate a directional alternative hypothesis. The following is a directional alternative:

$$H_A: \mu_1 < \mu_2$$

Another directional alternative is

$$H_A: \mu_1 > \mu_2$$

The following two examples illustrate situations where directional alternatives are appropriate.

---

**Example  
7.5.1**

**Niacin Supplementation** Consider a feeding experiment with lambs. The observation  $Y$  will be weight gain in a two-week trial. Ten animals will receive diet 1, and 10 animals will receive diet 2, where

$$\text{Diet 1} = \text{Standard ration + Niacin}$$

$$\text{Diet 2} = \text{Standard ration}$$

On biological grounds it is expected that niacin may increase weight gain; there is no reason to suspect that it could possibly decrease weight gain. An appropriate formulation would be

$$H_0: \text{Niacin is not effective in increasing weight gain } (\mu_1 = \mu_2).$$

$$H_A: \text{Niacin is effective in increasing weight gain } (\mu_1 > \mu_2). \quad \blacksquare$$

---

**Example  
7.5.2**

**Hair Dye and Cancer** Suppose a certain hair dye is to be tested to determine whether it is carcinogenic (cancer causing). The dye will be painted on the skins of 20 mice (group 1), and an inert substance will be painted on the skins of 20 mice (group 2) that will serve as controls. The observation  $Y$  will be the number of tumors appearing on each mouse. An appropriate formulation is

$$H_0: \text{The dye is not carcinogenic } (\mu_1 = \mu_2).$$

$$H_A: \text{The dye is carcinogenic } (\mu_1 > \mu_2). \quad \blacksquare$$

**Note:** If  $H_A$  is directional, then some people would rewrite  $H_0$  to include the “opposite direction.” For example, if  $H_A: \mu_1 > \mu_2$ , then we could write  $H_0$  as  $H_0: \mu_1 \leq \mu_2$ . Thus, the null hypothesis is stating that the mean of population 1 is not greater than the mean of population 2, whereas the alternative hypothesis asserts that the mean of population 1 is greater than the mean of population 2. Between these two hypotheses, all possibilities are covered.

## The One-Tailed Test Procedure

When the alternative hypothesis is directional, the *t* test procedure must be modified. The modified procedure is called a **one-tailed *t* test** and is carried out in two steps as follows:

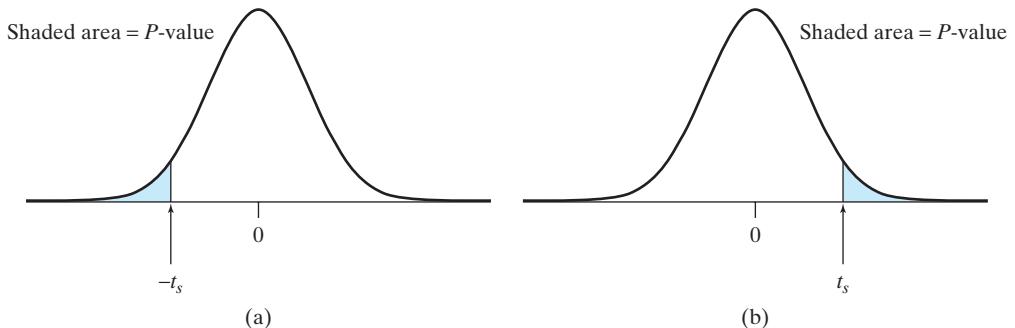
**Step 1** Check directionality—see if the data deviate from  $H_0$  in the direction specified by  $H_A$ :

- (a) If not, the *P*-value is greater than 0.50.
- (b) If so, proceed to step 2.

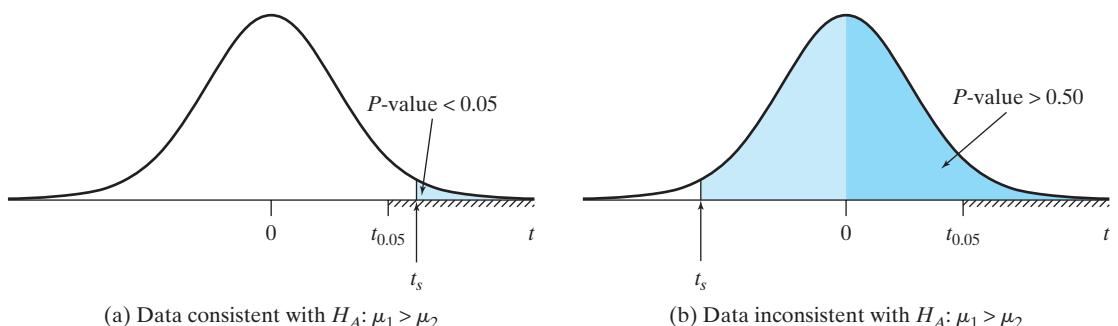
**Step 2** The *P*-value of the data is the *one-tailed* area beyond  $t_s$ .

To conclude the test, one can make a decision at a prespecified significance level  $\alpha$ :  $H_0$  is rejected if  $P\text{-value} \leq \alpha$ .

The rationale of the two-step procedure is that the *P*-value measures deviation from  $H_0$  in the direction specified by  $H_A$ . The one-tailed *P*-value is illustrated in Figure 7.5.1 for two cases in which the data deviate from  $H_0$  in the direction specified by  $H_A$ . Figure 7.5.2 illustrates the *P*-value for (a) a case in which the data are consistent with  $H_A$ :  $\mu_1 > \mu_2$  and (b) a case in which the data are inconsistent with  $H_A$ :  $\mu_1 > \mu_2$ . The two-step testing procedure is demonstrated in Example 7.5.3.



**Figure 7.5.1** One-tailed *P*-value for a *t* test, (a) if the alternative is  $H_A: \mu_1 < \mu_2$  and  $t_s$  is negative; (b) if the alternative is  $H_A: \mu_1 > \mu_2$  and  $t_s$  is positive



**Figure 7.5.2** One-tailed *P*-value for a *t* test, (a) in which the data are consistent with  $H_A: \mu_1 > \mu_2$ ; (b) in which the data are inconsistent with  $H_A: \mu_1 > \mu_2$

**Example  
7.5.3**

**Niacin Supplementation** Consider the lamb feeding experiment of Example 7.5.1. The alternative hypothesis is

$$H_A: \mu_1 > \mu_2$$

We will claim significant evidence for  $H_A$  if  $\bar{Y}_1$  is sufficiently greater than  $\bar{Y}_2$ . Suppose formula (6.7.1) yields  $df = 18$ . The critical values from Table 4 are reproduced in Table 7.5.1.

**Table 7.5.1** Critical values with  $df = 18$

Tail area	0.20	0.10	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
Critical value	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922

To illustrate the one-tailed test procedure, suppose that we have<sup>34</sup>

$$\text{SE}(\bar{Y}_1 - \bar{Y}_2) = 2.2 \text{ lb}$$

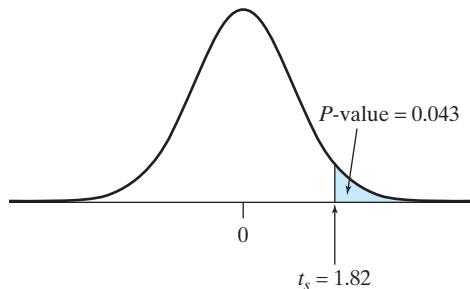
and that we choose  $\alpha = 0.05$ . Let us consider various possibilities for the two sample means.

- (a) Suppose the data give  $\bar{y}_1 = 10$  lb and  $\bar{y}_2 = 13$  lb. This deviation from  $H_0$  is opposite to the assertion of  $H_A$ ; we have  $\bar{y}_1 < \bar{y}_2$ , but  $H_A$  asserts that  $\mu_1 > \mu_2$ . Consequently,  $P\text{-value} > 0.50$ , so we would not find significant evidence for  $H_A$  at any significance level. (We would never use an  $\alpha$  greater than 0.50.) We conclude that the data provide no evidence that niacin is effective in increasing weight gain.
- (b) Suppose the data give  $\bar{y}_1 = 14$  lb and  $\bar{y}_2 = 10$  lb. This deviation from  $H_0$  is in the direction of  $H_A$  (because  $\bar{y}_1 > \bar{y}_2$ ), so we proceed to step 2. The value of  $t_s$  is

$$t_s = \frac{(14 - 10) - 0}{2.2} = 1.82$$

The (one-tailed)  $P$ -value for the test is the probability of getting a  $t$  statistic, with 18 degrees of freedom, that is as large or larger than 1.82. This upper tail probability (found with a computer) is 0.043, as shown in Figure 7.5.3.

**Figure 7.5.3** One-tailed  $P$ -value for the  $t$  test in Example 7.5.3



If we did not have a computer or graphing calculator available, we could use Table 4 to bracket the  $P$ -value. From Table 4, we see that the  $P$ -value would be bracketed as follows:

$$0.04 < \text{one-tailed } P\text{-value} < 0.05$$

Since  $P\text{-value} < \alpha$ , we reject  $H_0$  and conclude that there is some evidence that niacin is effective.

- (c) Suppose the data give  $\bar{y}_1 = 11$  lb and  $\bar{y}_2 = 10$  lb. Then, proceeding as in part (b), we compute the test statistic as  $t_s = 0.45$ . The  $P$ -value is 0.329.

If we did not have a computer or graphing calculator available, we could use Table 4 to bracket the  $P$ -value as

$$P\text{-value} > 0.20$$

Since  $P\text{-value} > \alpha$ , we do not find significant evidence for  $H_A$ ; we conclude that there is insufficient evidence to claim that niacin is effective. Thus, although these data deviate from  $H_0$  in the direction of  $H_A$ , the amount of deviation is not great enough to justify significant evidence for  $H_A$ . ■

Notice that what distinguishes a one-tailed from a two-tailed  $t$  test is the way in which the  $P$ -value is determined, but not the directionality or nondirectionality of the conclusion. If we find significant evidence for  $H_A$ , our conclusion may be considered directional even if our  $H_A$  is nondirectional.\* (For instance, in Example 7.2.4 we concluded that toluene increases NE concentration.)

## Directional versus Nondirectional Alternatives

The same data will give a different  $P$ -value depending on whether the alternative hypothesis is directional or nondirectional. Indeed, if the data deviate from  $H_0$  in the direction specified by  $H_A$ , the  $P$ -value for a directional alternative hypothesis will be 1/2 of the  $P$ -value for the test that uses a nondirectional alternative. It can happen that the same data will provide significant evidence for  $H_A$  using the one-tailed procedure but not using the two-tailed procedure, as Example 7.5.4 shows.

### Example 7.5.4

**Niacin Supplementation** Consider part (b) of Example 7.5.3. In that example we chose  $\alpha = 0.05$  and tested

$$H_0: \mu_1 = \mu_2$$

against the directional alternative hypothesis

$$H_A: \mu_1 > \mu_2$$

With  $\bar{y}_1 = 14$  lb and  $\bar{y}_2 = 10$  lb, the test statistic was  $t_s = 1.82$  and the  $P$ -value was 0.043, as indicated in Figure 7.5.3. Our conclusion was to claim there is significant evidence for  $H_A$ .

However, suppose we had wished to test

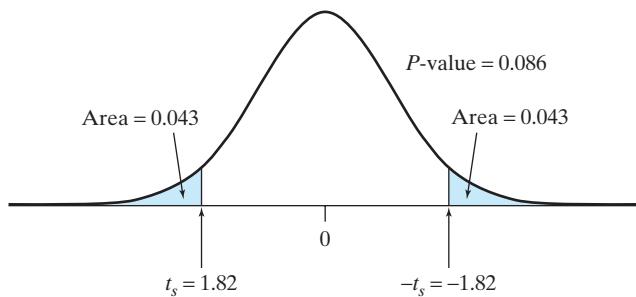
$$H_0: \mu_1 = \mu_2$$

against the nondirectional alternative hypothesis

$$H_A: \mu_1 \neq \mu_2$$

With the same data of  $\bar{y}_1 = 14$  lb and  $\bar{y}_2 = 10$  lb, the test statistic is still  $t_s = 1.82$ . The  $P$ -value, however, is 0.086, as shown in Figure 7.5.4. Thus,  $P$ -value  $> \alpha$  and we do not reject  $H_0$ .

**Figure 7.5.4** Two-tailed  $P$ -value for the  $t$  test in Example 7.5.4



\*Some authors prefer not to draw a directional conclusion if  $H_A$  is nondirectional.

Hence, the one-tailed procedure finds significant evidence for  $H_A$ , but the two-tailed procedure does not. In this sense, it is “easier” to claim that the evidence significantly supports  $H_A$  with the one-tailed procedure than with the two-tailed procedure. ■

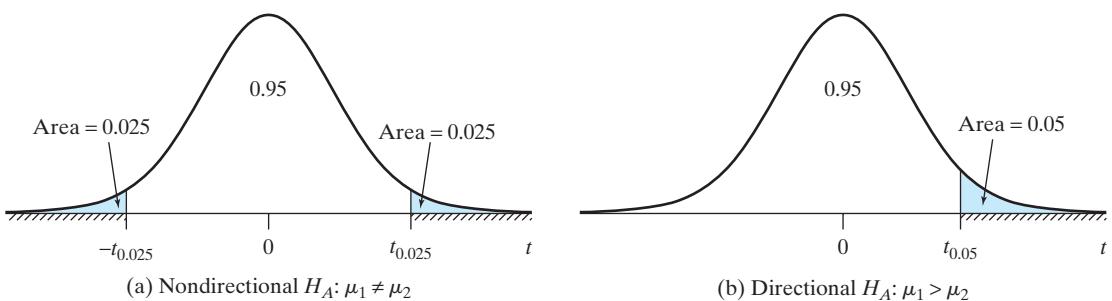
Why is the two-tailed  $P$ -value cut in half when the alternative hypothesis is directional? In Example 7.5.4, the researcher would conclude by saying, “The data suggest that niacin increases weight gain. But if niacin has no effect, then the kind of data I got in my experiment—having two sample means that differ by 1.82 SEs or more—would happen fairly often ( $P$ -value = 0.086). Sometimes the niacin diet would come out on top; sometimes the standard diet would come out on top. I cannot find significant evidence for  $H_A$  on the basis of what I have seen in these data.” In Example 7.5.3(b), the researcher would conclude by saying, “*Before the experiment was run*, I suspected that niacin increases weight gain. The data provide evidence in support of this theory. If niacin has no effect, then the kind of data I got in my experiment—having the niacin diet sample mean exceed the standard diet that differ by 1.82 SEs or more—would rarely happen ( $P$ -value 0.043). (Before the experiment was run I dismissed the possibility that the niacin diet mean could be less than the standard diet mean.) Thus, I can claim my evidence significantly supports  $H_A$ .” The researcher in Example 7.5.3(b) is using *two* sources of information to claim the significance of evidence for  $H_A$ : (1) what the data have to say (as measured by the tail area) and (2) previous expectations (which allow the researcher to ignore the lower tail area—the 0.043 area under the curve below  $-1.82$  in Figure 7.5.4).

Note that the modification in procedure, when going from a two-tailed to a one-tailed test, preserves the interpretation of significance level  $\alpha$  as given in Section 7.3, that is,

$$\alpha = \Pr\{\text{reject } H_0\} \text{ if } H_0 \text{ is true}$$

For instance, consider the case  $\alpha = 0.05$ . Figure 7.5.5 shows that the total shaded area—the probability of rejecting  $H_0$ —is equal to 0.05 in both a two-tailed test and a one-tailed test. This means that, if a great many investigators were to test a true  $H_0$ , then 5% of them would find significant evidence for  $H_A$  and commit a Type I error; this statement is true whether the alternative  $H_A$  is directional or nondirectional.

The crucial point in justification of the modified procedure for testing against a directional  $H_A$  is that *if* the direction of deviation of the data from  $H_0$  is *not* as specified by  $H_A$ , then we will not claim that the evidence significantly supports  $H_A$ . For example, in the carcinogenesis experiment of Example 7.5.2, if the mice exposed to the hair dye had *fewer* tumors than the control group, we might (1) simply conclude



**Figure 7.5.5** Two-tailed and one-tailed  $t$  test with  $\alpha = 0.05$ . The data provide significant evidence for  $H_A$  if  $t_s$  falls in the hatched region of the  $t$ -axis

that the data do not indicate a carcinogenic effect, or (2) if the exposed group had *substantially* fewer tumors, so that the test statistic  $t_s$  was very far in the wrong tail of the  $t$  distribution, we might look for methodological errors in the experiment—for example, mistakes in lab technique or in recording the data, nonrandom allocation of the mice to the two groups, and so on—but we would not claim significant evidence for  $H_A$ .

A one-tailed  $t$  test is especially natural when only one direction of deviation from  $H_0$  is believed to be plausible. However, one-tailed tests are also used in situations where deviation in both directions is possible, but only one direction is of interest. For instance, in the niacin experiment of Example 7.5.3, it is not necessary that the experimenter believe that it is *impossible* for niacin to reduce weight gain rather than increase it. Deviations in the wrong direction (less weight gain on niacin) would not lead to claiming there is significant evidence for  $H_A$ , and thus we would not make claims about the effect of niacin; this is the essential feature that distinguishes a directional from a nondirectional formulation.

## Choosing the Form of $H_A$

When is it legitimate to use a directional  $H_A$ , and so to perform a one-tailed test? The answer to this question is linked to the directionality check—step 1 of the two-step test procedure given previously. Clearly such a check makes sense only if  $H_A$  was formulated before the data were inspected. (If we were to formulate a directional  $H_A$  that was “inspired” by the data, then of course the data would always deviate from  $H_0$  in the “right” direction and the test procedure would always proceed to step 2.) This is the rationale for the following rule.

### Rule for Directional Alternatives

It is legitimate to use a directional alternative  $H_A$  only if  $H_A$  is formulated before seeing the data and there is no scientific interest in results that deviate in a manner opposite to that specified by  $H_A$ .

In research, investigators often get more pleasure from finding significant evidence for an alternative hypothesis than not finding evidence. In fact, research reports often contain phrases such as “we are unable to find significant evidence for the alternative hypothesis” or “the results failed to reach statistical significance.” Under these circumstances, one might wonder what the consequences would be if researchers succumbed to the natural temptation to ignore the preceding rule for using directional alternatives. After all, very often one can think of a rationale for an effect *ex post facto*—that is, after the effect has been observed. A return to the imaginary experiment on plants’ musical tastes will illustrate this situation.

### Example 7.5.5

**Music and Marigolds** Recall the imaginary experiment of Example 7.3.2, in which investigators measure the heights of marigolds exposed to Bach or Mozart. Suppose, as before, that the null hypothesis is true, that  $df = 60$ , and that the investigators all perform  $t$  tests at  $\alpha = 0.05$ . Now suppose in addition that all of the investigators violate the rule for use of directional alternatives, and that they formulate  $H_A$  after seeing the data. Half of the investigators would obtain data for which  $\bar{y}_1 > \bar{y}_2$ , and they would formulate the alternative

$$H_A: \mu_1 > \mu_2 \text{ (plants prefer Bach)}$$

The other half would obtain data for which  $\bar{y}_1 < \bar{y}_2$ , and they would formulate the alternative

$$H_A: \mu_1 < \mu_2 \text{ (plants prefer Mozart)}$$

Now envision what would happen. Since the investigators are using directional alternatives, they will all compute  $P$ -values using only one tail of the distribution. We would expect them to have the following experiences:

90% of them would get a  $t_s$  in the middle 90% of the distribution and would not find significant evidence for  $H_A$ .

5% of them would get a  $t_s$  in the top 5% of the distribution and would conclude that the plants prefer Bach.

5% of them would get a  $t_s$  in the bottom 5% of the distribution and would conclude that the plants prefer Mozart.

Thus, a total of 10% of the investigators would claim there is significant evidence for  $H_A$ . Of course each investigator individually never realizes that the overall percentage of Type I errors is 10% rather than 5%. And the conclusions that plants prefer Bach or Mozart could be supported by *ex post facto* rationales that would be limited only by the imagination of the investigators. ■

As Example 7.5.5 illustrates, a researcher who uses a directional alternative when it is not justified pays the price of a doubled risk of Type I error. Moreover, those who read the researcher's report will not be aware of this doubling of risk, which is why some scientists advocate never using a directional alternative.

## Exercises 7.5.1–7.5.13

**7.5.1** For each of the following data sets, use Table 4 to bracket the one-tailed  $P$ -value of the data as analyzed by the  $t$  test, assuming that the alternative hypothesis is  $H_A: \mu_1 > \mu_2$ .

(a)

	SAMPLE 1	SAMPLE 2
$n$	10	10
$\bar{y}$	10.8	10.5
$SE_{(\bar{Y}_1 - \bar{Y}_2)} = 0.23$ with df = 18		

(b)

	SAMPLE 1	SAMPLE 2
$n$	100	100
$\bar{y}$	750	730
$SE_{(\bar{Y}_1 - \bar{Y}_2)} = 11$ with df = 180		

**7.5.2** For each of the following data sets, use Table 4 to bracket the one-tailed  $P$ -value of the data as analyzed by the  $t$  test, assuming that the alternative hypothesis is  $H_A: \mu_1 > \mu_2$ .

(a)

	SAMPLE 1	SAMPLE 2
$n$	10	10
$\bar{y}$	3.24	3.00
$SE_{(\bar{Y}_1 - \bar{Y}_2)} = 0.61$ with df = 17		

(b)

	SAMPLE 1	SAMPLE 2
$n$	6	5
$\bar{y}$	560	500
$SE_{(\bar{Y}_1 - \bar{Y}_2)} = 45$ with df = 8		

(c)

	SAMPLE 1	SAMPLE 2
$n$	20	20
$\bar{y}$	73	79
$SE_{(\bar{Y}_1 - \bar{Y}_2)} = 2.8$ with df = 35		

**7.5.3** For each of the following situations, suppose  $H_0: \mu_1 = \mu_2$  is being tested against  $H_A: \mu_1 > \mu_2$ . State whether or not there is significant evidence for  $H_A$ .

- $t_s = 3.75$  with 19 degrees of freedom,  $\alpha = 0.01$ .
- $t_s = 2.6$  with 5 degrees of freedom,  $\alpha = 0.10$ .
- $t_s = 2.1$  with 7 degrees of freedom,  $\alpha = 0.05$ .
- $t_s = 1.8$  with 7 degrees of freedom,  $\alpha = 0.05$ .

**7.5.4** For each of the following situations, suppose  $H_0: \mu_1 = \mu_2$  is being tested against  $H_A: \mu_1 < \mu_2$ . State whether or not there is significant evidence for  $H_A$ .

- $t_s = -1.6$  with 23 degrees of freedom,  $\alpha = 0.05$ .
- $t_s = -2.3$  with 5 degrees of freedom,  $\alpha = 0.10$ .
- $t_s = 0.4$  with 16 degrees of freedom,  $\alpha = 0.10$ .
- $t_s = -2.8$  with 27 degrees of freedom,  $\alpha = 0.01$ .

**7.5.5** Ecological researchers measured the concentration of red cells in the blood of 27 field-caught lizards (*Sceloporus occidentalis*). In addition, they examined each lizard for infection by the malarial parasite *Plasmodium*. The red cell counts ( $10^{-3} \times$  cells per  $\text{mm}^3$ ) were as reported in the table.<sup>35</sup>

	INFECTED ANIMALS	NONINFECTED ANIMALS
$n$	12	15
$\bar{y}$	972.1	843.4
$s$	245.1	251.2

One might expect that malaria would reduce the red cell count, and in fact previous research with another lizard species had shown such an effect. Do the data support this expectation? Assume that the data are normally distributed. Test the null hypothesis of no difference against the alternative that the infected population has a lower red cell count. Use a  $t$  test at

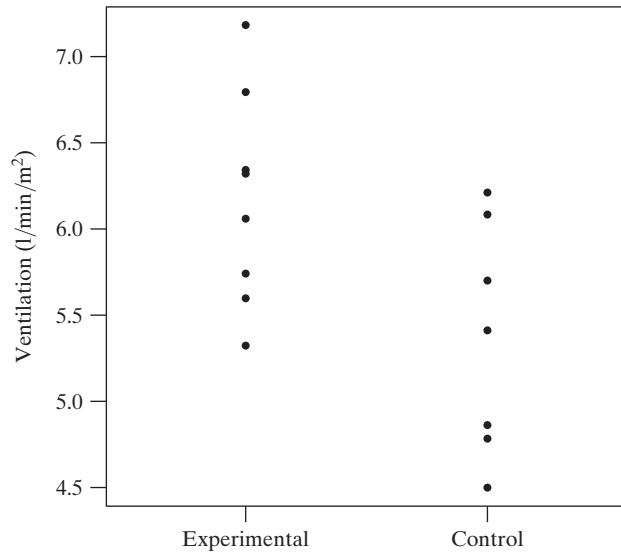
- $\alpha = 0.05$
- $\alpha = 0.10$

[Note: Formula (6.7.1) yields 24 df.]

**7.5.6** A study was undertaken to compare the respiratory responses of hypnotized and nonhypnotized subjects to certain instructions. The 16 male volunteers were allocated at random to an experimental group to be hypnotized or to a control group. Baseline measurements were taken at the start of the experiment. In analyzing the data, the researchers noticed that the baseline breathing patterns of the two groups were different; this was surprising, since all the subjects had been treated the same up to that time. One explanation proposed for this unexpected difference was that the experimental group were more excited in anticipation of the experience of being hypnotized. The accompanying table presents a summary of the baseline measurements of total ventilation (liters of air per

minute per square meter of body area). Parallel dotplots of the data are given in the following graph.<sup>36</sup> [Note: Formula (6.7.1) yields 14 df.]

	EXPERIMENTAL	CONTROL
	5.32	4.50
	5.60	4.78
	5.74	4.79
	6.06	4.86
	6.32	5.41
	6.34	5.70
	6.79	6.08
	7.18	6.21
$n$	8	8
$\bar{y}$	6.169	5.291
$s$	0.621	0.652



- Use a  $t$  test to test the hypothesis of no difference against a nondirectional alternative. Let  $\alpha = 0.05$ .
- Use a  $t$  test to test the hypothesis of no difference against the alternative that the experimental conditions produce a larger mean than the control conditions. Let  $\alpha = 0.05$ .
- Which of the two tests, that of part (a) or part (b), is more appropriate? Explain.

**7.5.7** In a study of lettuce growth, 10 seedlings were randomly allocated to be grown in either standard nutrient solution or in a solution containing extra nitrogen. After 22 days of growth, the plants were harvested and weighed, with the results given in the table.<sup>37</sup> Are the data sufficient to conclude that the extra nitrogen

enhances plant growth under these conditions? Use a  $t$  test at  $\alpha = 0.10$  against a directional alternative. (Assume that the data are normally distributed.) [Note: Formula (6.7.1) yields 7.7 df.]

NUTRIENT SOLUTION	LEAF DRY WEIGHT (GM)		
	n	MEAN	SD
Standard	5	3.62	0.54
Extra nitrogen	5	4.17	0.67

**7.5.8** Research has shown that for mammals giving birth to a son versus a daughter places a greater strain on mothers. Does this affect the health of their next child? A study compared the birthweights of humans born after a male versus after a female. Summary statistics for the sample of size 76 are given in the following table; the data appeared to be normally distributed.<sup>38</sup> Use a  $t$  test, with  $\alpha = 0.05$  and a directional alternative, to investigate the research hypothesis that birthweight is lower when the elder sibling is male. [Note: Formula (6.7.1) yields 69.5 df.]

SEX OF ELDER SIBLING	BIRTHWEIGHT (KG)		
	n	MEAN	SD
Male	33	3.32	0.62
Female	43	3.63	0.63

**7.5.9** An entomologist conducted an experiment to see if wounding a tomato plant would induce changes that improve its defense against insect attack. She grew larvae of the tobacco hornworm (*Manduca sexta*) on wounded plants or control plants. The accompanying table shows the weights (mg) of the larvae after seven days of growth.<sup>39</sup> (Assume that the data are normally distributed.) How strongly do the data support the researcher's expectation? Use a  $t$  test at the 5% significance level. Let  $H_A$  be that wounding the plant tends to diminish larval growth. [Note: Formula (6.7.1) yields 31.8 df.]

	WOUNDED	CONTROL
n	16	18
$\bar{y}$	28.66	37.96
s	9.02	11.14

**7.5.10** A pain-killing drug was tested for efficacy in 50 women who were experiencing uterine cramping pain following childbirth. Twenty-five of the women were randomly allocated to receive the drug, and the remain-

ing 25 received a placebo (inert substance). Capsules of drug or placebo were given before breakfast and again at noon. A pain relief score, based on hourly questioning throughout the day, was computed for each woman. The possible pain relief scores ranged from 0 (no relief) to 56 (complete relief for 8 hours). Summary results are shown in the table.<sup>40</sup> [Note: Formula (6.7.1) yields 47.2 df.]

TREATMENT	PAIN RELIEF SCORE		
	n	MEAN	SD
Drug	25	31.96	12.05
Placebo	25	25.32	13.78

- (a) Test for evidence of efficacy using a  $t$  test. Use a directional alternative and  $\alpha = 0.05$ .
- (b) If the alternative hypothesis were nondirectional, how would the answer to part (a) change?

**7.5.11** Postoperative ileus (POI) is a form of gastrointestinal dysfunction that commonly occurs after abdominal surgery and results in absent or delayed gastrointestinal motility. Does rocking in a chair after abdominal surgery reduce postoperative ileus (POI) duration? Sixty-six postoperative abdominal surgery patients were randomly divided into two groups. The experimental group ( $n = 34$ ) received standard care plus the use of a rocking chair while the control group ( $n = 32$ ) received only standard care. For each patient, the postoperative time until first flatus (days) (an indication that the POI has ended) was measured. The results are tabulated here.<sup>41</sup>

	TIME UNTIL FIRST FLATUS (DAYS)		
	n	MEAN (DAYS)	SD
Rocking	34	3.16	0.86
Control	32	3.88	0.80

- (a) Is there evidence that use of the rocking chair reduces POI duration (i.e., the time until first flatus)? Use a  $t$  test with a directional alternative with  $\alpha = 0.05$ .
- (b) While the researchers hypothesized that the use of a rocking chair could reduce POI duration, it is not unreasonable to hypothesize that the use of a rocking chair could increase POI duration. Based on this possibility, discuss the appropriateness of using a directional versus nondirectional test. (Hint: Consider what medical recommendations might be made based on this research.)

**7.5.12** In Example 7.2.6 we considered testing  $H_0: \mu_1 = \mu_2$  against the nondirectional alternative hypothesis  $H_A: \mu_1 \neq \mu_2$  and found that the  $P$ -value could be

bracketed as  $0.06 < P\text{-value} < 0.10$ . Recall that the sample mean for the group 1 (the control group) was 15.9, which was less than the sample mean of 11.0 for group 2 (the group treated with Ancymidol). However, Ancymidol is considered to be a growth inhibitor, which means that one would expect the control group to have a larger mean than the treatment group if ancy has any effect on the type of plant being studied (in this case, the Wisconsin Fast Plant). Suppose the researcher had expected ancy to retard growth—before conducting the experiment—and had conducted a test of  $H_0: \mu_1 = \mu_2$  against the nondirectional alternative hypothesis  $H_A: \mu_1 > \mu_2$ , using  $\alpha = 0.05$ . What would be the bounds on the  $P$ -value? Would  $H_0$  be rejected? Why or why not? What would be the conclusion of the experiment? (Note: This problem requires almost no calculation.)

**7.5.13 (Computer exercise)** An ecologist studied the habitat of a marine reef fish, the six bar wrasse (*Thalassoma hardwicke*), near an island in French Polynesia that is surrounded by a barrier reef. He examined 48 patch reef settlements at each of two distances from the reef crest: 250 meters from the crest and 800 meters from the crest. For each patch reef, he calculated the “settler density,” which is the number of settlers (juvenile fish) per unit of settlement habitat. Before collecting the data, he hypothesized that the settler density might decrease as distance from the reef crest increased, since the way that waves break over the reef crest causes resources (i.e., food) to tend to decrease as distance from the reef crest increases. Here are the data.<sup>42</sup>

250 METERS			800 METERS		
0.318	0.758	0.318	0.941	0.289	0.399
0.637	0.372	0.524	0.279	0.392	0.955
0.196	0.637	1.404	1.021	0.725	0.531
0.624	1.560	0.000	0.108	1.318	0.252
0.909	0.207	1.061	0.738	0.612	1.179
0.295	0.685	0.590	0.907	0.637	0.442
0.594	0.000	0.363	0.503	0.181	0.291
0.442	1.303	1.567	0.637	0.941	0.579
1.220	0.898	1.577	1.498	0.265	0.252
1.303	1.157	0.312	0.866	0.979	0.373
0.187	0.970	0.758	0.588	0.909	0.000
1.560	0.624	0.505	0.606	0.283	0.463
0.849	1.592	0.909	0.490	0.337	1.248
2.411	1.019	0.362	0.163	0.813	2.010
1.705	0.829	0.329	0.277	0.000	1.213
1.019	0.884	0.909	0.293	0.544	0.808

For 250 meters, the sample mean is 0.818 and the sample SD is 0.514. For 800 meters, the sample mean is 0.628 and the sample SD is 0.413. Do these data provide statistically significant evidence, at the 0.10 level, to support the ecologist’s theory? Investigate with an appropriate graph and test.

## 7.6 More on Interpretation of Statistical Significance

Ideally, statistical analysis should aid the researcher by helping to clarify whatever message is contained in the data. For this purpose, it is not enough that the statistical calculations be correct; the results must also be correctly interpreted. In this section we explore some principles of interpretation that apply not only to the  $t$  test, but also to other statistical tests to be discussed later.

### Significant Difference versus Important Difference

The term *significant* is often used in describing the results of a statistical analysis. For example, if an experiment to compare a drug against a placebo gave data with a very small  $P$ -value, then the conclusion might be stated as “The effect of the drug was highly significant.” As another example, if two fertilizers for wheat gave a yield comparison with a large  $P$ -value, then the conclusion might be stated as “The wheat yields did not differ significantly between the two fertilizers” or “The difference between the fertilizers was not significant.” As a third example, suppose a substance is tested for toxic effects by comparing exposed animals and control animals, and that the null hypothesis of no difference is not rejected. Then the conclusion might be stated as “No significant toxicity was found.”

Clearly such phraseology using the term *significant* can be seriously misleading. After all, in ordinary English usage, the word significant connotes “substantial” or “important.” In statistical jargon, however, the statement

“The difference was significant”

means nothing more or less than

“The null hypothesis of no difference was rejected.”

This is to say, “We found sufficient evidence that the difference in sample means was not caused by chance error alone.”

By the same token, the statement

“The difference was not significant”

means

“There was not sufficient evidence that the observed difference in means was due to anything other than chance variation.”

It would perhaps be preferable if a different word were used in place of “significant,” such as “discernible” (meaning that the test discerned a difference). Alas, the specialized usage of the word *significant* has become quite common in scientific writing and understandably is the source of much confusion.

It is essential to recognize that a statistical test provides information about only one question: Is the difference observed in the data large enough to infer that a difference in the same direction exists in the population? The question of whether a difference is *important*, as opposed to (statistically) significant, cannot be decided on the basis of the *P*-values alone but must also include an examination of the magnitude of the estimated population difference as well as specific expertise in the research area or practical situation. The following two examples illustrate this fact.

**Example  
7.6.1**

**Serum LD** Lactate dehydrogenase (LD) is an enzyme that may show elevated activity following damage to the heart muscle or other tissues. A large study of serum LD levels in healthy young people yielded the results shown in Table 7.6.1.<sup>43</sup>

<b>Table 7.6.1 Serum LD (U/l)</b>		
	Males	Females
<i>n</i>	270	264
$\bar{y}$	60	57
<i>s</i>	11	10

The difference between males and females is quite significant; in fact,  $t_s = 3.3$ , which gives a *P*-value  $\approx 0.001$ . However, this does not imply that the difference ( $60 - 57 = 3$  U/l) is large or important in any practical sense. ■

**Example  
7.6.2**

**Body Weight** Imagine that we are studying the body weight of men and women, and we obtain the fictitious but realistic data shown in Table 7.6.2.<sup>44</sup>

<b>Table 7.6.2</b> Body weight (lb)		
	Males	Females
<i>n</i>	2	2
$\bar{y}$	175	143
<i>s</i>	35	34

For these data the *t* test gives  $t_s = 0.93$  and a *P*-value  $\approx 0.45$ . The observed difference between males and females is not small (it is  $175 - 143 = 32$  lb), yet it is not statistically significant for any reasonable choice of  $\alpha$ . The lack of statistical significance does not imply that the sex difference in body weight is small or unimportant. It means only that the data are inadequate to characterize the difference in the population means. A sample difference of 32 lb could easily happen by chance if the two populations are identical, especially with such small sample sizes. ■

## Effect Size

The preceding examples show that the statistical significance or nonsignificance of a difference does not indicate whether the difference is important. Nevertheless, the question of “importance” can and should be addressed in most data analyses. To assess importance, one needs to consider the *magnitude* of the difference. In Example 7.6.1 the male versus female difference is “statistically significant,” but this is largely due to the sample sizes being quite large. A *t* test uses the test statistic

$$t_s = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

If  $n_1$  and  $n_2$  are large, then  $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$  will be small and the test statistic will tend to be large even when the difference in observed means ( $\bar{Y}_1 - \bar{Y}_2$ ) is very small. Thus, one might find significant evidence for  $H_A$  due to the sample size being large, even if  $\mu_1$  and  $\mu_2$  are nearly equal. The sample size acts like a magnifying glass: *The larger the sample size, the smaller the difference that can be detected in a hypothesis test.*

The **effect size** in a study is the difference between  $\mu_1$  and  $\mu_2$ , expressed relative to the standard deviation of one of the populations. If the two populations have the same standard deviation,  $\sigma$ , then the effect size is\*

$$\text{Effect size} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

Of course, when working with sample data we can only calculate an *estimated* effect size by using sample values in place of the unknown population values.

### Example 7.6.3

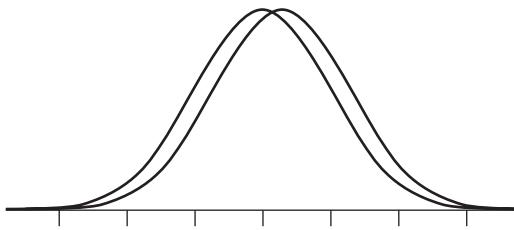
**Serum LD** For the data given in Example 7.6.1 the difference in sample means,  $60 - 57 = 3$ , is less than one-third of a standard deviation. Using the larger sample SD we can calculate a sample effect size of

$$\text{Effect size} = \frac{|\bar{y}_1 - \bar{y}_2|}{s} = \frac{60 - 57}{35} = 0.27$$

---

\*If the standard deviations are not equal, we can use the larger SD in defining the effect size.

**Figure 7.6.1** Overlap between two normally distributed populations when the effect size is 0.27



This indicates that there is a lot of overlap between the two groups. Figure 7.6.1 shows the extent of the overlap that occurs if two normally distributed populations differ on average by 0.27 SDs.

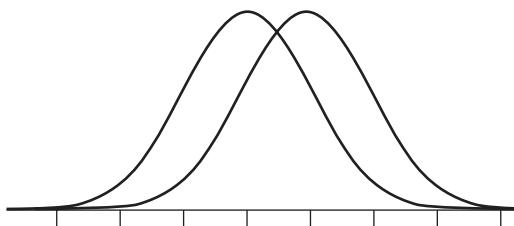
**Example  
7.6.4**

**Body Weight** For the data given in Example 7.6.2 the difference in sample means,  $175 - 143 = 32$ , is roughly one standard deviation. The sample effect size is

$$\text{Effect size} = \frac{|\bar{y}_1 - \bar{y}_2|}{s} = \frac{175 - 143}{35} = 0.91$$

Figure 7.6.2 shows the extent of the overlap that occurs if two normally distributed populations differ on average by 0.91 SD.

**Figure 7.6.2** Overlap between two normally distributed populations when the effect size is 0.91



The definition of effect size that we are using is probably unfamiliar to the biologically oriented reader. It is more common in biology to “standardize” a difference of two quantities by expressing it as a percentage of one of them. For example, the weight difference given in Table 7.6.2 between males and females, expressed as a percentage of mean female weight, is

$$\frac{\bar{y}_1 - \bar{y}_2}{\bar{y}_2} = \frac{175 - 143}{143} = 0.22 \text{ or } 22\%$$

Thus, the males are about 22% heavier than the females. However, from a statistical viewpoint it is often more relevant that the average weights for males and females are 0.91 SD apart.

### Confidence Intervals to Assess Importance

Calculating the effect size is one way to quantify how far apart two sample means are. Another reasonable approach is to use the observed difference ( $\bar{Y}_1 - \bar{Y}_2$ ) to construct a confidence interval for the population difference ( $\mu_1 - \mu_2$ ). In interpreting the confidence interval, the judgment of what is “important” is made on the basis of experience with the particular practical situation. The following three examples illustrate this use of confidence intervals.

**Example  
7.6.5**

**Serum LD** For the LD data of Example 7.6.1, a 95% confidence interval for  $(\mu_1 - \mu_2)$  is

$$3 \pm 1.8$$

or

$$(1.2, 4.8)$$

This interval implies (with 95% confidence) that the population difference in means between the sexes does not exceed 4.8 U/l. As an expert, a physician evaluating this information would know that typical day-to-day fluctuation in a person's LD level is around 6.5 U/l, which is higher than 4.8 U/l, the highest we estimate the mean sex difference to be, and therefore this difference is negligible from the medical standpoint. Consequently, the physician might conclude that it is unnecessary to differentiate between the sexes in establishing clinical thresholds for diagnosis of illness. In this case, the sex difference in LD may be said to be statistically significant but medically unimportant. To put this another way, the data suggest that men do in fact tend to have higher levels than women, but not higher in any clinically useful way. ■

**Example  
7.6.6**

**Body Weight** For the body-weight data of Example 7.6.2, a 95% confidence interval for  $(\mu_1 - \mu_2)$  is

$$32 \pm 149$$

or

$$(-117, 181)$$

From this confidence interval we cannot tell whether the true difference (between the population means) is large favoring females, is small, or is large favoring males. Because the confidence interval contains numbers of both small and large magnitude, it does not tell us whether the difference between the sexes is important or unimportant. With such a wide confidence interval a researcher would likely wish to conduct a larger study to better assess the importance of the difference. Suppose, for example, that the means and standard deviations were as given in Table 7.6.2, but that they were based on 2,000 rather than 2 people of each sex. Then the 95% confidence interval would be

$$32 \pm 2$$

or

$$(30, 34)$$

This interval would imply (with 95% confidence) that the difference is at least 30 lb, an amount that might reasonably be regarded as important, at least for some purposes. ■

**Example  
7.6.7**

**Yield of Tomatoes** Suppose a horticulturist is comparing the yields of two varieties of tomatoes; yield is measured as pounds of tomatoes per plant. On the basis of practical considerations, the horticulturist has decided that a difference between the varieties is "important" only if it exceeds 1 pound per plant, on the average. That is, the difference is important if

$$|\mu_1 - \mu_2| > 1.0 \text{ lb}$$

Suppose the horticulturist's data give the following 95% confidence interval:

$$(0.2, 0.3)$$

Because the largest estimate for the population difference is only 0.3 lb (all values in the interval are less than 1.0 lb), the data support (with 95% confidence) the assertion that the difference is *not* important, using the horticulturist's criterion. ■

In many investigations, statistical significance and practical importance are both of interest. The following example shows how the relationship between these two concepts can be visualized using confidence intervals.

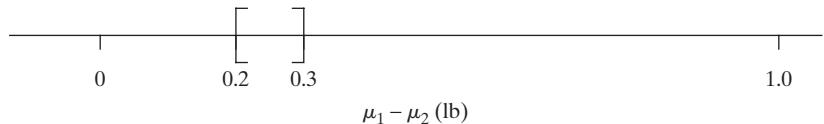
**Example  
7.6.8**

**Yield of Tomatoes** Let us return to the tomato experiment of Example 7.6.7. The confidence interval was

$$(0.2, 0.3)$$

Recall from Section 7.3 that the confidence interval can be interpreted in terms of a  $t$  test. Because all values within the confidence interval are positive, a  $t$  test (two-tailed) at  $\alpha = 0.05$  finds significant evidence for  $H_A$ . Thus, the difference between the two varieties is statistically significant, although it is not horticulturally important: The data indicate that variety 1 is better than variety 2, but also that it is not much better. The distinction between significance and importance for this example can be seen in Figure 7.6.3, which shows the confidence interval plotted on the  $(\mu_1 - \mu_2)$ -axis. Note that the confidence interval lies entirely to one side of zero and also entirely to one side of the “importance” threshold of 1.0.

**Figure 7.6.3** Confidence interval for Example 7.6.8



To further explore the relationship between significance and importance, let us consider other possible outcomes of the tomato experiment. Table 7.6.3 shows how the horticulturist would interpret various possible confidence intervals, still using the criterion that a difference must exceed 1.0 lb in order to be considered important.

**Table 7.6.3** Interpretation of confidence intervals

95% confidence interval	Is the difference significant?	important?
(0.2, 0.3)	Yes	No
(1.2, 1.3)	Yes	Yes
(0.2, 1.3)	Yes	Cannot tell
(-0.2, 0.3)	No	No
(-1.2, 1.3)	No	Cannot tell

Table 7.6.3 shows that a significant difference may or may not be important, and an important difference may or may not be significant. In practice, the assessment of importance using confidence intervals is a simple and extremely useful supplement to a test of hypothesis. ■

## Exercises 7.6.1–7.6.8

---

**7.6.1** A field trial was conducted to evaluate a new seed treatment that was supposed to increase soybean yield. When a statistician analyzed the data, the statistician found that the mean yield from the treated seeds was 40 lb/acre greater than that from control plots planted with untreated seeds. However, the statistician declared the difference to be “not (statistically) significant.” Proponents of the treatment objected strenuously to the statistician’s statement, pointing out that, at current market prices, 40 lb/acre would bring a tidy sum, which would be highly significant to the farmer. How would you answer this objection?<sup>45</sup>

**7.6.2** In a clinical study of treatments for rheumatoid arthritis, patients were randomly allocated to receive either a standard medication or a newly designed medication. After a suitable period of observation, statistical analysis showed that there was no significant difference in the therapeutic response of the two groups, but that the incidence of undesirable side effects was significantly lower in the group receiving the new medication. The researchers concluded that the new medication should be regarded as clearly preferable to the standard medication, because it had been shown to be equally effective therapeutically and to produce fewer side effects. In what respect is the researchers’ reasoning faulty? (Assume that the term “significant” refers to rejection of  $H_0$  at  $\alpha = 0.05$ .)

**7.6.3** There is an old folk belief that the sex of a baby can be guessed before birth on the basis of its heart rate. In an investigation to test this theory, fetal heart rates were observed for mothers admitted to a maternity ward. The results (in beats per minute) are summarized in the table.<sup>46</sup>

HEART RATE (bpm)			
	n	Mean	SE
Males	250	137.21	0.62
Females	250	137.18	0.53

Construct a 95% confidence interval for the difference in population means. Does the confidence interval support the claim that the population mean sex difference (if any) in fetal heart rates is small and unimportant? (Use your own “expert” knowledge of heart rate to make a judgment of what is “unimportant.”)

**7.6.4** Coumaric acid is a compound that may play a role in disease resistance in corn. A botanist measured the concentration of coumaric acid in corn seedlings grown

in the dark or in a light/dark photoperiod. The results (nmol acid per gm tissue) are given in the accompanying table.<sup>47</sup> [Note: Formula (6.7.1) yields 5.7 df.]

	DARK	PHOTOPERIOD
n	4	4
$\bar{y}$	106	102
s	21	27

Suppose the botanist considers the effect of lighting conditions to be “important” if the difference in means is 20%, that is, about 20 nmol/g. Based on a 95% confidence interval, do the preceding data indicate whether the true difference is “important”?

**7.6.5** Repeat Exercise 7.6.4, assuming that the means and standard deviations are as given in the table, but that the sample sizes are 10 times as large (that is,  $n = 40$  for “dark” and  $n = 40$  for “photoperiod”). [Note: Formula (6.7.1) yields 73.5 df.]

**7.6.6** Researchers measured the breadths, in mm, of the ankles of 460 youth (ages 11–16); the results are shown in the table.<sup>48</sup>

	MALES	FEMALES
n	244	216
$\bar{y}$	55.3	53.3
s	6.1	5.4

Calculate the sample effect size from these data.

**7.6.7** As part of a large study of serum chemistry in healthy people, the following data were obtained for the serum concentration of uric acid in men and women aged 18–55 years.<sup>49</sup>

	SERUM URIC ACID (mmol/l)	
	MEN	WOMEN
n	530	420
$\bar{y}$	0.354	0.263
s	0.058	0.051

Construct a 95% confidence interval for the true difference in population means. Suppose the investigators feel that the difference in population means is “clinically

important" if it exceeds 0.08 mmol/l. Does the confidence interval indicate whether the difference is "clinically important"? [Note: Formula (6.7.1) yields 934 df.]

**7.6.8** Repeat Exercise 7.6.7, assuming that the means and standard deviations are as given in the table, but that the sample sizes are only one-tenth as large (that is, 53 men and 42 women). [Note: Formula (6.7.1) yields 92 df.]

## 7.7 Planning for Adequate Power (Optional)

We have defined the power of a statistical test as

$$\text{Power} = \Pr\{\text{significant evidence for } H_A \text{ if } H_A \text{ is true}\}$$

To put this another way, the power of a test is the probability of obtaining data that provide statistically significant evidence for  $H_A$  when  $H_A$  is true.

Since the power is the probability of *not* making an error (of Type II), high power is desirable: If  $H_A$  is true, a researcher would like to find that out when conducting a study. But power comes at a price. All other things being equal, more observations (larger samples) bring more power, but observations cost time and money. In this section we explain how a researcher can rationally plan an experiment to have adequate power for the purposes of the research project and yet cost as little as possible.

Specifically, we will consider the power of the two-sample  $t$  test, conducted at significance level  $\alpha$ . We will assume that the populations are normal with equal SDs, and we denote the common value of the SD by  $\sigma$  (that is,  $\sigma_1 = \sigma_2 = \sigma$ ). It can be shown that in this case, for a given total sample size of  $2n$ , the power is maximized if the sample sizes are equal; thus we will assume that  $n_1$  and  $n_2$  are equal and denote the common value by  $n$  (that is,  $n_1 = n_2 = n$ ).

Under the above conditions, the power of the  $t$  test depends on the following factors: (a)  $\alpha$ ; (b)  $\sigma$ ; (c)  $n$ ; and (d)  $(\mu_1 - \mu_2)$ . After briefly discussing each of these factors, we will address the all-important question of choosing the value of  $n$ .

### Dependence of Power on $\alpha$

In choosing  $\alpha$ , one chooses a level of protection against Type I error. However, this protection is traded for vulnerability to Type II error. If, for example, one chooses  $\alpha = 0.01$  rather than  $\alpha = 0.05$ , then one is requiring stronger evidence for  $H_A$  before choosing to claim there is significant evidence for  $H_A$ , and so is (perhaps unwittingly) also choosing to increase the risk of Type II error and reduce the power. Thus, there is an unavoidable trade-off between the risk of Type I error and the risk of Type II error.

### Dependence on $\sigma$

The larger  $\sigma$ , the smaller the power (all other things being equal). Recall from Chapter 5 that the reliability of a sample mean is determined by the quantity

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

The larger  $\sigma$  is, the more variability there is in the sample mean. Thus, having a larger  $\sigma$  implies having samples that produce less reliable information about each

population mean, and so less power to discern a difference between them. In order to increase power, then, a researcher usually tries to design the investigation so as to have  $\sigma$  as small as possible. For example, a botanist will try to hold light conditions constant throughout a greenhouse area, a pharmacologist will use genetically identical experimental animals, and so on. Usually, however,  $\sigma$  cannot be reduced to zero; there is still considerable variation in the observations.

### Dependence on $n$

The larger  $n$ , the higher the power (all other things being equal). If we increase  $n$ , we decrease  $\sigma/\sqrt{n}$ ; this improves the precision of the sample means ( $\bar{Y}_1$  and  $\bar{Y}_2$ ). In addition, larger  $n$  gives more information about  $\sigma$ ; this is reflected in a reduced critical value for the test (reduced because of more df). Thus, increasing  $n$  increases the power of the test in two ways.

### Dependence on $(\mu_1 - \mu_2)$

In addition to the factors we have discussed, the power of the  $t$  test also depends on the actual difference between the population means, that is, on  $(\mu_1 - \mu_2)$ . This dependence is very natural, as illustrated by the following example.

---

**Example**  
**7.7.1**

**Heights of People** In order to clearly illustrate the concepts, we consider a familiar variable, body height of people. Imagine what would happen if an investigator were to measure the heights of two random samples of eleven people each ( $n = 11$ ), and then conduct a two-tailed  $t$  test at  $\alpha = 0.05$ .

- First, suppose that sample 1 consisted of 17-year-old males and sample 2 consisted of 17-year-old females. The two population means differ substantially; in fact,  $(\mu_1 - \mu_2)$  is about 5 inches ( $\mu_1 \approx 69.1$  and  $\mu_2 \approx 64.1$  inches).<sup>50</sup> It can be shown (as we will see) that in this case the investigator has about a 99% chance of obtaining significant evidence for a difference (i.e.,  $H_A$ ) and correctly concluding that the males in the population of 17-year-olds are taller (on average) than the females.
- By contrast, suppose that sample 1 consisted of 17-year-old females and sample 2 consisted of 14-year-old females. The two population means differ, but by a modest amount; the difference is  $(\mu_1 - \mu_2) = 0.6$  inches ( $\mu_1 \approx 64.1$  and  $\mu_2 \approx 63.5$  inches). It can be shown that in this case the investigator has less than a 10% chance of obtaining significant evidence of a difference (i.e.,  $H_A$ ); in other words, there is more than a 90% chance that the investigator will fail to detect the fact that 17-year-old girls are taller than 14-year-old girls. (In fact, it can be shown that there is a 29% chance that  $\bar{Y}_1$  will be less than  $\bar{Y}_2$ —that is, there is a 29% chance that eleven 17-year-old girls chosen at random will be shorter on the average than eleven 14-year-old girls chosen at random!)

The contrast between cases (a) and (b) is not due to any change in the SDs; in fact, for each of the three populations the value of  $\sigma$  is about 2.5 inches. Rather, the contrast is due to the simple fact that, with a fixed  $n$  and  $\sigma$ , it is easier to detect a large difference than a small difference. ■

## Planning a Study

Suppose an investigator is planning a study for which the  $t$  test will be appropriate. How shall she take into account all the factors that influence the power of the test? First consider the choice of significance level  $\alpha$ . A simple approach is to begin by determining the cost of an adequately powerful study using a somewhat liberal choice (say,  $\alpha = 0.05$  or  $0.10$ ). If that cost is not high, the investigator can consider reducing  $\alpha$  (say, to  $0.01$ ) and see if an adequately powerful study is still affordable.

Suppose, then, that the investigator has chosen a working value of  $\alpha$ . Suppose also that the experiment has been designed to reduce  $\sigma$  as far as practicable, and that the investigator has available an estimate or guess of the value of  $\sigma$ .

At this point, the investigator needs to ask herself about the magnitude of the difference she wants to detect. As we saw in Example 7.7.1, a given sample size may be adequate to detect a large difference in population means, but entirely inadequate to detect a small difference. As a more realistic example, an experiment using 5 rats in a treatment group and 5 rats in a control group might be large enough to detect a substantial treatment effect, while detection of a subtle treatment effect would require more rats (perhaps 30) in each group.

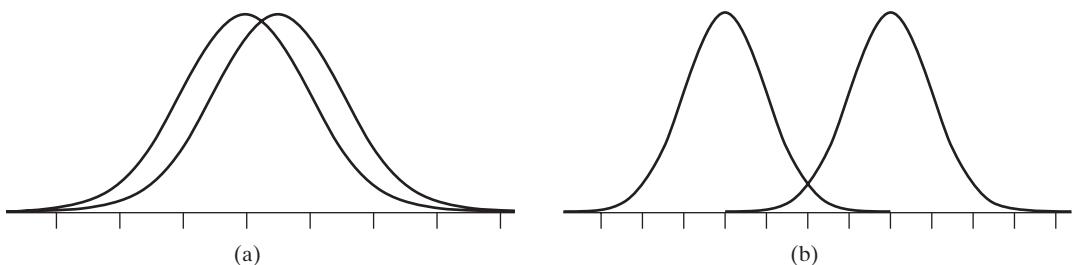
The preceding discussion suggests that choosing a sample size for adequate power is somewhat analogous to choosing a microscope: We need high resolving power if we want to see a very tiny structure; for large structures a hand lens will do. In order to proceed with planning the experiment, the investigator needs to decide how large an effect she is looking for.

Recall that in Section 7.7, we defined the effect size in a study as the difference between  $\mu_1$  and  $\mu_2$ , expressed relative to the standard deviation of one of the populations. If, as we are assuming here, the two populations have the same standard deviation,  $\sigma$ , then the effect size is

$$\text{Effect size} = \frac{|\mu_1 - \mu_2|}{\sigma}$$

That is, the effect size is the difference in population means expressed relative to the common population SD. The effect size is a kind of “signal to noise ratio,” where  $(\mu_1 - \mu_2)$  represents the signal we want to detect and  $\sigma$  represents the background noise that tends to obscure the signal. Figure 7.7.1(a) shows two normal curves for which the effect size is 0.5; Figure 7.7.1(b) shows two normal curves for which the effect size is 4. Clearly, at a fixed sample size it is easier to detect the difference between the curves in graph (b) than it is in graph (a).

If  $\alpha$  and the effect size have been specified, then the power of the  $t$  test depends only on the sample sizes ( $n$ ). Table 5 at the end of the book shows the value of  $n$



**Figure 7.7.1** Normal distributions with an effect size (a) of 0.5 and (b) of 4

required in order to achieve a specified power against a specified effect size. Let us see how Table 5 applies to our familiar example of body height.

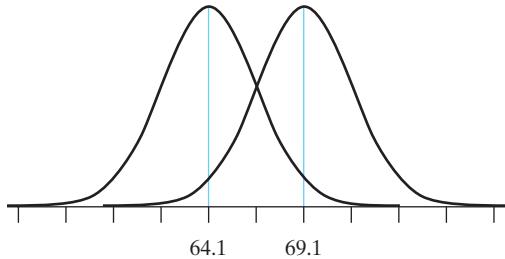
**Example  
7.7.2**

**Heights of People** In Example 7.7.1, case (a), we considered samples of 17-year-old males and 17-year-old females. The effect size is

$$\frac{|\mu_1 - \mu_2|}{\sigma} = \frac{|69.1 - 64.1|}{2.5} = \frac{5}{2.5} = 2.0$$

For a two-tailed  $t$  test at  $\alpha = 0.05$ , Table 5 shows that the sample size required for a power of 0.99 is  $n = 11$ ; this is the basis for the claim in Example 7.7.1 that the investigator has a 99% chance of detecting the difference between males and females. Figure 7.7.2 shows the two distributions being considered in Example 7.7.2. Suppose 100 researchers each conduct the following study. Take a random sample of eleven 17-year-old males and a random sample of eleven 17-year-old females, find the sample average heights of the two groups, and then conduct a two-tailed  $t$  test of  $H_0: \mu_1 = \mu_2$  using  $\alpha = 0.05$ . We would expect 99 of the 100 researchers to find statistically significant evidence that the average heights of 17-year-old males and females differ (i.e., significant evidence for  $H_A$ ). We would expect one of the 100 researchers to not find sufficient evidence for a difference, at the 0.05 level of significance. (So one researcher would make a Type II error.) ■

**Figure 7.7.2** Height distributions for Example 7.7.2



As we have seen, in order to choose a sample size the researcher needs to specify not only the size of the effect she wishes to detect, but also how certain she wants to be of detecting it; that is, it is necessary to specify how much power is wanted. Since the power measures the protection against Type II error, the choice of a desired power level depends on the consequences that would result from a Type II error. If the consequences of a Type II error would be very unfortunate (for example, if a promising but risky cancer treatment is being tested on humans and a negative result would discredit the treatment so that it would never be tested again), then the researcher might specify a high power, say 0.95 or 0.99. But of course high power is expensive in terms of  $n$ . For much research, a Type II error is not a disaster, and a lower power such as 0.80 is considered adequate.

The following example illustrates a typical use of Table 5 in planning an experiment.

**Example  
7.7.3**

**Postpartum Weight Loss** A group of scientists wished to investigate whether or not an Internet-based intervention program would help women lose weight after giving birth. One group of postpartum women was to be enrolled in an Internet-based program that provides weekly exercise and dietary guidance appropriate to their

time since giving birth, track their weight-loss progress, and establish an online forum for nutrition and exercise discussion with other recent mothers. Another group of postpartum women (the “control group”) was to be given traditional written dietary and exercise guidelines by their doctors. The response variable for the study was to be the amount of weight lost at 12 months postpartum in kg. Previous studies have shown that at 12 months postpartum, the mean weight loss is about 3.6 kg with a standard deviation of 4.0 kg. (*Note:* A negative weight loss is a weight gain). The research team wanted to show at least a 50% improvement in weight loss for the Internet-intervention group; that is, they would like to show that the Internet-based program women lose at least 1.8 kg (50% of 3.6kg) more weight than the controls. They planned to conduct a one-tailed *t*-test at the 5% significance level. The team had to decide how many women (*n*) to put in each group.

The effect size that the team wanted to consider is

$$\frac{|\mu_1 - \mu_2|}{\sigma} = \frac{1.8}{4.0} = 0.45$$

For this effect size, and for a power of 0.80 with a one-tailed test at the 5% significance level, Table 5 yields *n* = 62, which means 62 women were needed in each group.

At this point, the research team had to consider questions, such as (1) Is it feasible to enroll 124 postpartum women (62 in each group) in the study? If not, then (2) Would they perhaps be willing to redefine the size of the difference between the groups that they considered to be important, in order to reduce the required *n*? With questions such as these, and repeated use of Table 5, they could finally decide on a firm value for *n*, or possibly decide to abandon the project because an adequate study would be too costly.

Normally the story ends here, but there was an extra wrinkle in the planning of this study: The research team knew from experience that about 20% of the women enrolled in these types of studies would drop out, for one reason or another, before the study ended. (There is no formula or table that tells one how many subjects will drop out of a study such as this. Here the only guide is experience.) In this case, the research team planned to enroll 150 women (a little more than 20% extra, 13 women in each group), in order to allow for some attrition and still end up with enough data so that they would have the power they wanted.<sup>51</sup>

## Exercises 7.7.1–7.7.11

**7.7.1** One measure of the meat quality of pigs is backfat thickness. Suppose two researchers, Jones and Smith, are planning to measure backfat thickness in two groups of pigs raised on different diets. They have decided to use the same number (*n*) of pigs in each group, and to compare the mean backfat thickness using a two-tailed *t* test at the 5% significance level. Preliminary data indicate that the SD of backfat thickness is about 0.3 cm.

When the researchers approach a statistician for help in choosing *n*, she naturally asks how much difference they want to detect. Jones replies, “If the true difference is 1/4 cm or more, I want to be reasonably sure of rejecting  $H_0$ .” Smith replies, “If the true difference is 1/2 cm or more, I want to be very sure of rejecting  $H_0$ .”

If the statistician interprets “reasonably sure” as 80% power, and “very sure” as 95% power, what value of *n* will she recommend

- (a) to satisfy Jones’s requirement?
- (b) to satisfy Smith’s requirement?

**7.7.2** Refer to the brain NE data of Example 7.2.1. Suppose you are planning a similar experiment; you will study the effect of LSD (rather than toluene) on brain NE. You anticipate using a two-tailed *t* test at  $\alpha = 0.05$ . Suppose you have decided that a 10% effect (increase or decrease in mean NE) of LSD would be important, and so you want to have good power (80%) to detect a difference of this magnitude.

- (a) Using the data of Example 7.2.1 as a “pilot study,” determine how many rats you should have in each group. (The mean NE in the control group in Example 7.2.1 is 444.2 ng/g and the SD is = 69.6 ng/g.)

(b) If you were planning to use a one-tailed  $t$  test, what would be the required number of rats?

**7.7.3** Suppose you are planning a greenhouse experiment on growth of pepper plants. You will grow  $n$  individually potted seedlings in standard soil and another  $n$  seedlings in specially treated soil. After 21 days, you will measure  $Y = \text{total stem length (cm)}$  for each plant. If the effect of the soil treatment is to increase the population mean stem length by 2 cm, you would like to have a 90% chance of rejecting  $H_0$  with a one-tailed  $t$  test. Data from a pilot study (such as the data in Exercise 2.62) on 15 plants grown in standard soil give  $\bar{y} = 12.5$  cm and  $s = 0.8$  cm.

- (a) Suppose you plan to test at  $\alpha = 0.05$ . Use the pilot information to determine what value of  $n$  you should use.
  - (b) What conditions are necessary for the validity of the calculation in part (a)? Which of these can be checked (roughly) from the data of the pilot study?
  - (c) Suppose you decide to adopt a more conservative posture and test at  $\alpha = 0.01$ . What value of  $n$  should you use?

**7.7.4** Diastolic blood pressure measurements on American men aged 18–44 years follow approximately a normal curve with  $\mu = 81$  mm Hg and  $\sigma = 11$  mm Hg. The distribution for women aged 18–44 is also approximately normal with the same SD but with a lower mean:  $\mu = 75$  mm Hg.<sup>52</sup> Suppose we are going to measure the diastolic blood pressure of  $n$  randomly selected men and  $n$  randomly selected women in the age group 18–44 years. Let  $E$  be the event that the difference between men and women will be found statistically significant by a  $t$  test. How large must  $n$  be in order to have  $\Pr[E] = 0.9$

- (a) if we use a two-tailed test at  $\alpha = 0.05$ ?
  - (b) if we use a two-tailed test at  $\alpha = 0.01$ ?
  - (c) if we use a one-tailed test (in the correct direction) at  $\alpha = 0.05$ ?

**7.7.5** Suppose you are planning an experiment to test the effect of a certain drug treatment on drinking behavior in the rat. You will use a two-tailed  $t$  test to compare a treated group of rats against a control group; the observed variable will be  $Y$  = one-hour water consumption after 23-hour deprivation. You have decided that, if the effect of the drug is to shift the population mean consumption by 2 ml or more, then you want to have at least an 80% chance of finding significant evidence for  $H_A$  at the 5% significance level.

- (a) Preliminary data indicate that the SD of  $Y$  under control conditions is approximately 2.5 ml. Using this as a guess of  $\sigma$ , determine how many rats you should have in each group.

- (b) Suppose that, because the calculation of part (a) indicates a rather large number of rats, you consider modifying the experiment so as to reduce  $\sigma$ . You find that, by switching to a better supplier of rats and by improving lab procedures, you could cut the SD in half; however, the cost of each observation would be doubled. Would these measures be cost-effective; that is, would the modified experiment be less costly?

**7.7.6** Data from a large study indicate that the serum concentration of lactate dehydrogenase (LD) is higher in men than in women. (The data are summarized in Example 7.6.1.) Suppose Dr. Sanchez proposes to conduct his own study to replicate this finding; however, because of limited resources Sanchez can enlist only 35 men and 35 women for his study. Supposing that the true difference in population means is 4 U/l and each population SD is 10 U/l, what is the probability that Sanchez will be successful? Specifically, find the probability that Sanchez will reject  $H_0$  with a one-tailed  $t$  test at the 5% significance level.

**7.7.7** Refer to the painkiller study of Exercise 7.5.10. That study included 25 observations in each treatment group and showed an effect size of about 0.5. If this is the true population effect size, what is the (approximate) chance of finding a significant difference between the mean effectiveness of the two drugs in an experiment of this size (i.e., with samples of 25 each)?

**7.7.8** Refer to the painkiller study of Exercise 7.5.10. In that study, the evidence favoring the drug was marginally significant ( $0.025 < P < 0.05$ ). Suppose Dr. Williams is planning a new study on the same drug in order to try to replicate the original findings, that is, to show the drug to be effective. She will consider this study successful if she rejects  $H_0$  with a one-tailed test at  $\alpha = 0.05$ . In the original study, the difference between the treatment means was about half a standard deviation  $[(32 - 25)/13 \approx 0.5]$ . Taking this as a provisional value for the effect size, determine how many patients Williams should have in each group in order for her chance of success to be

- (a) 80% (b) 90%

(Note: This problem illustrates that surprisingly large sample sizes may be required to make a replication study worthwhile, especially if the original findings were only marginally significant.)

**7.7.9** Consider comparing two normally distributed distributions for which the effect size of the difference is

- (a) 3 (b) 1

In each case, draw a sketch that shows how the distributions overlap. (See Figure 7.2.1.)

**7.7.10** An animal scientist is planning an experiment to evaluate a new dietary supplement for beef cattle. One group of cattle will receive a standard diet and a second group will receive the standard diet plus the supplement.

The researcher wants to have 90% power to detect an increase in mean weight gain of 20 kg, using a one-tailed *t* test at  $\alpha = 0.05$ . Based on previous experience, he expects the SD to be 17 kg. How many cattle does he need for each group?

**7.7.11** A researcher is planning to conduct a study that will be analyzed with a two-tailed *t* test at the 5% significance level. She can afford to collect 20 observations in each of the two groups in her study. What is the smallest effect size for which she has at least 95% power?

## 7.8 Student's *t*: Conditions and Summary

In the preceding sections we have discussed the comparison of two means using classical methods based on Student's *t* distribution. In this section we describe the conditions on which these methods are based. In addition, we summarize the methods for convenient reference.

### Conditions

The *t* test and confidence interval procedures we have described are appropriate if the following conditions\* hold:

1. *Conditions on the design of the study*
  - (a) It must be reasonable to regard the data as random samples from their respective populations. The populations must be large relative to their sample sizes. The observations within each sample must be independent.
  - (b) The two samples must be independent of each other.
2. *Condition on the form of the population distributions*

The sampling distributions of  $\bar{Y}_1$  and  $\bar{Y}_2$  must be (approximately) normal. This can be achieved via normality of the populations or by appealing to the Central Limit Theorem (recall Section 6.5) if the populations are nonnormal but the sample sizes are large, where “largeness” depends on the degree of nonnormality of the populations. In many practical situations, moderate sample sizes (say,  $n_1 = 20$ ,  $n_2 = 20$ ) are quite “large” enough. However, we always need to be aware that one or two extreme outliers can have a great effect on the results of any statistical procedure, including the *t* test.

### Verification of Conditions

A check of the preceding conditions should be a part of every data analysis.

A check of condition 1(a) would proceed as for a confidence interval (Section 6.5), with the researcher looking for biases in the experimental design and verifying that there is no hierarchical structure within each sample.

Condition 1(b) means that there must be no pairing or dependency between the two samples. The full meaning of this condition will become clear in Chapters 8 and 9.

Sometimes it is known from previous studies whether the populations can be considered to be approximately normal. In the absence of such information, the normality requirement can be checked by making histograms, normal probability plots,

---

\*Many authors use the word “assumptions” where we are using the word “conditions.”

or Shapiro–Wilk normality tests for each sample separately. Fortunately, the  $t$  test is fairly robust against departures from normality.<sup>53</sup> Usually, only a rather conspicuous departure from normality (outliers, or long straggly tails) should be cause for concern. Moderate skewness has very little effect on the  $t$  test, even for small samples.

## Consequences of Inappropriate Use of Student's $t$

Our discussion of the  $t$  test and confidence interval (in Sections 7.3–7.8) was based on the conditions (1) and (2). Violation of the conditions may render the methods inappropriate.

If the conditions are not satisfied, then the  $t$  test may be inappropriate in two possible ways:

1. It may be invalid in the sense that the actual risk of Type I error is larger than the nominal significance level  $\alpha$ . (To put this another way, the  $P$ -value yielded by the  $t$  test procedure may be inappropriately small.)
2. The  $t$  test may be valid, but less powerful than a more appropriate test.

If the design includes hierarchical structures that are ignored in the analysis, the  $t$  test may be seriously invalid. If the samples are not independent of each other, the usual consequence is a loss of power.

One fairly common type of departure from the condition of normality is for one or both populations to have long straggly tails. The effect of this form of nonnormality is to inflate the SE, and thus to rob the  $t$  test of power.

Inappropriate use of confidence intervals is analogous to that for  $t$  tests. If the conditions are violated, then the confidence interval may not be valid (i.e., too narrow for the prescribed level of confidence), or it may be valid but wider than necessary.

## Other Approaches

Because methods based on Student's  $t$  distribution are not always the most appropriate, statisticians have devised other methods that serve similar purposes. One of these is the Wilcoxon–Mann–Whitney test, which we will describe in Section 7.10. Another approach to the difficulty is to transform the data, for instance, to analyze  $\log(Y)$  or  $\ln(Y)$  instead of  $Y$  itself.

---

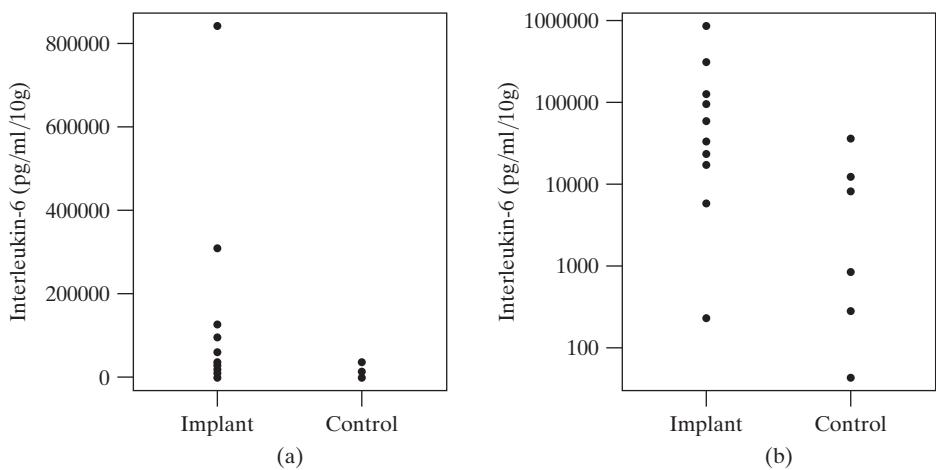
### Example 7.8.1

**Tissue Inflammation** Researchers took skin samples from 10 patients who had breast implants and from a control group of 6 patients. They recorded the level of interleukin-6 (in pg/ml/10 g of tissue), a measure of tissue inflammation, after each tissue sample was cultured for 24 hours. Table 7.8.1 shows the data.<sup>54</sup> Parallel dotplots of these data shown in Figure 7.8.1(a) and normal probability plots shown in Figure 7.8.2(a) indicate that the distributions are severely skewed, so a transformation is needed before Student's  $t$  procedure can be used. Taking the base 10 logarithm of each observation produces the values shown in the right-hand columns of Table 7.8.1 and in Figure 7.8.1(b). The normal probability plots in Figure 7.8.2(b) show that the condition of normality is met after the data have been transformed to log scale. Thus, we will conduct an analysis of the data in log scale. That is, we will test

$$H_0: \mu_1 = \mu_2$$

**Table 7.8.1** Interleukin-6 levels of breast implant patients and control patients

Original data		Log scale	
Breast implant patients	Control patients	Breast implant patients	Control patients
231	35,324	2.364	4.548
308,287	12,457	5.489	4.095
33,291	8,276	4.522	3.918
124,550	44	5.095	1.643
17,075	278	4.232	2.444
22,955	840	4.361	2.924
95,102		4.978	
5,649		3.752	
840,585		5.925	
58,924		4.770	
$\bar{y}$	150,665	4.549	3.262
$s$	259,189	0.992	1.111

**Figure 7.8.1** Dotplots of tissue inflammation data from Example 7.8.1 (a) in the original scale; (b) in log scale

against

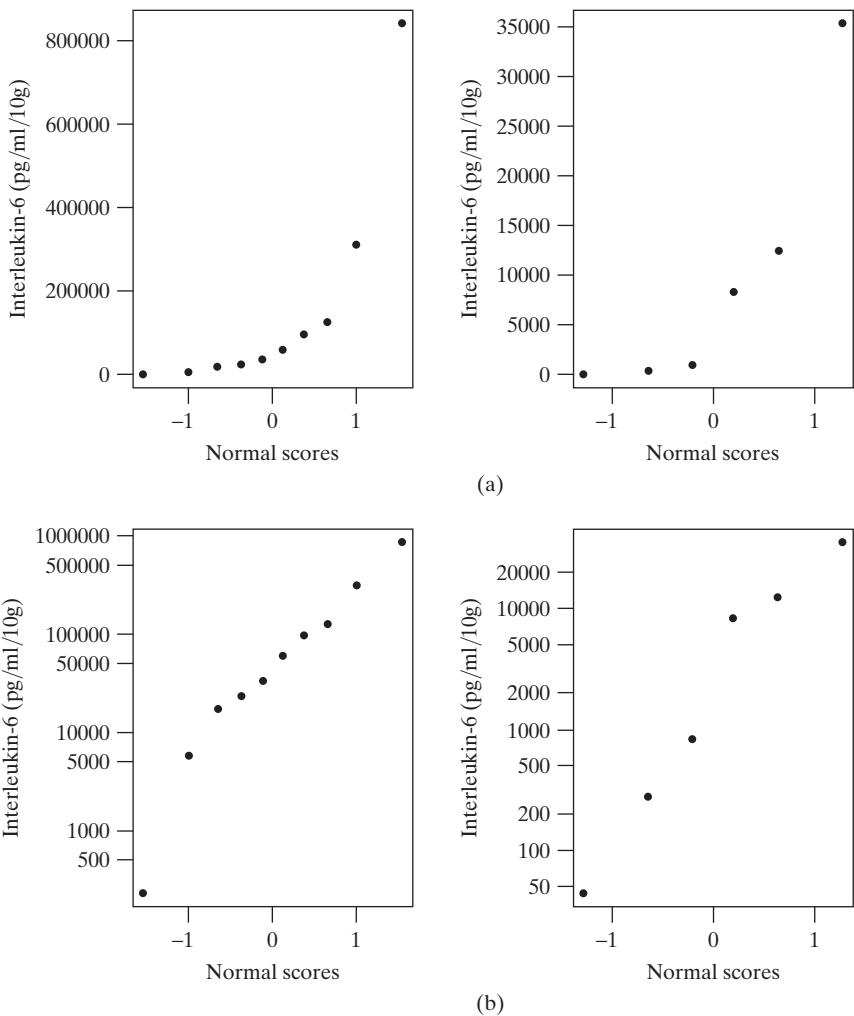
$$H_A: \mu_1 \neq \mu_2$$

where  $\mu_1$  is the population mean of the log of interleukin-6 level for breast implant patients and  $\mu_2$  is the population mean of the log of interleukin-6 level for control patients. Suppose we choose  $\alpha = 0.10$ . The test statistic is

$$t_s = \frac{(4.549 - 3.262)}{0.553} = 2.33$$

Formula (6.7.1) yields  $df = 9.7$ . The *P*-value for the test is 0.045. Thus, we have evidence, at the 0.10 level of significance (and at the 0.05 level, as well), that the mean log interleukin-6 level is higher in the breast implant population than in the control population. ■

**Figure 7.8.2** Normal probability plots of tissue inflammation data from Example 7.8.1 (a) in the original scale and (b) in log scale



## Summary of $t$ Test Mechanics

For convenient reference, we summarize the mechanics for Student's  $t$  test of equality of the means of independent samples.

### $t$ Test

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2 \text{ (nondirectional)}$$

$$H_A: \mu_1 < \mu_2 \text{ (directional)}$$

$$H_A: \mu_1 > \mu_2 \text{ (directional)}$$

$$\text{Test statistic: } t_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}}$$

$P$ -value = tail area under Student's  $t$  curve with

$$\text{df} = \frac{(\text{SE}_1^2 + \text{SE}_2^2)^2}{\text{SE}_1^4/(n_1 - 1) + \text{SE}_2^4/(n_2 - 1)}$$

Nondirectional  $H_A$ :  $P$ -value = two-tailed area beyond  $t_s$  and  $-t_s$

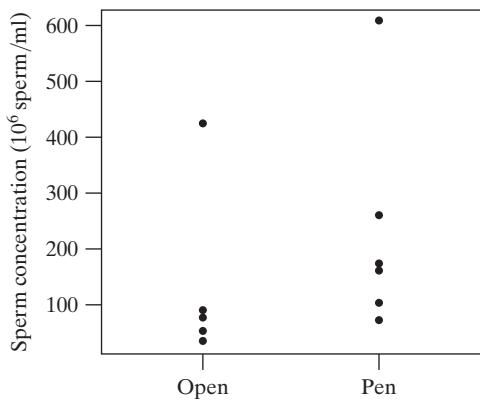
Directional  $H_A$ : Step 1. Check directionality.

Step 2.  $P$ -value = single-tail area beyond  $t_s$

Decision: Significant evidence for  $H_A$  if  $P$ -value  $\leq \alpha$

## Exercises 7.8.1–7.8.2

**7.8.1** To determine if the environment can affect sperm quality and production in cattle, a researcher randomly assigned 13 bulls to one of two environments. Six were raised in an open range environment while 7 were reared in a smaller penned environment. The following plot displays the sperm concentrations (millions of sperm/ml) of semen samples from the 13 bulls.<sup>55</sup>



- (a) Using the preceding graph to justify your answer, would the use of Student's  $t$  method be appropriate to compare the mean sperm concentrations under these two experimental conditions?
- (b) How would your answer to (a) change if the data consisted of 60 and 70 specimens rather than 6 and 7?
- (c) The Shapiro-Wilk test of normality yields  $P$ -values of 0.0012 and 0.0139 for the Open and Pen data, respectively. How do these results support or refute your response to part (a)?
- (d) How might a transformation help you analyze these data?

**7.8.2** Refer to the serotonin data of Exercise 7.2.7. On what grounds might an objection be raised to the use of the  $t$  test on these data? (Hint: For each sample, calculate the SD and compare it to the sample mean.)

## 7.9 More on Principles of Testing Hypotheses

Our study of the  $t$  test has illustrated some of the general principles of statistical tests of hypotheses. In the remainder of this book we will introduce several other types of tests besides the  $t$  test.

### A General View of Hypothesis Tests

A typical statistical test involves a null hypothesis  $H_0$ , an alternative hypothesis, or research hypothesis,  $H_A$ , and a test statistic that measures deviation or discrepancy of the data from  $H_0$ . The sampling distribution of the test statistic, under the assumption that  $H_0$  is true, is called the **null distribution** of the test statistic. (If we are conducting a randomization test as in Section 7.1, then the null distribution is the distribution of all possible differences in sample means due to random assignment of observations to groups, such as that shown in Table 7.1.2; as another example, if we are conducting a  $t$  test, then the null distribution of the  $t$  statistic  $t_s$  is—under certain conditions—a Student's  $t$  distribution.) The null distribution indicates how much the test statistic can be expected to deviate from  $H_0$  because of chance alone.

In testing a hypothesis, we assess the evidence against  $H_0$  (and in favor of  $H_A$ ) by locating the test statistic within the null distribution; the  $P$ -value is a measure of

this location, which indicates the degree of compatibility between the data and  $H_0$ . The dividing line between compatibility and incompatibility is specified by an arbitrarily chosen significance level  $\alpha$ . The decision whether to claim there is significant evidence for  $H_A$  is made according to the following rule:

Reject  $H_0$  if  $P\text{-value} \leq \alpha$ .

When a computer is not available, we will not be able to calculate the  $P$ -value exactly but will bracket it using a table of critical values. If  $H_A$  is directional, the bracketing of  $P$ -value is a two-step procedure.

Every test of a null hypothesis  $H_0$  has its associated risks of Type I error (finding significant evidence for  $H_A$  when  $H_0$  is true) and Type II error (not finding significant evidence for  $H_A$  when  $H_A$  is true). The risk of Type I error is always limited by the chosen significance level,  $\alpha$ :

$$\Pr\{\text{reject } H_0\} \leq \alpha \text{ if } H_0 \text{ is true}$$

Thus, the hypothesis testing procedure treats the Type I error as the one to be most stringently guarded against. By contrast, the power of a test can be quite low, and equivalently the risk of Type II error can be quite large, if the samples are small.

## How Are $H_0$ and $H_A$ Chosen?

A common difficulty when first studying hypothesis testing is figuring out what the null and alternative hypotheses should be. In general, the null hypothesis represents the status quo—what one would believe, by default, unless the data showed otherwise.\* Typically the alternative hypothesis is a statement that the researcher is trying to establish; thus  $H_A$  is also referred to as the research hypothesis. For example, if we are testing a new drug against a standard drug, the research hypothesis is that the new drug is better than the standard drug, while the null hypothesis is that the new drug is no different than the standard—in the absence of evidence, we would expect the two drugs to be equally effective. The typical null hypothesis,  $H_0: \mu_1 = \mu_2$ , states that the two population means are equal and that any difference between the sample means is simply due to chance error in the sampling process. The alternative hypothesis is that there *is* a difference between the drugs, so that any observed difference in sample means is due to a real effect, rather than being due to chance error alone. We conclude that we have statistically significant evidence for the research hypothesis if the data show a difference in sample means beyond what can reasonably be attributed to chance.

Here are other examples: If we are comparing men and women on some attribute, the usual null hypothesis is that there is no difference, on average, between men and women; if we are studying a measure of biodiversity in two environments, the usual null hypothesis is that the biodiversities of the two environments are equal, on average; if we are studying two diets, the usual null hypothesis is that the diets produce the same average response.

## Another Look at $P$ -Value

In order to place  $P$ -value in a general setting, let us consider some verbal interpretations of  $P$ -value.

---

\*This general rule is not always true; it is provided only as a guideline.

First we revisit the randomization test. For a nondirectional  $H_A$  the  $P$ -value is the proportion of all randomizations that results in a difference of sample means that is as large, or larger than, the difference that was observed in the actual study. Thus we can define the  $P$ -value as follows:

The  $P$ -value of the data is the probability (assuming  $H_0$  is true) of getting a result as extreme as, or more extreme than, the result that was actually observed.

To put this another way,

The  $P$ -value is the probability that, if  $H_0$  were true, a result would be obtained that would deviate from  $H_0$  as much as (or more than) the actual data do.

Now consider the  $t$  test. For a nondirectional  $H_A$ , we have defined the  $P$ -value to be the two-tailed area under the Student's  $t$  curve beyond the observed value of  $t_s$ .

Actually, these descriptions of  $P$ -value are a bit too limited. The  $P$ -value actually depends on the nature of the alternative hypothesis. When we are performing a  $t$  test against a *directional* alternative, the  $P$ -value of the data is (if the observed deviation is in the direction of  $H_A$ ) only a *single-tailed* area beyond the observed value of  $t_s$ . The more general definition of  $P$ -value is the following:

The  $P$ -value of the data is the probability (assuming  $H_0$  is true) of getting a result as deviant as, or more deviant than, the result actually observed—where deviance is measured as discrepancy from  $H_0$  in the direction of  $H_A$ .

The  $P$ -value measures how easily the observed deviation could be explained as chance variation rather than by the alternative explanation provided by  $H_A$ . For example, if the  $t$  test yields a  $P$ -value of  $P = 0.036$  for our data, then we may say that *if  $H_0$  were true* we would expect data to deviate from  $H_0$  as much as our data did only 3.6% of the time (in the meta-study).

Another definition of  $P$ -value that is worth thinking about is the following:

The  $P$ -value of the data is the value of  $\alpha$  for which  $H_0$  would just barely be rejected, using those data.

To interpret this definition, imagine that a research report that includes a  $P$ -value is read by a number of interested scientists. The scientists who are quite skeptical of  $H_A$  might require very strong evidence before being convinced and thus would use a very conservative decision threshold, such as  $\alpha = 0.001$ ; the scientists who are more favorably disposed toward  $H_A$  might require only weak evidence and thus use a liberal value such as  $\alpha = 0.10$ . The  $P$ -value of the data determines the point, within this spectrum of opinion, that separates those who find the data to be convincing in favor of  $H_A$  and those who do not. Of course, if the  $P$ -value is large, for instance  $P = 0.40$ , then presumably no reasonable person would reject  $H_0$  and be convinced of  $H_A$ .

As the preceding discussion shows, the  $P$ -value does not describe all facets of the data, but relates only to a test of a particular null hypothesis against a particular alternative. In fact, we will see that the  $P$ -value of the data also depends on which statistical test is used to test a given null hypothesis. For this reason, when describing in a scientific report the results of a statistical test, it is best to report the  $P$ -value (exactly, if possible), the name of the statistical test, and whether the alternative hypothesis was directional or nondirectional.

We repeat here, because it applies to any statistical test, the principle expounded in Section 7.6: The  $P$ -value is a measure of the strength of the evidence against

$H_0$ , but the  $P$ -value does *not* reflect the *magnitude* of the discrepancy between the data and  $H_0$ . The data may deviate from  $H_0$  only slightly, yet if the samples are large, the  $P$ -value may be quite small. By the same token, data that deviate substantially from  $H_0$  can nevertheless yield a large  $P$ -value. The  $P$ -value alone does *not* indicate whether a scientific finding is important.

## Interpretation of Error Probabilities

A common mistake is to interpret the  $P$ -value as the probability that the null hypothesis is true. A related misconception is the belief that, if we find significant evidence for  $H_A$  (for example) at the 5% significance level, then the probability that  $H_0$  is true is 5%. These interpretations are not correct.\* This point can be illustrated by an analogy with medical diagnosis.

In applying a diagnostic test for an illness, the null hypothesis is that the person is healthy—this is what we will believe unless the medical test indicates otherwise. Two types of error are possible: A healthy individual may be diagnosed as ill (false positive) or an ill individual may be diagnosed as healthy (false negative). Trying out a diagnostic test on individuals *known* to be healthy or ill will enable us to estimate the proportions of these groups who will be misdiagnosed; yet this information alone will not tell us what proportion of all positive diagnoses are false diagnoses. These ideas are illustrated numerically in the next example.

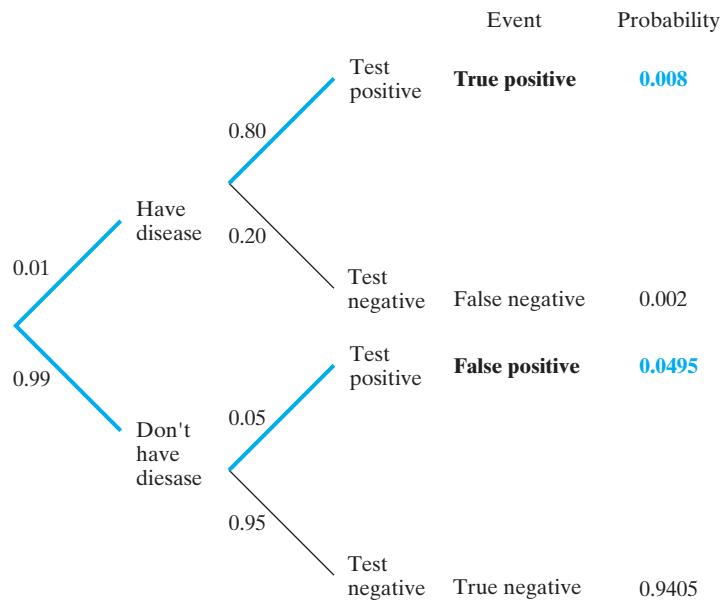
### Example 7.9.1

**Medical Testing** Suppose a medical test is conducted to detect an illness. Further, suppose that 1% of the population has the illness in question. If the test indicates that the disease is present, we reject the null hypothesis that the person is healthy. If  $H_0$  is true, then this is a Type I error—a false positive. If the test indicates that the disease is not present, we have a lack of significant evidence for  $H_A$  (illness). Suppose that the test has an 80% chance of detecting the disease if the person has it (this is analogous to the power of a hypothesis test being 80%) and a 95% chance of correctly indicating that the disease is absent if the person really does not have the disease (this is analogous to a 5% Type I error rate). Figure 7.9.1 shows a probability tree for this situation, with bold lines indicating the two ways in which the test result can be positive (i.e., the two ways that  $H_0$  can be rejected).

Now suppose that 100,000 persons are tested and that 1,000 of them (1%) actually have the illness. Then we would expect results like those given in Table 7.9.1, with 5,750 persons testing positive (which is like finding significant evidence for  $H_A$  5,750 times). Of these, 4,950 are false positives. Put another way, the proportion of the time that  $H_0$  is true, given that we found significant evidence for  $H_A$ , is  $\frac{4,950}{5,750} \approx 0.86$ , which is quite different from 0.05; this startlingly high proportion of false positives is due to the rarity of the disease. (The proportion of times that there is significant evidence for  $H_A$ , given that  $H_0$  is true, is  $\frac{4,950}{99,000} = 0.05$ , as expected, but that is a different conditional probability.  $\Pr\{A \text{ given } B\} \neq \Pr\{B \text{ given } A\}$ : The probability of rainfall, given that there are thunder and lightning, is not the same as the probability of thunder and lightning, given that it is raining.) ■

\*In fact, the probability that  $H_0$  is true cannot be calculated at all within the standard, “frequentist” approach to hypothesis testing.  $\Pr\{H_0 \text{ is true}\}$  can be calculated if one uses what are known as Bayesian methods, which are beyond the scope of this book.

**Figure 7.9.1** Probability tree for medical testing example



**Table 7.9.1** Hypothetical results of medical test of 100,000 persons

TEST RESULT	True situation			Total
		Healthy ( $H_0$ true)	Ill ( $H_A$ true)	
Negative (lack of significant evidence for $H_A$ )		94,050	200	94,250
Positive (significant evidence for $H_A$ )		4,950	800	5,750
Total		99,000	1,000	100,000

The risk of Type I error is a probability computed *under the assumption that  $H_0$  is true*; similarly, the risk of a Type II error is computed assuming that  $H_A$  is true. If we have a well-designed study with adequate sample sizes, both of these probabilities will be small. We then have a good test procedure in the same sense that the medical test is a good diagnostic procedure. But this does not in itself guarantee that most of the null hypotheses we reject are in fact false, or that most of those we do not reject are in fact true. The validity or nonvalidity of such guarantees would depend on an unknown and unknowable quantity—namely, the proportion of true null hypotheses among all null hypotheses that are tested (which is analogous to the incidence of the illness in the medical test scenario).

## Perspective

We should mention that the philosophy of statistical hypothesis testing that we have explained in this chapter is not shared by all statisticians. The view presented here, which is called the **frequentist view**, is widely used in scientific research. An alternative view, the **Bayesian view**, incorporates not only the data observed in the study at hand, but also the information that the researcher has from previous, related studies.

In the past, many Bayesian techniques were not practical due to the complexity of the mathematics that they require. However, greater computing power and improved software have made Bayesian methods more popular in recent years.

### Exercise 7.9.1

**7.9.1** Suppose we have conducted a  $t$  test, with  $\alpha = 0.05$ , and the  $P$ -value is 0.04. For each of the following statements, say whether the statement is true or false and explain why.

- (a) There is a 4% chance that  $H_0$  is true.
- (b) We reject  $H_0$  with  $\alpha = 0.05$ .

- (c) We should reject  $H_0$ , and if we repeated the experiment, there is a 4% chance that we would reject  $H_0$  again.
- (d) If  $H_0$  is true, the probability of getting a test statistic at least as extreme as the value of the  $t_s$  that was actually obtained is 4%.

## 7.10 The Wilcoxon-Mann-Whitney Test

The **Wilcoxon-Mann-Whitney test** is used to compare two independent samples.\* It is a competitor to the  $t$  test, but unlike the  $t$  test, the Wilcoxon-Mann-Whitney test is valid even if the population distributions are not normal. The Wilcoxon-Mann-Whitney test is therefore called a **distribution-free** type of test. In addition, the Wilcoxon-Mann-Whitney test does not focus on any particular parameter such as a mean or a median; for this reason it is called a **nonparametric** type of test.

### Statement of $H_0$ and $H_A$

Let us denote the observations in the two samples by  $Y_1$  and  $Y_2$ . A general statement of the null and alternative hypotheses of a Wilcoxon-Mann-Whitney test are

$H_0$ : The population distributions of  $Y_1$  and  $Y_2$  are the same.

$H_A$ : The population distribution of  $Y_1$  is shifted from the population distribution of  $Y_2$  (i.e.,  $Y_1$  tends to be either greater or less than  $Y_2$ ).

In practice, it is more natural to state  $H_0$  and  $H_A$  in words suitable to the particular application, as illustrated in Example 7.10.1.

#### Example 7.10.1

**Soil Respiration** Soil respiration is a measure of microbial activity in soil, which affects plant growth. In one study, soil cores were taken from two locations in a forest: (1) under an opening in the forest canopy (the “gap” location) and (2) at a nearby area under heavy tree growth (the “growth” location). The amount of carbon dioxide given off by each soil core was measured (in mol CO<sub>2</sub>/g soil/hr). Table 7.10.1 contains the data.<sup>56</sup>

An appropriate null hypothesis could be stated as

$H_0$ : The populations from which the two samples were drawn have the same distribution of soil respiration.

---

\*The test presented here was developed by Wilcoxon in a 1945 article. Mann and Whitney, in a 1947 article, elaborated on the test, which can be conducted in two mathematically equivalent ways. Thus, some books and some computer programs implement the test in a different fashion than the way it is presented here. Also note that some books refer to this as the Wilcoxon test, some as the Mann-Whitney test, and some (including this text) as the Wilcoxon-Mann-Whitney test.

**Table 7.10.1** Soil respiration data (mol CO<sub>2</sub>/g soil/hr) from Example 7.10.1

Growth	Gap
17 20 170 315	22 29 13 16
22 190 64	15 18 14 6

or, more informally, as

$H_0$ : The gap and growth areas do not differ with respect to soil respiration.

A nondirectional alternative could be stated as

$H_A$ : The distribution of soil respiration rates tends to be higher in one of the two populations.

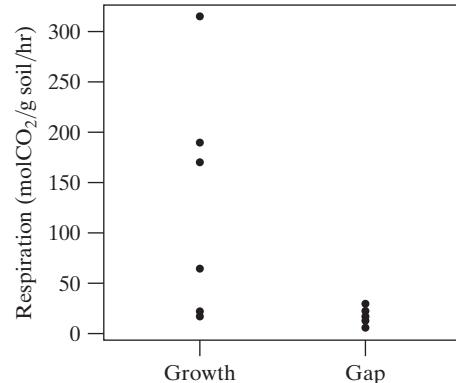
or the alternative hypothesis might be directional, for example,

$H_A$ : Soil respiration rates tend to be greater in the growth area than there are in the gap area. ■

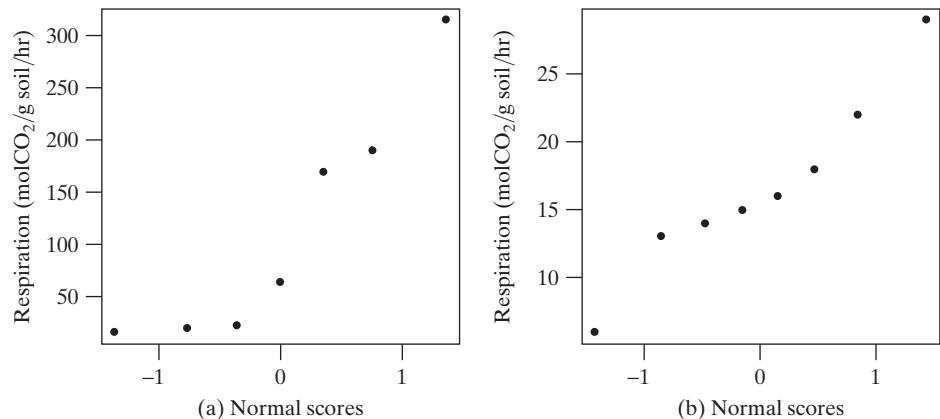
## Applicability of the Wilcoxon-Mann-Whitney Test

Figure 7.10.1 shows dotplots of the soil respiration data from Example 7.10.1; Figure 7.10.2 shows normal probability plots of these data. The growth distribution

**Figure 7.10.1** Dotplots of the soil respiration data from Example 7.10.1



**Figure 7.10.2** Normal probability plots of (a) the growth data and (b) the gap data from Example 7.10.1



is skewed to the right, whereas the gap distribution is slightly skewed to the left. If both distributions were skewed to the right, we could apply a transformation to the data. However, any attempt to transform the growth distribution, such as taking logarithms of the data, will make the skewness of the gap distribution worse. Hence, the  $t$  test is not applicable here. The Wilcoxon-Mann-Whitney test does not require normality of the distributions.

## Method

The Wilcoxon-Mann-Whitney test statistic, which is denoted  $U_s$ , measures the degree of separation or shift between two samples. A large value of  $U_s$  indicates that the two samples are well separated, with relatively little overlap between them. Critical values for the Wilcoxon-Mann-Whitney test are given in Table 6 at the end of this book. The following example illustrates the Wilcoxon-Mann-Whitney test.

### Example 7.10.2

**Soil Respiration** Let us carry out a Wilcoxon-Mann-Whitney test on the biodiversity data of Example 7.10.1.

1. The value of  $U_s$  depends on the relative positions of the  $Y_1$ 's and the  $Y_2$ 's. The first step in determining  $U_s$  is to arrange the observations in increasing order, as is shown in Table 7.10.2.
2. We next determine two counts,  $K_1$  and  $K_2$ , as follows:
  - (a) *The  $K_1$  count* For each observation in sample 1, we count the number of observations in sample 2 that are smaller in value (that is, to the left). We count 1/2 for each tied observation. In the above data, there are five  $Y_2$ 's less than the first  $Y_1$ , there are six  $Y_2$ 's less than the second  $Y_1$ , there are six  $Y_2$ 's less than the third  $Y_1$  and one equal to it, so we count 6 1/2. So far we have counts of 5, 6, and 6.5. Continuing in a similar way, we get further counts of 8, 8, 8, and 8. All together there are seven counts, one for each  $Y_1$ . The sum of all seven counts is  $K_1 = 49.5$ .
  - (b) *The  $K_2$  count* For each observation in sample 2, we count the number of observations in sample 1 that are smaller in value, counting 1/2 for ties.

**Table 7.10.2** Wilcoxon-Mann-Whitney calculations for Example 7.10.2

Number of gap observations that are smaller	$Y_1$ Growth data	$Y_2$ Gap data	Number of growth observations that are smaller
5	17	6	0
6	20	13	0
6.5	22	14	0
8	64	15	0
8	170	16	0
8	190	18	1
8	315	22	2.5
		29	3
$K_1 = 49.5$		$K_2 = 6.5$	

This gives counts of 0, 0, 0, 0, 0, 1, 2.5, and 3. The sum of these counts is  $K_2 = 6.5$ .

- (c) *Check* If the work is correct, the sum of  $K_1$  and  $K_2$  should be equal to the product of the sample sizes:

$$\begin{aligned}K_1 + K_2 &= n_1 n_2 \\49.5 + 6.5 &= 7 \times 8\end{aligned}$$

3. The test statistic  $U_s$  is the larger of  $K_1$  and  $K_2$ . In this example,  $U_s = 49.5$ .
4. To determine the  $P$ -value, we consult Table 6 with  $n =$  the larger sample size, and  $n' =$  the smaller sample size. In the present case,  $n = 8$  and  $n' = 7$ . Values from Table 6 are reproduced in Table 7.10.3.

**Table 7.10.3** Values from Table 6 for  $n = 8, n' = 7$

<b>40</b>	0.189	<b>44</b>	0.093	<b>46</b>	0.054	<b>47</b>	0.040	<b>48</b>	0.021	<b>49</b>	0.014	<b>50</b>	0.009
-----------	-------	-----------	-------	-----------	-------	-----------	-------	-----------	-------	-----------	-------	-----------	-------

Let us test  $H_0$  against a nondirectional alternative at significance level  $\alpha = 0.05$ . From Table 7.10.3, we note that when  $U_s = 49$ , the  $P$ -value is 0.014 and when  $U_s = 50$ , the  $P$ -value is 0.009; since  $49 < U_s < 50$ , the  $P$ -value is between 0.009 and 0.014 and thus there is significant evidence for  $H_A$ . There is sufficient evidence to conclude that soil respiration rates are different in the gap and growth areas. ■

As Example 7.10.2 illustrates, Table 6 can be used to bracket the  $P$ -value for the Wilcoxon-Mann-Whitney test just as Table 4 is used for the  $t$  test. If the observed  $U_s$  value is not given, then one simply locates the values that bracket the observed  $U_s$ . One then brackets the  $P$ -value by the corresponding column headings.

**Directionality** For the  $t$  test, one determines the directionality of the data by seeing whether  $\bar{Y}_1 > \bar{Y}_2$  or  $\bar{Y}_1 < \bar{Y}_2$ . Similarly, one can check directionality for the Wilcoxon-Mann-Whitney test by comparing  $K_1$  and  $K_2$ :  $K_1 > K_2$  indicates a trend for the  $Y_1$ 's to be larger than the  $Y_2$ 's, while  $K_1 < K_2$  indicates the opposite trend. Often, however, this formal comparison is unnecessary; a glance at a graph of the data is enough.

**Directional Alternative** If the alternative hypothesis  $H_A$  is directional rather than nondirectional, the Wilcoxon-Mann-Whitney procedure must be modified. As with the  $t$  test, the modified procedure has two steps and the second step involves halving the nondirectional  $P$ -value to obtain the directional  $P$ -value.

**Step 1** Check directionality—see if the data deviate from  $H_0$  in the direction specified by  $H_A$ .

- (a) If not, the  $P$ -value is greater than 0.50.
- (b) If so, proceed to step 2.

**Step 2** The  $P$ -value of the data is half as much as it would be if  $H_A$  were nondirectional.

To make a decision at a prespecified significance level  $\alpha$ , one claims significant evidence for  $H_A$  if  $P$ -value  $\leq \alpha$ .

The following example illustrates the two-step procedure.

**Example  
7.10.3**

**Directional  $H_A$**  Suppose  $n = 8$ ,  $n' = 7$ , and  $H_A$  is directional. Suppose further that the data do deviate from  $H_0$  in the direction specified by  $H_A$ . The values shown in Table 7.10.3 can be used to find the  $P$ -value as follows:

If  $U_s = 40$ , then  $P$ -value =  $0.189/2 = 0.0945$ .

If  $U_s = 46$ , then  $P$ -value =  $0.054/2 = 0.027$ .

If  $U_s = 49.5$ , then  $0.009/2 < P$ -value  $< 0.014/2$  so  $0.0045 < P$ -value  $< 0.007$ .

If  $U_s = 50$  (or larger), then  $P$ -value  $< 0.009/2 = 0.0045$ . ■

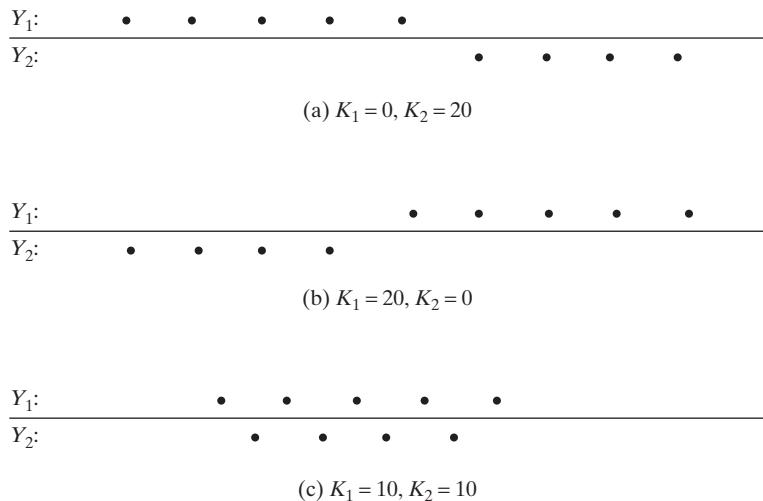
## Rationale

Let us see why the Wilcoxon-Mann-Whitney test procedure makes sense. To take a specific case, suppose the sample sizes are  $n_1 = 5$  and  $n_2 = 4$ , so that there are  $5 \times 4 = 20$  comparisons that can be made between a data point in the first sample and a data point in the second sample. Thus, regardless of what the data look like, we must have

$$K_1 + K_2 = 5 \times 4 = 20$$

The relative magnitudes of  $K_1$  and  $K_2$  indicate the amount of overlap of the  $Y_1$ 's and the  $Y_2$ 's. Figure 7.10.3 shows how this works. For the data of Figure 7.10.3(a), the two samples do not overlap at all; the data are *least* compatible with  $H_0$  and show the *strongest* evidence for  $H_A$  and thus  $U_s$  has its maximum value,  $U_s = 20$ . Similarly,  $U_s = 20$  for Figure 7.10.3(b). On the other hand, the arrangement *most* compatible with  $H_0$  and shows a lack of evidence for  $H_A$  is the one with maximal overlap, shown in Figure 7.10.3(c); for this arrangement  $K_1 = 10$ ,  $K_2 = 10$ , and  $U_s = 10$ .

**Figure 7.10.3** Three data arrays for a Wilcoxon-Mann-Whitney Test



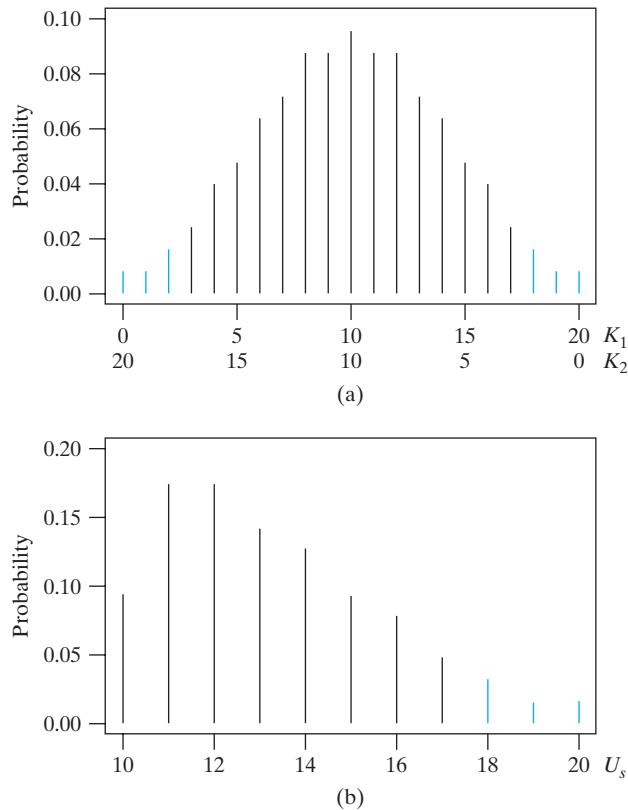
All other possible arrangements of the data lie somewhere between the three arrangements shown in Figure 7.10.3; those with much overlap have  $U_s$  close to 10, and those with little overlap have  $U_s$  closer to 20. Thus, large values of  $U_s$  indicate evidence for the research hypothesis,  $H_A$ , or equivalently the incompatibility of the data with  $H_0$ .

We now briefly consider the null distribution of  $U_s$  and indicate how the critical values of Table 6 were determined. (Recall from Section 7.10 that, for any statistical test, the reference distribution for critical values is always the null distribution of the test statistic—that is, its sampling distribution under the condition that  $H_0$  is true.) To determine the null distribution of  $U_s$ , it is necessary to calculate the probabilities associated with various arrangements of the data, assuming that all the  $Y$ 's were actually drawn from the same population.\* (The method for calculating the probabilities is briefly described in Appendix 7.2.)

Figure 7.10.4(a) shows the null distribution of  $K_1$  and  $K_2$  for the case  $n = 5, n' = 4$ . For example, it can be shown that, if  $H_0$  is true, then

$$\Pr\{K_1 = 0, K_2 = 20\} = 0.008$$

**Figure 7.10.4** Null distributions for the Wilcoxon-Mann-Whitney test when  $n = 5, n' = 4$ . (a) Null distribution of  $K_1$  and  $K_2$ ; (b) Null distribution of  $U_s$ . Shading corresponds to the  $P$ -value when  $U_s = 18$ .



This is the first probability plotted in Figure 7.10.4(a). Note that Figure 7.10.4(a) is roughly analogous to a  $t$  distribution; large values of  $K_1$  (right tail) represent evidence that the  $Y_1$ 's tend to be larger than the  $Y_2$ 's and large values of  $K_2$  (left tail) represent evidence that the  $Y_2$ 's tend to be larger than the  $Y_1$ 's.

Figure 7.10.4(b) shows the null distribution of  $U_s$ , which is derived directly from the distribution in Figure 7.10.4(a). For instance, if  $H_0$  is true, then

$$\Pr\{K_1 = 0, K_2 = 20\} = 0.008$$

---

\*In calculating the probabilities used in this section, it has been assumed that the chance of tied observations is negligible. This will be true for a continuous variable that is measured with high precision. If the number of ties is large, a correction can be made; see Noether (1967).<sup>57</sup>

and

$$\Pr\{K_1 = 20, K_2 = 0\} = 0.008$$

so that

$$\Pr\{U_s = 20\} = 0.008 + 0.008 = 0.016$$

which is the rightmost probability plotted in Figure 7.10.4(b). Thus, both tails of the  $K$  distribution have been “folded” into the upper tail of the  $U$  distribution; for instance, the one-tailed shaded area in Figure 7.10.4(b) is equal to the two-tailed shaded area in Figure 7.10.4(a).

$P$ -values for the Wilcoxon-Mann-Whitney test are upper-tail areas in the  $U_s$  distribution. For instance, it can be shown that the blue shaded area in Figure 7.10.4(b) is equal to 0.064; this means that if  $H_0$  is true, then

$$\Pr\{U_s \geq 18\} = 0.064$$

Thus, a data set that yielded  $U_s = 18$  would have an associated  $P$ -value 0.064 (assuming a nondirectional  $H_A$ ).

The values in Table 6 have been determined from the null distribution of  $U_s$ . Because the  $U_s$  distribution is discrete, only a few  $P$ -values are possible for any given sample sizes  $n_1$  and  $n_2$ . Table 6 shows selected values of  $U_s$  in bold type, with the  $P$ -value given in italics. For example, if the sample sizes are 5 and 4, then a  $U_s$  value of 17 gives a  $P$ -value of 0.111, a  $U_s$  value of 18 gives a  $P$ -value of 0.064, and a  $U_s$  value of 19 gives a  $P$ -value of 0.032. Thus, to achieve statistical significance at the  $\alpha = 0.05$  level requires a test statistic ( $U_s$ ) value of 19. The smallest possible  $P$ -value when the sample sizes are 5 and 4 is 0.016, when  $U_s = 20$ , which means that statistical significance at the  $\alpha = 0.01$  level cannot be obtained with a nondirectional test.

## Conditions for Use of the Wilcoxon-Mann-Whitney Test

In order for the Wilcoxon-Mann-Whitney test to be applicable, it must be reasonable to regard the data as random samples from their respective populations, with the observations within each sample being independent, and the two samples being independent of each other. Under these conditions, the Wilcoxon-Mann-Whitney test is valid no matter what the form of the population distributions, provided that the observed variable  $Y$  is continuous.<sup>58</sup>

The critical values given in Table 6 have been calculated assuming that ties do not occur. If the data contain only a few ties, then the  $P$ -values are approximately correct.\*

## The Wilcoxon-Mann-Whitney Test versus the $t$ Test and the Randomization Test

While the Wilcoxon-Mann-Whitney test and the  $t$  test are aimed at answering the same basic question—Are the locations of the two population distributions different or does one population tend to have larger (or smaller) values than the other?—

---

\*Actually, the Wilcoxon-Mann-Whitney test need not be restricted to continuous variables; it can be applied to any ordinal variable. However, if  $Y$  is discrete or categorical, then the data may contain many ties, and the test should not be used without appropriate modification of the critical values.

they treat the data in very different ways. Unlike the  $t$  test, the Wilcoxon-Mann-Whitney test does not use the actual values of the  $Y$ 's but only their relative positions in a rank ordering. This is both a strength and a weakness of the Wilcoxon-Mann-Whitney test. On the one hand, the test is distribution free because the null distribution of  $U_s$  relates only to the various rankings of the  $Y$ 's, and therefore does not depend on the form of the population distribution. On the other hand, the Wilcoxon-Mann-Whitney test can be inefficient: It can lack power because it does not use all the information in the data. This inefficiency is especially evident for small samples.

The randomization test is similar in spirit to the Wilcoxon-Mann-Whitney test in that it does not depend on normality, yet the power of the randomization test is often similar to that of the  $t$  test. Conducting a randomization test can be difficult, which is a primary reason that randomization tests were not more widely used until computing power became more prevalent.

None of the competitors—the randomization test, the  $t$  test, or the Wilcoxon-Mann-Whitney test—is clearly superior to the others. If the population distributions are not approximately normal, the  $t$  test may not even be valid. In addition, the Wilcoxon-Mann-Whitney test can be much more powerful than the  $t$  test, especially if the population distributions are highly skewed. If the population distributions are approximately normal with equal standard deviations, then the  $t$  test is best, but its properties are similar to those of the randomization test. For moderate sample sizes, the Wilcoxon-Mann-Whitney test can be nearly as powerful as the  $t$  test.<sup>59</sup>

There is a confidence interval procedure for population medians that is associated with the Wilcoxon-Mann-Whitney test in the same way that the confidence interval for  $(\mu_1 - \mu_2)$  is associated with the  $t$  test. The procedure is beyond the scope of this book.

## Exercises 7.10.1–7.10.9

**7.10.1** Consider two samples of sizes  $n_1 = 5, n_2 = 7$ . Use Table 6 to find the  $P$ -value, assuming that  $H_A$  is nondirectional and that

- (a)  $U_s = 26$
- (b)  $U_s = 30$
- (c)  $U_s = 35$

**7.10.2** Consider two samples of sizes  $n_1 = 4, n_2 = 8$ . Use Table 6 to find the  $P$ -value, assuming that  $H_A$  is nondirectional and that

- (a)  $U_s = 25$
- (b)  $U_s = 31$
- (c)  $U_s = 32$

**7.10.3** In a pharmacological study, researchers measured the concentration of the brain chemical dopamine in six rats exposed to toluene and six control rats. (This is the same study described in Example 7.2.1.) The concentra-

tions in the striatum region of the brain were as shown in the table.<sup>4</sup>

DOPAMINE (ng/gm)	
TOLUENE	CONTROL
3,420	1,820
2,314	1,843
1,911	1,397
2,464	1,803
2,781	2,539
2,803	1,990

- (a) Use a Wilcoxon-Mann-Whitney test to compare the treatments at  $\alpha = 0.05$ . Use a nondirectional alternative.
- (b) Proceed as in part (a), but let the alternative hypothesis be that toluene increases dopamine concentration.

**7.10.4** In a study of hypnosis, breathing patterns were observed in an experimental group of subjects and in a control group. The measurements of total ventilation (liters of air per minute per square meter of body area) are shown.<sup>60</sup> (These are the same data that were summarized in Exercise 7.5.6.) Use a Wilcoxon-Mann-Whitney test to compare the two groups at  $\alpha = 0.10$ . Use a nondirectional alternative.

EXPERIMENTAL	CONTROL
5.32	4.50
5.60	4.78
5.74	4.79
6.06	4.86
6.32	5.41
6.34	5.70
6.79	6.08
7.18	6.21

**7.10.5** In an experiment to compare the effects of two different growing conditions on the heights of greenhouse chrysanthemums, all plants grown under condition 1 were found to be taller than any of those grown under condition 2 (that is, the two height distributions did not overlap). Calculate the value of  $U_s$  and find the  $P$ -value if the number of plants in each group was

- (a) 3
- (b) 4
- (c) 5

(Assume that  $H_A$  is nondirectional.)

**7.10.6** In a study of preening behavior in the fruitfly *Drosophila melanogaster*, a single experimental fly was observed for three minutes while in a chamber with 10 other flies of the same sex. The observer recorded the timing of each episode ("bout") of preening by the experimental fly. This experiment was replicated 15 times with male flies and 15 times with female flies (different flies each time). One question of interest was whether there is a sex difference in preening behavior. The observed preening times (average time per bout, in seconds) were as follows:<sup>61</sup>

---

Male: 1.2, 1.2, 1.3, 1.9, 1.9, 2.0, 2.1, 2.2, 2.2, 2.3, 2.3, 2.4, 2.7, 2.9, 3.3

---

$$\bar{y} = 2.127 \quad s = 0.5936$$

---

Female: 2.0, 2.2, 2.4, 2.4, 2.4, 2.8, 2.8, 2.8, 2.9, 3.2, 3.7, 4.0, 5.4, 10.7, 11.7

---

$$\bar{y} = 4.093 \quad s = 3.014$$

- (a) For these data, the value of the Wilcoxon-Mann-Whitney statistic is  $U_s = 189.5$ . Use a Wilcoxon-Mann-Whitney test to investigate the sex difference in preening behavior. Let  $H_A$  be nondirectional and let  $\alpha = 0.01$ .
- (b) For these data, the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$  is  $SE = 0.7933$  sec. Use a  $t$  test to investigate the sex difference in preening behavior. Let  $H_A$  be nondirectional and let  $\alpha = 0.01$ .
- (c) What condition is required for the validity of the  $t$  test but not for the Wilcoxon-Mann-Whitney test? What feature or features of the data suggest that this condition may not hold in this case?
- (d) Verify the value of  $U_s$  given in part (a).

**7.10.7** Substances to be tested for cancer-causing potential are often painted on the skin of mice. The question arose whether mice might get an additional dose of the substance by licking or biting their cagemates. To answer this question, the compound benzo(a)pyrene was applied to the backs of 10 mice: Five were individually housed and 5 were group-housed in a single cage. After 48 hours, the concentration of the compound in the stomach tissue of each mouse was determined. The results (nmol/gm) were as follows:<sup>62</sup>

SINGLY HOUSED	GROUP-HOUSED
3.3	3.9
2.4	4.1
2.5	4.8
3.3	3.9
2.4	3.4

- (a) Use a Wilcoxon-Mann-Whitney test to compare the two distributions at  $\alpha = 0.01$ . Let the alternative hypothesis be that benzo(a)pyrene concentrations tend to be higher in group-housed mice than in singly housed mice.

- (b) Why is a directional alternative valid in this case?

**7.10.8** Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress. An exercise physiologist measured the resting (unstressed) blood concentration of HBE in two groups of men: Group 1 consisted of 11 men who had been jogging regularly for some time, and group 2 consisted of 15 men who had just entered a physical fitness program. The results are given in the following table.<sup>63</sup>

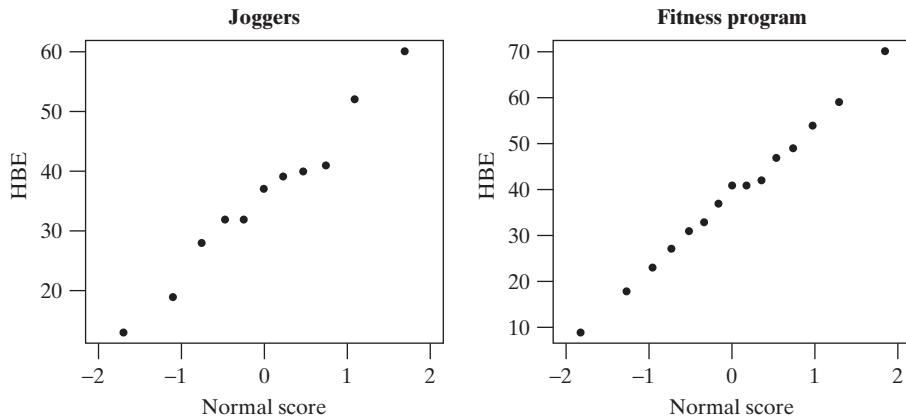
		FITNESS PROGRAM			
		JOGGERS		ENTRANTS	
39	40	32	60	70	47
19	52	41	32	42	37
13	37	28		33	23
				49	41
				59	

Use a Wilcoxon-Mann-Whitney test to compare the two distributions at  $\alpha = 0.10$ . Use a nondirectional alternative.

**7.10.9 (Continuation of 7.10.8)** Below are normal probability plots of the HBE data from Exercise 7.10.8.

- Using the plots to support your answer, is there evidence of abnormality in either of the samples?
- Considering your answer to (a) and the preceding plots, should we conclude that the data are indeed normally distributed? Explain.

- If the data are indeed normally distributed, explain in the context of this problem what the drawback would be with using the Wilcoxon-Mann-Whitney test over the two-sample  $t$  test to analyze these data.
- If the data are not normally distributed, explain in the context of this problem what the drawback would be with using the two-sample  $t$  test over the Wilcoxon-Mann-Whitney test to analyze this data.
- Considering your answers to the above, argue which test should be used with these data. Note there is more than one correct answer.



## 7.11 Perspective

In this chapter we have discussed several techniques—confidence intervals and hypothesis tests—for comparing two independent samples when the observed variable is quantitative. In coming chapters we will introduce confidence interval and hypothesis testing techniques that are applicable in various other situations. Before proceeding, we pause to reconsider the methods of this chapter.

## An Implicit Assumption

In discussing the tests of this chapter—the  $t$  test and the Wilcoxon-Mann-Whitney test—we have made an unspoken assumption, which we now bring to light. When interpreting the comparison of two distributions, we have assumed that the relationship between the two distributions is relatively simple—that if the distributions differ, then one of the two variables has a consistent tendency to be larger than the other. For instance, suppose we are comparing the effects of two diets on the weight gain of mice, with

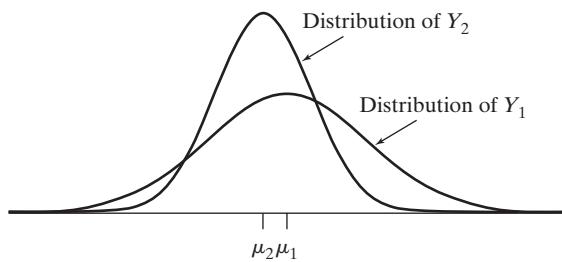
$$Y_1 = \text{Weight gain of mice on diet 1}$$

$$Y_2 = \text{Weight gain of mice on diet 2}$$

Our implicit assumption has been that, if the two diets differ at all, then that difference is in a consistent direction for all individual mice. To appreciate the meaning

of this assumption, suppose the two distributions are as pictured in Figure 7.11.1. In this case, even though the mean weight gain is higher on diet 1, it would be an oversimplification to say that mice tend to gain more weight on diet 1 than on diet 2; apparently *some* mice gain *less* on diet 1. Paradoxical situations of this kind do occasionally occur, and then the simple analysis typified by the *t* test and the Wilcoxon-Mann-Whitney test may be inadequate.

**Figure 7.11.1** Weight gain distributions on two diets



It is relatively easy to compare two distributions that have the same general shape and similar standard deviations. However, if either the shapes or the SDs of two distributions are very different from one another, then making a meaningful comparison of the distributions is difficult. In particular, a comparison of the two means might not be appropriate.

### Which Method to Use When

If we are comparing samples from two normally distributed populations, a *t* test can be used to infer whether the population means differ and a confidence interval can be used to estimate how much the two population means might differ, if at all. A confidence interval generally provides more information than does a test, since the test is restricted to a narrow question (“Might the difference between the sample be reasonably attributed to chance?”), whereas the confidence interval addresses a larger question (“How much larger is  $\mu_1$  than  $\mu_2$ ?”).

Both the confidence interval and the *t* test depend on the condition that the populations are normally distributed. If this condition is not met, then a transformation might be used to make the distributions approximately normal before proceeding. If, despite considering transformations, the normality condition is questionable, then the Wilcoxon-Mann-Whitney test can be used. (Indeed, the Wilcoxon-Mann-Whitney test can be used if the data are normal, although it is less powerful than the *t* test). When in doubt, a good piece of advice is to conduct both a *t* test and a Wilcoxon-Mann-Whitney test. If the two tests give similar, clear, conclusions (i.e., if the *P*-values for the tests are similar and both are considerably larger than  $\alpha$  or both are considerably smaller than  $\alpha$ ), then we can feel comfortable with the conclusion. However, if one test yields a *P*-value somewhat larger than  $\alpha$  and the other gives a *P*-value smaller than  $\alpha$ , then we might well declare that the tests are inconclusive.

Sometimes an outlier will be present in a data set, calling into question the result of a *t* test. It is not legitimate to simply ignore the outlier. A sensible procedure is to conduct the analysis with the outlier included and then delete the outlier and repeat the analysis. If the conclusion is unchanged when the outlier is removed, then we can feel confident that no single observation is having undue influence on the inferences we draw from the data. If the conclusion changes when the outlier is

removed, then we cannot be confident in the inferences we draw. For example, if the  $P$ -value for a test is small with the outlier present but large when the outlier is deleted, then we might state, “There is evidence that the populations differ from one another, but this evidence is largely due to a single observation.” Such a statement warns the reader that not too much should be read into any differences that were observed between the samples.

## Comparison of Variability

It sometimes happens that the variability of  $Y$ , rather than its average value, is of primary interest. For instance, in comparing two different lab techniques for measuring the concentration of an enzyme, a researcher might want primarily to know whether one of the techniques is more precise than the other, that is, whether its measurement error distribution has a smaller standard deviation. There are techniques available for testing the hypothesis  $H_0: \sigma_1 = \sigma_2$ , and for using a confidence interval to compare  $\sigma_1$  and  $\sigma_2$ . Most of these techniques are very sensitive to the condition that the underlying distributions are normal, which limits their use in practice. The implementation of these techniques is beyond the scope of this book.

## Supplementary Exercises 7.S.1–7.S.30

(Note: Exercises preceded by an asterisk refer to optional sections.)

Answers to hypothesis testing questions should include a statement of the conclusion in the context of the setting. (See Examples 7.2.4 and 7.2.5.)

**7.S.1** For each of the following pairs of samples, compute the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$ .

(a)

	SAMPLE 1	SAMPLE 2
$n$	12	13
$\bar{y}$	42	47
$s$	9.6	10.2

(b)

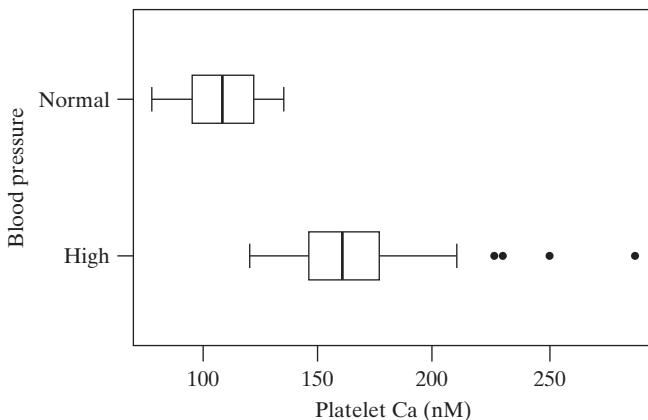
	SAMPLE 1	SAMPLE 2
$n$	22	19
$\bar{y}$	112	126
$s$	2.7	1.9

(c)

	SAMPLE 1	SAMPLE 2
$n$	5	7
$\bar{y}$	14	16
SE	1.2	1.4

**7.S.2** To investigate the relationship between intracellular calcium and blood pressure, researchers measured the free calcium concentration in the blood platelets of 38 people with normal blood pressure and 45 people with high blood pressure. The results are given in the table and the distributions are shown in the boxplots.<sup>64</sup> Use the  $t$  test to compare the means. Let  $\alpha = 0.01$  and let  $H_A$  be nondirectional. [Note: Formula (6.7.1) yields 67.5 df.]

PLATELET CALCIUM (nM)			
BLOOD PRESSURE	$n$	MEAN	SD
Normal	38	107.9	16.1
High	45	168.2	31.7



**7.S.3** Refer to Exercise 7.S.2. Construct a 95% confidence interval for the difference between the population means.

**7.S.4** Refer to Exercise 7.S.2. The boxplot for the high blood pressure group is skewed to the right and includes outliers. Does this mean that the *t* test is not valid for these data? Why or why not?

**7.S.5** In a study of methods of producing sheep's milk for use in cheese manufacture, ewes were randomly allocated to either a mechanical or a manual milking method. The investigator suspected that the mechanical method might irritate the udder and thus produce a higher concentration of somatic cells in the milk. The accompanying data show the average somatic cell count for each animal.<sup>65</sup>

SOMATIC COUNT ( $10^{-3} \times \text{cells/ml}$ )	
MECHANICAL MILKING	MANUAL MILKING
2,966	186
269	107
59	65
1,887	126
3,452	123
189	164
93	408
618	324
130	548
2,493	139
<i>n</i>	10
Mean	1,215.6
SD	1,342.9

- (a) Do the data support the investigator's suspicion? Use a *t* test against a directional alternative at  $\alpha = 0.05$ . The standard error of  $(\bar{Y}_1 - \bar{Y}_2)$  is  $SE = 427.54$  and formula (6.7.1) yields 9.2 df.
- (b) Do the data support the investigator's suspicion? Use a Wilcoxon-Mann-Whitney test against a directional alternative at  $\alpha = 0.05$ . (The value of the Wilcoxon-Mann-Whitney statistic is  $U_s = 69$ .) Compare with the result of part (a).
- (c) What condition is required for the validity of the *t* test but not for the Wilcoxon-Mann-Whitney test? What features of the data cast doubt on this condition?
- (d) Verify the value of  $U_s$  given in part (b).

**7.S.6** A plant physiologist conducted an experiment to determine whether mechanical stress can retard the growth of soybean plants. Young plants were randomly allocated to two groups of 13 plants each. Plants in one group were mechanically agitated by shaking for 20 minutes twice daily, while plants in the other group were not agitated. After 16 days of growth, the total stem length (cm) of each plant was measured, with the results given in the accompanying table.<sup>66</sup>

Use a *t* test to compare the treatments at  $\alpha = 0.01$ . Let the alternative hypothesis be that stress tends to retard growth. [Note: Formula (6.7.1) yields 23 df.]

	CONTROL	STRESS
<i>n</i>	13	13
$\bar{y}$	30.59	27.78
<i>s</i>	2.13	1.73

**7.S.7** Refer to Exercise 7.S.6. Construct a 95% confidence interval for the population mean reduction in stem length. Does the confidence interval indicate whether the effect of stress is "horticulturally important," if "horticulturally important" is defined as a reduction in population mean stem length of at least

- (a) 1 cm
- (b) 2 cm
- (c) 5 cm

**7.S.8** Refer to Exercise 7.S.6. The observations (cm), in increasing order, are shown. Compare the treatments using a Wilcoxon-Mann-Whitney test at  $\alpha = 0.01$ . Let the alternative hypothesis be that stress tends to retard growth.

	CONTROL	STRESS
	25.2	24.7
	29.5	25.7
	30.1	26.5
	30.1	27.0
	30.2	27.1
	30.2	27.2
	30.3	27.3
	30.6	27.7
	31.1	28.7
	31.2	28.9
	31.4	29.7
	33.5	30.0
	34.3	30.6

**7.S.9** One measure of the impact of pollution along a river is the diversity of species in the river floodplain. In one study, two rivers, the Black River and the Vermilion River, were compared. Random 50-m  $\times$  20-m plots were sampled along each river and the number of species of trees in each plot was recorded. The following table contains the data.<sup>67</sup>

VERMILION RIVER	BLACK RIVER
9 9 16 13 12	13 10 6 9
13 13 13 8 11	10 7 6 18
9 9 10	6

The Black River was considered to have been polluted quite a bit more than the Vermilion River, and this was expected to lead to lower biodiversity along the Black River. Conduct a Wilcoxon-Mann-Whitney test, with  $\alpha = 0.10$ , of the null hypothesis that the populations from which the two samples were drawn have the same biodiversity (distribution of tree species per plot) versus an appropriate directional alternative.

**7.S.10** A developmental biologist removed the oocytes (developing egg cells) from the ovaries of 24 frogs (*Xenopus laevis*). For each frog the oocyte pH was determined. In addition, each frog was classified according to its response to a certain stimulus with the hormone progesterone. The pH values were as follows.<sup>68</sup>

*Positive response:*

7.06, 7.18, 7.30, 7.30, 7.31, 7.32, 7.33, 7.34, 7.36, 7.36, 7.40, 7.41, 7.43, 7.48, 7.49, 7.53, 7.55, 7.57

*No response:*

7.55, 7.70, 7.73, 7.75, 7.75, 7.77

Investigate the relationship of oocyte pH to progesterone response using a Wilcoxon-Mann-Whitney test at  $\alpha = 0.05$ . Use a nondirectional alternative.

**7.S.11** Refer to Exercise 7.S.10. Summary statistics for the pH measurements are given in the following table. Investigate the relationship of oocyte pH to progesterone response using a *t* test at  $\alpha = 0.05$ . Use a nondirectional alternative. [Note: Formula (6.7.1) yields 14.1 df.]

	POSITIVE RESPONSE	NO RESPONSE
<i>n</i>	18	6
$\bar{y}$	7.373	7.708
<i>s</i>	0.129	0.081

**7.S.12** A proposed new diet for beef cattle is less expensive than the standard diet. The proponents of the new diet have conducted a comparative study in which one group of cattle was fed the new diet and another group

was fed the standard. They found that the mean weight gains in the two groups were not statistically significantly different at the 5% significance level, and they stated that this finding supported the claim that the new cheaper diet was as good (for weight gain) as the standard diet. Criticize this statement.

**\*7.S.13** Refer to Exercise 7.S.12. Suppose you discover that the study used 25 animals on each of the two diets, and that the coefficient of variation of weight gain under the conditions of the study was about 20%. Using this additional information, write an expanded criticism of the proponents' claim, indicating how likely such a study would be to detect a 10% deficiency in weight gain on the cheaper diet (using a two-tailed test at the 5% significance level).

**7.S.14** In a study of hearing loss, endolymphatic sac tumors (ELSTs) were discovered in 13 patients. These 13 patients had a total of 15 tumors (i.e., more patients had a single tumor, but two of the patients had 2 tumors each). Ten of the tumors were associated with the loss of functional hearing in an ear, but for 5 of the ears with tumors the patient had no hearing loss.<sup>69</sup> A natural question is whether hearing loss is more likely with large tumors than with small tumors. Thus, the sizes of the tumors were measured. Suppose that the sample means and standard deviations were given and that a comparison of average tumor size (hearing loss versus no hearing loss) was being considered.

- Explain why a *t* test to compare average tumor size is not appropriate here.
- If the raw data were given, could a Wilcoxon-Mann-Whitney test be used?

**7.S.15 (Computer exercise)** In an investigation of the possible influence of dietary chromium on diabetic symptoms, 14 rats were fed a low-chromium diet and 10 were fed a normal diet. One response variable was activity of the liver enzyme GITH, which was measured using a radioactively labeled molecule. The accompanying table shows the results, expressed as thousands of counts per minute per gram of liver.<sup>70</sup> Use a *t* test to compare the diets at  $\alpha = 0.05$ . Use a nondirectional alternative. [Note: Formula (6.7.1) yields 21.9 df.]

LOW-CHROMIUM DIET		NORMAL DIET	
42.3	52.8	53.1	53.6
51.5	51.3	50.7	47.8
53.7	58.5	55.8	61.8
48.0	55.4	55.1	52.6
56.0	38.3	47.5	53.7
55.7	54.1		
54.8	52.1		

**7.S.16 (Computer exercise)** Refer to Exercise 7.S.15. Use a Wilcoxon-Mann-Whitney test to compare the diets at  $\alpha = 0.05$ . Use a nondirectional alternative.

**7.S.17 (Computer exercise)** Refer to Exercise 7.S.15.

- Construct a 95% confidence interval for the difference in population means.
- Suppose the investigators believe that the effect of the low-chromium diet is “unimportant” if it shifts mean GITH activity by less than 15% — that is, if the population mean difference is less than about 8 thousand cpm/gm. According to the confidence interval of part (a), do the data support the conclusion that the difference is “unimportant”?
- How would you answer the question in part (b) if the criterion were 4 thousand rather than 8 thousand cpm/gm?

**7.S.18 (Computer exercise)** In a study of the lizard *Sceloporus occidentalis*, researchers examined field-caught lizards for infection by the malarial parasite *Plasmodium*. To help assess the ecological impact of malarial infection, the researchers tested 15 infected and 15 noninfected lizards for stamina, as indicated by the distance each animal could run in two minutes. The distances (meters) are shown in the table.<sup>71</sup>

INFECTED ANIMALS		UNINFECTED ANIMALS	
16.4	36.7	22.2	18.4
29.4	28.7	34.8	27.5
37.1	30.2	42.1	45.5
23.0	21.8	32.9	34.0
24.1	37.1	26.4	45.5
24.5	20.3	30.6	24.5
16.4	28.3	32.9	28.7
29.1		37.5	

Do the data provide evidence that the infection is associated with decreased stamina? Investigate this question using

- a  $t$  test.
- a Wilcoxon-Mann-Whitney test.

Let  $H_A$  be directional and  $\alpha = 0.05$ .

**7.S.19** In a study of the effect of amphetamine on water consumption, a pharmacologist injected four rats with amphetamine and four with saline as controls. She measured the amount of water each rat consumed in 24 hours. The following are the results, expressed as ml water per kg body weight:<sup>72</sup>

AMPHETAMINE	CONTROL
118.4	122.9
124.4	162.1
169.4	184.1
105.3	154.9

- Use a  $t$  test to compare the treatments at  $\alpha = 0.10$ . Let the alternative hypothesis be that amphetamine tends to suppress water consumption.
- Use a Wilcoxon-Mann-Whitney test to compare the treatments at  $\alpha = 0.10$ , with the directional alternative that amphetamine tends to suppress water consumption.
- Why is it important that some of the rats received saline injections as a control? That is, why didn’t the researchers simply compare rats receiving amphetamine injections to rats receiving no injection?

**7.S.20** Nitric oxide is sometimes given to newborns who experience respiratory failure. In one experiment, nitric oxide was given to 114 infants. This group was compared to a control group of 121 infants. The length of hospitalization (in days) was recorded for each of the 235 infants. The mean in the nitric oxide sample was  $\bar{y}_1 = 36.4$ ; the mean in the control sample was  $\bar{y}_2 = 29.5$ . A 95% confidence interval for  $\mu_1 - \mu_2$  is  $(-2.3, 16.1)$ , where  $\mu_1$  is the population mean length of hospitalization for infants who get nitric oxide and  $\mu_2$  is the mean length of hospitalization for infants in the control population.<sup>73</sup> For each of the following, say whether the statement is true or false and say why.

- We are 95% confident that  $\mu_1$  is greater than  $\mu_2$ , since most of the confidence interval is greater than zero.
- We are 95% confident that the difference between  $\mu_1$  and  $\mu_2$  is between  $-2.3$  days and  $16.1$  days.
- We are 95% confident that the difference between  $\bar{y}_1$  and  $\bar{y}_2$  is between  $-2.3$  days and  $16.1$  days.
- 95% of the nitric oxide infants were hospitalized longer than the average control infant.

**7.S.21** Consider the confidence interval for  $\mu_1 - \mu_2$  from Exercise 7.S.20:  $(-2.3, 16.1)$ . True or false: If we tested  $H_0: \mu_1 = \mu_2$  against  $H_A: \mu_1 \neq \mu_2$ , using  $\alpha = 0.05$ , we would reject  $H_0$ .

**7.S.22** Researchers studied subjects who had pneumonia and classified them as being in one of two groups: those who were given medical therapy that is consistent with American Thoracic Society (ATS) guidelines and those who were given medical therapy that is inconsistent with ATS guidelines. Subjects in the “consistent” group were generally able to return to work sooner than were subjects in the “inconsistent” group. A Wilcoxon-Mann-

Whitney test was applied to the data; the  $P$ -value for the test was 0.04.<sup>74</sup> For each of the following, say whether the statement is true or false and say why.

- There is a 4% chance that the “consistent” and “inconsistent” population distributions really are the same.
- If the “consistent” and “inconsistent” population distributions really are the same, then a difference between the two samples as large as the difference that these researchers observed would only happen 4% of the time.
- If a new study were done that compared the “consistent” and “inconsistent” populations, there is a 4% probability that  $H_0$  would be rejected again.

**7.S.23** A student recorded the number of calories in each of 56 entrees—28 vegetarian and 28 nonvegetarian—served at a college dining hall.<sup>75</sup> The following table summarizes the data. Graphs of the data (not given here) show that both distributions are reasonably symmetric and bell shaped. A 95% confidence interval for  $\mu_1 - \mu_2$  is  $(-27, 85)$ . For each of the following, say whether the statement is true or false and say why.

	n	MEAN	SD
Vegetarian	28	351	119
Nonvegetarian	28	322	87

- 95% of the data are between  $-27$  and  $85$  calories.
- We are 95% confident that  $\mu_1 - \mu_2$  is between  $-27$  and  $85$  calories.
- 95% of the time  $\bar{Y}_1 - \bar{Y}_2$  will be between  $-27$  and  $85$  calories.
- 95% of the vegetarian entrees have between  $27$  fewer calories and  $85$  more calories than the average nonvegetarian entree.

**7.S.24** Refer to Exercise 7.S.23. True or false (and say why): 95% of the time, when conducting a study of this size, the difference in sample means ( $\bar{Y}_1 - \bar{Y}_2$ ) will be within approximately  $\frac{(85 - (-27))}{2} = 56$  calories of the difference in population means ( $\mu_1 - \mu_2$ ).

**7.S.25 (Computer exercise)** Lianas are woody vines that grow in tropical forests. Researchers measured liana abundance (stems/ha) in several plots in the central Amazon region of Brazil. The plots were classified into two types: plots that were near the edge of the forest (less than 100 meters from the edge) or plots far from the edge of the forest. The raw data are given and are summarized in the table.<sup>76</sup>

	n	MEAN	SD
Near	34	438	125
Far	34	368	114

NEAR	FAR
639	601
605	581
535	531
437	423
376	362
349	346
320	317
285	271
250	450
436	432
419	407
702	676
600	339
555	395
466	252
380	215
350	228
337	445
310	267
265	493
441	275
420	502
407	181
648	418
294	540
352	250
495	425
338	590

- Make normal probability plots of the data to confirm that the distributions are mildly skewed.
- Conduct a  $t$  test to compare the two types of plots at  $\alpha = 0.05$ . Use a nondirectional alternative.
- Apply a logarithm transformation to the data and repeat parts (a) and (b).
- Compare the  $t$  tests from parts (b) and (c). What do these results indicate about the effect on a  $t$  test of mild skewness when the sample sizes are fairly large?

**7.S.26** Androstenedione (andro) is a steroid that is thought by some athletes to increase strength. Researchers investigated this claim by giving andro to one group of men and a placebo to a control group of men. One of the variables measured in the experiment was the increase in “lat pulldown” strength (in pounds) of each subject after four weeks. (A lat pulldown is a type of weightlifting exercise.) The raw data are given below and are summarized in the table.<sup>77</sup>

	n	MEAN	SD
Control	9	14.4	13.3
Andro	10	20.0	12.5

CONTROL				ANDRO			
30	10	10	30	0	10	0	10
40	20	30	20	10	40	20	10
10	0			30			

- (a) Conduct a  $t$  test to compare the two groups at  $\alpha = 0.10$ . Use a nondirectional alternative. [Note: Formula (6.7.1) yields 16.5 df.]
- (b) Prior to the study it was expected that andro would increase strength, which means that a directional alternative might have been used. Redo the analysis in part (a) using the appropriate directional alternative.

**7.S.27** The following is a sample of computer output from a study.<sup>78</sup> Describe the problem and the conclusion, based on the computer output.

$Y = \text{number of drinks in the previous 7 days}$

Two-sample T for treatment vs. control:

	n	Mean	SD
Treatment	244	13.62	12.39
Control	238	16.86	13.49

95% CI for  $\mu_1 - \mu_2$ : (-5.56, -0.92)

T-test  $\mu_1 = \mu_2$  (vs <):

T = -2.74 P = .0031 DF = 474.3

**7.S.28** In a controversial study to determine the effectiveness of AZT, a group of HIV-positive pregnant women were randomly assigned to get either AZT or a placebo. Some of the babies born to these women were HIV-positive, while others were not.<sup>79</sup>

- (a) What is the explanatory variable?  
 (b) What is the response variable?  
 (c) What are the experimental units?

**7.S.29** Patients suffering from acute respiratory failure were randomly assigned to either be placed in a prone (face down) position or a supine (face up) position. In the prone group, 21 out of 152 patients died. In the supine group, 25 out of 152 patients died.<sup>80</sup>

- (a) What is the explanatory variable?  
 (b) What is the response variable?  
 (c) What are the experimental units?

**7.S.30** A study of postmenopausal women on hormone replacement therapy (H.R.T.) reported that they had a reduced heart attack rate, but had even greater reductions in death from homicide and accidents—two causes of death that cannot be linked to H.R.T. It seems that the women on H.R.T. differ from others in many other aspects of their lives—for instance, they exercise more; they also tend to be wealthier and to be better educated.<sup>81</sup> Use the language of statistics to discuss what these data say about the relationships between H.R.T., heart attack risk, and variables such as exercise, wealth, and education. Use a schematic diagram similar to Figure 7.4.1 or Figure 7.4.2 to support your explanation.

# COMPARISON OF PAIRED SAMPLES

## Objectives

In this chapter we study comparisons of paired samples. We will

- demonstrate how to conduct a paired  $t$  test.
- demonstrate how to construct and interpret a confidence interval for the mean of a paired difference.
- discuss ways in which paired data arise and how pairing can be advantageous.
- consider the conditions under which a paired  $t$  test is valid.
- show how paired data may be analyzed using the sign test and the Wilcoxon signed-rank test.

### 8.1 Introduction

In Chapter 7 we considered the comparison of two independent samples when the response variable  $Y$  is a quantitative variable. In the present chapter we consider the comparison of two samples that are not independent but are paired. In a **paired design**, the observations ( $Y_1, Y_2$ ) occur in pairs; the observational units in a pair are linked in some way, so that they have more in common with each other than with members of another pair. The following is an example of a paired design.

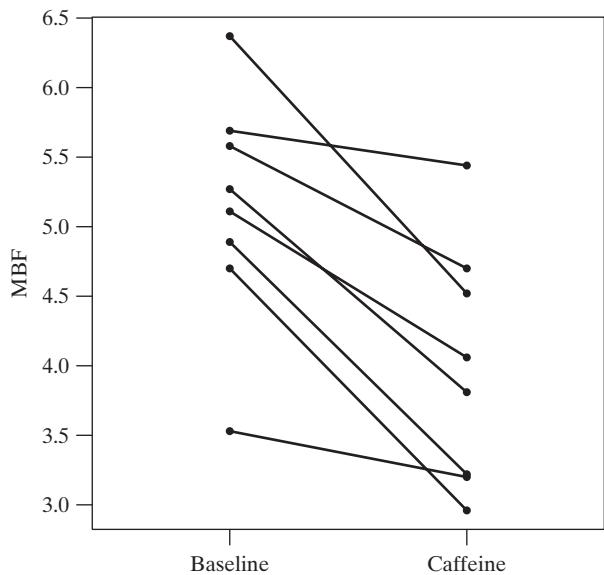
#### Example 8.1.1

**Blood Flow** Does drinking coffee affect blood flow, particularly during exercise? Doctors studying healthy subjects measured myocardial blood flow (MBF)\* during bicycle exercise before and after giving the subjects a dose of caffeine that was equivalent to drinking two cups of coffee. Table 8.1.1 shows the MBF levels before (baseline) and after (caffeine) the subjects took a tablet containing 200 mg of caffeine.<sup>1</sup> Figure 8.1.1 shows parallel dotplots of these data, with line segments that connect the baseline and caffeine readings for each subject so that the change from “before” to “after” is evident for each subject. ■

In Example 8.1.1 the data arise in pairs; the data in a pair are linked by virtue of being measurements on the same person. A suitable analysis of the data should take advantage of this pairing. That is, we could imagine an experiment in which some subjects are studied after being given caffeine and others are studied without ever being given caffeine; such an experiment would provide two independent samples of data and could be analyzed using the methods of Chapter 7. But the current experiment used a paired design. Myocardial blood flow varies from person to person, with some subjects having high MBF levels both before and after consuming caffeine and others having low MBF levels. Knowing a subject’s MBF level at baseline

\*MBF was measured by taking positron emission tomography (PET) images after oxygen-15 labeled water was infused in the patients.

<b>Table 8.1.1</b> Myocardial blood flow (ml/min/g) for eight subjects		
Subject	MBF	
	Baseline $y_1$	Caffeine $y_2$
1	6.37	4.52
2	5.69	5.44
3	5.58	4.70
4	5.27	3.81
5	5.11	4.06
6	4.89	3.22
7	4.70	2.96
8	3.53	3.20
Mean	5.14	3.99
SD	0.83	0.86



**Figure 8.1.1** Dotplots of MBF readings before and after caffeine consumption, with line segments connecting readings on each subject

tells us something about how the subject did on caffeine, and vice versa. We want to use this information when we analyze the data.

In Section 8.2 we show how to analyze paired data using methods based on Student's  $t$  distribution. In Sections 8.4 and 8.5 we describe two nonparametric tests for paired data. Sections 8.3, 8.6, and 8.7 contain more examples and discussion of the paired design.

## 8.2 The Paired-Sample $t$ Test and Confidence Interval

In this section we discuss the use of Student's  $t$  distribution to obtain tests and confidence intervals for paired data.

### Analyzing Differences

In Chapter 7 we considered how to analyze data from two independent samples. When we have paired data, we make a simple shift of viewpoint: Instead of considering  $Y_1$  and  $Y_2$  separately, we consider the *difference*  $D$ , defined as

$$D = Y_1 - Y_2$$

Note that it is often natural to consider a difference as the response variable of interest in a study. For example, if we were studying the growth rates of plants, we might grow plants under control conditions for a while at the beginning of a study and then apply a treatment for one week. We would measure the growth that takes place during the week after the treatment is introduced as  $D = Y_1 - Y_2$ , where  $Y_1$  = height one week after applying the treatment and  $Y_2$  = height before the treatment is applied.\* Sometimes data are paired in a way that is less obvious, but whenever we have paired data, it is the observed differences that we wish to analyze.

\*Exercises 7.2.11 and 7.2.12 both involve such "before versus after" data.

Let us denote the mean of sample  $D$ 's as  $\bar{D}$ . The quantity  $\bar{D}$  is related to the individual sample means as follows:

$$\bar{D} = (\bar{Y}_1 - \bar{Y}_2)$$

The relationship between population means is analogous:

$$\mu_D = \mu_1 - \mu_2$$

Thus, we may say that *the mean of the difference is equal to the difference of the means*. Because of this simple relationship, a comparison of two paired means can be carried out by concentrating entirely on the  $D$ 's.

The standard error for  $\bar{D}$  is easy to calculate. Because  $\bar{D}$  is just the mean of a single sample, we can apply the SE formula of Chapter 6 to obtain the following formula:

$$\text{SE}_{\bar{D}} = \frac{s_D}{\sqrt{n_D}}$$

where  $s_D$  is the standard deviation of the  $D$ 's and  $n_D$  is the number of  $D$ 's. The following example illustrates the calculation.

**Example  
8.2.1**

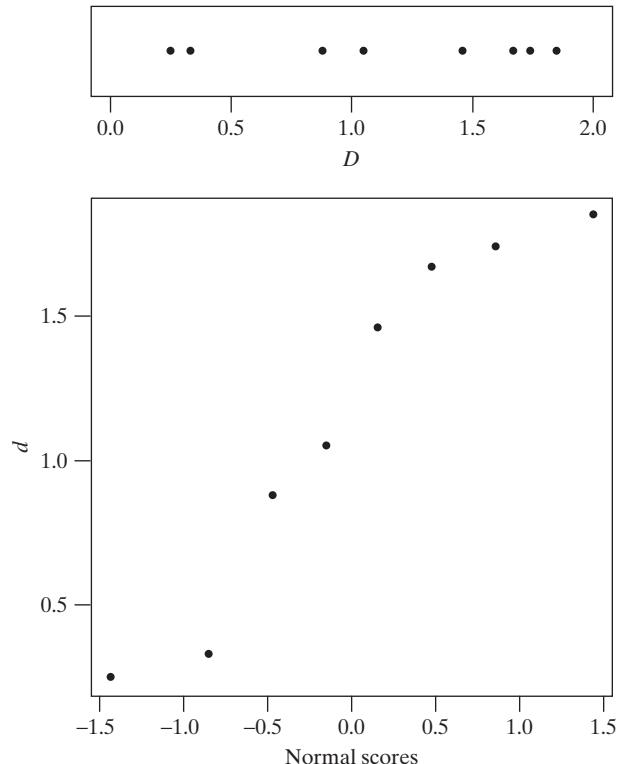
**Blood Flow** Table 8.2.1 shows the blood flow data of Example 8.1.1 and the differences  $d$ .

Note that the mean of the difference is equal to the difference of the means:

$$\bar{d} = 1.15 = 5.14 - 3.99$$

Figure 8.2.1 shows the distribution of the 8 sample differences.

MBF			
Subject	Baseline $y_1$	Caffeine $y_2$	Difference $d = y_1 - y_2$
1	6.37	4.52	1.85
2	5.69	5.44	0.25
3	5.58	4.70	0.88
4	5.27	3.81	1.46
5	5.11	4.06	1.05
6	4.89	3.22	1.67
7	4.70	2.96	1.74
8	3.53	3.20	0.33
Mean	5.14	3.99	1.15
SD	0.83	0.86	0.63



**Figure 8.2.1** Dotplot of differences in MBF at baseline and after taking caffeine, along with a normal probability plot of the data

We calculate the standard error of the mean difference as follows:

$$s_D = 0.63$$

$$n_D = 8$$

$$\text{SE}_{\bar{D}} = \frac{0.63}{\sqrt{8}} = 0.22$$

While the mean of the difference is the same as the difference of the means, note that the standard error of the mean difference is *not* the difference of standard errors of the means.

## Confidence Interval and Test of Hypothesis

The standard error described previously is the basis for the **paired-sample *t* method** of analysis, which can take the form of a confidence interval or a test of hypothesis.

A 95% confidence interval for  $\mu_D$  is constructed as

$$\bar{d} \pm t_{n_D-1, 0.025} \text{SE}_{\bar{D}}$$

where the constant  $t_{n_D-1, 0.025}$  is determined from Student's *t* distribution with

$$\text{df} = n_D - 1$$

Intervals with other confidence coefficients (such as 90%, 99%, etc.) are constructed analogously (using  $t_{0.05}$ ,  $t_{0.005}$ , etc.). The following example illustrates the confidence interval.

---

### Example 8.2.2

**Blood Flow** For the blood flow data, we have  $\text{df} = 8 - 1 = 7$ . From Table 4 we find that  $t_{7, 0.025} = 2.365$ ; thus, the 95% confidence interval for  $\mu_D$  is

$$1.15 \pm (2.365) \left( \frac{0.63}{\sqrt{8}} \right)$$

or

$$1.15 \pm 0.53$$

or

$$(0.62, 1.68)$$

We can also conduct a *t* test. To test the null hypothesis

$$H_0: \mu_D = 0$$

we use the test statistic

$$t_s = \frac{\bar{d} - 0}{\text{SE}_{\bar{D}}}$$

Critical values are obtained from Student's *t* distribution (Table 4) with  $\text{df} = n_D - 1$ . The following example illustrates the *t* test.

---

### Example 8.2.3

**Blood Flow** For the blood flow data, let us formulate the null hypothesis and nondirectional alternative:

$H_0$ : Mean myocardial blood flow is the same at baseline as it is after taking caffeine.

$H_A$ : Mean myocardial blood flow is different after taking caffeine than at baseline.

or, in symbols,

$$H_0: \mu_D = 0$$

$$H_A: \mu_D \neq 0$$

Let us test  $H_0$  against  $H_A$  at significance level  $\alpha = 0.05$ . The test statistic is

$$t_s = \frac{1.15 - 0}{0.63/\sqrt{8}} = 5.16$$

From Table 4,  $t_{7, 0.005} = 3.499$  and  $t_{7, 0.0005} = 5.408$ . We reject  $H_0$  and find that there is sufficient evidence ( $0.001 < P < 0.01$ ) to conclude that mean myocardial blood flow is decreased after taking caffeine. (Using a computer gives the  $P$ -value as  $P = 0.0013$ .) (Note that even though there is significant evidence for a decrease in MBF after taking the caffeine, we cannot conclude that caffeine caused the decrease. For example, it may be that blood flow decreased due to the passage of time.) ■

## Result of Ignoring Pairing

Suppose that a study is conducted using a paired design, but that the pairing is ignored in the analysis of the data. Such an analysis is not valid because it assumes that the samples are independent when in fact they are not. The incorrect analysis can be misleading, as the following example illustrates.

---

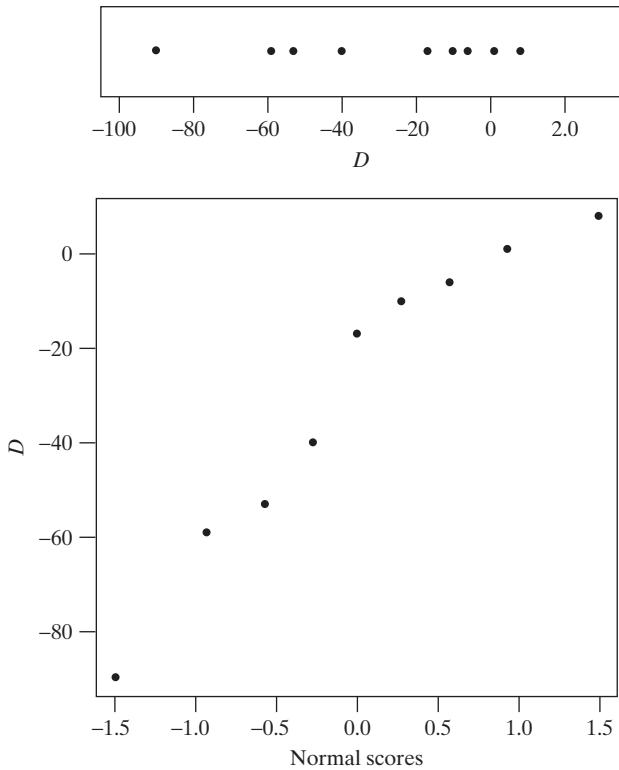
**Example  
8.2.4**

**Hunger Rating** During a weight loss study each of nine subjects was given either the active drug *m*-chlorophenylpiperazine (mCPP) for two weeks and then a placebo for another two weeks, or else was given the placebo for the first two weeks and then mCPP for the second two weeks. As part of the study the subjects were asked to rate how hungry they were at the end of each two-week period. The hunger rating data are shown in Table 8.2.2.<sup>2</sup>

**Table 8.2.2** Hunger Rating for Nine Women

Subject	Hunger rating		
	Drug (mCPP) $y_1$	Placebo $y_2$	Difference $d = y_1 - y_2$
1	79	78	1
2	48	54	-6
3	52	142	-90
4	15	25	-10
5	61	101	-40
6	107	99	8
7	77	94	-17
8	54	107	-53
9	5	64	-59
Mean	55	85	-30
SD	32	34	33

**Figure 8.2.2** Dotplot of differences in hunger rating when on the drug and when on placebo, along with a normal probability plot of the data



For the hunger rating data, the SE for the mean difference is

$$\text{SE}_{\bar{D}} = \frac{33}{\sqrt{9}} = 11$$

Figure 8.2.2 shows the distribution of the nine sample differences.

A test of

$$H_0: \mu_D = 0$$

versus

$$H_A: \mu_D \neq 0$$

gives a test statistic of

$$t_s = \frac{-30 - 0}{11} = -2.72$$

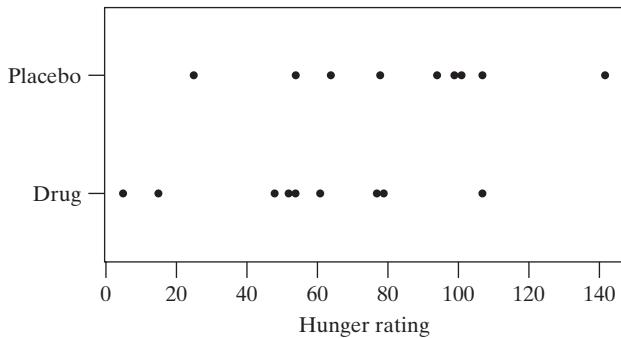
This test statistic has 8 degrees of freedom. Using a computer gives the *P*-value as *P* = 0.027.

Figure 8.2.3 displays the drug and placebo data separately. There is considerable overlap in the two distributions. This plot does not show compelling evidence that the drug lowers hunger ratings (as determined from the paired analysis above) because this plot does not take into account the paired nature of these data.

Looking at the drug and placebo data separately, the two sample SDs are  $s_1 = 32$  and  $s_2 = 34$ . If we proceed improperly as if the samples were independent and apply the SE formula of Chapter 7, we obtain

$$\begin{aligned}\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)} &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= \sqrt{\frac{32^2}{9} + \frac{34^2}{9}} = 15.6\end{aligned}$$

**Figure 8.2.3** Parallel dotplots of hunger rating when on the drug and when on placebo



This SE is quite a bit larger than the value ( $SE_{\bar{D}} = 11$ ) that we calculated using the pairing.

Continuing to (wrongly) proceed as if the samples were independent, the test statistic is

$$t_s = \frac{55 - 85}{15.6} = -1.92$$

The *P*-value for this test is 0.075, which is much greater than the *P*-value for the correct test, 0.027.

To further compare the paired and unpaired analyses, let us consider the 95% confidence interval for  $(\mu_1 - \mu_2)$ . For the unpaired analysis, formula (6.7.1) yields  $15.9 \approx 16$  degrees of freedom; this gives a *t* multiplier of  $t_{16, 0.025} = 2.121$  and yields a confidence interval of

$$(55 - 85) \pm (2.121)(15.6)$$

or

$$-30 \pm 33.1$$

or

$$(-63.1, 3.1)$$

This erroneous confidence interval is wider than the correct confidence interval from a paired analysis. A paired analysis yields the narrower interval

$$-30 \pm (2.306)(11)$$

or

$$-30 \pm 25.4$$

or

$$(-55.4, -4.6)$$

The paired-sample interval is narrower because it uses a smaller SE; this effect is slightly offset by a larger value of  $t_{0.025}$  (2.306 versus 2.121).

Why is the paired-sample SE smaller than the independent-samples SE calculated from the same data ( $SE = 11$  versus  $SE = 15.6$ )? Table 8.2.2 reveals the reason. The data show that there is large variation from one subject to the next. For instance, subject 4 has low hunger ratings (both when on the drug and when on placebo) and subject 6 has high values. The independent-samples SE formula incorporates all this variation (expressed through  $s_1$  and  $s_2$ ); in the paired-sample approach, intersubject variation in hunger rating has no influence on the calculations because only the  $D$ 's are used. By using each subject as her own control, the experimenter has increased the precision of the experiment. But if the pairing is ignored in the analysis, the extra precision is wasted. ■

The preceding example illustrates the gain in precision that can result from a paired design coupled with a paired analysis. The choice between a paired and an unpaired design will be discussed in Section 8.3.

## Conditions for Validity of Student's *t* Analysis

The conditions for validity of the paired-sample *t* test and confidence interval are as follows:

1. It must be reasonable to regard the *differences* (the *D*'s) as a random sample from some large population.
2. The population distribution of the *D*'s must be normal. The methods are approximately valid if the population distribution is approximately normal or if the sample size ( $n_D$ ) is large.

The preceding conditions are the same as those given in Chapter 6; in the present case, the conditions apply to the *D*'s because the analysis is based on the *D*'s. Verification of the conditions can proceed as described in Chapter 6. First, the design should be checked to assure that the *D*'s are independent of each other, and especially that there is no hierarchical structure within the *D*'s. (Note, however, that the  $Y_1$ 's are not independent of the  $Y_2$ 's because of the pairing.) Second, a histogram or dotplot of the *D*'s can provide a rough check for approximate normality. A normal probability plot can also be used to assess normality.

Notice that normality of the  $Y_1$ 's and  $Y_2$ 's is not required, because the analysis depends only on the *D*'s. The following example shows a case in which the  $Y_1$ 's and  $Y_2$ 's are not normally distributed, but the *D*'s are.

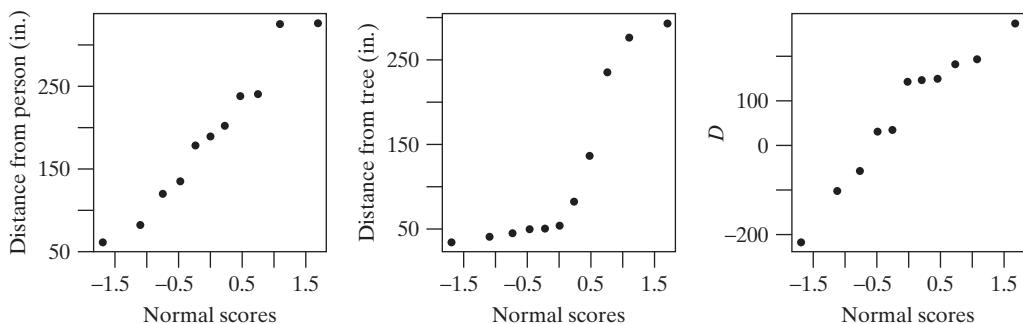
### Example 8.2.5

**Squirrels** If you walk toward a squirrel that is on the ground, it will eventually run to the nearest tree for safety. A researcher wondered whether he could get closer to the squirrel than the squirrel was to the nearest tree before the squirrel would start to run. He made 11 observations, which are given in Table 8.2.3. Figure 8.2.4 shows

**Table 8.2.3** Distances (in inches) from person and from tree when squirrel started to run

Squirrel	From person $y_1$	From tree $y_2$	Difference $d = y_1 - y_2$
1	81	137	-56
2	178	34	144
3	202	51	151
4	325	50	275
5	238	54	184
6	134	236	-102
7	240	45	195
8	326	293	33
9	60	277	-217
10	119	83	36
11	189	41	148
Mean	190	118	72
SD	89	101	148

**Figure 8.2.4** Normal probability plots of distance from squirrel to person and from squirrel to tree



that the distribution of distances from squirrel to person appear to be reasonably normal, but that the distances from squirrel to tree are far from being normally distributed. However, panel (c) of Figure 8.2.4 shows that the 11 differences do meet the normality condition. Since a paired *t* test analyzes the differences, a *t* test (or confidence interval) is valid here.<sup>3</sup>

## Summary of Formulas

For convenient reference, we summarize the formulas for the paired-sample methods based on Student's *t*.

### Standard Error of $\bar{D}$

$$\text{SE}_{\bar{D}} = \frac{s_D}{\sqrt{n_D}}$$

### *t* Test

$$H_0: \mu_D = 0$$

$$t_s = \frac{\bar{d} - 0}{\text{SE}_{\bar{D}}}$$

### 95% Confidence Interval for $\mu_d$

$$\bar{d} \pm t_{0.025} \text{SE}_{\bar{D}}$$

Intervals with other confidence levels (e.g., 90%, 99%) are constructed analogously (e.g., using  $t_{0.05}$ ,  $t_{0.005}$ ).

## Exercises 8.2.1–8.2.11

**8.2.1** In an agronomic field experiment, blocks of land were subdivided into two plots of 346 square feet each. Each block provided two paired observations: one for each of the varieties of wheat. The plot yields (lb) of wheat are given in the table.<sup>4</sup>

- (a) Calculate the standard error of the mean difference between the varieties.
- (b) Test for a difference between the varieties using a paired *t* test at  $\alpha = 0.05$ . Use a nondirectional alternative.
- (c) Test for a difference between the varieties the wrong way, using an independent-samples test. Compare with the result of part (b).

BLOCK	VARIETY		DIFFERENCE
	1	2	
1	32.1	34.5	-2.4
2	30.6	32.6	-2.0
3	33.7	34.6	-0.9
4	29.7	31.0	-1.3
Mean	31.52	33.17	-1.65
SD	1.76	1.72	0.68

**8.2.2** In an experiment to compare two diets for fattening beef steers, nine pairs of animals were chosen from the herd; members of each pair were matched as closely as possible with respect to hereditary factors. The members of each pair were randomly allocated, one to each diet. The following table shows the weight gains (lb) of the animals over a 140-day test period on diet 1 ( $Y_1$ ) and on diet 2 ( $Y_2$ ).<sup>5</sup>

PAIR	DIET 1	DIET 2	DIFFERENCE
1	596	498	98
2	422	460	-38
3	524	468	56
4	454	458	-4
5	538	530	8
6	552	482	70
7	478	528	-50
8	564	598	-34
9	556	456	100
Mean	520.4	497.6	22.9
SD	57.1	47.3	59.3

- (a) Calculate the standard error of the mean difference.
- (b) Test for a difference between the diets using a paired  $t$  test at  $\alpha = 0.10$ . Use a nondirectional alternative.
- (c) Construct a 90% confidence interval for  $\mu_D$ .
- (d) Interpret the confidence interval from part (c) in the context of this setting.

**8.2.3** Cyclic adenosine monophosphate (cAMP) is a substance that can mediate cellular response to hormones. In a study of maturation of egg cells in the frog *Xenopus laevis*, oocytes from each of four females were divided into two batches; one batch was exposed to progesterone and the other was not. After two minutes, each batch was assayed for its cAMP content, with the results given in the table.<sup>6</sup> Use a  $t$  test to investigate the effect of progesterone on cAMP. Let  $H_A$  be nondirectional and let  $\alpha = 0.10$ .

FROG	cAMP (pmol/oocyte)		
	CONTROL	PROGESTERONE	$d$
1	6.01	5.23	0.78
2	2.28	1.21	1.07
3	1.51	1.40	0.11
4	2.12	1.38	0.74
Mean	2.98	2.31	0.68
SD	2.05	1.95	0.40

**8.2.4** The following table shows the amount of weight loss (kg) for the nine subjects from Example 8.2.4 when taking the drug mCPP and when taking a placebo.<sup>2</sup> (Note that if a subject gained weight, then the recorded weight loss is negative, as is the case for subject 2 who gained 0.3 kg when on the placebo.) Use a  $t$  test to investigate the claim that mCPP affects weight loss. Let  $H_A$  be nondirectional and let  $\alpha = 0.01$ .

SUBJECT	WEIGHT CHANGE		
	MCPP	PLACEBO	DIFFERENCE
1	1.1	0.0	1.1
2	1.3	-0.3	1.6
3	1.0	0.6	0.4
4	1.7	0.3	1.4
5	1.4	-0.7	2.1
6	0.1	-0.2	0.3
7	0.5	0.6	-0.1
8	1.6	0.9	0.7
9	-0.5	-2.0	1.5
Mean	0.91	-0.09	1.00
SD	0.74	0.88	0.72

**8.2.5** Refer to Exercise 8.2.4.

- (a) Construct a 99% confidence interval for  $\mu_D$ .
- (b) Interpret the confidence interval from part (a) in the context of this setting.

**8.2.6** Under certain conditions, electrical stimulation of a beef carcass will improve the tenderness of the meat. In one study of this effect, beef carcasses were split in half; one side (half) was subjected to a brief electrical current and the other side was an untreated control. For each side, a steak was cut and tested in various ways for tenderness. In one test, the experimenter obtained a specimen of connective tissue (collagen) from the steak and determined the temperature at which the tissue would shrink; a tender piece of meat tends to yield a low collagen shrinkage temperature. The data are given in the following table.<sup>7</sup>

- (a) Construct a 95% confidence interval for the mean difference between the treated side and the control side.
- (b) Construct a 95% confidence interval the wrong way, using the independent-samples method. How does this interval differ from the one you obtained in part (a)?

CARCASS	COLLAGEN SHRINKAGE TEMPERATURE ( $^{\circ}\text{C}$ )		
	TREATED SIDE	CONTROL SIDE	DIFFERENCE
1	69.50	70.00	-0.50
2	67.00	69.00	-2.00
3	70.75	69.50	1.25
4	68.50	69.25	-0.75
5	66.75	67.75	-1.00
6	68.50	66.50	2.00
7	69.50	68.75	0.75
8	69.00	70.00	-1.00
9	66.75	66.75	0.00
10	69.00	68.50	0.50
11	69.50	69.00	0.50
12	69.00	69.75	-0.75
13	70.50	70.25	0.25
14	68.00	66.25	1.75
15	69.00	68.25	0.75
Mean	68.750	68.633	0.117
SD	1.217	1.302	1.118

**8.2.7** Refer to Exercise 8.2.6. Use a  $t$  test to test the null hypothesis of no effect against the alternative hypothesis that the electrical treatment tends to reduce the collagen shrinkage temperature. Let  $\alpha = 0.10$ .

**8.2.8** Trichotillomania is a psychiatric illness that causes its victims to have an irresistible compulsion to pull their own hair. Two drugs were compared as treatments for trichotillomania in a study involving 13 women. Each woman took clomipramine during one time period and desipramine during another time period in a double-blind experiment. Scores on a trichotillomania-impairment scale, in which high scores indicate greater impairment, were measured on each woman during each time period. The average of the 13 measurements for clomipramine was 6.2; the average of the 13 measurements for desipramine was 4.2.<sup>8</sup> A paired  $t$  test gave a value of  $t_s = 2.47$  and a two-tailed  $P$ -value of 0.03. Interpret the result of the  $t$  test. That is, what does the test indicate about clomipramine, desipramine, and hair pulling?

**8.2.9** A scientist conducted a study of how often her pet parakeet chirps. She recorded the number of distinct chirps the parakeet made in a 30-minute period, sometimes when the room was silent and sometimes when music was playing. The data are shown in the following table.<sup>9</sup> Construct a 95% confidence interval for the mean increase in chirps (per 30 minutes) when music is playing over when music is not playing.

DAY	CHIRPS IN 30 MINUTES		
	WITH MUSIC	WITHOUT MUSIC	DIFFERENCE
1	12	3	9
2	14	1	13
3	11	2	9
4	13	1	12
5	20	5	15
6	14	3	11
7	10	0	10
8	12	2	10
9	8	6	2
10	13	3	10
11	14	2	12
12	15	4	11
13	12	3	9
14	13	2	11
15	8	0	8
16	18	5	13
17	15	3	12
18	12	2	10
19	17	2	15
20	15	4	11
21	11	3	8
22	22	4	18
23	14	2	12
24	18	4	14
25	15	5	10
26	8	1	7
27	13	2	11
28	16	3	13
Mean	13.7	2.8	10.9
SD	3.4	1.5	3.0

**8.2.10** Consider the data in Exercise 8.2.9. There are two outliers among the 28 differences: the smallest value, which is 2, and the largest value, which is 18. Delete these two observations and construct a 95% confidence interval for the mean increase, using the remaining 26 observations. Do the outliers have much of an effect on the confidence interval?

**8.2.11** Invent a paired data set, consisting of five pairs of observations, for which  $\bar{y}_1$  and  $\bar{y}_2$  are not equal, and  $\text{SE}_{\bar{Y}_1} > 0$  and  $\text{SE}_{\bar{Y}_2} > 0$ , but  $\text{SE}_{\bar{D}} = 0$ .

## 8.3 The Paired Design

Ideally, in a paired design the members of a pair are relatively similar to each other—that is, more similar to each other than to members of other pairs—with respect to extraneous variables. The advantage of this arrangement is that, when members of a pair are compared, the comparison is free of the extraneous variation that originates in between-pair differences. We will expand on this theme after giving some examples.

### Examples of Paired Designs

Paired designs can arise in a variety of ways, including the following:

- Experiments in which similar experimental units form pairs
- Observational studies of identical twins
- Repeated measurements on the same individual at two different times
- Pairing by time

**Experiments with Pairs of Units** Often researchers who wish to compare two treatments will first form pairs of experimental units (pairs of animals, pairs of plots of land, etc.) that are similar (e.g., animals of the same age and sex or plots of land with the same type of soil and exposure to wind, rain, and sun). Then one member of a pair is randomly chosen to receive the first treatment and the other member is given the second treatment. The following is an example.

---

**Example  
8.3.1**

**Fertilizers for Eggplants** In a greenhouse experiment to compare two fertilizer treatments for eggplants, individually potted plants are arranged on the greenhouse bench in pairs, such that two plants in the same pair are subject to the same amount of sunlight, the same temperature, and so on. Within each pair, one (randomly chosen) plant will receive treatment 1 and the other will receive treatment 2. ■

**Observational Studies** As noted in Section 7.4, randomized experiments are preferred over observational studies, due to the many confounding variables that can arise within an observational study. An observational study may tell us that  $X$  and  $Y$  are *associated*, but only an experiment can address the question of whether  $X$  *causes*  $Y$ . If no experiment is possible and an observational study must be carried out, then it is preferable (although rarely possible) to study identical twins as the observational units. For example, in a study of the effect of “secondhand smoke” it would be ideal to enroll several sets of nonsmoking twins for which, in each pair, one of the twins lived with a smoker and the other twin did not. Because sets of twins are rarely, if ever, available, **matched-pair designs**, in which two groups are matched with respect to various extraneous variables, are often used.<sup>10</sup> Here is an example.

---

**Example  
8.3.2**

**Smoking and Lung Cancer** In a case-control study of lung cancer, 100 lung cancer patients were identified. For each case, a control was chosen who was individually matched to the case with respect to age, sex, and education level. The smoking habits of the cases and controls were compared. ■

**Repeated Measurements** Many biological investigations involve repeated measurements made on the same individual at different times. These include studies of growth and development, studies of biological processes, and studies in which measurements are made before and after application of a certain treatment. When only two times are involved, the measurements are paired, as in Example 8.1.1. The following is another example.

**Example  
8.3.3**

**Exercise and Serum Triglycerides** Triglycerides are blood constituents that are thought to play a role in coronary artery disease. To see whether regular exercise could reduce triglyceride levels, researchers measured the concentration of triglycerides in the blood serum of seven male volunteers, before and after participation in a 10-week exercise program. The results are shown in Table 8.3.1.<sup>11</sup> Note that there is considerable variation from one participant to another. For instance, participant 1 had relatively low triglyceride levels both before and after, while participant 3 had relatively high levels. ■

<b>Table 8.3.1</b> Serum triglycerides (mmol/L)		
Participant	Before	After
1	0.87	0.57
2	1.13	1.03
3	3.14	1.47
4	2.14	1.43
5	2.98	1.20
6	1.18	1.09
7	1.60	1.51

**Pairing by Time** In some situations, pairs are formed implicitly when replicate measurements are made at different times. The following is an example.

**Example  
8.3.4**

**Growth of Viruses** In a series of experiments on a certain virus (mengovirus), a microbiologist measured the growth of two strains of the virus—a mutant strain and a nonmutant strain—on mouse cells in petri dishes. Replicate experiments were run on 19 different days. The data are shown in Table 8.3.2. Each number represents the total growth in 24 hours of the viruses in a single dish.<sup>12</sup>

Note that there is considerable variation from one run to another. For instance, run 1 gave relatively large values (160 and 97), whereas run 2 gave relatively small values (36 and 55). This variation between runs arises from unavoidable small variations in the experimental conditions. For instance, both the growth of the viruses and the measurement technique are highly sensitive to environmental conditions such as the temperature and CO<sub>2</sub> concentration in the incubator. Slight fluctuations in the environmental conditions cannot be prevented, and these fluctuations cause the variation that is reflected in the data. In this kind of situation the advantage of running the two strains concurrently (that is, in pairs) is particularly striking. ■

Examples 8.3.3 and 8.3.4 both involve measurements at different times. But notice that the pairing structure in the two examples is entirely different. In Example 8.3.3 the members of a pair are measurements on the same individual at two times, whereas in Example 8.3.4 the members of a pair are measurements on

**Table 8.3.2** Virus growth at twenty-four hours

Run	Nonmutant strain	Mutant strain	Run	Nonmutant strain	Mutant strain
1	160	97	11	61	15
2	36	55	12	14	10
3	82	31	13	140	150
4	100	95	14	68	44
5	140	80	15	110	31
6	73	110	16	37	14
7	110	100	17	95	57
8	180	100	18	64	70
9	62	6	19	58	45
10	43	7			

two petri dishes at the same time. Nevertheless, in both examples the principle of pairing is the same: Members of a pair are similar to each other with respect to extraneous variables. In Example 8.3.4 time is an extraneous variable, whereas in Example 8.3.3 the comparison between two times (before and after) is of primary interest and interperson variation is extraneous.

## Purposes of Pairing

Pairing in an experimental design can serve to reduce bias, to increase precision, or both. Usually the primary purpose of pairing is to increase precision.

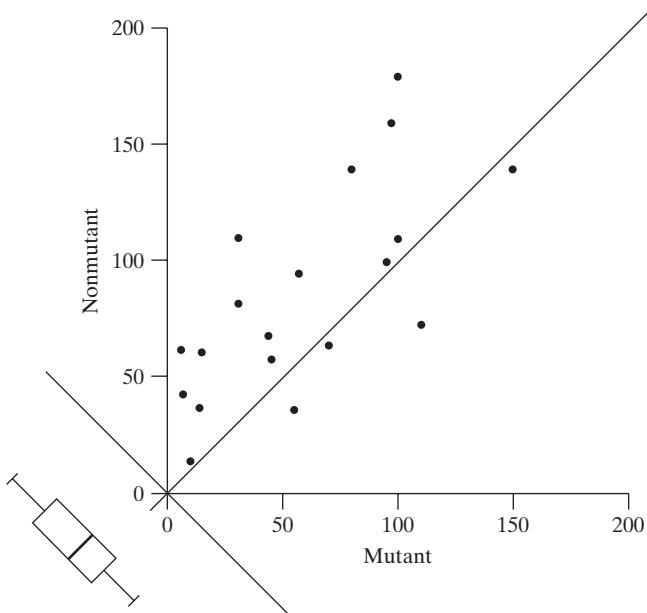
We noted in Section 7.4 that pairing or matching can reduce bias by controlling variation due to extraneous variables. The variables used in the matching are necessarily balanced in the two groups to be compared and therefore cannot distort the comparison. For instance, if two groups are composed of age-matched pairs of people, then a comparison between the two groups is free of any bias due to a difference in age distribution.

In randomized experiments, where bias can be controlled by randomized allocation, a major reason for pairing is to increase precision. Effective pairing increases precision by increasing the information available in an experiment. An appropriate analysis, which extracts this extra information, leads to more powerful tests and narrower confidence intervals. Thus, an effectively paired experiment is more efficient; it yields more information than an unpaired experiment with the same number of observations.

We saw an instance of effective pairing in the hunger rating data of Example 8.2.4. The pairing was effective because much of the variation in the measurements was due to variation between subjects, which did not enter the comparison between the treatments. As a result, the experiment yielded more precise information about the treatment difference than would a comparable unpaired experiment—that is, an experiment that would compare hunger ratings of nine women given mCPP to hunger ratings of nine different control women who were given the placebo.

The effectiveness of a given pairing can be displayed visually in a scatterplot of  $Y_2$  against  $Y_1$ ; each point in the scatterplot represents a single pair ( $Y_1, Y_2$ ). Figure 8.3.1 shows a scatterplot for the virus growth data of Example 8.3.4, together with a boxplot of the differences; each point in the scatterplot represents a single run. Notice that the points in the scatterplot show a definite upward

**Figure 8.3.1** Scatterplot for the virus growth data, with boxplot of the differences



trend. This upward trend indicates the effectiveness of the pairing: Measurements on the same run (i.e., the same day) have more in common than measurements on different runs, so that a run with a relatively high value of  $Y_1$  tends to have a relatively high value of  $Y_2$ , and similarly for low values.

Note that pairing is a strategy of *design*, not of analysis, and is therefore carried out *before* the  $Y$ 's are observed. It is not correct to use the observations themselves to form pairs. Such a data manipulation could severely distort the experimental results and could be considered scientific fraud.

### Randomized Pairs Design versus Completely Randomized Design

In planning a randomized experiment, the experimenter may need to decide between a paired design and a design that uses random assignment without any pairing, called a completely randomized design. We have said that effective pairing can greatly enhance the precision of an experiment. On the other hand, pairing in an experiment may not be effective, if the observed variable  $Y$  is not related to the factors used in the pairing. For instance, suppose pairs were matched on age only, but in fact  $Y$  turned out not to be age related. It can be shown that ineffective pairing can actually yield less precision than no pairing at all. For instance, in relation to a  $t$  test, ineffective pairing would not tend to reduce the SE, but it would reduce the degrees of freedom, and the net result would be a loss of power.

The choice of whether to use a paired design depends on practical considerations (pairing may be expensive or unwieldy) and on precision considerations. With respect to precision, the choice depends on how effective the pairing is expected to be. The following example illustrates this issue.

#### Example 8.3.5

**Fertilizers for Eggplants** A horticulturist is planning a greenhouse experiment with individually potted eggplants. Two fertilizer treatments are to be compared, and the observed variable is to be  $Y$  = yield of eggplants (pounds). The experimenter knows that  $Y$  is influenced by such factors as light and temperature, which vary somewhat from place to place on the greenhouse bench. The allocation of pots to

positions on the bench could be carried out according to a completely randomized design, or according to a paired design, as in Example 8.3.1. In deciding between these options, the experimenter must use her knowledge of how effective the pairing would be—that is, whether two pots sitting adjacent on the bench would be very much more similar in yield than pots farther apart. If she judges that the pairing would not be very effective, she may opt for the completely randomized design. ■

Note that effective pairing is *not* the same as simply holding experimental conditions constant. Pairing is a way of *organizing* the unavoidable variation that still remains after experimental conditions have been made as constant as possible. The ideal pairing organizes the variation in such a way that the variation within each pair is minimal and the variation between pairs is maximal.

## Choice of Analysis

The analysis of data should fit the design of the study. If the design is paired, a paired-sample analysis should be used; if the design is unpaired, an independent-samples analysis (as in Chapter 7) should be used.

Note that the extra information made available by an effectively paired design is *entirely wasted* if an unpaired analysis is used. (We saw an illustration of this in Example 8.2.4.) Thus, the paired design does not increase efficiency unless it is accompanied by a paired analysis.

## Exercises 8.3.1–8.3.4

**8.3.1 (Sampling exercise)** This exercise illustrates the application of a matched-pairs design to the population of 100 ellipses (shown with Exercise 3.1.1). The accompanying table shows a grouping of the 100 ellipses into 50 pairs.

PAIR	ELLIPSE ID NUMBERS		PAIR	ELLIPSE ID NUMBERS		PAIR	ELLIPSE ID NUMBERS	
01	20	45	18	11	46	35	16	66
02	03	49	19	09	29	36	18	58
03	07	27	20	19	39	37	30	50
04	42	82	21	00	10	38	76	86
05	81	91	22	40	55	39	17	83
06	38	72	23	21	56	40	04	52
07	60	70	24	08	62	41	12	64
08	31	61	25	24	78	42	23	57
09	77	89	26	67	93	43	98	99
10	01	41	27	35	80	44	36	96
11	14	48	28	74	88	45	44	84
12	59	87	29	94	97	46	06	51
13	22	68	30	02	28	47	85	90
14	47	79	31	26	71	48	37	63
15	05	95	32	25	65	49	43	69
16	53	73	33	15	75	50	34	54
17	13	33	34	32	92			

To better appreciate this exercise, imagine the following experimental setting. We want to investigate the effect of a certain treatment, T, on the organism *C. ellipticus*. We will observe the variable  $Y = \text{length}$ . We can measure each individual only once, and so we will compare  $n$  treated individuals with  $n$  untreated controls. We know that the individuals available for the experiment are of various ages, and we know that age is related to length, so we have formed 50 age-matched pairs, some of which will be used in the experiment. The purpose of the pairing is to increase the power of the experiment by eliminating the random variation due to age. (Of course, the ellipses do not actually have ages, but the pairing shown in the table has been constructed in a way that *simulates* age matching.)

- (a) Use random digits (from Table 1 or your calculator) to choose a random sample of five pairs from the list.
- (b) For each pair, use random digits (or toss a coin) to randomly allocate one member to treatment (T) and the other to control (C).
- (c) Measure the lengths of all 10 ellipses. Then, to simulate a treatment effect, add 6 mm to each length in the T group.
- (d) Apply a paired-sample  $t$  test to the data. Use a nondirectional alternative and let  $\alpha = 0.05$ .
- (e) Did the analysis of part (d) lead you to a Type II error?

**8.3.2 (Continuation of Exercise 8.3.1)** Apply an independent-samples  $t$  test to your data. Use a nondirectional

alternative and let  $\alpha = 0.05$ . Does this analysis lead you to a Type II error?

**8.3.3 (Sampling exercise)** Refer to Exercise 8.3.1. Imagine that a matched-pairs experiment is not practical (perhaps because the ages of the individuals cannot be measured), so we decide to use a completely randomized design to evaluate the treatment T.

(a) Use random digits (from Table 1 or your calculator) to choose a random sample of 10 individuals from the ellipse population (shown with Exercise 3.1.1). From these 10, randomly allocate 5 to T and 5 to C. (Or, equivalently, just randomly select 5 from the population to receive T and 5 to receive C.)

(b) Measure the lengths of all 10 ellipses. Then, to simulate a treatment effect, add 6 mm to each length in the T group.

(c) Apply an independent-samples  $t$  test to the data. Use a nondirectional alternative and let  $\alpha = 0.05$ .

(d) Did the analysis of part (c) lead you to a Type II error?

**8.3.4** Refer to each of the following exercises. Construct a scatterplot of the data. Does the appearance of the scatterplot indicate that the pairing was effective?

(a) Exercise 8.2.1

(b) Exercise 8.2.2

(c) Exercise 8.2.6

## 8.4 The Sign Test

The **sign test** is a nonparametric test that can be used to compare two paired samples. It is not particularly powerful, but it is very flexible in application and is especially simple to use and understand—a blunt but handy tool.

### Method

Like the paired-sample  $t$  test, the sign test is based on the differences

$$D = Y_1 - Y_2$$

The only information used by the sign test is the *sign* (positive or negative) of each difference. If the differences are preponderantly of one sign, this is taken as evidence for the alternative hypothesis. The following examples illustrate the sign test.

---

**Example**  
**8.4.1**

**Skin Grafts** Skin from cadavers can be used to provide temporary skin grafts for severely burned patients. The longer such a graft survives before its inevitable rejection by the immune system, the more the patient benefits. A medical team investigated the usefulness of matching graft to patient with respect to the HL-A (Human Leukocyte Antigen) antigen system. Each patient received two grafts, one with close HL-A compatibility and the other with poor compatibility. The survival times (in days) of the skin grafts are shown in the Table 8.4.1.<sup>13</sup>

Notice that a  $t$  test could not be applied here because two of the observations are incomplete; patient 3 died with a graft still surviving and the observation on patient 10 was incomplete for an unspecified reason. Nonetheless, we can proceed with a sign test, since the sign test depends only on the sign of the difference for each patient and we know that  $Y_1 - Y_2$  is positive for both of these patients.

Let us carry out a sign test to compare the survival times of the two sets of skin grafts using  $\alpha = 0.05$ . A directional research (alternative) hypothesis is appropriate for this experiment:

$H_A$ : Skin grafts tend to last longer when the HL-A compatibility is close.

The null hypothesis is

$H_0$ : The survival time distribution is the same for close compatibility as it is for poor compatibility.

**Table 8.4.1** Skin graft survival times

Patient	HL-A COMPATIBILITY		
	Close $y_1$	Poor $y_2$	Sign of $d = y_1 - y_2$
1	37	29	+
2	19	13	+
3	57+	15	+
4	93	26	+
5	16	11	+
6	23	18	+
7	20	26	-
8	63	43	+
9	29	18	+
10	60+	42	+
11	18	19	-

The first step is to determine the following counts:

$$N_+ = \text{Number of positive differences}$$

$$N_- = \text{Number of negative differences}$$

Because  $H_A$  is directional and it predicts that most of the differences will be positive, the test statistic  $B_s$  is

$$B_s = N_+$$

For the present data, we have

$$N_+ = 9$$

$$N_- = 2$$

$$B_s = 9$$

The next step is to find the  $P$ -value. We use the letter  $B$  in labeling the test statistic  $B_s$  because the distribution of  $B_s$  is based on the binomial distribution. Let  $p$  represent the probability that a difference will be positive. If the null hypothesis is true, then  $p = 0.5$ . Thus, the null distribution of  $B_s$  is a binomial with  $n = 11$  and  $p = 0.5$ . That is, the null hypothesis implies that the sign of each difference is like the result of a coin toss, with heads corresponding to a positive difference and tails to a negative difference.

For the skin graft data, the  $P$ -value for the test is the probability of getting 9 or more positive differences in 11 patients if  $p = 0.5$ . This is the probability that a binomial random variable with  $n = 11$  and  $p = 0.5$  will be greater than or equal to 9. Using the binomial formula from Chapter 3, or a computer, we find that this probability is 0.03272.\*

Because the  $P$ -value is less than  $\alpha$ , we find significant evidence for  $H_A$  that skin grafts tend to last longer when the HL-A compatibility is close than when it is poor. ■

\*Later in this section we shall learn how to use a table to compute these  $P$ -values; however, if you have covered the optional section on the binomial distribution, you can compute this probability using the binomial formula

$${}_{11}C_9(0.5)^9(0.5)^2 + {}_{11}C_{10}(0.5)^{10}(0.5)^1 + {}_{11}C_{11}(0.5)^{11} = 0.02686 + 0.00537 + 0.00049 = 0.03272$$

**Example 8.4.2**

**Growth of Viruses** Table 8.4.2 shows the virus growth data of Example 8.3.4, together with the signs of the differences.

**Table 8.4.2** Virus growth at twenty-four hours

Run	Nonmutant strain		Sign of $d = y_1 - y_2$	Run	Nonmutant strain		Sign of $d = y_1 - y_2$
	$y_1$	$y_2$			$y_1$	$y_2$	
1	160	97	+	11	61	15	+
2	36	55	-	12	14	10	+
3	82	31	+	13	140	150	-
4	100	95	+	14	68	44	+
5	140	80	+	15	110	31	+
6	73	110	-	16	37	14	+
7	110	100	+	17	95	57	+
8	180	100	+	18	64	70	-
9	62	6	+	19	58	45	+
10	43	7	+				

Let's carry out a sign test to compare the growth of the two strains, using  $\alpha = 0.10$ . The null hypothesis and nondirectional alternative are

$H_0$ : The two strains of virus grow equally well.

$H_A$ : One of the strains grows better than the other.

For these data,

$$N_+ = 15$$

$$N_- = 4$$

When the alternative is nondirectional,  $B_s$  is defined as

$$B_s = \text{Larger of } N_+ \text{ and } N_-$$

so for the virus growth data,

$$B_s = 15$$

The  $P$ -value for the test is the probability of getting 15 or more successes, plus the probability of getting 4 or fewer successes, in a binomial experiment with  $n = 19$ . We could use the binomial formula to calculate the  $P$ -value. As an alternative, critical values and  $P$ -values for the sign test are given in Table 7 (at the end of the book). Using Table 7 with  $n_D = 19$ , we obtain the critical values and corresponding  $P$ -values shown in Table 8.4.3:

**Table 8.4.3** Critical values and  $P$ -values for the sign test when  $n_D = 19$ 

$n_D$	0.20	0.10	0.05	0.02	0.01	0.002	0.001
19	<b>13</b> 0.167	<b>14</b> 0.064	<b>15</b> 0.019	<b>15</b> 0.019	<b>16</b> 0.004	<b>17</b> 0.0007	<b>17</b> 0.0007

From the table we see that for  $B_s = 15$  the  $P$ -value is 0.019, so there is significant evidence for  $H_A$ . That is, we reject  $H_0$  and conclude that the data provide significant evidence that the nonmutant strain grows better (at 24 hours) than the mutant strain of the virus. ■

**Bracketing the P-Value** Like the Wilcoxon-Mann-Whitney test, the sign test has a discrete null distribution. Certain critical value entries in Table 7 are blank, for in some cases the most extreme data possible do not lead to a small  $P$ -value. Table 7 has another peculiarity that is not shared by the Wilcoxon-Mann-Whitney test: Some critical values appear more than once in the same row due to the discreteness of the null distribution.

**Directional Alternative** To use Table 7 if the alternative hypothesis is directional, we proceed with the familiar two-step procedure:

**Step 1.** Check directionality (see if the data deviate from  $H_0$  in the direction specified by  $H_A$ ).

- (a) If not, the  $P$ -value is greater than 0.50.
- (b) If so, proceed to step 2.

**Step 2.** The  $P$ -value is half what it would be if  $H_A$  were nondirectional.

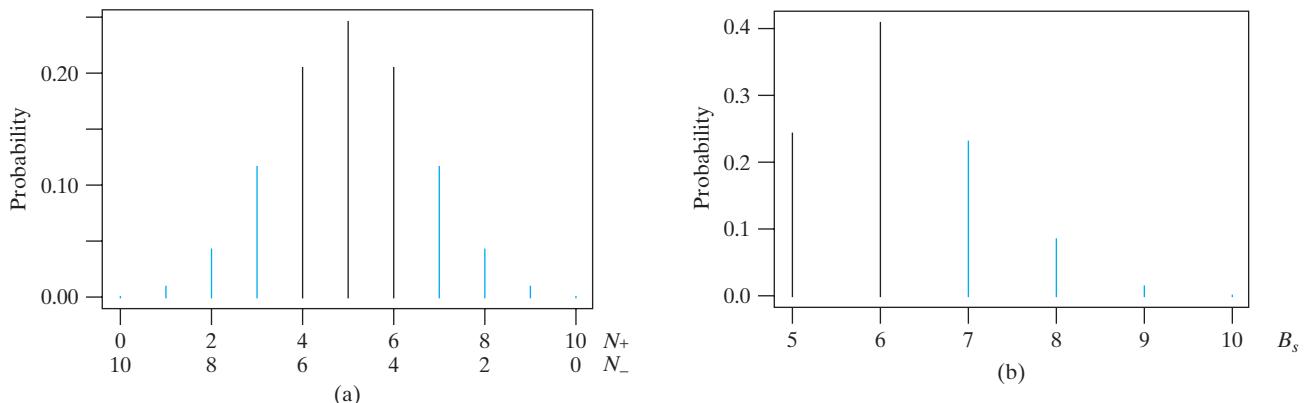
**Caution** Note that Table 7, for the sign test, and Table 4, for the  $t$  test, are organized differently: Table 7 is entered with  $n_D$ , while Table 4 is entered with  $(df = n_D - 1)$ .

**Treatment of Zeros** It may happen that some of the differences ( $Y_1 - Y_2$ ) are equal to zero. Should these be counted as positive or negative in determining  $B_s$ ? A recommended procedure is to drop the corresponding pairs from the analysis and reduce the sample size  $n_D$  accordingly. In other words, each pair whose difference is zero is ignored entirely; such pairs are regarded as providing no evidence against  $H_0$  in either direction. Notice that this procedure has no parallel in the  $t$  test; the  $t$  test treats differences of zero the same as any other value.

**Example  
8.4.3**

**Null Distribution** Consider an experiment with 10 pairs, so that  $n_D = 10$ . If  $H_0$  is true, then the probability distribution of  $N_+$  is a binomial distribution with  $n = 10$  and  $p = 0.5$ . Figure 8.4.1(a) shows this binomial distribution, together with the associated values of  $N_+$  and  $N_-$ . Figure 8.4.1(b) shows the null distribution of  $B_s$ , which is a “folded” version of Figure 8.4.1(a). (We saw a similar relationship between parts (a) and (b) of Figure 7.10.4.)

If  $N_+$  is 7 and  $H_A$  is directional (and predicts that positive differences are more likely than negative differences), then the  $P$ -value is the probability of 7 or more (+) signs in 10 trials. Using the binomial formula from Chapter 3, or a computer, we find



**Figure 8.4.1** Null distributions for the sign test when  $n_d = 10$ . (a) Distribution of  $N_+$  and  $N_-$  and (b) distribution of  $B_s$ .

that this probability is 0.17188.\* This value (0.17188) is the sum of the shaded bars in the right-hand tail in Figure 8.4.1(a). If  $H_A$  is nondirectional, then the  $P$ -value is the sum of the shaded bars in the left-hand tail and of the right-hand tail of Figure 8.4.1(a). The two shaded areas are both equal to 0.17188; consequently, the total shaded area, which is the  $P$ -value, is

$$P = 2(0.17188) = 0.34376 \approx 0.34$$

In terms of the null distribution of  $B_s$ , the  $P$ -value is an upper-tail probability; thus, the sum of the shaded bars in Figure 8.4.1(b) is equal to 0.34. ■

**How Table 7 Is Calculated** Throughout your study of statistics you are asked to take on faith the critical values given in various tables. Table 7 is an exception; the following example shows how you could (if you wished to) calculate the critical values yourself. Understanding the example will help you to appreciate how the other tables of critical values have been obtained.

**Example  
8.4.4**

Suppose  $n_D = 10$ . We saw in Example 8.4.3 that

If  $B_s = 7$ , the  $P$ -value of the data is 0.34376.

Similar calculations using the binomial formula show that

If  $B_s = 8$ , the  $P$ -value of the data is 0.10938.

If  $B_s = 9$ , the  $P$ -value of the data is 0.02148.

If  $B_s = 10$ , the  $P$ -value of the data is 0.00195.

For  $n_D = 10$ , the critical values from Table 7 are reproduced in Table 8.4.4.

**Table 8.4.4** Critical values and  $P$ -values for the sign test when  $n_D = 10$

$n_D$	0.20	0.10	0.05	0.02	0.01	0.002	0.001
10	<b>8</b> 0.109	<b>9</b> 0.021	<b>9</b> 0.021	<b>10</b> 0.002	<b>10</b> 0.002	<b>10</b> 0.0020	

The smallest value of  $B_s$  that gives a  $P$ -value less than 0.20 is  $B_s = 8$ , so this is the entry in the 0.20 column. For  $\alpha = 0.10$  or  $\alpha = 0.05$ ,  $B_s = 9$  is needed. The most extreme possibility,  $B_s = 10$ , gives a  $P$ -value of 0.00195, which is rounded to 0.0020 in the table. It is not possible to obtain a nondirectional  $P$ -value as small as 0.001, so that entry is left blank. ■

### Applicability of the Sign Test

The sign test is valid in any situation where the  $D$ 's are independent of each other and the null hypothesis can be appropriately translated as

$$H_0: \Pr\{D \text{ is positive}\} = 0.5$$

Thus, the sign test is distribution free; its validity does not depend on any conditions about the form of the population distribution of the  $D$ 's. This broad validity is bought at a price: If the population distribution of the  $D$ 's is indeed normal, then the sign test is much less powerful than the  $t$  test.

\*Applying the binomial formula we have

$$\begin{aligned} {}_{10}C_7(0.5)^7(0.5)^3 + {}_{10}C_8(0.5)^8(0.5)^2 + {}_{10}C_9(0.5)^9(0.5)^1 + {}_{10}C_{10}(0.5)^{10} \\ = 0.11719 + 0.04394 + 0.00977 + 0.00098 = 0.17188 \end{aligned}$$

The sign test is useful because it can be applied quickly and in a wide variety of settings. In fact, sometimes the sign test can be applied to data that do not permit a  $t$  test at all, as was shown in Example 8.4.1. There is another test for paired data, the Wilcoxon signed-ranks test, which is presented in Section 8.5, that is generally more powerful than the sign test and yet is distribution free. However, the Wilcoxon signed-ranks test is more difficult to carry out than the sign test and, like the  $t$  test, there are situations in which it cannot be conducted. The following is another example in which only a sign test is possible.

**Example  
8.4.5**

**THC and Chemotherapy** Chemotherapy for cancer often produces nausea and vomiting. The effectiveness of THC (tetrahydrocannabinol—the active ingredient of marijuana) in preventing these side effects was compared with the standard drug Compazine. Of the 46 patients who tried both drugs (but were not told which was which), 21 expressed no preference, while 20 preferred THC and 5 preferred Compazine. Since “preference” indicates a sign for the difference, but not a magnitude, a  $t$  test is impossible in this situation. For a sign test, we have  $n_d = 25$  and  $B_s = 20$ , so that the  $P$ -value is 0.004; even at  $\alpha = 0.005$  we would reject  $H_0$  and find that the data provide sufficient evidence to conclude that THC is preferred to Compazine.<sup>14</sup> ■

## Exercises 8.4.1–8.4.11

**8.4.1** Use Table 7 to find the  $P$ -value for a sign test (against a nondirectional alternative), assuming that  $n_D = 9$  and

- |               |               |
|---------------|---------------|
| (a) $B_s = 6$ | (b) $B_s = 7$ |
| (c) $B_s = 8$ | (d) $B_s = 9$ |

**8.4.2** Use Table 7 to find the  $P$ -value for a sign test (against a nondirectional alternative), assuming that  $n_D = 15$  and

- |                |                |
|----------------|----------------|
| (a) $B_s = 10$ | (b) $B_s = 11$ |
| (c) $B_s = 12$ | (d) $B_s = 13$ |
| (e) $B_s = 14$ | (f) $B_s = 15$ |

**8.4.3** A group of 30 postmenopausal women were given oral conjugated estrogen for one month. Plasma levels of plasminogen-activator inhibitor type 1 (PAI-1) went down for 22 of the women, but went up for 8 women.<sup>15</sup> Use a sign test to test the null hypothesis that oral conjugated estrogen has no effect on PAI-1 level. Use  $\alpha = 0.10$  and use a nondirectional alternative.

**8.4.4** Can mental exercise build “mental muscle”? In one study of this question, 12 littermate pairs of young male rats were used; one member of each pair, chosen at random, was raised in an “enriched” environment with toys and companions, while its littermate was raised alone in

an “impoverished” environment. After 80 days, the animals were sacrificed and their brains were dissected by a researcher who did not know which treatment each rat had received. One variable of interest was the weight of the cerebral cortex, expressed relative to total brain weight. For 10 of the 12 pairs, the relative cortex weight was greater for the “enriched” rat than for his “impoverished” littermate; in the other 2 pairs, the “impoverished” rat had the larger cortex. Use a sign test to compare the environments at  $\alpha = 0.05$ ; let the alternative hypothesis be that environmental enrichment tends to increase the relative size of the cortex.<sup>16</sup>

**8.4.5** Twenty institutionalized epileptic patients participated in a study of a new anticonvulsant drug, valproate. Ten of the patients (chosen at random) were started on daily valproate and the remaining 10 received an identical placebo pill. During an eight-week observation period, the numbers of major and minor epileptic seizures were counted for each patient. After this, all patients were “crossed over” to the other treatment, and seizure counts were made during a second eight-week observation period. The numbers of minor seizures are given in the accompanying table.<sup>17</sup> Test for efficacy of valproate using the sign test at  $\alpha = 0.05$ . Use a directional alternative. (Note that this analysis ignores the possible effect of time—that is, first versus second observation period.)

PATIENT NUMBER	PLACEBO PERIOD	VALPROATE PERIOD	PATIENT NUMBER	PLACEBO PERIOD	VALPROATE PERIOD
1	37	5	11	7	8
2	52	22	12	9	8
3	63	41	13	65	30
4	2	4	14	52	22
5	25	32	15	6	11
6	29	20	16	17	1
7	15	10	17	54	31
8	52	25	18	27	15
9	19	17	19	36	13
10	12	14	20	5	5

**8.4.6** An ecological researcher studied the interaction between birds of two subspecies, the Carolina Junco and the Northern Junco. He placed a Carolina male and a Northern male, matched by size, together in an aviary and observed their behavior for 45 minutes beginning at dawn. This was repeated on different days with different pairs of birds. The table shows counts of the episodes in which one bird displayed dominance over the other—for instance, by chasing it or displacing it from its perch.<sup>18</sup> Use a sign test to compare the subspecies. Use a nondirectional alternative and let  $\alpha = 0.01$ .

PAIR	NUMBER OF EPISODES IN WHICH	
	NORTHERN WAS DOMINANT	CAROLINA WAS DOMINANT
1	0	9
2	0	6
3	0	22
4	2	16
5	0	17
6	2	33
7	1	24
8	0	40

#### 8.4.7

- (a) Suppose a paired data set has  $n_D = 4$  and  $B_s = 4$ . Calculate the exact  $P$ -value of the data as analyzed by the sign test (against a nondirectional alternative).

- (b) Explain why, in Table 7 with  $n_D = 3$ , no critical values are given in any column.

**8.4.8** Suppose a paired data set has  $n_D = 15$ . Calculate the exact  $P$ -value of the data as analyzed by the sign test (against a nondirectional alternative) if  $B_s = 15$ .

**8.4.9** The study described in Example 8.2.4, involving the compound mCPP, included a group of men. The men were asked to rate how hungry they were at the end of each two-week period and differences were computed (hunger rating when taking mCPP—hunger rating when taking the placebo). The distribution of the differences was not normal. Nonetheless, a sign can be conducted using the following information: Out of eight men who recorded hunger ratings, three reported greater hunger on mCPP than on the placebo and five reported lower hunger on mCPP than on the placebo.<sup>2</sup> Conduct a sign test at the  $\alpha = 0.10$  level; use a nondirectional alternative.

**8.4.10** Refer to Exercise 8.4.9. Calculate the exact  $P$ -value of the data as analyzed by the sign test. (*Note:*  $H_A$  is nondirectional.)

**8.4.11 (Power)** A researcher is planning to conduct an experiment to compare two treatments in which matched pairs of subjects will be given the treatments and a sign test will be used, with a nondirectional alternative, to analyze the difference in responses.

Suppose the researcher believes that one treatment will always do better than the other. How many pairs does he need to have in the experiment if he wants to be able to reject  $H_0$  when  $\alpha = 0.05$ ? If one treatment “wins” in every pair, what will be the  $P$ -value from the resulting test?

## 8.5 The Wilcoxon Signed-Rank Test

The **Wilcoxon signed-rank test**, like the sign test, is a nonparametric method that can be used to compare paired samples. Conducting a Wilcoxon signed-rank test is somewhat more complicated than conducting a sign test, but the Wilcoxon test is more powerful than the sign test. Like the sign test, the Wilcoxon signed-rank test does *not* require that the data be a sample from a normally distributed population.

The Wilcoxon signed-rank test is based on the set of differences,  $D = Y_1 - Y_2$ . It combines the main idea of the sign test—“look at the signs of the differences”—with the main idea of the paired  $t$  test—“look at the magnitudes of the differences.”

## Method

The Wilcoxon signed-rank test proceeds in several steps, which we present here in the context of an example.

### Example 8.5.1

**Nerve Cell Density** For each of nine horses, a veterinary anatomist measured the density of nerve cells at specified sites in the intestine. The results for site I (mid-region of jejunum) and site II (mesenteric region of jejunum) are given in the accompanying table.<sup>19</sup> Each density value is the average of counts of nerve cells in five equal sections of tissue. The null hypothesis of interest is that in the population of all horses there is no difference between the two sites.

1. The first step in the Wilcoxon signed-rank test is to calculate the differences, as shown in Table 8.5.1.

**Table 8.5.1** Nerve cell density at each of two sites

Animal	Site I	Site II	Difference
1	50.6	38.0	12.6
2	39.2	18.6	20.6
3	35.2	23.2	12.0
4	17.0	19.0	-2.0
5	11.2	6.6	4.6
6	14.2	16.4	-2.2
7	24.2	14.4	9.8
8	37.4	37.6	-0.2
9	35.2	24.4	10.8

2. Next we find the absolute value of each difference.
3. We then rank these absolute values, from smallest to largest, as shown in Table 8.5.2.

**Table 8.5.2**

Animal	Difference, $d$	$ d $	Rank of $ d $
1	12.6	12.6	8
2	20.6	20.6	9
3	12.0	12.0	7
4	-2.0	2.0	2
5	4.6	4.6	4
6	-2.2	2.2	3
7	9.8	9.8	5
8	-0.2	0.2	1
9	10.8	10.8	6

4. Next we restore the + and – signs to the ranks of the absolute differences to produce signed ranks, as shown in Table 8.5.3.

**Table 8.5.3**

Animal	Difference, $d$	Rank of $ d $	Signed rank
1	12.6	8	8
2	20.6	9	9
3	12.0	7	7
4	-2.0	2	-2
5	4.6	4	4
6	-2.2	3	-3
7	9.8	5	5
8	-0.2	1	-1
9	10.8	6	6

5. We sum the positive signed ranks to get  $W_+$ ; we sum the absolute values of the negative signed ranks to get  $W_-$ . For the nerve cell data,  $W_+ = 8 + 9 + 7 + 4 + 5 + 6 = 39$  and  $W_- = 2 + 3 + 1 = 6$ . The test statistic,  $W_s$  is defined as

$$W_s = \text{Larger of } W_+ \text{ and } W_-$$

For the nerve cell data,  $W_s = 39$ .

6. To find the  $P$ -value, we consult Table 8 (at the end of the book). Part of Table 8 is reproduced in Table 8.5.4.

**Table 8.5.4** Critical values for the Wilcoxon signed-rank test when  $n_D = 9$ 

$n$	0.20	0.10	0.05	0.02	0.01	0.002	0.001
9	<b>35</b> 0.164	<b>37</b> 0.098	<b>40</b> 0.039	<b>42</b> 0.020	<b>44</b> 0.0078		

From Table 8.5.4, we see that for  $W_s = 37$  the  $P$ -value is 0.098. There is weak but suggestive evidence ( $P = 0.098$ ) that there is a difference in nerve cell density in the two regions. (We reject  $H_0$  if  $\alpha$  is 0.10 or larger.) ■

**Bracketing the  $P$ -Value** Like the sign test, the Wilcoxon signed-rank test has a discrete null distribution. Certain critical value entries in Table 8 are blank; this situation is familiar from our study of the Wilcoxon-Mann-Whitney test and the sign test. For example, if  $n_D = 9$ , then the strongest possible evidence against  $H_0$  occurs when all 9 differences are positive (or when all 9 differences are negative), in which case  $W_s = 45$ . But the chance that  $W_s$  will equal 45 when  $H_0$  is true is  $(1/2)^9 + (1/2)^9$ , which is approximately 0.0039. Thus, it is not possible to have a two-tailed  $P$ -value smaller than 0.002, let alone 0.001. This is why the last two entries are blank in the  $n_D = 9$  row of Table 8. Also note that if  $W_s = 34$ , for example, then the table only tells us that  $P > 0.20$ .

**Directional Alternative** To use Table 8 if the alternative hypothesis is directional, we proceed with the familiar two-step procedure:

**Step 1.** Check directionality (see if the data deviate from  $H_0$  in the direction specified by  $H_A$ ).

(a) If not, the  $P$ -value is greater than 0.50.

(b) If so, proceed to step 2.

**Step 2.** The  $P$ -value is half what it would be if  $H_A$  were nondirectional.

**Treatment of Zeros** If any of the differences ( $Y_1 - Y_2$ ) are zero, then those data points are deleted and the sample size is reduced accordingly. For example, if one of the 9 differences in Example 8.5.1 had been zero, we would have deleted that point when conducting the Wilcoxon test, so that the sample size would have become 8.

**Treatment of Ties** If there are ties among the absolute values of the differences (in step 3) we average the ranks of the tied values. If there are ties, then the  $P$ -value given by the Wilcoxon signed-rank test is only approximate.

## Applicability of the Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test can be used in any situation in which the  $D$ 's are independent of each other and come from a symmetric distribution; the distribution need not be normal.\* The null hypothesis of “no treatment effect” or “no difference between populations” can be stated as

$$H_0: \mu_D = 0$$

Sometimes the Wilcoxon signed-rank test can be carried out even with incomplete information. For example, a Wilcoxon test is possible for the skin graft data of Example 8.4.1. It is true that an exact value of  $d$  cannot be calculated for two of the patients, but for both of these patients the difference is positive and is larger than either of the negative differences. The data in Table 8.5.5 show that there only are two negative differences. The smaller of these is  $-1$ , for patient 11. This is the smallest difference in absolute value, so it has signed rank  $-1$ . The only other negative signed rank is for patient 7; all of the other signed ranks are positive. (The rest of this example is left as an exercise.)

**Table 8.5.5** Skin graft survival times

Patient	HL-A COMPATIBILITY		
	Close $y_1$	Poor $y_2$	$d = y_1 - y_2$
1	37	29	8
2	19	13	6
3	57+	15	42+
4	93	26	67
5	16	11	5
6	23	18	5
7	20	26	-6
8	63	43	20
9	29	18	11
10	60+	42	18+
11	18	19	-1

As with the Wilcoxon-Mann-Whitney test for independent samples, there is a procedure associated with the Wilcoxon signed-rank test that can be used to construct a confidence interval for  $\mu_D$ . The procedure is beyond the scope of this book.

---

\*Strictly speaking, the distribution must be continuous, which means that the probability of a tie is zero.

In summary, when dealing with paired data we have three inference procedures: the paired  $t$  test, the Wilcoxon signed-rank test, and the sign test. The  $t$  test requires that the data come from a normally distributed population; if this condition is met then the  $t$  test is recommended, as it is more powerful than the Wilcoxon test or sign test. The Wilcoxon test does not require normality but does require that the differences come from a symmetric distribution and that they can be ranked; it has more power than the sign test. The sign test is the least powerful of the three methods, but the most widely applicable, since it only requires that we determine whether each difference is positive or negative.

## Exercises 8.5.1–8.5.7

**8.5.1** Use Table 8 to find the  $P$ -value for a Wilcoxon signed-rank test (against a nondirectional alternative), assuming that  $n_D = 7$  and

- (a)  $W_s = 22$
- (b)  $W_s = 25$
- (c)  $W_s = 26$
- (d)  $W_s = 28$

**8.5.2** Use Table 8 to find the  $P$ -value for a Wilcoxon signed-rank test (against a nondirectional alternative), assuming that  $n_D = 12$  and

- (a)  $W_s = 55$
- (b)  $W_s = 65$
- (c)  $W_s = 71$
- (d)  $W_s = 73$

**8.5.3** The study described in Example 8.2.4, involving the compound mCPP, included a group of nine men. The men were asked to rate how hungry they were at the end of each two-week period and differences were computed (hunger rating when taking mCPP – hunger rating when taking the placebo). Data for one of the subjects are not available; the data for the other eight subjects are given in the accompanying table.<sup>2</sup> Analyze these data with a Wilcoxon signed-rank test at the  $\alpha = 0.10$  level; use a nondirectional alternative.

SUBJECT	HUNGER RATING		
	MCPP $y_1$	PLACEBO $y_2$	DIFFERENCE $d = y_1 - y_2$
1	64	69	-5
2	119	112	7
3	0	28	-28
4	48	95	-47
5	65	145	-80
6	119	112	7
7	149	141	8
8	NA	NA	NA
9	99	119	-20

**8.5.4** As part of the study described in Example 8.2.4 (and in Exercise 8.5.3), involving the compound mCPP, weight change was measured for nine men. For each man two measurements were made: weight change when taking mCPP and weight change when taking the placebo. The data are given in the accompanying table.<sup>2</sup> Analyze these data with a Wilcoxon signed-rank test at the  $\alpha = 0.05$  level; use a nondirectional alternative.

SUBJECT	WEIGHT CHANGE		
	MCPP $y_1$	PLACEBO $y_2$	DIFFERENCE $d = y_1 - y_2$
1	0.0	-1.1	1.1
2	-1.1	0.5	-1.6
3	-1.6	0.5	-2.1
4	-0.3	0.0	-0.3
5	-1.1	-0.5	-0.6
6	-0.9	1.3	-2.2
7	-0.5	-1.4	0.9
8	0.7	0.0	0.7
9	-1.2	-0.8	-0.4

**8.5.5** Consider the skin graft data of Example 8.4.1. Table 8.5.5, at the end of Section 8.5, shows the first steps in conducting a Wilcoxon signed-rank test of the null hypothesis that HL-A compatibility has no effect on graft survival time. Complete this test. Use  $\alpha = 0.05$  and use the directional alternative that survival time tends to be greater when compatibility score is close.

**8.5.6** In an investigation of possible brain damage due to alcoholism, an X-ray procedure known as a computerized tomography (CT) scan was used to measure brain densities in 11 chronic alcoholics. For each alcoholic, a nonalcoholic control was selected who matched the alcoholic on age, sex, education, and other factors. The brain density measurements on the alcoholics and the matched controls are reported in the accompanying table.<sup>20</sup> Use a Wilcoxon signed-rank test to test the null hypothesis of no difference against the alternative that alcoholism reduces brain density. Let  $\alpha = 0.01$ .

PAIR	ALCOHOLIC	CONTROL	DIFFERENCE
1	40.1	41.3	-1.2
2	38.5	40.2	-1.7
3	36.9	37.4	-0.5
4	41.4	46.1	-4.7
5	40.6	43.9	-3.3
6	42.3	41.9	0.4
7	37.2	39.9	-2.7
8	38.6	40.4	-1.8
9	38.5	38.6	-0.1
10	38.4	38.1	0.3
11	38.1	39.5	-1.4
Mean	39.14	40.66	-1.52
SD	1.72	2.56	1.58

**8.5.7** The study described in Example 8.1.1, on the effect of caffeine on myocardial blood flow, had another component in which 10 subjects had their blood flow measured before and after consuming caffeine, but under different environmental conditions than those for the

subjects in Example 8.1.1.<sup>21</sup> For this setting the differences do not follow a normal distribution, so a *t* test would not be valid. Use a Wilcoxon signed-rank test to test the null hypothesis of no difference against the alternative that caffeine has an effect on myocardial blood flow. Let  $\alpha = 0.01$ .

SUBJECT	BASELINE	CAFFEINE	DIFFERENCE
1	3.43	2.72	0.71
2	3.08	2.94	0.14
3	3.07	1.76	1.31
4	2.65	2.16	0.49
5	2.49	2	0.49
6	2.33	2.37	-0.04
7	2.31	2.35	-0.04
8	2.24	2.26	-0.02
9	2.17	1.72	0.45
10	1.34	1.22	0.12
Mean	2.51	2.15	0.36
SD	0.59	0.50	0.43

## 8.6 Perspective

In this section we consider some limitations to the analysis of paired data.

### Before–After Studies

Many studies in the life sciences compare measurements before and after some experimental intervention, which can present another limitation. These studies can be difficult to interpret, because the effect of the experimental intervention may be confounded with other changes over time. For example, in Example 8.2.3 we found significant evidence for a decrease in myocardial blood flow after taking caffeine, but we noted that it is possible that blood flow would have decreased with the passage of time even if the subjects had not taken caffeine. One way to protect against this difficulty is to use randomized concurrent controls, as in the following example.

#### Example 8.6.1

**Biofeedback and Blood Pressure** A medical research team investigated the effectiveness of a biofeedback training program designed to reduce high blood pressure. Volunteers were randomly allocated to a biofeedback group or a control group. All volunteers received health education literature and a brief lecture. In addition, the biofeedback group received eight weeks of relaxation training, aided by biofeedback, meditation, and breathing exercises. The results for systolic blood pressure, before and after the eight weeks, are shown in Table 8.6.1.<sup>22</sup>

Let us analyze the before–after changes by paired *t* tests at  $\alpha = 0.05$ . In the biofeedback group, the mean systolic blood pressure fell by 13.8 mm Hg. To evaluate the statistical significance of this drop, the test statistic is

$$t_s = \frac{13.8}{1.34} = 10.3$$

**Table 8.6.1** Results of biofeedback experiment

Group	<i>n</i>	Systolic blood pressure (mm Hg)			Difference Mean	SE
		Before Mean	After Mean	Difference Mean		
Biofeedback	99	145.2	131.4	13.8	1.34	
Control	93	144.2	140.2	4.0	1.30	

which is highly significant ( $P$ -value  $\ll 0.0001$ ). However, this result alone does not demonstrate the effectiveness of the biofeedback training; the drop in blood pressure might be partly or entirely due to other factors, such as the health education literature or the special attention received by all the participants. Indeed, a paired  $t$  test applied to the control group gives

$$t_s = \frac{4.0}{1.30} = 3.08 \quad 0.001 < P\text{-value} < 0.01$$

Thus, the people who received *no* biofeedback training *also* experienced a statistically significant drop in blood pressure.

To isolate the effect of the biofeedback training, we can compare the experience of the two treatment groups, using an independent-samples  $t$  test *on the two samples of differences*. We again choose  $\alpha = 0.05$ . The difference between the mean changes in the two groups is

$$13.8 - 4.0 = 9.8 \text{ mm Hg}$$

and the standard error of this difference is

$$\sqrt{1.34^2 + 1.30^2} = 1.87$$

Thus, the  $t$  statistic is

$$t_s = \frac{9.8}{1.87} = 5.24$$

This test provides strong evidence ( $P < 0.0001$ ) that the biofeedback program is effective. If the experimental design had not included the control group, then this last crucial comparison would not have been possible, and the support for efficacy of biofeedback would have been shaky indeed. ■

In analyzing real data, it is wise to keep in mind that the statistical methods we have been considering address only limited questions.

The paired  $t$  test is limited in two ways:

1. It is limited to questions concerning  $\bar{D}$ .
2. It is limited to questions about *aggregate* differences.

The second limitation is very broad; it applies not only to the methods of this chapter but also to those of Chapter 7 and to many other elementary statistical techniques. We will discuss these two limitations separately.

## Limitation of $\bar{D}$

One limitation of the paired  $t$  test and confidence interval is simple, but too often overlooked: When some of the  $D$ 's are positive and some are negative, the magnitude of  $\bar{D}$  does not reflect the “typical” magnitude of the  $D$ 's. The following example shows how misleading  $\bar{D}$  can be.

**Example  
8.6.2**

**Measuring Serum Cholesterol** Suppose a clinical chemist wants to compare two methods of measuring serum cholesterol; she is interested in how closely the two methods agree with each other. She takes blood specimens from 400 patients, splits each specimen in half, and assays one half by method A and the other by method B. Table 8.6.2 shows fictitious data, exaggerated to clarify the issue.

<b>Table 8.6.2</b> Serum cholesterol (mg/dl)			
Specimen	Method A	Method B	$d = A - B$
1	200	234	-34
2	284	272	+12
3	146	153	-7
4	263	250	+13
5	258	232	+26
:	:	:	:
400	176	190	-14
Mean	215.2	214.5	0.7
SD	45.6	59.8	18.8

In Table 8.6.2, the sample mean difference is small ( $\bar{d} = 0.7$ ). Furthermore, the data indicate that the population mean difference is small (a 95% confidence interval is  $-1.1 \text{ mg/dl} < \mu_D < 2.5 \text{ mg/dl}$ ). But such discussion of  $\bar{D}$  or  $\mu_D$  does not address the central question, which is: How closely do the methods agree? In fact, Table 8.6.2 indicates that the two methods do not agree well; the individual differences between method A and method B are not small *in magnitude*. The mean  $\bar{d}$  is small because the positive and negative differences tend to cancel each other. A graph similar to Figure 8.3.1 would be very helpful in visually determining how well the methods agree. We would examine such a graph to see how closely the points cluster around the  $y = x$  line as well as to see the spread in the boxplot of differences. To make a numerical assessment of agreement between the methods we should not focus on the mean difference,  $\bar{D}$ . It would be far more relevant to analyze the absolute (unsigned) magnitudes of the  $d$ 's (that is, 34, 12, 7, 13, 26, and so on). These magnitudes could be analyzed in various ways: We could average them, we could count how many are “large” (say, more than 10 mg/dl), and so on. ■

### Limitation of the Aggregate Viewpoint

Consider a paired experiment in which two treatments, say A and B, are applied to the same person. If we apply a  $t$  test, a sign test, or a Wilcoxon signed-rank test, we are viewing the people as an ensemble rather than individually. This is appropriate if we are willing to assume that the difference (if any) between A and B is in a consistent direction for all people—or, at least, that the important features of the difference are preserved even when the people are viewed *en masse*. The following example illustrates the issue.

**Example  
8.6.3**

**Treatment of Acne** Consider a clinical study to compare two medicated lotions for treating acne. Twenty patients participate. Each patient uses lotion A on one (randomly chosen) side of his face and lotion B on the other side. After three weeks, each side of the face is scored for total improvement.

First, suppose that the A side improves more than the B side in 10 patients, while in the other 10 the B side improves more. According to a sign test, this result is in perfect agreement with the null hypothesis. And yet, two very different interpretations are logically possible:

*Interpretation 1:* Treatments A and B are in fact completely equivalent; their action is indistinguishable. The observed differences between A and B sides of the face were entirely due to chance variation.

*Interpretation 2:* Treatments A and B are in fact completely different. For some people (about 50% of the population), treatment A is more effective than treatment B, whereas in the remaining half of the population treatment B is more effective. The observed differences between A and B sides of the face were biologically meaningful.\*

The same ambiguity of interpretation arises if the results favor one treatment over another. For instance, suppose the A side improved more than the B side in 18 of the 20 cases, while B was favored in 2 patients. This result, which is statistically significant ( $P < 0.001$ ), could again be interpreted in two ways. It could mean that treatment A is in fact superior to B for everybody, but chance variation obscured its superiority in two of the patients; or it could mean that A is superior to B for most people, but for about 10% of the population ( $2/10 = 0.10$ ) B is superior to A. ■

The difficulty illustrated by Example 8.6.3 is not confined to experiments with randomized pairs. In fact, it is particularly clear in another type of paired experiment—the measurement of change over time. Consider, for instance, the blood pressure data of Example 8.6.1. Our discussion of that study hinged on an aggregate measure of blood pressure: the mean. If some patients' pressures rose as a result of biofeedback and others fell, these details were ignored in the analysis based on Student's  $t$ ; only the average change was analyzed.

The difficulties described previously aren't only confined to human experiments either. Suppose, for instance, that two fertilizers, A and B, are to be compared in an agronomic field experiment using a paired design, with the data to be analyzed by a paired  $t$  test. If treatment A is superior to B on acid soils, but B is better than A on alkaline soils, this fact would be obscured in an experiment that included soils of both types.

The issue raised by the preceding examples is a very general one. Simple statistical methods such as the sign test and the  $t$  test are designed to evaluate treatment effects *in the aggregate*—that is, *collectively*—for a population of people, or of mice, or of plots of ground. The segregation of differential treatment effects in subpopulations requires more delicate handling, both in design and analysis.

This confinement to the aggregate point of view applies to Chapter 7 (independent samples) even more forcefully than to the present chapter. For instance, if treatment A is given to one group of mice and treatment B to another, it is quite impossible to know how a mouse in group A would have responded if it had received treatment B; the only possible comparison is an aggregate one. In Section 7.11 we

---

\*This may seem farfetched, but phenomena of this kind do occur; as an obvious example, consider the response of patients to blood transfusions of type A or type B blood.

stated that the statistical comparison of independent samples depends on an “implicit assumption”; essentially, the assumption is that the phenomenon under study can be adequately perceived from an aggregate viewpoint.

In many, perhaps most, biological investigations the phenomena of interest are reasonably universal, so that this issue of submerging the individual in the aggregate does not cause a serious problem. Nevertheless, one should not lose sight of the fact that aggregation may obscure important individual detail.

## Reporting of Data

In communicating experimental results, it is desirable to choose a form of reporting that conveys the extra information provided by pairing. With small samples, a graphical approach can be used, as in Figure 8.1.1, where the line segments gave clear visual evidence that blood flow decreased for each subject.

In published reports of biological research, the crucial information related to pairing is often omitted. For instance, a common practice is to report the means and standard deviations of  $Y_1$  and  $Y_2$  but to omit the standard deviation of the difference,  $D$ ! This is a serious error. It is best to report some description of  $D$ , using either a display like Figure 8.1.1, a histogram of the  $D$ 's, or at least the standard deviation of the  $D$ 's.

## Exercises 8.6.1–8.6.4

**8.6.1** Thirty-three men with high serum cholesterol, all regular coffee drinkers, participated in a study to see whether abstaining from coffee would affect their cholesterol level. Twenty-five of the men (chosen at random) drank no coffee for five weeks, while the remaining 8 men drank coffee as usual. The accompanying table shows the serum cholesterol levels (in mg/dl) at baseline (at the beginning of the study) and the change from baseline after five weeks.<sup>23</sup>

NO COFFEE ( $n = 25$ )		USUAL COFFEE ( $n = 8$ )		
	MEAN	SD	MEAN	SD
Baseline	341	37	331	30
Change from baseline	-35	27	+26	56

For the following  $t$  tests use nondirectional alternatives and let  $\alpha = 0.05$ .

- (a) The no-coffee group experienced a 35 mg/dl drop in mean cholesterol level. Use a  $t$  test to assess the statistical significance of this drop.
- (b) The usual-coffee group experienced a 26 mg/dl rise in mean cholesterol level. Use a  $t$  test to assess the statistical significance of this rise.

- (c) Use a  $t$  test to compare the no-coffee mean change (-35) to the usual-coffee mean change (+26).

**8.6.2** Eight young women participated in a study to investigate the relationship between the menstrual cycle and food intake. Dietary information was obtained every day by interview; the study was double-blind in the sense that the participants did not know its purpose and the interviewer did not know the timing of their menstrual cycles. The table shows, for each participant, the average caloric intake for the 10 days preceding and the 10 days following the onset of the menstrual period (these data are for one cycle only). For these data, prepare a display like that of Figure 8.1.1.<sup>24</sup>

PARTICIPANT	FOOD INTAKE (CAL)	
	PREMENSTRUAL	POSTMENSTRUAL
1	2,378	1,706
2	1,393	958
3	1,519	1,194
4	2,414	1,682
5	2,008	1,652
6	2,092	1,260
7	1,710	1,239
8	1,967	1,758

**8.6.3** For each of 29 healthy dogs, a veterinarian measured the glucose concentration in the anterior chamber of the left eye and the right eye, with the results shown in the table.<sup>25</sup>

ANIMAL NUMBER	GLUCOSE (mg/dl)		ANIMAL NUMBER	GLUCOSE (mg/dl)	
	RIGHT EYE	LEFT EYE		RIGHT EYE	LEFT EYE
1	79	79	16	80	80
2	81	82	17	78	78
3	87	91	18	112	110
4	85	86	19	89	91
5	87	92	20	87	91
6	73	74	21	71	69
7	72	74	22	92	93
8	70	66	23	91	87
9	67	67	24	102	101
10	69	69	25	116	113
11	77	78	26	84	80
12	77	77	27	78	80
13	84	83	28	94	95
14	83	82	29	100	102
15	74	75			

Using the paired *t* method, a 95% confidence interval for the mean difference is  $-1.1 \text{ mg/dl} < \mu_D < 0.7 \text{ mg/dl}$ . Does this result suggest that, for the typical dog in the population, the difference in glucose concentration between the two eyes is less than 1.1 mg/dl? Explain.

**8.6.4** Tobramycin is a powerful antibiotic. To minimize its toxic side effects, the dose can be individualized for each patient. Thirty patients participated in a study of the accuracy of this individualized dosing. For each patient, the predicted peak concentration of Tobramycin in the blood serum was calculated, based on the patient's age, sex, weight, and other characteristics. Then Tobramycin was administered and the actual peak concen-

tration ( $\mu\text{g/ml}$ ) was measured. The results were reported as in the table.<sup>26</sup>

	PREDICTED	ACTUAL
Mean	4.52	4.40
SD	0.90	0.85
<i>n</i>	30	30

Does the reported summary give enough information for you to judge whether the individualized dosing is, on the whole, accurate in its prediction of peak concentration? If so, describe how you would make this judgment. If not, describe what additional information you would need and why.

## Supplementary Exercises 8.S.1–8.S.23

**8.S.1** A volunteer working at an animal shelter conducted a study of the effect of catnip on cats at the shelter. She recorded the number of “negative interactions” each of 15 cats made in 15-minute periods before and after being given a teaspoon of catnip. The paired measurements were collected on the same day within 30 minutes of one another; the data are given in the accompanying table.<sup>27</sup>

- (a) Construct a 95% confidence interval for the difference in mean number of negative interactions.
- (b) Construct a 95% confidence interval the wrong way, using the independent-samples method. How does this interval differ from the one obtained in part (a)?

CAT	BEFORE ( $Y_1$ )	AFTER ( $Y_2$ )	DIFFERENCE
Amelia	0	0	0
Bathsheba	3	6	-3
Boris	3	4	-1
Frank	0	1	-1
Jupiter	0	0	0
Lupine	4	5	-1
Madonna	1	3	-2
Michelangelo	2	1	1
Oregano	3	5	-2
Phantom	5	7	-2
Posh	1	0	1
Sawyer	0	1	-1
Scary	3	5	-2
Slater	0	2	-2
Tucker	2	2	0
Mean	1.8	2.8	-1
SD	1.66	2.37	1.20

**8.S.2** Refer to Exercise 8.S.1. Compare the before and after populations using a  $t$  test at  $\alpha = 0.05$ . Use a nondirectional alternative.

**8.S.3** Refer to Exercise 8.S.1.

Compare the before and after populations using a sign test at  $\alpha = 0.05$ . Use a nondirectional alternative.

**8.S.4** Refer to Exercise 8.S.1. Construct a scatterplot of the data. Does the appearance of the scatterplot indicate that the pairing was effective? Explain.

**8.S.5** As part of a study of the physiology of wheat maturation, an agronomist selected six wheat plants at random from a field plot. For each plant, she measured the moisture content in two batches of seeds: one batch from the “central” portion of the wheat head, and one batch from the “top” portion, with the results shown in the following table.<sup>28</sup> Construct a 90% confidence interval for the mean difference in moisture content of the two regions of the wheat head.

PLANT	PERCENT MOISTURE	
	CENTRAL	TOP
1	62.7	59.7
2	63.6	61.6
3	60.9	58.2
4	63.0	60.5
5	62.7	60.6
6	63.7	60.8

**8.S.6** Biologists noticed that some stream fishes are most often found in pools, which are deep, slow-moving parts of the stream, while others prefer riffles, which are shallow, fast-moving regions. To investigate whether these two habitats support equal levels of diversity (i.e., equal numbers of species), they captured fish at 15 locations along a river. At each location, they recorded the number of species captured in a riffle and the number captured in an adjacent pool. The following table contains the data.<sup>29</sup> Construct a 90% confidence interval for the difference in mean diversity between the types of habitats.

LOCATION	POOL	RIFFLE	DIFFERENCE
1	6	3	3
2	6	3	3
3	3	3	0
4	8	4	4
5	5	2	3
6	2	2	0
7	6	2	4
8	7	2	5
9	1	2	-1
10	3	2	1
11	4	3	1
12	5	1	4
13	4	3	1
14	6	2	4
15	4	3	1
Mean	4.7	2.5	2.2
SD	1.91	0.74	1.86

**8.S.7** Refer to Exercise 8.S.6. What conditions are necessary for the confidence interval to be valid? Are those conditions satisfied? How do you know?

**8.S.8** Refer to Exercise 8.S.6. Compare the habitats using a  $t$  test at  $\alpha = 0.10$ . Use a nondirectional alternative.

**8.S.9** Refer to Exercise 8.S.6.

- Compare the habitats using a sign test at  $\alpha = 0.10$ . Use a nondirectional alternative.
- Use the binomial formula to calculate the exact  $P$ -value for part (a).

**8.S.10** Refer to Exercise 8.S.6. Analyze these data using a Wilcoxon signed-rank test.

**8.S.11** Refer to the Wilcoxon signed-rank test from Exercise 8.S.10. On what grounds could it be argued that the

*P*-value found in this test might not be accurate? This is, why might it be argued that the Wilcoxon test *P*-value is not a completely accurate measure of the strength of the evidence against  $H_0$  in this case?

**8.S.12** In a study of the effect of caffeine on muscle metabolism, nine male volunteers underwent arm exercise tests on two separate occasions. On one occasion, the volunteer took a placebo capsule an hour before the test; on the other occasion he received a capsule containing pure caffeine. (The time order of the two occasions was randomly determined.) During each exercise test, the subject's respiratory exchange ratio (RER) was measured. The RER is the ratio of carbon dioxide produced to oxygen consumed and is an indicator of whether energy is being obtained from carbohydrates or from fats. The results are presented in the accompanying table.<sup>30</sup> Use a *t* test to assess the effect of caffeine. Use a nondirectional alternative and let  $\alpha = 0.05$ .

SUBJECT	RER (%)	
	PLACEBO	CAFFEINE
1	105	96
2	119	99
3	92	89
4	97	95
5	96	88
6	101	95
7	94	88
8	95	93
9	98	88

**8.S.13** For the data of Exercise 8.S.12, construct a display like that of Figure 8.1.1.

**8.S.14** Refer to Exercise 8.S.12. Analyze these data using a sign test.

**8.S.15** Certain types of nerve cells have the ability to regenerate a part of the cell that has been amputated. In an early study of this process, measurements were made on the nerves in the spinal cord in rhesus monkeys. Nerves emanating from the left side of the cord were cut, while nerves from the right side were kept intact. During the regeneration process, the content of creatine phosphate (CP) was measured in the left and the right portion of the spinal cord. The following table shows the data for the right (control) side ( $Y_1$ ), and for the left (regenerating) side ( $Y_2$ ). The units of measurement are mg CP per 100 gm tissue.<sup>31</sup> Use a *t* test to compare the two sides at  $\alpha = 0.05$ . Use a nondirectional alternative.

ANIMAL	RIGHT SIDE (CONTROL)	LEFT SIDE (REGENERATING)	DIFFERENCE
1	16.3	11.5	4.8
2	4.8	3.6	1.2
3	10.9	12.5	-1.6
4	14.2	6.3	7.9
5	16.3	15.2	1.1
6	9.9	8.1	1.8
7	29.2	16.6	12.6
8	22.4	13.1	9.3
Mean	15.50	10.86	4.64
SD	7.61	4.49	4.89

**8.S.16** Aldosterone is a hormone involved in maintaining fluid balance in the body. In a veterinary study, six dogs with heart failure were treated with the drug Captopril, and plasma concentrations of aldosterone were measured before and after the treatment. The results are given in the following table.<sup>32</sup> Use a sign test at  $\alpha = 0.10$ , and a nondirectional alternative, to investigate the claim that Captopril affects aldosterone level.

ANIMAL	BEFORE	AFTER	DIFFERENCE
1	749	374	375
2	469	300	169
3	343	146	197
4	314	134	180
5	286	69	217
6	223	20	203
Mean	397.3	173.8	223.5
SD	190.5	136.4	76.1

**8.S.17** Refer to Exercise 8.S.16. Analyze these data using a Wilcoxon signed-rank test.

**8.S.18** Refer to Exercise 8.S.16. Note that the dogs in this study are not compared to a control group. How does this weaken any inference that might be made about the effectiveness of Captopril?

**8.S.19 (Computer exercise)** For an investigation of the mechanism of wound healing, a biologist chose a paired design, using the left and right hindlimbs of the salamander *Notophthalmus viridescens*. After amputating each limb, she made a small wound in the skin and then kept the limb for 4 hours in either a solution containing benzamil or a control solution. She theorized that the benzamil would impair the healing. The accompanying table shows the amount of healing, expressed as the area ( $\text{mm}^2$ ) covered with new skin after 4 hours.<sup>33</sup>

ANIMAL	CONTROL LIMB	BENZAMIL LIMB	ANIMAL	CONTROL LIMB	BENZAMIL LIMB
1	0.55	0.14	10	0.42	0.21
2	0.15	0.08	11	0.49	0.11
3	0.00	0.00	12	0.08	0.03
4	0.13	0.13	13	0.32	0.14
5	0.26	0.10	14	0.18	0.37
6	0.07	0.08	15	0.35	0.25
7	0.20	0.11	16	0.03	0.05
8	0.16	0.00	17	0.24	0.16
9	0.03	0.05			

- (a) Assess the effect of benzamil using a  $t$  test at  $\alpha = 0.05$ . Let the alternative hypothesis be that the researcher's expectation is correct.
- (b) Proceed as in part (a) but use a sign test.
- (c) Construct a 95% confidence interval for the mean effect of benzamil.
- (d) Construct a scatterplot of the data. Does the appearance of the scatterplot indicate that the pairing was effective? Explain.

**8.S.20 (Computer exercise)** In a study of hypnotic suggestion, 16 male volunteers were randomly allocated to an experimental group and a control group. Each subject participated in a two-phase experimental session. In the first phase, respiration was measured while the subject was awake and at rest. (These measurements were also described in Exercises 7.5.6 and 7.10.4.) In the second phase, the subject was told to imagine that he was performing muscular work, and respiration was measured again.

For subjects in the experimental group, hypnosis was induced between the first and second phases; thus, the suggestion to imagine muscular work was "hypnotic suggestion" for experimental subjects and "waking suggestion" for control subjects. The accompanying table shows the measurements of total ventilation (liters of air per minute per square meter of body area) for all 16 subjects.<sup>34</sup>

- (a) Use a  $t$  test to compare the mean resting values in the two groups. Use a nondirectional alternative and let  $\alpha = 0.05$ . This is the same as Exercise 7.5.6(a).
- (b) Use suitable paired and unpaired  $t$  tests to investigate (i) the response of the experimental group to suggestion; (ii) the response of the control group to suggestion; (iii) the difference between the responses of the experimental and control groups. Use directional alternatives (suggestion increases ventilation, and hypnotic suggestion increases it more than waking suggestion) and let  $\alpha = 0.05$  for each test.

SUBJECT	EXPERIMENTAL GROUP		CONTROL GROUP		
	REST	WORK	SUBJECT	REST	WORK
1	5.74	6.24	9	6.21	5.50
2	6.79	9.07	10	4.50	4.64
3	5.32	7.77	11	4.86	4.61
4	7.18	16.46	12	4.78	3.78
5	5.60	6.95	13	4.79	5.41
6	6.06	8.14	14	5.70	5.32
7	6.32	11.72	15	5.41	4.54
8	6.34	8.06	16	6.08	5.98

- (c) Repeat the investigations of part (b) using suitable nonparametric tests (sign and Wilcoxon-Mann-Whitney tests).
- (d) Use suitable graphs to investigate the reasonableness of the normality condition underlying the  $t$  tests of part (b). How does this investigation shed light on the discrepancies between the results of parts (b) and (c)?

**8.S.21** Suppose we want to test whether an experimental drug reduces blood pressure more than does a placebo. We are planning to administer the drug or the placebo to some subjects and record how much their blood pressures are reduced. We have 20 subjects available.

- (a) We could form 10 matched pairs, where we form a pair by matching subjects, as best we can, on the basis of age and sex, and then randomly assign one subject in each pair to the drug and the other subject in the pair to the placebo. Explain why using a matched pairs design might be a good idea.
- (b) Briefly explain why a matched pairs design might *not* be a good idea. That is, how might such a design be inferior to a completely randomized design?

**8.S.22** A group of 20 postmenopausal women were given transdermal estradiol for one month. Plasma levels of

plasminogen-activator inhibitor type 1 (PAI-1) went down for 10 of the women and went up for the other 10 women.<sup>35</sup> Use a sign test to test the null hypothesis that transdermal estradiol has no effect on PAI-1 level. Use  $\alpha = 0.05$  and use a nondirectional alternative.

**8.S.23** Six patients with renal disease underwent plasmapheresis. Urinary protein excretion (grams of protein per gram of creatinine) was measured for each patient before and after plasmapheresis. The data are given in the following table.<sup>36</sup> Use these data to investigate whether or not plasmapheresis affects urinary protein excretion in patients with renal disease. (*Hint:* Graph the data and consider whether a *t* test is appropriate in the original scale.)

PATIENT	BEFORE	AFTER	DIFFERENCE
1	20.3	0.8	19.5
2	9.3	0.1	9.2
3	7.6	3.0	4.6
4	6.1	0.6	5.5
5	5.8	0.9	4.9
6	4.0	0.2	3.8
Mean	8.9	0.9	7.9
SD	5.9	1.1	6.0

# CATEGORICAL DATA: ONE-SAMPLE DISTRIBUTIONS

## Objectives

In this chapter we study categorical data. We will

- explore sampling distributions for estimators that describe dichotomous populations.
- demonstrate how to make and interpret confidence intervals for proportions.
- provide a method for finding an optimal sample size for estimating a proportion.
- show how and when to conduct a chi-square goodness-of-fit test.

### 9.1 Dichotomous Observations

In Chapter 5 we worked with problems involving numeric variables and examined the sampling distribution of the sample mean. In Chapter 6 we used the sampling distribution to explain how the sample mean tends to vary from the population mean and we constructed confidence intervals for the population mean. We begin this chapter by proceeding in a similar manner by first considering a simple dichotomous categorical variable (i.e., a categorical variable that has only two possible values) and the sampling distribution of the sample proportion. In Section 9.2 we will use the sampling distribution of the sample proportion to construct a confidence interval for a population proportion.

#### The Wilson-Adjusted Sample Proportion, $\tilde{p}$

When sampling from a large dichotomous population, a natural estimate of the population proportion,  $p$ , is the sample proportion,  $\hat{p} = y/n$ , where  $y$  is the number of observations in the sample with the attribute of interest and  $n$  is the sample size.

**Example**  
**9.1.1**

**Contaminated Soda** At any given time, soft-drink dispensers may harbor bacteria such as *Chryseobacterium meningosepticum* that can cause illness.<sup>1</sup> To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum*. Thus the sample proportion of contaminated dispensers is

$$\hat{p} = \frac{5}{30} = 0.167$$

The estimate,  $\hat{p} = 0.167$ , given in Example 9.1.1 is a good estimate of the population proportion of contaminated soda dispensers, but it is not the only possible estimate. The Wilson-adjusted sample proportion,  $\tilde{p}$ , is another estimate of the population proportion and is given by the formula in the following box.

### Wilson-Adjusted Sample Proportion, $\tilde{p}$

$$\tilde{p} = \frac{y + 2}{n + 4}$$

#### Example 9.1.2

**Contaminated Soda** The Wilson-adjusted sample proportion of contaminated dispensers is

$$\tilde{p} = \frac{5 + 2}{30 + 4} = 0.206^*$$

As the previous example illustrates,  $\tilde{P}$  is equivalent to computing the ordinary sample proportion  $\hat{P}$  on an augmented sample: one that includes four extra observations of soft-drink dispensers—two that are contaminated and two that are not. This augmentation has the effect of biasing the estimate towards the value 1/2. Generally speaking we would like to avoid biased estimates, but as we shall see in Section 9.2, confidence intervals based on this biased estimate,  $\tilde{P}$ , actually are more reliable than those based on  $\hat{P}$ .

### The Sampling Distribution of $\tilde{P}$

For random sampling from a large dichotomous population, we saw in Chapter 3 how to use the binomial distribution to calculate the probabilities of all the various possible sample compositions. These probabilities in turn determine the sampling distribution of  $\tilde{P}$ . An example follows.

#### Example 9.1.3

**Contaminated Soda** Suppose that in a certain region of the United States, 17% of all soft-drink dispensers are contaminated with *Chryseobacterium meningosepticum*. If we were to examine a random sample of two drink dispensers from this population of dispensers, then we will get either zero, one, or two contaminated machines. The probability that both dispensers are contaminated is  $0.17 \times 0.17 = 0.0289$ . The probability that neither are contaminated is  $(1 - 0.17) \times (1 - 0.17) = 0.6889$ . There are two ways to get a sample in which one machine is contaminated and one is not: The first could be contaminated, but not the second, or vice versa. Thus, the probability that exactly one machine is contaminated is

$$0.17 \times (1 - 0.17) + 0.17 \times (1 - 0.17) = 0.2822$$

If we let  $\tilde{P}$  represent the Wilson-adjusted sample proportion of contaminated dispensers, then a sample that contains no contaminated dispensers has  $\tilde{p} = \frac{0 + 2}{2 + 4} = 0.33$ , which occurs with probability 0.6889. A sample that contains one contaminated machine has  $\tilde{p} = \frac{1 + 2}{2 + 4} = 0.50$ ; this happens with probability 0.2822. Finally, a sample that contains two contaminated machines has  $\tilde{p} = \frac{2 + 2}{2 + 4} = 0.67$ , which occurs with probability 0.0289.<sup>†</sup> Thus, there is roughly a 69% chance that  $\tilde{P}$  will equal 0.33, a 28% chance that  $\tilde{P}$  will equal 0.50, and a 3% chance that  $\tilde{P}$  will equal 0.67.

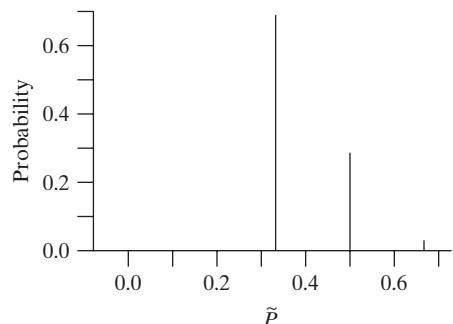
\*In keeping with our convention,  $\tilde{P}$  denotes a random variable, whereas  $\tilde{p}$  denotes a particular number (such as 0.206 in this example).

<sup>†</sup>It is worth noting that with a small sample size ( $n = 2$ ) the possible values of  $\tilde{p}$  are 0.33, 0.50, and 0.67 while the possible values of  $\hat{p}$  are 0.00, 0.50, and 1.00. This sheds some light as to why  $\tilde{p}$  is a sensible estimator of the population proportion, particularly for small samples. With a small sample it is quite likely that one could obtain no contaminated machines even if a reasonable proportion of the population is contaminated. It would be unwise, with such a small sample, to assert that the population proportion of contaminated machines is 0.

This sampling distribution is given in Table 9.1.1 and Figure 9.1.1. ■

**Table 9.1.1** Sampling distribution of  $Y$  (the number of contaminated dispensers) and of  $\tilde{P}$  (the Wilson-adjusted proportion of contaminated dispensers) for samples of size  $n = 2$  for a population with 17% of the dispensers contaminated

$Y$	$\tilde{P}$	Probability
0	0.33	0.6889
1	0.50	0.2822
2	0.67	0.0289



**Figure 9.1.1** Sampling distribution of  $\tilde{P}$  for  $n = 2$  and  $p = 0.17$

**Example  
9.1.4**

**Contaminated Soda and a Larger Sample** Suppose we were to examine a sample of 20 dispensers from a population in which 17% are contaminated. How many contaminated dispensers might we expect to find in the sample? As was true in Example 9.1.3, this question can be answered in the language of probability. However, since  $n = 20$  is rather large, we will not list each possible sample. Rather, we will make calculations using the binomial distribution with  $n = 20$  and  $p = 0.17$ . For instance, let us calculate the probability that 5 dispensers in the sample would be contaminated and 15 would not:

$$\begin{aligned}\Pr\{5 \text{ contaminated, } 15 \text{ not contaminated}\} &= {}_{20}C_5(0.17)^5(0.83)^{15} \\ &= 15,504(0.17)^5(0.83)^{15} \\ &= 0.1345\end{aligned}$$

Letting  $\tilde{P}$  represent the Wilson-adjusted sample proportion of contaminated dispensers, a sample that contains 5 contaminated dispensers has  $\tilde{p} = \frac{5 + 2}{20 + 4} = 0.2917$ . Thus, we have found that

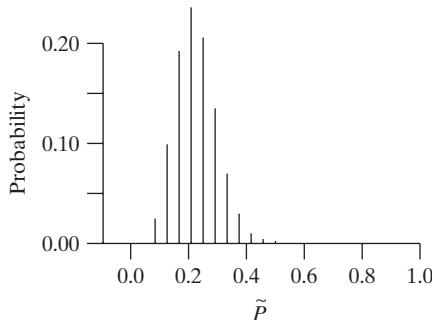
$$\Pr\{\tilde{P} = 0.2917\} = 0.1345$$

The binomial distribution can be used to determine the entire sampling distribution of  $\tilde{P}$ . The distribution is displayed in Table 9.1.2 and as a probability histogram in Figure 9.1.2.

**Table 9.1.2** Sampling distribution of  $Y$ , the number of successes, and of  $\tilde{P}$ , the Wilson-adjusted proportion of successes, when  $n = 20$  and  $p = 0.17$

$Y$	$\tilde{P}$	Probability	$Y$	$\tilde{P}$	Probability
0	0.0833	0.0241	11	0.5417	0.0001
1	0.1250	0.0986	12	0.5833	0.0000
2	0.1667	0.1919	13	0.6250	0.0000
3	0.2083	0.2358	14	0.6667	0.0000
4	0.2500	0.2053	15	0.7083	0.0000
5	0.2917	0.1345	16	0.7500	0.0000
6	0.3333	0.0689	17	0.7917	0.0000
7	0.3750	0.0282	18	0.8333	0.0000
8	0.4167	0.0094	19	0.8750	0.0000
9	0.4583	0.0026	20	0.9167	0.0000
10	0.5000	0.0006			

**Figure 9.1.2** Sampling distribution of  $\tilde{P}$  when  $n = 20$  and  $p = 0.17$



We can use this distribution to answer questions such as “If we take a random sample of size  $n = 20$ , what is the probability that no more than 5 will be contaminated?” Notice that this question can be asked in two equivalent ways: “What is  $\Pr\{Y \leq 5\}$ ?” and “What is  $\Pr\{\tilde{P} \leq 0.2917\}$ ?” The answer to either question is found by adding the first six probabilities in Table 9.1.2:

$$\begin{aligned}\Pr\{Y \leq 5\} &= \Pr\{\tilde{P} \leq 0.2917\} \\ &= 0.0241 + 0.0986 + 0.1919 + 0.2358 + 0.2053 + 0.1345 \\ &= 0.8902\end{aligned}$$

■

## Relationship to Statistical Inference

In making a statistical inference from a sample to the population, it is reasonable to use  $\tilde{P}$  as our estimate of  $p$ . The sampling distribution of  $\tilde{P}$  can be used to predict how much sampling error to expect in this estimate. For example, suppose we want to know whether the sampling error will be less than 5 percentage points, in other words, whether  $\tilde{P}$  will be within  $\pm 0.05$  of  $p$ . We cannot predict for certain whether this event will occur, but we can find the probability of it happening, as illustrated in the following example.

### Example 9.1.5

**Contaminated Soda** In the soda-dispenser example with  $n = 20$ , we see from Table 9.1.2 that

$$\begin{aligned}\Pr\{0.12 \leq \tilde{P} \leq 0.22\} &= 0.0986 + 0.1919 + 0.2358 \\ &= 0.5263 \approx 0.53\end{aligned}$$

Thus, there is a 53% chance that, for a sample of size 20,  $\tilde{P}$  will be within  $\pm 0.05$  of  $p$ .

■

## Dependence on Sample Size

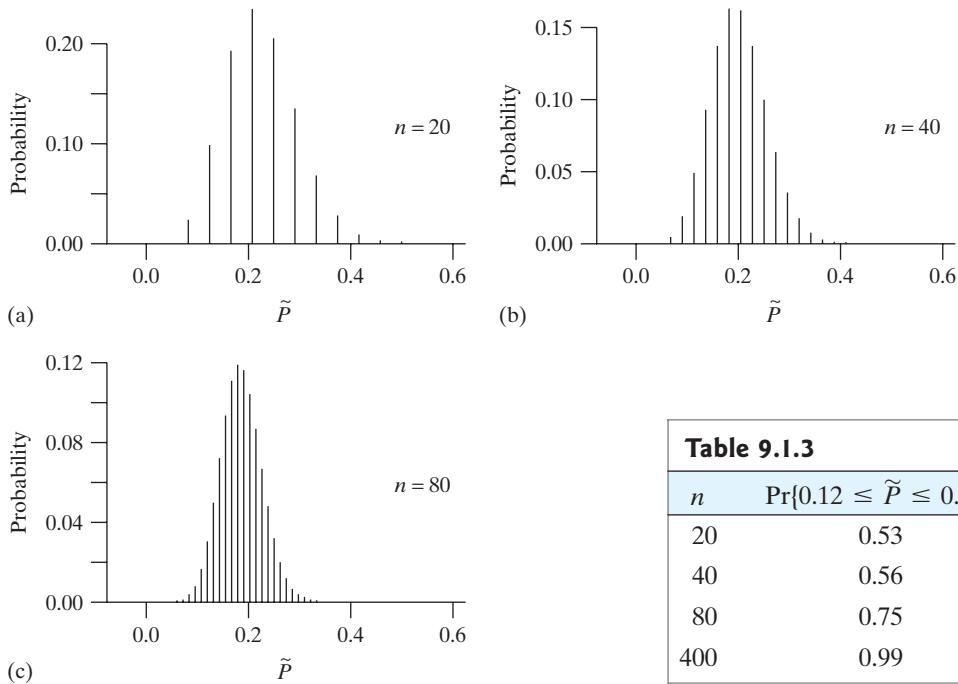
Just as the sampling distribution of  $\bar{Y}$  depends on  $n$ , so does the sampling distribution of  $\tilde{P}$ . The larger the value of  $n$ , then the more likely it is  $\tilde{P}$  will be close to  $p$ .\* The following example illustrates this effect.

### Example 9.1.5

**Contaminated Soda** Figure 9.1.3 shows the sampling distribution of  $\tilde{P}$ , for three different values of  $n$ , for the soft-drink dispenser population of Example 9.1.1. (Each sampling distribution is determined by a binomial distribution with  $p = 0.17$ .)

\*This statement should not be interpreted too literally. As a function of  $n$ , the probability that  $\tilde{P}$  is close to  $p$  has an overall increasing trend, but it can fluctuate somewhat.

**Figure 9.1.3** Sampling distributions of  $\tilde{P}$  for  $p = 0.17$  and various values of  $n$



**Table 9.1.3**

$n$	$\Pr\{0.12 \leq \tilde{P} \leq 0.22\}$
20	0.53
40	0.56
80	0.75
400	0.99

You can see from the figure that as  $n$  increases, the sampling distribution becomes more compressed around the value  $p = 0.17$ ; thus, the probability that  $\tilde{P}$  is close to  $p$  tends to increase as  $n$  increases. For example, consider the probability that  $\tilde{P}$  is within  $\pm 5$  percentage points of  $p$ . We saw in Example 9.1.5 that for  $n = 20$  this probability is equal to 0.53; Table 9.1.3 and Figure 9.1.3 shows how the probability depends on  $n$ .

**Note:** A larger sample improves the probability that  $\tilde{P}$  will be close to  $p$ . We should be mindful, however, that the probability that  $\tilde{P}$  is exactly *equal* to  $p$  is very small for large  $n$ . In fact,

$$\Pr\{\tilde{P} = 0.17\} = 0.110 \text{ for } n = 80^*$$

The value  $\Pr\{0.12 \leq \tilde{P} \leq 0.22\} = 0.75$  is the sum of many small probabilities, the largest of which is 0.110; you can see this effect clearly in Figure 9.1.3(c). ■

## Exercises 9.1.1–9.1.10

**9.1.1** Consider taking a random sample of size 3 from a population of persons who smoke and recording how many of them, if any, have lung cancer. Let  $\tilde{P}$  represent the Wilson-adjusted proportion of persons in the sample with lung cancer. What are the possible values in the sampling distribution of  $\tilde{P}$ ?

**9.1.2** Suppose we are to draw a random sample of three individuals from a large population in which 37% of the individuals are mutants (as in Example 3.6.4). Let  $\tilde{P}$  represent the Wilson-adjusted proportion of mutants in the sample. Calculate the probability that  $\tilde{P}$  will be equal to

(a)  $2/7$

(b)  $3/7$

Is it possible to obtain a sample of three individuals for which  $\tilde{P}$  is zero? Explain.

**9.1.3** Suppose we are to draw a random sample of five individuals from a large population in which 37% of the individuals are mutants (as in Example 3.6.4). Let  $\tilde{P}$  represent the Wilson-adjusted proportion of mutants in the sample.

(a) Use the results in Table 3.6.3 to determine the probability that  $\tilde{P}$  will be equal to

- (i)  $2/9$
- (ii)  $3/9$
- (iii)  $4/9$
- (iv)  $5/9$
- (v)  $6/9$
- (vi)  $7/9$

(b) Display the sampling distribution of  $\tilde{P}$  in a graph similar to Figure 9.1.1.

\*For  $n = 80$ ,  $\tilde{P} = 0.1677$  when  $y = 12$ , is the closest possible value to 0.17.

**9.1.4** A new treatment for acquired immune deficiency syndrome (AIDS) is to be tested in a small clinical trial on 15 patients. The Wilson-adjusted proportion  $\tilde{P}$  who respond to the treatment will be used as an estimate of the proportion  $p$  of (potential) responders in the entire population of AIDS patients. If in fact  $p = 0.2$ , and if the 15 patients can be regarded as a random sample from the population, find the probability that

- (a)  $\tilde{P} = 5/19$       (b)  $\tilde{P} = 2/19$

**9.1.5** In a certain forest, 25% of the white pine trees are infected with blister rust. Suppose a random sample of four white pine trees is to be chosen, and let  $\tilde{P}$  be the Wilson-adjusted sample proportion of infected trees.

- (a) Compute the probability that  $\tilde{P}$  will be equal to  
 (i)  $2/8$     (ii)  $3/8$     (iii)  $4/8$     (iv)  $5/8$     (v)  $6/8$   
 (b) Display the sampling distribution of  $\tilde{P}$  in a graph similar to Figure 9.1.1.

**9.1.6** Refer to Exercise 9.1.5.

- (a) Determine the sampling distribution of  $\tilde{P}$  for samples of size  $n = 8$  white pine trees from the same forest.  
 (b) Construct graphs of the sampling distributions of  $\tilde{P}$  for  $n = 4$  and for  $n = 8$ , using the same horizontal and vertical scales for both. Compare the two distributions visually. How do they differ?

**9.1.7** The shell of the land snail *Limocolaria mafensisiana* has two possible color forms: streaked and pallid. In a certain population of these snails, 60% of the individuals

have streaked shells (as in Exercise 3.6.4). Suppose a random sample of six snails is to be chosen from the population; let  $\tilde{p}$  be the Wilson-adjusted sample proportion of streaked snails. Find

- (a)  $\Pr\{\tilde{P} = 0.5\}$       (b)  $\Pr\{\tilde{P} = 0.6\}$   
 (c)  $\Pr\{\tilde{P} = 0.7\}$       (d)  $\Pr\{0.5 \leq \tilde{P} \leq 0.7\}$   
 (e) the percentage of samples for which  $\tilde{P}$  is within  $\pm 0.10$  of  $p$ .

**9.1.8** In a certain community, 17% of the soda dispensers are contaminated (as in Example 9.1.5). Suppose a random sample of five dispensers is to be chosen and the contamination observed. Let  $\tilde{P}$  represent the Wilson-adjusted sample proportion contaminated dispensers.

- (a) Compute the sampling distribution of  $\tilde{P}$ .  
 (b) Construct a histogram of the distribution found in part (a) and compare it visually with Figure 9.1.3. How do the two distributions differ?

**9.1.9** Consider random sampling from a dichotomous population; let  $E$  be the event that  $\tilde{P}$  is within  $\pm .05$  of  $p$ . In Example 9.1.5, we found that  $\Pr\{E\} = 0.53$  for  $n = 20$  and  $p = 0.17$ . Calculate  $\Pr\{E\}$  for  $n = 20$  and  $p = 0.25$ . (Perhaps surprisingly, the two probabilities are roughly equal.)

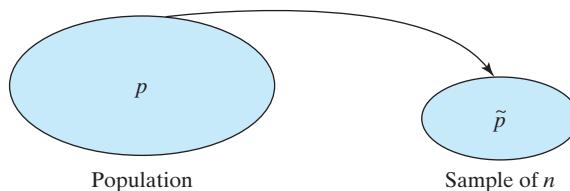
**9.1.10** Consider taking a random sample of size 10 from the population of students at a certain college and asking each of the 10 students whether or not they smoke. In the context of this setting, explain what is meant by the sampling distribution of  $\hat{P}$ , the ordinary sample proportion.

## 9.2 Confidence Interval for a Population Proportion

In Section 6.3 we described confidence intervals when the observed variable is quantitative. Similar ideas can be used to construct confidence intervals in situations in which the variable is *categorical* and the parameter of interest is a population *proportion*. We assume that the data can be regarded as a random sample from some population. In this section we discuss construction of a confidence interval for a population proportion.

Consider a random sample of  $n$  categorical observations, and let us fix attention on one of the categories. For instance, suppose a geneticist observes  $n$  guinea pigs whose coat color can be either black, sepia, cream, or albino; let us fix attention on the category “black.” Let  $p$  denote the population proportion of the category of interest, and let  $\tilde{p}$  denote the Wilson-adjusted sample proportion (as in Section 9.1), which is our estimate of  $p$ . The situation is schematically represented in Figure 9.2.1.

**Figure 9.2.1** Notation for population and sample proportion



How close to  $p$  is  $\tilde{P}$  likely to be? We saw in Section 9.1 that this question can be answered in terms of the sampling distribution of  $\tilde{P}$  (which in turn is computed from the binomial distribution). As we shall see, by using properties of the sampling distribution of  $\tilde{P}$ , such as the standard error and  $\tilde{P}$ 's approximately normal behavior under certain situations, we will be able to construct confidence statements for  $p$ . To construct the intervals we will use the same rationale used for numeric data in Section 6.3 where we constructed confidence statements for  $\mu$  based on the properties of the sampling distribution of  $\bar{Y}$ .

Although a confidence interval for  $p$  can be constructed directly from the binomial distribution, for many practical situations a simple approximate method can be used instead. When the sample size,  $n$ , is large, the sampling distribution of  $\tilde{P}$  is approximately normal; this approximation is related to the Central Limit Theorem. If you review Figure 9.1.2, you will see that the sampling distributions resemble normal curves, especially the distribution with  $n = 80$ . (The approximation is described in detail in optional Section 5.4.) In Section 6.3 we stated that when the data come from a normal population, a 95% confidence interval for a population mean  $\mu$  is constructed as

$$\bar{y} \pm t_{0.025} \text{SE}_{\bar{Y}}$$

A confidence interval for a population proportion  $p$  is constructed analogously. We will use  $\tilde{P}$  as the center of a 95% confidence interval for  $p$ . In order to proceed we need to calculate the standard error for  $\tilde{P}$ .

## Standard Error of $\tilde{P}$

The standard error of the estimate is found using the following formula.

### Standard Error of $\tilde{P}$ (for a 95% Confidence Interval) —

$$\text{SE}_{\tilde{P}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

This formula for the standard error of the estimate looks similar to the formula for the standard error of a mean, but with  $\sqrt{\tilde{p}(1 - \tilde{p})}$  playing the role of  $s$  and with  $n + 4$  in place of  $n$ .

---

#### Example 9.2.1

**Smoking during Pregnancy** As part of the National Survey of Family Growth, 496 women aged 20 to 24 who had given birth were asked about their smoking habits.<sup>2</sup> Smoking during pregnancy was reported by 78 of those sampled, which is 15.7 percent ( $78/496 = 0.157$  or 15.7%). Thus,  $\tilde{p}$  is  $\frac{78 + 2}{496 + 4} = \frac{80}{500} = 0.16$ ; the standard error is  $\sqrt{\frac{0.16(1 - 0.16)}{500}} = 0.016$  or 1.6%. A sample value of  $\tilde{P}$  is typically within  $\pm 2$  standard errors of the population proportion  $p$ . Based on this standard error, we can expect that the proportion,  $p$ , of all women aged 20 to 24 who smoked during pregnancy is in the interval (0.128, 0.192) or (12.8%, 19.2%). A confidence interval for  $p$  makes this idea more precise. ■

## 95% Confidence Interval for $p$

Once we have the standard error of  $\tilde{P}$ , we need to know how likely it is that  $\tilde{P}$  will be close to  $p$ . The general process of constructing a confidence interval for a proportion is similar to that used in Section 6.3 to construct a confidence interval for a

mean. However, when constructing a confidence interval for a mean, we multiplied the standard error by a  $t$  multiplier. This was based on having a sample from a normal distribution. When dealing with proportion data we know that the population is not normal—there only are two values in the population!—but the Central Limit Theorem tells us that the sampling distribution of  $\tilde{P}$  is approximately normal if the sample size,  $n$ , is large. Moreover, it turns out that even for moderate or small samples, intervals based on  $\tilde{P}$  and  $Z$  multipliers do a very good job of estimating the population proportion,  $p$ .<sup>3</sup>

For a 95% confidence interval, the appropriate  $Z$  multiplier is  $z_{0.025} = 1.960$ . Thus, the approximate 95% confidence interval for a population proportion  $p$  is constructed as shown in the following box.\*

### 95% Confidence Interval for $p$

$$95\% \text{ confidence interval: } \tilde{p} \pm 1.96 \text{SE}_{\tilde{p}}$$

#### Example 9.2.2

**Breast Cancer**  $BRCA1$  is a gene that has been linked to breast cancer. Researchers used DNA analysis to search for  $BRCA1$  mutations in 169 women with family histories of breast cancer. Of the 169 women tested, 27 (16%) had  $BRCA1$  mutations.<sup>4</sup> Let  $p$  denote the probability that a woman with a family history of breast cancer will have a  $BRCA1$  mutation. For these data,  $\tilde{p} = \frac{27 + 2}{169 + 4} = 0.168$ . The standard error for  $\tilde{P}$  is  $\sqrt{\frac{0.168(1 - 0.168)}{169 + 4}} = 0.028$ . Thus, a 95% confidence interval for  $p$  is

$$0.168 \pm (1.96)(0.028)$$

or

$$0.168 \pm 0.055$$

or

$$0.113 < p < 0.223$$

Thus, we are 95% confident that the probability of a  $BRCA1$  mutation in a woman with a family history of breast cancer is between 0.113 and 0.223 (i.e., between 11.3% and 22.3%). ■

Note that the size of the standard error is inversely proportional to  $\sqrt{n}$ , as illustrated in the following example.

#### Example 9.2.3

**Breast Cancer** Suppose, as in Example 9.2.2, that a sample of  $n$  women with family histories of breast cancer contains 16% with  $BRCA1$  mutations. Then  $\tilde{p} \approx 0.168$  and

$$\text{SE}_{\tilde{p}} \approx \sqrt{\frac{0.168(0.832)}{n + 4}}$$

We saw in Example 9.2.2 that if  $n = 169$ , then

$$\text{SE}_{\tilde{p}} = 0.028$$

If  $n = 4 \times 169 = 676$ , then

$$\text{SE}_{\tilde{p}} = 0.014$$

---

\*Many statistics books present the confidence interval for a proportion as  $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$  where  $\hat{p} = y/n$ .

This commonly used interval is similar to the interval we present, particularly if  $n$  is large. For small or moderate sample sizes, the interval we present is more likely to cover the population proportion  $p$ . A technical discussion of the Wilson-interval using  $\tilde{P}$  is given in Appendix 9.1.

Thus, a sample with the same composition (that is, 16% with *BRCA1* mutations) but four times as large, would yield twice as much precision in the estimation of  $p$ . ■

The Wilson-adjusted sample proportion can be used to construct a confidence interval for  $p$  even when the sample size is small, as the following example illustrates.

**Example  
9.2.4**

**ECMO** Extracorporeal membrane oxygenation (ECMO) is a potentially life-saving procedure that is used to treat newborn babies who suffer from severe respiratory failure. An experiment was conducted in which 15 babies were treated with ECMO; none of the 15 babies died.<sup>5</sup> Let  $p$  denote the probability of death for a baby treated with ECMO. The fact that none of the babies in the experiment died should not lead us to believe that the probability of death,  $p$ , is precisely zero—only that it is close to zero. The estimate given by  $\tilde{p}$  is  $2/15 = 0.133$ . The standard error of  $\tilde{p}$  is

$$\sqrt{\frac{0.133(0.867)}{15}} = 0.088^*$$

Thus, a 95% confidence interval for  $p$  is

$$0.133 \pm (1.96)(0.088)$$

or

$$0.133 \pm 0.172$$

or

$$-0.039 < p < 0.305$$

We know that  $p$  cannot be negative, so we state the confidence interval as  $(0, 0.305)$ .

Thus, we are 95% confident that the probability of death in a newborn with severe respiratory failure who is treated with ECMO is between 0 and 0.305 (i.e., between 0% and 30.5%). ■

## One-Sided Confidence Intervals

Most confidence intervals are of the form “estimate  $\pm$  margin of error”; these are known as two-sided intervals. However, it is possible to construct a one-sided confidence interval, which is appropriate when only a lower bound, or only an upper bound, is of interest. The following example provides an illustration.

**Example  
9.2.5**

**ECMO—One-Sided** Consider the ECMO data from Example 9.2.4, which are used to estimate the probability of death,  $p$ , in a newborn with severe respiratory failure. We know that  $p$  cannot be less than zero, but we might want to know how large  $p$  might be. Whereas a two-sided confidence interval is based on capturing the middle 95% of a standard normal distribution and thus uses the  $Z$  multipliers of  $\pm 1.96$ , a one-sided 95% (upper) confidence interval uses the fact that  $\Pr(-\infty < Z < 1.645) = 0.95$ . Thus, the upper limit of the confidence interval is  $\tilde{p} + 1.645 \times \text{SE}_{\tilde{p}}$  and the lower limit of the interval is negative infinity. In this case we get

$$0.133 + (1.645)(0.088) = 0.133 + 0.145 = 0.278$$

as the upper limit. The resulting interval is  $(-\infty, 0.278)$ , but since  $p$  cannot be negative, we state the confidence interval as  $(0, 0.278)$ . That is, we are 95% confident that the probability of death is at most 27.8%. ■

---

\*Note that if we used the commonly presented method of  $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$  we would find that the standard error is zero, leading to a confidence interval of  $0 \pm 0$ . Such an interval would not seem to be very useful in practice!

## Planning a Study to Estimate $p$

In Section 6.4 we discussed a method for choosing the sample size  $n$  so that a proposed study would have sufficient precision for its intended purpose. The approach depended on two elements: (1) a specification of the desired  $\text{SE}_{\bar{Y}}$  and (2) a preliminary guess of the SD. In the present context, when the observed variable is categorical, a similar approach can be used. If a desired value of  $\text{SE}_{\tilde{p}}$  is specified, and if a rough informed guess of  $\tilde{p}$  is available, then the required sample size  $n$  can be determined from the following equation:

$$\text{Desired SE} = \sqrt{\frac{(\text{Guessed } \tilde{p})(1 - \text{Guessed } \tilde{p})}{n + 4}}$$

The following example illustrates the use of the method.

### Example 9.2.6

**Left-Handedness** In a survey of English and Scottish college students, 40 of 400 male students were found to be left-handed.<sup>6</sup>

The sample estimate of the proportion is

$$\tilde{p} = \frac{40 + 2}{400 + 4} \approx 0.104$$

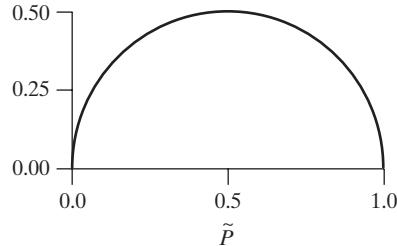
Suppose we regard these data as a pilot study and we now wish to plan a study large enough to estimate  $p$  with a standard error of one percentage point, that is, 0.01. We choose  $n$  to satisfy the following relation:

$$\sqrt{\frac{0.104(0.896)}{n + 4}} \leq 0.01$$

This equation is easily solved to give  $n + 4 \geq 931.8$ . We should plan a sample of 928 students. ■

**Planning in Ignorance** Suppose no preliminary informed guess of  $p$  is available. Remarkably, in this situation it is still possible to plan an experiment to achieve a desired value of  $\text{SE}_{\tilde{p}}$ .\* Such a “blind” plan depends on the fact that the crucial quantity  $\sqrt{\tilde{p}(1 - \tilde{p})}$  is largest when  $\tilde{p} = 0.5$ ; you can see this in the graph of Figure 9.2.2. It follows that a value of  $n$  calculated using “guessed  $\tilde{p}$ ” = 0.5 will be *conservative*—that is, it will certainly be large enough. (Of course, it will be much larger than necessary if  $\tilde{p}$  is really very different from 0.5.) The following example shows how such “worst-case” planning is used.

**Figure 9.2.2** How  $\sqrt{\tilde{p}(1 - \tilde{p})}$  depends on  $\tilde{p}$



### Example 9.2.7

**Left-Handedness** Suppose, as in Example 9.2.6, that we are planning a study of left-handedness and that we want  $\text{SE}_{\tilde{p}}$  to be 0.01, but suppose that we have no preliminary

\*By contrast, it would not be possible if we were planning a study to estimate a population mean  $\mu$  and we had no information whatsoever about the value of the SD.

information whatsoever. We can proceed as in Example 9.2.6, but using a guessed value of  $\tilde{p}$  of 0.5. Then we have

$$\sqrt{\frac{0.5(0.5)}{n+4}} \leq 0.01$$

which means that  $n+4 \geq 2500$ , so we need  $n = 2,496$ . Thus, a sample of 2,496 students would be adequate to estimate  $p$  with a standard error of 0.01, regardless of the actual value of  $p$ . (Of course, if  $p = 0.1$ , this value of  $n$  is much larger than is necessary.) ■

## Exercises 9.2.1–9.2.13

---

**9.2.1** A series of patients with bacterial wound infections were treated with the antibiotic Cefotaxime. Bacteriologic response (disappearance of the bacteria from the wound) was considered “satisfactory” in  $\tilde{p} = .84$  of the patients.<sup>7</sup> Determine the standard error of  $\tilde{P}$ , the Wilson-adjusted observed proportion of “satisfactory” responses, if the series contained

- (a) 50 patients of whom 42 were considered “satisfactory.”
- (b) 200 patients of whom 168 were considered “satisfactory.”

**9.2.2** In an experiment with a certain mutation in the fruitfly *Drosophila*,  $n$  individuals were examined; of these, 20% were found to be mutants. Determine the standard error of  $\tilde{P}$  if

- (a)  $n = 100$  (20 mutants).
- (b)  $n = 400$  (80 mutants).

**9.2.3** Refer to Exercise 9.2.2. In each case ( $n = 100$  and  $n = 400$ ) construct a 95% confidence interval for the population proportion of mutants.

**9.2.4** In a natural population of mice (*Mus musculus*) near Ann Arbor, Michigan, the coats of some individuals are white spotted on the belly. In a sample of 580 mice from the population, 28 individuals were found to have white-spotted bellies.<sup>8</sup> Construct a 95% confidence interval for the population proportion of this trait.

**9.2.5** To evaluate the policy of routine vaccination of infants for whooping cough, adverse reactions were monitored in 339 infants who received their first injection of vaccine. Reactions were noted in 69 of the infants.<sup>9</sup>

- (a) Construct a 95% confidence interval for the probability of an adverse reaction to the vaccine.
- (b) Interpret the confidence interval from part (a). What does the interval say about whooping cough vaccinations?
- (c) Using your interval from part (a), can we be confident that the probability of an adverse reaction to the vaccine is less than 0.25?
- (d) What level of confidence is associated with your answer to part (c)? (Hint: What is the associated one-sided interval confidence level?)

**9.2.6** In a study of human blood types in nonhuman primates, a sample of 71 orangutans were tested and 14 were found to be blood type B.<sup>10</sup> Construct a 95% confidence interval for the relative frequency of blood type B in the orangutan population.

**9.2.7** In populations of the snail *Cepaea*, the shells of some individuals have dark bands, while other individuals have unbanded shells.<sup>11</sup> Suppose that a biologist is planning a study to estimate the percentage of banded individuals in a certain natural population, and that she wants to estimate the percentage—which she anticipates will be in the neighborhood of 60%—with a standard error not to exceed 4 percentage points. How many snails should she plan to collect?

**9.2.8 (Continuation of Exercise 9.2.7)** What would the answer be if the anticipated percentage of banded snails were 50% rather than 60%?

**9.2.9** The ability to taste the compound phenylthiocarbamide (PTC) is a genetically controlled trait in humans. In Europe and Asia, about 70% of people are “tasters.”<sup>12</sup> Suppose a study is being planned to estimate the relative frequency of tasters in a certain Asian population, and it is desired that the standard error of the estimated relative frequency should be 0.01. How many people should be included in the study?

**9.2.10** Refer to Exercise 9.2.9. Suppose a study is being planned for a part of the world for which the percentage of tasters is completely unknown, so that the 70% figure used in Exercise 9.2.9 is not applicable. What sample size is needed so that the standard error will be no larger than 0.01?

**9.2.11** Refer to Exercise 9.2.9. Suppose the SE requirement is relaxed by a factor of 2—from 0.01 to 0.02. Would this reduce the required sample size by a factor of 2? Explain.

**9.2.12** The “Luso” variety of wheat is resistant to the Hessian fly. In order to understand the genetic mechanism controlling this resistance, an agronomist plans to examine the progeny of a certain cross involving Luso and a nonresistant variety. Each progeny plant will be classified as resistant or susceptible and the agronomist

will estimate the proportion of progeny that are resistant.<sup>13</sup> How many progeny does he need to classify in order to guarantee that the standard error of his estimate of this proportion will not exceed 0.05?

**9.2.13 (Continuation of Exercise 9.2.12)** Suppose the agronomist is considering two possible genetic mechanisms for the inheritance of resistance; the population

ratio of resistant to susceptible progeny would be 1:1 under one mechanism and 3:1 under the other. If the agronomist uses the sample size determined in Exercise 9.2.12, can he be sure that a 95% confidence interval will exclude at least one of the mechanisms? That is, can he be sure that the confidence interval will *not* contain both 0.50 and 0.75? Explain.

## 9.3 Other Confidence Levels (Optional)

The procedure outlined in Section 9.2 can be used to construct 95% confidence intervals. In order to construct intervals with other confidence coefficients, some modifications to the procedure are needed. The first modification concerns  $\tilde{p}$ . For a 95% confidence interval we defined  $\tilde{p}$  to be  $\frac{y + 2}{n + 4}$ . In general, for a confidence interval of level  $100(1 - \alpha)\%$ ,  $\tilde{p}$  is defined as

$$\tilde{p} = \frac{y + 0.5(z_{\alpha/2}^2)}{n + z_{\alpha/2}^2}$$

For a 95% confidence interval  $z_{\alpha/2}$  is 1.96, so that  $\tilde{p} = \frac{y + 0.5(1.96^2)}{n + 1.96^2}$ . This is equal to  $\frac{y + 1.92}{n + 3.84}$ , which we rounded off as  $\frac{y + 2}{n + 4}$ . However, any confidence level can be used. As an example, for a 90% confidence interval,  $\tilde{p} = \frac{y + 0.5(1.645^2)}{n + 1.645^2}$ ; this is equal to  $\frac{y + 1.35}{n + 2.7}$ .

The second modification concerns the standard error. For a 95% confidence interval we used  $\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$  as the standard error term. In general, we use  $\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + z_{\alpha/2}^2}}$  as the standard error term.

Finally, the Z multiplier must match the confidence level (1.645 for a 90% confidence interval, etc.). These can be found most easily from Table 4 with  $df = \infty$ . (Recall from Section 6.3 that the *t* distribution with  $df = \infty$  is a normal (*Z*) distribution.) The following example illustrates these modifications.

---

### Example 9.3.1

**Left-Handedness** In Example 9.2.6 we considered a survey of English and Scottish college students where 40 of 400 male students were found to be left-handed. Let us construct a 90% confidence interval for the proportion,  $p$ , of left-handed individuals in the population.<sup>6</sup>

The sample estimate of the proportion is

$$\tilde{p} = \frac{40 + 0.5(1.645^2)}{400 + 1.645^2} = \frac{40 + 1.35}{400 + 2.7} \approx 0.103$$

and the SE is

$$\sqrt{\frac{0.103(0.897)}{402.7}} = 0.015$$

A 90% confidence interval for  $p$  is

$$0.103 \pm (1.645)(0.015)$$

or

$$0.078 < p < 0.128$$

*Thus, we are 90% confident that between 7.8% and 12.8% of the population that was sampled are left-handed.* ■

## Exercises 9.3.1–9.3.4

---

**9.3.1** In a sample of 848 children aged 3 to 5 it was found that 3.7% of them had iron deficiency.<sup>14</sup> Use these data to construct a 90% confidence interval for the proportion of all 3- to 5-year-old children with iron deficiency.

**9.3.2** Researchers tested patients with cardiac pacemakers to see if use of a cellular telephone interferes with the operation of the pacemaker. There were 959 tests conducted for one type of cellular telephone; interference with the pacemaker (detected with electrocardiographic monitoring) was found in 15.7% of these tests.<sup>15</sup>

- (a) Use these data to construct an appropriate 90% confidence interval.
- (b) The confidence interval from part (a) is a confidence interval for what quantity? Answer in the context of the setting.

**9.3.3** Gene mutations have been found in patients with muscular dystrophy. In one study, it was found that there were defects in the gene coding of sarcoglycan proteins in 23 of 180 patients with limb-girdle muscular dystrophy.<sup>16</sup> Use these data to construct a 99% confidence interval for the corresponding population proportion.

**9.3.4** In an ecological study of the Carolina Junco, 53 birds were captured from a certain population; of these, 40 were male.<sup>17</sup> Use these data to construct a 90% confidence interval for the proportion of male birds in the Carolina Junco population.

## 9.4 Inference for Proportions: The Chi-Square Goodness-of-Fit Test

In Section 9.2 we described methods for constructing confidence intervals when the observed variable is categorical. We now turn our attention to hypothesis testing for categorical data. We will begin by considering analysis of a single sample of categorical data. We assume that the data can be regarded as a random sample from some population and we will test a null hypothesis,  $H_0$ , that specifies the population proportions, or probabilities, of the various categories. Here is an example.

---

**Example  
9.4.1**

**Deer Habitat and Fire** Does fire affect deer behavior? Six months after a fire burned 730 acres of homogenous deer habitat, researchers surveyed a 3,000-acre parcel surrounding the area, which they divided into four regions: the region near the heat of the burn (1), the inside edge of the burn (2), the outside edge of the burn (3), and the area outside of the burned area (4); see Figure 9.4.1 and Table 9.4.1.<sup>18</sup> The null hypothesis is that that deer show no preference to any particular type of burned/unburned habitat—that they are randomly distributed over the 3,000 acres. The alternative hypothesis is that the deer do show a preference for some of the regions—that they are not randomly distributed across all 3,000 acres.

Under the null hypothesis, if deer were randomly distributed over the 3,000 acres, then we would expect the counts of deer in the regions to be in proportion to the sizes of the regions. Expressing the null hypothesis numerically we have the following probabilities of sighting deer:

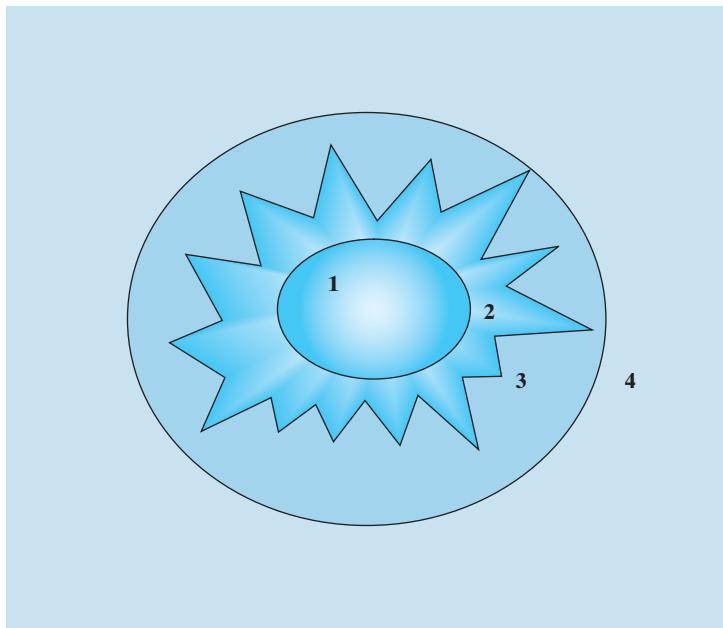
$$H_0: \Pr\{\text{inner burn}\} = \frac{520}{3,000} = 0.173$$

$$\Pr\{\text{inner edge}\} = \frac{210}{3,000} = 0.070$$

$$\Pr\{\text{outer edge}\} = \frac{240}{3,000} = 0.080$$

$$\Pr\{\text{outer unburned}\} = \frac{2,030}{3,000} = 0.677$$

**Figure 9.4.1** Schematic of three thousand-acre parcel with an interior 730-acre fire (not to scale)



**Table 9.4.1** Deer distribution

Region	Acres	Proportion
1. Inner burn	520	0.173
2. Inner edge	210	0.070
3. Outer edge	240	0.080
4. Outer unburned	2,030	0.677
	3,000	1.000

Because the alternative hypothesis is not specific (it only states that the deer prefer some regions over others but doesn't indicate the nature of the preference), there is no simple symbolic way to express the alternative hypothesis. Thus, typically we do not use a symbolic representation. If we chose to express the alternative symbolically we could write:

$$H_A: \Pr\{\text{inner burn}\} \neq 0.173, \text{and/or } \Pr\{\text{inner edge}\} \neq 0.070, \text{and/or}$$

$$\Pr\{\text{outer edge}\} \neq 0.080, \text{and/or } \Pr\{\text{outer unburned}\} \neq 0.677$$

■

Given a random sample of  $n$  categorical observations, how can one judge whether they provide evidence against a null hypothesis  $H_0$  that specifies the probabilities of the categories? There are two complementary approaches to this question: The first considers an examination of the observed relative frequencies of each category while the second examines the frequencies directly. Considering the first method, the observed relative frequencies serve as estimates of the probabilities of the categories. The following notation for relative frequencies is useful: When a probability  $\Pr\{E\}$  is estimated from observed data, the estimate is denoted by a hat ("^"); thus,

$$\hat{\Pr}\{E\} = \text{the estimated probability of event } E$$

---

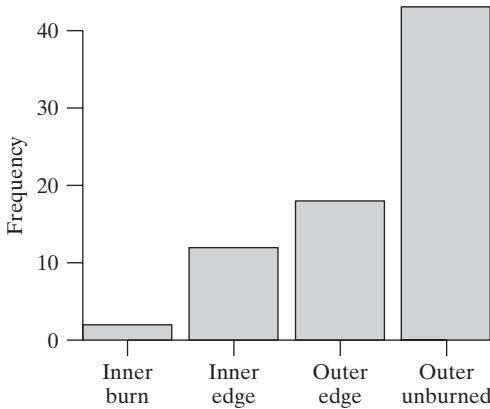
**Example  
9.4.2**

**Deer Habitat and Fire** Researchers observed a total of 75 deer in the 3,000-acre parcel described in Example 9.4.1: Two were in the region near the heat of the burn (Region 1),

12 were on the inside edge of the burn (Region 2), 18 were on the outside edge of the burn (Region 3), and 43 were outside of the burned area (Region 4).

These data are shown in Figure 9.4.2.

**Figure 9.4.2** Bar chart of deer distribution data



The estimated category probabilities are

$$\hat{P}\{inner\ burn\} = \frac{2}{75} = 0.027$$

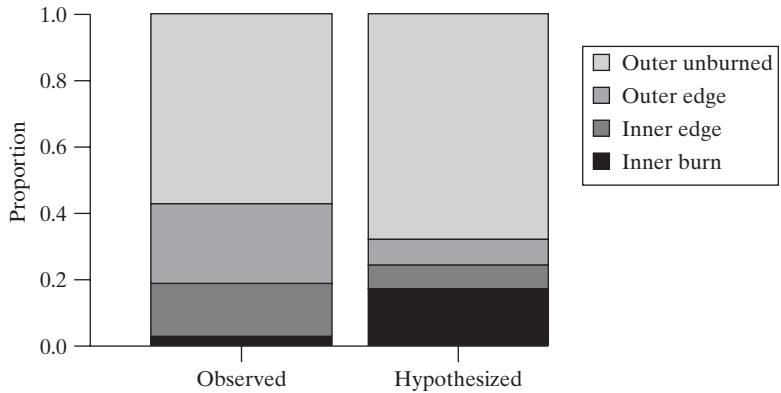
$$\hat{P}\{inner\ edge\} = \frac{12}{75} = 0.160$$

$$\hat{P}\{outer\ edge\} = \frac{18}{75} = 0.240$$

$$\hat{P}\{outer\ unburned\} = \frac{43}{75} = 0.573$$

These estimated probabilities differ quite a bit from those in the model that is specified by  $H_0$ . Figure 9.4.3 shows stacked bar charts of the observed and hypothesized proportions. ■

**Figure 9.4.3** Stacked bar charts of the deer proportions



## The Chi-Square Statistic

The second approach, which considers the actual frequencies, is to use a statistical test, called a **goodness-of-fit test**, to assess the compatibility of the data with  $H_0$ . The most widely used goodness-of-fit test is the **chi-square test** or  $\chi^2$  test ( $\chi$  is the Greek letter “chi”).

The calculation of the chi-square test statistic is done in terms of the absolute, rather than the relative, frequencies of the categories. For each category level,  $i$ , let  $o_i$  represent the **observed frequency** of the category and let  $e_i$  represent the **expected frequency**—that is, the frequency that would be expected according to  $H_0$ . The  $e_i$ 's are calculated by multiplying each probability specified in  $H_0$  by  $n$ , as shown in Example 9.4.3.

**Example  
9.4.3**

**Deer Habitat and Fire** Consider the null hypothesis specified in Example 9.4.1 and the data from Example 9.4.2. If the null hypothesis is true, then we expect 17.3% of the 75 deer to be in the inner burn region; 17.3% of 75 is 13.0:

$$\text{Inner burn: } e_1 = (0.173)(75) = 13.00$$

The corresponding expected frequencies for the other regions are

$$\text{Inner edge: } e_2 = (0.070)(75) = 5.25$$

$$\text{Outer edge: } e_3 = (0.080)(75) = 6.00$$

$$\text{Outer unburned: } e_4 = (0.677)(75) = 50.75 \quad \blacksquare$$

The test statistic for the chi-square goodness-of-fit test is then calculated from the  $o_i$ 's and the  $e_i$ 's using the formula given in the accompanying box with  $k$  equal to the number of category levels. Example 9.4.4 illustrates the calculation of the chi-square statistic.

**The Chi-Square Statistic**

$$\chi^2_s = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

where the summation is over all  $k$  categories.

**Example  
9.4.4**

**Deer Habitat and Fire** The observed frequencies of 75 deer locations are

Region	Inner Burn	Inner Edge	Outer Edge	Outer Unburned	Total
Observed ( $o_i$ )	2	12	18	43	75

The expected frequencies are

Region	Inner Burn	Inner Edge	Outer Edge	Outer Unburned	Total
Expected ( $e_i$ )	13	5.25	6	50.75	75

Note that the sum of the expected frequencies is the same as the sum of the observed frequencies (75). The  $\chi^2$  statistic is

$$\begin{aligned} \chi^2_s &= \frac{(2 - 13)^2}{13} + \frac{(12 - 5.25)^2}{5.25} + \frac{(18 - 6)^2}{6} + \frac{(43 - 50.75)^2}{50.75} \\ &= 43.2 \end{aligned} \quad \blacksquare$$

**Computational Note:** In calculating a chi-square statistic the  $o_i$ 's must be *absolute*, rather than relative, frequencies.

## The $\chi^2$ Distribution

From the way in which  $\chi_s^2$  is defined, it is clear that small values of  $\chi_s^2$  would indicate that the data agree with  $H_0$ , while large values of  $\chi_s^2$  would indicate disagreement. In order to base a statistical test on this agreement or disagreement, we need to know how much  $\chi_s^2$  may be affected by sampling variation.

We consider the null distribution of  $\chi_s^2$ —that is, the sampling distribution that  $\chi_s^2$  follows if  $H_0$  is true. It can be shown (using the methods of mathematical statistics) that, if the sample size is large enough, then the null distribution of  $\chi_s^2$  can be approximated by a distribution known as a  **$\chi^2$  distribution**. The form of a  $\chi^2$  distribution depends on a parameter called “degrees of freedom” (df). Figure 9.4.4 shows the  $\chi^2$  distribution with df = 5.

**Figure 9.4.4** The  $\chi^2$  distribution with df = 5

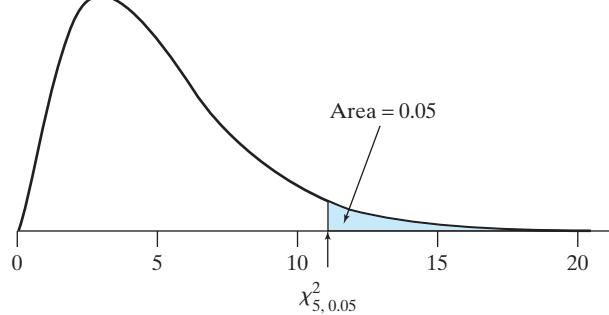


Table 9 (at the end of this book) gives critical values for the  $\chi^2$  distribution. For instance, for df = 5, the 5% critical value is  $\chi_{5, 0.05}^2 = 11.07$ . This critical value corresponds to an area of 0.05 in the upper tail of the  $\chi^2$  distribution, as shown in Figure 9.4.4.

## The Goodness-of-Fit Test

For the chi-square goodness-of-fit test we have presented, the null distribution of  $\chi_s^2$  is approximately a  $\chi^2$  distribution with\*

$$\text{df} = k - 1, \text{ where } k \text{ equals the number of categories}$$

For example, for the setting presented in Example 9.4.4 there are four categories so  $k = 4$ . The null hypothesis specifies the probabilities for each of the four categories. However, once the first three probabilities are specified, the last one is determined, since the four probabilities must sum to 1. There are four categories, but only three of them are “free”; the last one is constrained by the first three.

The test of  $H_0$  is carried out using critical values from Table 9, as illustrated in the following example.

---

\*The chi-square test can be extended to more general situations in which parameters are estimated from the data before the expected frequencies are calculated. In general, the degrees of freedom for the test are (number of categories) – (number of parameters estimated) – 1. We are considering only the case in which there are no parameters to be estimated from the data.

**Example  
9.4.5**

**Deer Habitat and Fire** For the deer habitat data of Example 9.4.4, the observed chi-square statistic was  $\chi_s^2 = 43.2$ . Because there are four categories, the degrees of freedom for the null distribution are calculated as

$$df = 4 - 1 = 3$$

From Table 9 with  $df = 3$  we find that  $\chi_{3, 0.0001}^2 = 21.11$ . Since  $\chi_s^2 = 43.2$  is greater than 21.11, the upper tail area beyond 43.2 is less than 0.0001. Thus, the  $P$ -value is less than 0.0001 and we have strong evidence against  $H_0$  and in favor of the alternative hypothesis that the deer show preference for some areas over others. Upon comparing the observed and expected frequencies (or equivalently the hypothesized and estimated probabilities), we note that deer moved away from the burned and unburned regions (1) and (4) to be near the edge regions (2) and (3) (where there is likely to be new growth of vegetation yet proximity to old-growth shelter). ■

The chi-square test can be used with any number of categories. In Example 9.4.6 the test is applied to a variable with six categories.

**Example  
9.4.6**

**Flax Seeds** Researchers studied a mutant type of flax seed that they hoped would produce oil for use in margarine and shortening. The amount of palmitic acid in the flax seed was an important factor in this research; a related factor was whether the seed was brown or was variegated. The seeds were classified into six combinations of palmitic acid and color, as shown in Table 9.4.2.<sup>19</sup> According to a hypothesized (Mendelian) genetic model, the six combinations should occur in a 3:6:3:1:2:1 ratio. That is, brown and low acid level should occur with probability 3/16, brown and intermediate acid level should occur with probability 6/16, and so on. The null hypothesis is that the model is correct; the alternative hypothesis is that the model is incorrect. The  $\chi^2$  statistic is

$$\begin{aligned}\chi_s^2 &= \frac{(15 - 13.5)^2}{13.5} + \frac{(26 - 27)^2}{27} + \frac{(15 - 13.5)^2}{13.5} + \frac{(0 - 4.5)^2}{4.5} + \frac{(8 - 9)^2}{9} + \frac{(8 - 4.5)^2}{4.5} \\ &= 7.71\end{aligned}$$

**Table 9.4.2** Flax seed distribution

Color	Acid level	Observed ( $o_i$ )	Expected ( $e_i$ )
Brown	Low	15	13.5
Brown	Intermediate	26	27
Brown	High	15	13.5
Variegated	Low	0	4.5
Variegated	Intermediate	8	9
Variegated	High	8	4.5
Total		72	72

The  $\chi^2$  test has  $6 - 1 = 5$  degrees of freedom. From Table 9 with  $df = 5$ , we find that  $\chi_{5, 0.20}^2 = 7.29$  and  $\chi_{5, 0.10}^2 = 9.24$ . Thus, the  $P$ -value is bracketed as  $0.10 < P\text{-value} < 0.20$ . If the level of  $\alpha$  chosen for the test is 0.10 or smaller, then

the  $P$ -value is larger than  $\alpha$  and we would not reject  $H_0$ . We conclude that there is no significant evidence that the data are inconsistent with the Mendelian model. (Note that we have not necessarily demonstrated that the Mendelian model is correct, only that we cannot reject this model.) ■

Note that the critical values for the chi-square test do not depend on the sample size,  $n$ . However, the test procedure is affected by  $n$ , through the value of the chi-square statistic. If we change the size of a sample while keeping its percentage composition fixed, then  $\chi_s^2$  varies directly as the sample size,  $n$ . For instance, imagine appending a replicate of a sample to the sample itself. Then the expanded sample would have twice as many observations as the original, but they would be in the same relative proportions. The value of each  $o_i$  would be doubled, the value of each  $e_i$  would be doubled, and so the value of  $\chi^2$  would be doubled [because in each term of  $\chi_s^2$  the numerator  $(o_i - e_i)^2$  would be multiplied by 4, and the denominator  $e_i$  would be multiplied by 2]. That is, the value of  $\chi_s^2$  would go up by a factor of 2, despite the fact that the pattern in the data stayed the same! In this way, an increased sample size magnifies any discrepancy between what is observed and what is expected under the null hypothesis.

## Compound Hypotheses and Directionality

Let us examine the goodness-of-fit null hypothesis more closely. In a two-sample comparison such as a  $t$  test, the null hypothesis contains exactly one assertion—for instance, that two population means are equal. By contrast, a goodness-of-fit null hypothesis can contain more than one assertion. Such a null hypothesis may be called a **compound null hypothesis**. An example follows.

---

### Example 9.4.7

**Deer Habitat and Fire** The null hypothesis of Example 9.4.1 is

$$H_0: \Pr\{\text{inner burn}\} = 0.173, \Pr\{\text{inner edge}\} = 0.070, \Pr\{\text{outer edge}\} = 0.080, \\ \Pr\{\text{outer unburned}\} = 0.677$$

This is a compound hypothesis because it makes three independent assertions, namely

$$\Pr\{\text{inner burn}\} = 0.173, \Pr\{\text{inner edge}\} = 0.070, \text{and } \Pr\{\text{outer edge}\} = 0.080$$

Note that the fourth assertion ( $\Pr\{\text{outer unburned}\} = 0.677$ ) is not an independent assertion because it follows from the other three. ■

When the null hypothesis is compound, the chi-square test has two special features. First, the alternative hypothesis is necessarily nondirectional. Second, if  $H_0$  is rejected, the test does not yield a directional conclusion. (However, if  $H_0$  is rejected, then an examination of the observed proportions will sometimes show an interesting pattern of departure from  $H_0$ , as in Example 9.4.5.)

When  $H_0$  is compound, the chi-square test is nondirectional in nature (perhaps “omnidirectional” would be a better term) because the chi-square statistic measures deviations from  $H_0$  in all directions. Statistical methods are available that do yield directional conclusions and that can handle directional alternatives, but such methods are beyond the scope of this book.

## Dichotomous Variables

If the categorical variable analyzed by a goodness-of-fit test is dichotomous, then the null hypothesis is not compound, and directional alternatives and directional conclusions do not pose any particular difficulty.\*

**Directional Conclusion** The following example illustrates the directional conclusion.

**Example  
9.4.8**

**Deer Habitat, Fire, and Two Regions** Suppose that the deer habitat data of Example 9.4.1 had been presented as being from only two regions, A and B, where region A is the area at the edge of the fire, which combines regions (2) and (3), and region B is the remainder of the parcel, combining regions (1) and (4). There were 30 deer seen in region A and 45 deer seen in region B. Is this evidence that deer prefer one region over the other?

An appropriate null hypothesis is

$$H_0: \Pr\{\text{region A}\} = \frac{450}{3,000} = 0.15, \Pr\{\text{region B}\} = \frac{2,550}{3,000} = 0.85$$

This hypothesis is not compound because it contains only one independent assertion. (Note that the second assertion— $\Pr\{\text{region B}\} = 0.85$ —is redundant; it follows from the first.)

Let us test  $H_0$  against the nondirectional alternative

$$H_A: \Pr\{\text{region A}\} \neq 0.15$$

The observed and expected frequencies are shown in Table 9.4.3.

<b>Table 9.4.3</b> Deer habitat data for two regions			
	A	B	Total
Observed	30	45	75
Expected	11.25	63.75	75

The data yield  $\chi^2_s = 36.8$  and from Table 9 we find that  $P < 0.0001$ . Even at  $\alpha = 0.0001$  we would reject  $H_0$  and find that there is sufficient evidence to conclude that the population of deer prefers one region over the other. Comparing the observed and expected counts we observe that they prefer region A over region B. ■

To recapitulate, the directional conclusion in Example 9.4.8 is legitimate because we know that if  $H_0$  is false, then necessarily either  $\Pr\{\text{region A}\} < 0.15$  or  $\Pr\{\text{region A}\} > 0.15$ . By contrast, in Example 9.4.7  $H_0$  may be false but  $\Pr\{\text{outer unburned}\}$  may still be equal to 0.677; the chi-square analysis does not determine which of the probabilities are not as specified by  $H_0$ .

---

\*When the data are dichotomous, there is an alternative to the goodness-of-fit test that is known as the Z test for a single proportion. The calculations used in the Z test look quite different from those of the goodness-of-fit test but, in fact, the two tests are mathematically equivalent. However, unlike the goodness-of-fit test, which can handle any number of categories, the Z test can be used only when the data are limited to two categories. Thus, we do not present it here.

**Directional Alternative** A chi-square goodness-of-fit test against a directional alternative (when the observed variable is dichotomous) uses the familiar two-step procedure:

**Step 1** Check directionality (see if the data deviate from  $H_0$  in the direction specified by  $H_A$ ).

- (a) If not, the  $P$ -value is greater than 0.50.
- (b) If so, proceed to step 2.

**Step 2** The  $P$ -value is half what it would be if  $H_A$  were nondirectional.

The following example illustrates the procedure.

**Example  
9.4.9**

**Harvest Moon Festival** Can people who are close to death postpone dying until after a symbolically meaningful occasion? Researchers studied death from natural causes among elderly Chinese women (over age 75) living in California. They chose to study the time around the Harvest Moon Festival because (1) the date of the traditional Chinese festival changes somewhat from year to year, making it less likely that a time-of-year effect would be confounded with the effect they were studying and (2) it is a festival in which the role of the oldest woman in the family is very important.

Previous research had suggested that there might be a decrease in the mortality rate among elderly Chinese women immediately prior to the festival, with a corresponding increase afterwards. The researchers found that over a period of several years there were 33 deaths in the group in the week preceding the Harvest Moon Festival and 70 deaths in the week following the festival.<sup>20</sup> How strongly does this support the interpretation that people can prolong life until a symbolically meaningful event?

We may formulate null and alternative hypotheses as follows:

$H_0$ : Given that an elderly Chinese woman dies within one week of the Harvest Moon Festival, she is equally likely to die before the festival or after the festival.

$H_A$ : Given that an elderly Chinese woman dies within one week of the Harvest Moon Festival, she is more likely to die after the festival than before the festival.

These hypotheses can be translated as

$$H_0: \Pr\{\text{die after festival}\} = \frac{1}{2}$$

$$H_A: \Pr\{\text{die after festival}\} > \frac{1}{2}$$

where it is understood that  $\Pr\{\text{die after festival}\}$  is the probability of death after the festival, given that the woman dies within one week before or after the festival. The observed and expected frequencies are shown in Table 9.4.4.

<b>Table 9.4.4</b> Harvest Moon Festival data			
	Before	After	Total
Observed	33	70	103
Expected	51.5	51.5	103

From the data on the 103 deaths, we first note that the data do, indeed, deviate from  $H_0$  in the direction specified by  $H_A$ , because the observed relative frequency of deaths after the festival is 70/103, which is greater than 1/2. The value of the chi-square statistic is  $\chi_s^2 = 13.3$ ; from Table 9 we see that the  $P$ -value would have been bracketed between 0.0001 and 0.001 had  $H_A$  been nondirectional. However, for the directional alternative hypothesis specified in this test, we bracket the  $P$ -value as  $0.00005 < P\text{-value} < 0.0005$ . We conclude that the evidence is very strong that the death rate among elderly Chinese women goes up after the festival.\* ■

## Exercises 9.4.1–9.4.13

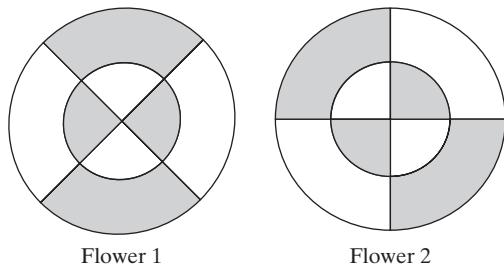
**9.4.1** A cross between white and yellow summer squash gave progeny of the following colors:<sup>21</sup>

COLOR	WHITE	YELLOW	GREEN
Number of progeny	155	40	10

Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model? Use a chi-square test at  $\alpha = 0.10$ .

**9.4.2** Refer to Exercise 9.4.1. Suppose the sample had the same composition but was 10 times as large: 1,550 white, 400 yellow, and 100 green progeny. Would the data be consistent with the 12:3:1 model?

**9.4.3** How do bees recognize flowers? As part of a study of this question, researchers used the following two artificial “flowers”:<sup>22</sup>



The experiment was conducted as a series of trials on individual bees; each trial consisted of presenting a bee with both flowers and observing which flower it landed on first. (Flower 1 was sometimes on the left and sometimes on the right.) During the “training” trials, flower 1 contained a sucrose solution and flower 2 did not; thus, the bee was trained to prefer flower 1. During the testing trials, neither flower contained sucrose. In 25 testing trials with a particular bee, the bee chose flower 1 twenty times and flower 2 five times.

Use a goodness-of-fit test to assess the evidence that the bee could remember and distinguish the flower patterns. Use a directional alternative and let  $\alpha = 0.05$ .

**9.4.4** At a midwestern hospital there was a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends.<sup>23</sup> Do these data reveal more than chance deviation from random timing of the births? (Test for goodness of fit, with two categories of births: weekday and weekend. Use a nondirectional alternative and let  $\alpha = 0.05$ .)

**9.4.5** In a breeding experiment, white chickens with small combs were mated and produced 190 offspring of the types shown in the accompanying table.<sup>24</sup> Are these data consistent with the Mendelian expected ratios of 9:3:3:1 for the four types? Use a chi-square test at  $\alpha = 0.10$ .

TYPE	NUMBER OF OFFSPRING
White feathers, small comb	111
White feathers, large comb	37
Dark feathers, small comb	34
Dark feathers, large comb	8
Total	190

**9.4.6** Among  $n$  babies born in a certain city, 51% were boys.<sup>25</sup> Suppose we want to test the hypothesis that the true probability of a boy is  $\frac{1}{2}$ . Calculate the value of  $\chi_s^2$ , and bracket the  $P$ -value for testing against a nondirectional alternative, if

- (a)  $n = 1,000$
- (b)  $n = 5,000$
- (c)  $n = 10,000$

\*Based on these results, one might jump to the conclusion that this festival should be canceled to protect elderly Chinese women. As this study is only observational, however, we must not jump to causal conclusions!

**9.4.7** In an agronomy experiment peanuts with shriveled seeds were crossed with normal peanuts. The genetic model that the agronomists were considering predicted that the ratio of normal to shriveled progeny would be 3:1. They obtained 95 normal and 54 shriveled progeny.<sup>26</sup> Do these data support the hypothesized model?

Conduct a chi-square test with  $\alpha = 0.05$ . Use a nondirectional alternative.

**9.4.8** An experimental design using litter-matching was employed to test a certain drug for cancer-causing potential. From each of 50 litters of rats, three females were selected; one of these three, chosen at random, received the test drug, and the other two were kept as controls. During a two-year observation period, the time of occurrence of a tumor, and/or death from various causes, was recorded for each animal. One way to analyze the data is to note simply which rat (in each triplet) developed a tumor first. Some triplets were uninformative on this point because either (a) none of the three littermates developed a tumor, or (b) a rat developed a tumor after its littermate had died from some other cause. The results for the 50 triplets are shown in the table.<sup>27</sup> Use a goodness-of-fit test to evaluate the evidence that the drug causes cancer. Use a directional alternative and let  $\alpha = 0.01$ . State your conclusion from part (a) in the context of this setting. (*Hint:* Use only the 20 triplets that provide complete information.)

	NUMBER OF TRIPLETS
Tumor first in the treated rat	12
Tumor first in one of the two control rats	8
No tumor	23
Death from another cause	7
Total	50

**9.4.9** A study of color vision in squirrels used an apparatus containing three small translucent panels that could be separately illuminated. The animals were trained to choose, by pressing a lever, the panel that appeared different from the other two. (During these “training” trials, the panels differed in brightness, rather than color.) Then the animals were tested for their ability to discriminate between various colors. In one series of “testing” trials on a single animal, one of the panels was red and the other two were white; the location of the red panel was varied randomly from trial to trial. In 75 trials, the animal chose correctly 45 times and incorrectly 30 times.<sup>28</sup> How strongly does this support the interpretation that the animal can discriminate between the two colors?

- (a) Test the null hypothesis that the animal cannot discriminate red from white. Use a directional alternative and let  $\alpha = 0.02$ .

(b) Why is a directional alternative appropriate in this case?

**9.4.10** Scientists have used Mongolian gerbils when conducting neurological research. A certain breed of these gerbils was crossed and gave progeny of the following colors:<sup>29</sup>

COLOR	BLACK	BROWN	WHITE
Number of progeny	40	59	42

Are these data consistent with the 1:2:1 ratio predicted by a certain genetic model? Use a chi-square test at  $\alpha = 0.05$ .

**9.4.11** Each of 36 men was asked to touch the foreheads of three women, one of whom was their romantic partner, while blindfolded. The two “decoy” women were the same age, height, and weight as the man’s partner. Of the 36 men tested, 18 were able to correctly identify their partner.<sup>30</sup> Do the data provide sufficient evidence to conclude that men can do better than they would do by merely guessing?

Conduct an appropriate test.

**9.4.12** Geneticists studying the inheritance pattern of cowpea plants classified the plants in one experiment according to the nature of their leaves. The data follow:<sup>31</sup>

TYPE	I	II	III
Number	179	44	23

Test the null hypothesis that the three types occur with probabilities 12/16, 3/16, and 1/16. Use a chi-square test with  $\alpha = 0.10$ .

**9.4.13** In the snapdragon (*Antirrhinum majus*), individual plants can be red flowered, pink flowered, or white flowered. According to a certain Mendelian genetic model, self-pollination of pink-flowered plants should produce progeny that are red, pink, and white in the ratio 1:2:1. A geneticist self-pollinated pink-flowered snapdragon plants and produced 234 progeny with the following colors:<sup>32</sup>

TYPE	RED	PINK	WHITE
Number	54	122	58

Test the null hypothesis that the three colors occur with probabilities 1/4, 1/2, and 1/4. Use a chi-square test with  $\alpha = 0.10$ .

## 9.5 Perspective and Summary

In this chapter we have discussed inference for categorical data, including confidence intervals and hypothesis tests. The procedures that we have developed, which are summarized next, can be applied if (1) the data can be regarded as a random sample from a large population and (2) the observations are independent.

### Summary of Inference Methods for Categorical Data

#### 95% Confidence interval for $p$

$$\tilde{p} \pm 1.96 \times \text{SE}_{\tilde{p}}$$

where

$$\tilde{p} = \frac{y + 2}{n + 4}$$

and

$$\text{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

#### General confidence interval for $p$

$$\tilde{p} \pm z_{\alpha/2} \times \text{SE}_{\tilde{p}}$$

where

$$\tilde{p} = \frac{y + 0.5(z_{\alpha/2}^2)}{n + z_{\alpha/2}^2}$$

$$\text{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + z_{\alpha/2}^2}}$$

#### Goodness-of-fit test

*Data:*

$o_i$  = the observed frequency of category  $i$

*Null hypothesis:*

$H_0$  specifies the probability of each category.\*

*Calculation of expected frequencies:*

$e_i = n \times \text{Probability specified for category } i \text{ by } H_0$

*Test statistic:*

$$\chi_s^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

---

\*A slightly modified form of the goodness-of-fit test can be used to test a hypothesis that merely constrains the probabilities rather than specifying them exactly. An example would be testing the fit of a binomial distribution to data (see optional Section 3.9). The details of this test are beyond the scope of this text.

*Null distribution (approximate):*

$\chi^2$  distribution with  $df = k - 1$

where  $k$  = the number of categories

This approximation is adequate if  $e_i \geq 5$  for every category.

## Supplementary Exercises 9.S.1–9.S.22

**9.S.1** In a certain population, 83% of the people have Rh-positive blood type.<sup>33</sup> Suppose a random sample of  $n = 10$  people is to be chosen from the population and let  $\tilde{P}$  represent the Wilson-adjusted proportion of Rh-positive people in the sample. Find

- (a)  $\Pr\{\tilde{P} = 0.714\}$
- (b)  $\Pr\{\tilde{P} = 0.786\}$

**9.S.2** In a population of flatworms (*Planaria*) living in a certain pond, one in five individuals is adult and four are juvenile.<sup>34</sup> An ecologist plans to count the adults in a random sample of 16 flatworms from the pond; she will then use  $\tilde{P}$ , the Wilson-adjusted sample proportion of adults in the sample, as her estimate of  $p$ , the proportion of adults in the pond population. Find

- (a)  $\Pr\{\tilde{P} = p\}$
- (b)  $\Pr\{p - 0.05 \leq \tilde{P} \leq p + 0.05\}$

**9.S.3** In a study of environmental effects upon reproduction, 123 female adult white-tailed deer from the central Adirondack area were captured and 97 were found to be pregnant.<sup>35</sup> Construct a 95% confidence interval for the proportion of females pregnant in this deer population.

**9.S.4** Refer to Exercise 9.S.3. Which of the conditions for validity of the confidence interval might have been violated in this study?

**9.S.5** A sample of 32 breastfed infants found that 2 of them developed iron deficiency by age 5.5 months.<sup>36</sup>

- (a) Use these data to construct an appropriate 90% confidence interval.
- (b) What conditions are necessary for the confidence interval from part (a) to be valid?
- (c) Interpret your confidence interval from part (a) in the context of this setting. That is, what do the numbers in the confidence interval tell us about iron deficiency in breastfed infants?

**9.S.6** A certain California winery produces 720,000 bottles of wine each year. Suppose you want to estimate the proportion of those bottles that have cork taint (i.e., the wine is spoiled due to a failure of the cork). Suppose that 4% of all corked wine has cork taint. Using this as a preliminary guess of  $p$ , how many bottles of wine would need to be included in a random sample if you want the standard error of your estimate to be less than or equal to 1 percentage point?<sup>37</sup>

**9.S.7** Refer to Exercise 9.S.6. Suppose you do not trust that the 4% taint rate for wines in general is a useful guess for this particular winery.

- (a) Suppose that, based on previous years of data at this winery, about 10% of the wines have had cork taint. How many bottles would need to be included in a random sample if you want the standard error of your estimate to be less than or equal to 1 percentage point?
- (b) How many bottles would need to be included in a random sample if you want the standard error of your estimate to be less than or equal to 1 percentage point, no matter what the value of  $p$  is?

**9.S.8** When male mice are grouped, one of them usually becomes dominant over the others. In order to see how a parasitic infection might affect the competition for dominance, male mice were housed in groups, three mice to a cage; two mice in each cage received a mild dose of the parasitic worm *H. polygyrus*. Two weeks later, criteria such as the relative absence of tail wounds were used to identify the dominant mouse in each cage. It was found that the uninfected mouse had become dominant in 15 of 30 cages.<sup>38</sup> Is this evidence that the parasitic infection tends to inhibit the development of dominant behavior? Use a goodness-of-fit test against a directional alternative. Let  $\alpha = 0.05$ . (*Hint:* The observational unit in this experiment is not an individual mouse, but a cage of three mice.)

**9.S.9** Are mice right-handed or left-handed? In a study of this question, 320 mice of a highly inbred strain were tested for paw preference by observing which forepaw—right or left—they used to retrieve food from a narrow tube. Each animal was tested 50 times, for a total of  $320 \times 50 = 16,000$  observations. The results were as follows:<sup>39</sup>

	RIGHT	LEFT
Number of observations	7,871	8,129

Suppose we assign an expected frequency of 8,000 to each category and perform a goodness-of-fit test; we find that  $\chi^2_s = 4.16$ , so that at  $\alpha = 0.05$  we would reject the hypothesis of a 1:1 ratio and find that there is sufficient

evidence to conclude that mice of this strain are (slightly) biased toward use of the left paw. This analysis contains a fatal flaw. What is it?

**9.S.10** As part of the study of the inheritance pattern of cowpea plants, geneticists classified the plants in one experiment according to whether the plants had one leaf or three. The data follow:<sup>40</sup>

NUMBER OF LEAVES	1	3
Number of plants	74	61

Test the null hypothesis that the two types of plants occur with equal probabilities. Use a nondirectional alternative and let  $\alpha = 0.05$ .

**9.S.11** People who harvest wild mushrooms sometimes accidentally eat the toxic “death cap” mushroom, *Amanita phalloides*. In reviewing 205 European cases of death-cap poisoning from 1971 through 1980, researchers found that 45 of the victims had died.<sup>41</sup> Conduct a test to compare this mortality to the 30% mortality that was recorded before 1970. Let the alternative hypothesis be that mortality has decreased with time and let  $\alpha = 0.05$ .

**9.S.12** The appearance of leaf pigment glands in the seedling stage of cotton plants is genetically controlled. According to one theory of the control mechanism, the population ratio of glandular to glandless plants resulting from a certain cross should be 11:5; according to another theory it should be 13:3. In one experiment, the cross produced 89 glandular and 36 glandless plants.<sup>42</sup> Use goodness-of-fit tests (at  $\alpha = 0.10$ ) to determine whether these data are consistent with

- (a) the 11:5 theory
- (b) the 13:3 theory

### 9.S.13 (Continuation of 9.S.12)

- (a) If the 11:5 and 13:3 ratios are the only two reasonable theories to consider, would you have compelling evidence that the theory you selected in Exercise 9.S.12 is the correct theory? Explain.
- (b) If there are also other possible theoretical ratios that weren’t considered, would you have compelling evidence that the theory you selected in Exercise 9.S.12 is the correct theory? Explain.

**9.S.14** When fleeing a predator, the minnow *Fundulus notti* will often head for shore and jump onto the bank. In a study of spatial orientation in this fish, individuals were caught at various locations and later tested in an artificial pool to see which direction they would choose when released: Would they swim in a direction which, at their place of capture, would have led toward shore? The following are the directional choices ( $\pm 45^\circ$ ) of 50 fish tested under cloudy skies:<sup>43</sup>

Toward shore	18
Away from shore	12
Along shore to the right	13
Along shore to the left	7

Use chi-square tests at  $\alpha = .05$  to test the hypothesis that directional choice under cloudy skies is random,

- (a) using the four categories listed in the table.
- (b) collapsing to two categories—“toward shore” and “away from or along shore”—and using a directional  $H_A$ .

(Note: Although the chi-square test is valid in this setting, it should be noted that more powerful tests are available for analysis of orientation data.)<sup>44</sup>

### 9.S.15 Refer to the cortex-weight data of Exercise 8.4.4.

- (a) Use a goodness-of-fit test to test the hypothesis that the environmental manipulation has no effect. As in Exercise 8.4.4, use a directional alternative and let  $\alpha = 0.05$ . (This exercise shows how, by a shift of viewpoint, the sign test can be reinterpreted as a goodness-of-fit test. Of course, the chi-square goodness-of-fit test described in this chapter can be used only if the number of observations is large enough.)
- (b) Is the number of observations large enough for the test in part (a) to be valid?

**9.S.16** A biologist wanted to know if the cowpea weevil has a preference for one type of bean over others as a place to lay eggs. She put equal amounts of four types of seeds into a jar and added adult cowpea weevils. After a few days she observed the following data:<sup>45</sup>

TYPE OF BEAN	NUMBER OF EGGS
Pinto	167
Cowpea	176
Navy beans	174
Northern beans	194

Do these data provide evidence of a preference for some types of beans over others? That is, are the data consistent with the claim that the eggs are distributed randomly among the four types of bean?

**9.S.17** An experiment was conducted in which two types of acorn squash were crossed. According to a genetic model, 1/2 of the resulting plants should have dark stems and dark fruit, 1/4 should have light stems and light fruit, and 1/4 should have light stems and plain fruit. The actual data were 220, 129, and 105 for these three categories.<sup>46</sup> Do these data refute this model? Conduct a chi-square test with  $\alpha = 0.10$ .

**9.S.18** Each of 36 men was asked to touch the backs of the hands of three women, one of whom was the man's romantic partner, while blindfolded. The two "decoy" women were the same age, height, and weight as the man's partner.<sup>30</sup> Of the 36 men tested, 16 were able to correctly identify their partner. Do the data provide sufficient evidence that the men are able to sense their partners better than guessing would predict? Conduct a goodness-of-fit test of the data, using  $\alpha = 0.05$ .

**9.S.19** In a study of resistance to a certain soybean virus, biologists cross fertilized two soybean cultivars. They expected to get a 3:1 ratio of resistant to susceptible plants. The observed data were 58 resistant and 26 susceptible plants.<sup>47</sup> Are these data significantly inconsistent with the expected 3:1 ratio? Conduct a test, using  $\alpha = 0.10$ ; use a nondirectional alternative.

**9.S.20** A group of 1,438 sexually active patients were counseled on condom use and the risk of contracting a

sexually transmitted disease (STD). After six months, 103 of the patients had new STDs.<sup>48</sup> Construct a 95% confidence interval for the probability of contracting an STD within six months after being part of a counseling program like the one used in this study.

**9.S.21 (Continuation of 9.S.20)** Suppose that for (uncounseled) sexually active individuals the probability of acquiring an STD in a six-month period is 10%.

- Using your interval computed in Exercise 9.S.21, is there compelling evidence that the six-month STD probability is different for those who receive counseling?
- Using the data from Exercise 9.S.21, conduct a nondirectional chi-square test to determine if the six-month STD rate is different for counseled individuals compared to the uncounseled population.
- Do your answers to parts (a) and (b) agree? Explain.



# Chapter

# 10

## CATEGORICAL DATA: RELATIONSHIPS

### Objectives

In this chapter we extend our study of categorical data to several populations.

We will

- discuss independence and association for categorical variables.
- describe a chi-square test to assess the independence between two categorical variables.
- consider the conditions under which a chi-square test is valid.
- describe Fisher's exact test of independence between two categorical variables.
- present McNemar's test to analyze paired categorical data.
- calculate relative risk, the odds ratio, and its associated confidence interval.

### 10.1 Introduction

In Chapter 9 we considered the analysis of a single sample of categorical data. The basic techniques we employed were estimation of category probabilities and comparison of observed category frequencies with frequencies “expected” according to a null hypothesis. In this chapter we will extend these basic techniques to more complicated situations. To set the stage, here are two examples, the first of which presents an experiment; the second, an observational study.

**Example**  
**10.1.1**

**Migraine Headache** Patients who suffered from moderate to severe migraine headache took part in a double-blind clinical trial to assess an experimental surgery. A group of 75 patients were randomly assigned to receive either the real surgery on migraine trigger sites ( $n = 49$ ) or a sham surgery ( $n = 26$ ) in which an incision was made but no further procedure was performed. The surgeons hoped that patients would experience “a substantial reduction\* in migraine headaches,” which we will label as “success.” Table 10.1.1 shows the results of the experiment.<sup>1</sup>

Table 10.1.1 Response to migraine surgery			
		Surgery	
		Real	Sham
Substantial reduction in migraine headaches?	Success	41	15
	No success	8	11
	Total	49	26

\*“Substantial reduction” means at least a 50 percent reduction in migraine headache frequency, intensity, or duration when compared with baseline (presurgery) values.

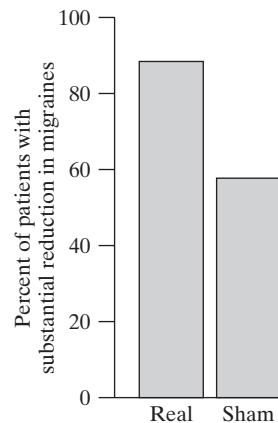
A natural way to express the results is in terms of percentages, as follows:

Of the real surgeries,  $\frac{41}{49}$  or 83.7% were successful.

Of sham surgeries,  $\frac{15}{26}$  or 57.7% were successful.

In this study successful reduction in migraine headache was more common among patients who received the real surgery than among those who received the sham surgery—83.7% versus 57.7%. Table 10.1.2 provides a summary of the data; Figure 10.1.1 is a bar chart showing the percentages of successful surgeries for the two groups. ■

<b>Table 10.1.2</b> Response to migraine surgery		
	Surgery	
	Real	Sham
<i>n</i>	49	26
Success	41	15
Percent	83.7%	57.7%



**Figure 10.1.1** Bar chart of migraine surgery data

### Example 10.1.2

**HIV Testing** A random sample of 120 college students found that 9 of the 61 women in the sample had taken an HIV test, compared to 8 of the 59 men.<sup>2</sup> These data are shown in Table 10.1.3.

<b>Table 10.1.3</b> HIV testing data		
	Female	Male
HIV test	9	8
No HIV test	52	51
Total	61	59

Of the women  $\frac{9}{61} = 0.148$  or 14.8% had been tested for HIV. Of the men  $\frac{8}{59} = 0.136$  or 13.6% had been tested for HIV.

These two percentages are nearly identical. ■

Tables such as Tables 10.1.1 and 10.1.3 are called **contingency tables**. The focus of interest in a contingency table is the dependence or association between the column variable and the row variable—for instance, between treatment and response in Tables 10.1.1 and 10.1.3. (The word *contingent* means “dependent.”) In particular, Tables 10.1.1 and 10.1.3 are called **2 × 2** (“two-by-two”) **contingency tables**, because they consist of two rows (excluding the “total” row) and two columns. Each category in the contingency table is called a **cell**; thus, a  $2 \times 2$  contingency table has four cells.

We will consider the analysis and interpretation of  $2 \times 2$  contingency tables before extending the discussion to larger tables.

## 10.2 The Chi-Square Test for the $2 \times 2$ Contingency Table

When analyzing a  $2 \times 2$  contingency table it is natural to think of the probability of an event under either of two conditions being compared. We will find it is useful to extend the language of probability to include a new concept: conditional probability.\*

### Conditional Probability

Recall that the probability of an event predicts how often the event will occur. A **conditional probability** predicts how often an event will occur under specified conditions. The notation for a conditional probability is

$$\Pr\{E|C\}$$

which is read “probability of  $E$ , given  $C$ .” When a conditional probability is estimated from observed data, the estimate is denoted by a hat (“ $\hat{\cdot}$ ”); thus,

$$\hat{\Pr}\{E|C\}$$

The following example illustrates these ideas.

#### Example 10.2.1

**Migraine Headache** Consider the migraine headache data from Example 10.1.1. The conditional probabilities of interest are as follows:

$$\begin{aligned} \Pr\{\text{substantial reduction in migraines}|\text{real surgery}\} &= \Pr\{\text{Success}|\text{Real}\} \\ &= \text{probability that a patient will have a substantial reduction in headache if given the real surgery} \\ \Pr\{\text{substantial reduction in migraines}|\text{sham surgery}\} &= \Pr\{\text{Success}|\text{Sham}\} \\ &= \text{probability that a patient will have a substantial reduction in headache if given the sham surgery} \end{aligned}$$

The estimates of these conditional probabilities from the data of Table 10.1.1 are

$$\hat{\Pr}(\text{Success}|\text{Real}) = \frac{41}{49} = 0.837$$

and

$$\hat{\Pr}(\text{Success}|\text{Sham}) = \frac{15}{26} = 0.577$$

The natural hypothesis to test is that the conditional probabilities associated with a  $2 \times 2$  table are equal, which is to say that the probability of the event  $E$  does not depend on whether the first condition,  $C$ , is present or the second condition, “not  $C$ ,” is present.

$$H_0: \Pr\{E|C\} = \Pr\{E|\text{not } C\}$$

The following example illustrates this null hypothesis.

---

\*Conditional probability is also discussed in optional Section 3.3.

**Example  
10.2.2**

**Migraine Headache** For the migraine study of Example 10.1.1, the null hypothesis is

$$H_0: \Pr\{\text{Success} | \text{Real}\} = \Pr\{\text{Success} | \text{Sham}\}$$

or equivalently

$$H_0: \Pr\{\text{Success} | \text{Real}\} = \Pr\{\text{Success} | \text{not Real}\}$$



## The Chi-Square Statistic

Clearly, a natural way to test the preceding null hypothesis would be to reject  $H_0$  if  $\hat{\Pr}\{E | C\}$  and  $\hat{\Pr}\{E | \text{not } C\}$  are different by a sufficient amount. We describe a test procedure that compares  $\hat{\Pr}\{E | C\}$  and  $\hat{\Pr}\{E | \text{not } C\}$  indirectly, rather than directly. The procedure is a chi-square test, based on the test statistic  $\chi_s^2$  that was introduced in Section 9.4:

$$\chi_s^2 = \sum_{i=1}^4 \frac{(o_i - e_i)^2}{e_i}$$

In the formula, the sum is taken over all four cells in the contingency table. Each  $o$  represents an observed frequency and each  $e$  represents the corresponding expected frequency according to  $H_0$ . We now describe how to calculate the  $e$ 's.

The first step in determining the  $e$ 's for a contingency table is to calculate the row and column total frequencies (these are called the **marginal frequencies**) and also the grand total of all the cell frequencies. The  $e$ 's then follow from a simple rationale, as illustrated in Example 10.2.3.

**Example  
10.2.3**

**Migraine Headache** Table 10.2.1 shows the migraine data of Example 10.1.1, together with the marginal frequencies.

Table 10.2.1 Observed frequencies for migraine study			
	Surgery		Total
	Real	Sham	
Success	41	15	56
No success	8	11	19
Total	49	26	75

The  $e$ 's should agree exactly with the null hypothesis. Because  $H_0$  asserts that the probability of success does not depend on the treatment, we can generate an estimate of this probability by pooling the two treatment groups; from Table 10.2.1, the pooled estimate, based on the marginal totals, is  $\frac{56}{75}$ . That is, if  $H_0$  is true, then the two columns “Real” and “Sham” are equivalent and we can pool them together. Our best estimate of  $\Pr\{\text{successful outcome}\}$  is then the pooled estimate  $\frac{56}{75}$ . We can then apply this estimate to each treatment group to yield the number of successful outcomes expected according to  $H_0$ , as follows:

$$\text{Real surgery group: } \frac{56}{75} \times 49 = 36.59 \text{ successful outcomes expected}$$

$$\text{Sham surgery group: } \frac{56}{75} \times 26 = 19.41 \text{ successful outcomes expected}$$

Likewise, the pooled estimate of  $\Pr\{\text{a surgery will } \textit{not} \text{ be successful}\}$  is  $\frac{19}{75}$ . Applying this probability to the two treatment groups gives

$$\text{Real surgery group: } \frac{19}{75} \times 49 = 12.41 \text{ unsuccessful outcomes expected}$$

$$\text{Sham surgery group: } \frac{19}{75} \times 26 = 6.59 \text{ unsuccessful outcomes expected}$$

The expected frequencies are shown in parentheses in Table 10.2.2. Note that the marginal totals for the  $e$ 's are the same as for the  $o$ 's.

**Table 10.2.2** Observed and expected frequencies for migraine study

	Surgery		Total
	Real	Sham	
Success	41 (36.59)	15 (19.41)	56
No success	8 (12.41)	11 (6.59)	19
Total	49	26	75

In practice, it is not necessary to proceed through a chain of reasoning to obtain the expected frequencies for a contingency table. The procedure for calculating the  $e$ 's can be condensed into a simple formula. The expected frequency for each cell is calculated from the marginal total frequencies for the same row and column, as follows:

### Expected Frequencies in a Contingency Table

$$e = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Grand total}}$$

The formula produces the same calculation as does the rationale given in Example 10.2.2, as the following example shows.

#### Example 10.2.4

**Migraine Headache** We will apply the preceding formula to the migraine data of Example 10.1.1. The expected frequency of successful outcomes for the real surgery is calculated from the marginal totals as

$$e = \frac{56 \times 49}{75} = 36.59$$

Note that this is the same answer obtained in Example 10.2.2. Proceeding similarly for each cell in the contingency table, we would obtain all the  $e$ 's shown in Table 10.2.2.

**Note:** Although the formula for  $\chi^2$  for contingency tables is the same as given for goodness-of-fit tests in Section 9.4, the method of calculating the  $e$ 's is quite different for contingency tables because the null hypothesis is different.

## The Test Procedure

Other than the differences noted previously when computing expected counts, the chi-square test for a contingency table is carried out similarly to the chi-square goodness-of-fit test. Large values of  $\chi^2_s$  indicate evidence against  $H_0$ . Critical values

are determined from Table 9; the number of degrees of freedom for a  $2 \times 2$  contingency table is

$$df = 1$$

The chi-square test for a  $2 \times 2$  table has 1 degree of freedom because, in a sense, there only is one free cell in the table. Table 10.2.2 has four cells, but once we have determined that the expected cell frequency for the top-left cell is 36.59, the expected frequency for the top-right cell is constrained to be 19.41, since the top row adds across to a total of 56. Likewise, the bottom-left cell is constrained to be 12.41, since the left column adds down to a total of 49. Once these three cells are determined, the remaining cell, on the bottom right, is constrained as well. Thus, there are four cells in the table, but only one of them is “free”; once we have used the null hypothesis to determine the expected frequency for one of the cells, the other cells are constrained.

For a  $2 \times 2$  contingency table, the alternative hypothesis can be directional or nondirectional. Directional alternatives are handled by the familiar two-step procedure, cutting the nondirectional  $P$ -value in half if the data deviate from  $H_0$  in the direction specified by  $H_A$  (or reporting that the  $P$ -value is  $>0.50$  if the data deviate from  $H_0$  in the direction opposite to the direction specified by  $H_A$ ). Note that  $\chi_s^2$  itself does not express directionality; to determine the directionality of the data, one must calculate and compare the estimated probabilities.

The following example illustrates the chi-square test.

**Example  
10.2.5**

**Migraine Headache** For the migraine experiment of Example 10.1.1, let us apply a chi-square test. Given that the experiment involves cranial surgery, so that a Type I error would be quite serious, a conservative choice of  $\alpha$  is called for; we will use  $\alpha = 0.01$ . We may state the null hypothesis and a directional alternative informally as follows:

$H_0$ : The real surgery is no better than the sham surgery for reducing migraine headache.

$H_A$ : The real surgery is better than the sham surgery for reducing migraine headache.

Using the notation of conditional probability, the statements are

$$H_0: \Pr\{\text{Success} | \text{Real}\} = \Pr\{\text{Success} | \text{Sham}\}$$

$$H_A: \Pr\{\text{Success} | \text{Real}\} > \Pr\{\text{Success} | \text{Sham}\}$$

To check the directionality of the data, we calculate the estimated probabilities of response:

$$\hat{\Pr}\{\text{Success} | \text{Real}\} = \frac{41}{49} = 0.837$$

$$\hat{\Pr}\{\text{Success} | \text{Sham}\} = \frac{15}{26} = 0.577$$

and we note that

$$\hat{\Pr}\{\text{Success} | \text{Real}\} > \hat{\Pr}\{\text{Success} | \text{Sham}\}$$

Thus, the data do deviate from  $H_0$  in the direction specified by  $H_A$ . We proceed to calculate the chi-square statistic from Table 10.2.2 as

$$\begin{aligned} \chi_s^2 &= \frac{(41 - 36.59)^2}{36.59} + \frac{(15 - 19.41)^2}{19.41} + \frac{(8 - 12.41)^2}{12.41} + \frac{(11 - 6.59)^2}{6.59} \\ &= 6.06 \end{aligned}$$

From Table 9 with  $df = 1$ , we find that  $\chi_{1,0.02}^2 = 5.41$  and  $\chi_{1,0.01}^2 = 6.63$ , and so we have  $0.005 < P\text{-value} < 0.01$ . Thus, we reject  $H_0$  and find that the data provide

sufficient evidence to conclude that the real surgery is better than the sham surgery for reducing migraine headache.

Note that, even though  $\hat{\Pr}\{\text{Reduction}|\text{Real}\}$  and  $\hat{\Pr}\{\text{Reduction}|\text{Sham}\}$  do not enter into the calculation of  $\chi_s^2$ , the calculation of  $\hat{\Pr}\{\text{Reduction}|\text{Real}\}$  and  $\hat{\Pr}\{\text{Reduction}|\text{Sham}\}$  is an important part of the test procedure; the information provided by the quantities  $\hat{\Pr}\{\text{Success}|\text{Real}\}$  and  $\hat{\Pr}\{\text{Success}|\text{Sham}\}$  is essential for meaningful interpretation of the results.\* ■

**Computational Notes** The following tips are helpful in analyzing a  $2 \times 2$  contingency table:

1. The contingency table format is convenient for computations. For presenting the data in a report, however, it is usually better to use a more readable form of display such as Table 10.1.2; some additional examples are shown in the exercises.
2. For calculating  $\chi_s^2$ , the observed frequencies ( $o$ 's) must be *absolute*, rather than relative, frequencies; also, *the table must contain all four cells*, so that the sum of the  $o$ 's is equal to the total number of observations.

### Illustration of the Null Hypothesis

The chi-square statistic measures discrepancy between the data and the null hypothesis in an indirect way; the sample conditional probabilities are involved indirectly in the calculation of the expected frequencies. If sample conditional probabilities are equal, then the value of  $\chi_s^2$  is zero. Here is an example.

---

**Example  
10.2.6**

**Fictitious Migraine Study** Table 10.2.3 shows fictitious data for a migraine study similar to that described in Example 10.1.1.

For the data of Table 10.2.3, the estimated probabilities of successful surgery are *equal*:

$$\hat{\Pr}\{\text{Success}|\text{Real}\} = \frac{30}{150} = 0.20$$

$$\hat{\Pr}\{\text{Success}|\text{Sham}\} = \frac{20}{100} = 0.20$$

You can easily verify that, for Table 10.2.3, the expected frequencies are equal to observed frequencies, so that the value of  $\chi_s^2$  is zero. Also notice that the columns of the table are proportional to each other:

$$\frac{30}{120} = \frac{20}{80}$$

**Table 10.2.3** Fictitious data for migraine study

		Surgery		Total
		Real	Sham	
Success	Real	30	20	50
	No success	120	80	200
Total		150	100	250

\*It is natural to wonder why we do not use a more direct comparison of  $\hat{\Pr}\{E|C\}$  and  $\hat{\Pr}\{E|\text{not } C\}$ . In fact, there is a test procedure based on a *t*-type statistic, calculated by dividing  $(\hat{\Pr}\{E|C\} - \hat{\Pr}\{E|\text{not } C\})$  by its standard error. This *t*-type procedure is equivalent to the chi-square test. We have chosen to present the chi-square test instead, for two reasons: (1) It can be extended to contingency tables larger than  $2 \times 2$ ; (2) in certain applications the chi-square statistic is more natural than the *t*-type statistic; some of these applications appear in Section 10.3.

As the preceding example suggests, an “eyeball” analysis of a contingency table is based on checking for proportionality of the columns. If the columns are nearly proportional, then the data agree fairly well with  $H_0$ ; if they are highly nonproportional, then the data disagree with  $H_0$ . The following example shows a case in which the data agree quite well with the expected frequencies under  $H_0$ .

**Example  
10.2.7**

**HIV Testing** The data from Example 10.1.2 show similar percentages of men and women who had been tested for HIV. The natural null hypothesis is that  $\Pr\{\text{HIV test}|\text{Female}\} = \Pr\{\text{HIV test}|\text{Male}\}$  and that the sample proportions differ only due to chance error in the sampling process. The expected frequencies are shown in parentheses in Table 10.2.4. The chi-square test statistic is  $\chi_s^2 = 0.035$ . From Table 8 with  $df = 1$ , we find that  $\chi_{1,0.20}^2 = 1.64$ . Thus, the  $P$ -value is greater than 0.20 (using a computer yields  $P$ -value = 0.85) and we do not reject the null hypothesis. Our conclusion is that the data provide no significant evidence that there is a difference in the rates which which men and women (at the college where the study was conducted) have been tested for HIV. ■

<b>Table 10.2.4</b> Observed and expected frequencies for HIV study			
	Female	Male	Total
HIV test	9 (8.64)	8 (8.36)	17
No HIV test	52 (52.36)	51 (50.64)	103
Total	61	59	120

Note that the actual value of  $\chi_s^2$  depends on the sample sizes as well as the degree of nonproportionality; as discussed in Section 9.4, the value of  $\chi_s^2$  varies directly with the number of observations if the percentage composition of the data is kept fixed and the number of observations is varied. This reflects the fact that a given percentage deviation from  $H_0$  is less likely to occur by chance with a larger number of observations.

### Exercises 10.2.1–10.2.14

**10.2.1** The accompanying partially complete contingency table shows the responses to two treatments:

		TREATMENT		Response	Success	Failure	Total
		1	2				
Response	Success	70					
	Failure						
	Total	100	200				

**10.2.2** Proceed as in Exercise 10.2.1 for the following contingency table:

		TREATMENT		Response	Success	Failure	Total
		1	2				
Response	Success	30					
	Failure						
	Total	300	100				

**10.2.3** Proceed as in Exercise 10.2.1 for the following contingency table:

		TREATMENT		Response	Success	Failure	Total
		1	2				
Response	Success	5	20				
	Failure						
	Total	10					

- (a) Invent a fictitious data set that agrees with the table and for which  $\chi_s^2 = 0$ .
- (b) Calculate the estimated probabilities of success ( $\hat{\Pr}(\text{Success}|\text{Treatment 1})$  and  $\hat{\Pr}(\text{Success}|\text{Treatment 2})$ ) a for your data set. Are they equal?

**10.2.4** Most salamanders of the species *P. cinereus* are red striped, but some individuals are all red. The all-red form is thought to be a mimic of the salamander *N. viridescens*, which is toxic to birds. In order to test whether the mimic form actually survives more successfully, 163 striped and 41 red individuals of *P. cinereus* were exposed to predation by a natural bird population. After two hours, 65 of the striped and 23 of the red individuals were still alive.<sup>3</sup> Use a chi-square test to assess the evidence that the mimic form survives more successfully. Use a directional alternative and let  $\alpha = 0.05$ .

- State the null hypothesis in words.
- State the null hypothesis in symbols.
- Compute the sample survival proportions for each group and display the values in a table similar to Table 10.1.2.
- Find the value of the test statistic and the  $P$ -value.
- State the conclusion of the test in the context of this setting.

**10.2.5** Can attack of a plant by one organism induce resistance to subsequent attack by a different organism? In a study of this question, individually potted cotton (*Gossypium*) plants were randomly allocated to two groups. Each plant in one group received an infestation of spider mites (*Tetranychus*); the other group were kept as controls. After two weeks the mites were removed and all plants were inoculated with *Verticillium*, a fungus that causes wilt disease. The accompanying table shows the numbers of plants that developed symptoms of wilt disease.<sup>4</sup> Do the data provide sufficient evidence to conclude that infestation with mites induces resistance to wilt disease? Use a chi-square test against a directional alternative following the five steps (a–e) in Problem 10.2.4. Let  $\alpha = 0.01$ .

		TREATMENT	
		MITES	NO MITES
Response	Wilt disease	11	17
	No wilt disease	15	4
	Total	26	21

**10.2.6** It has been suspected that prolonged use of a cellular telephone increases the chance of developing brain cancer due to the microwave-frequency signal that is transmitted by the cell phone. According to this theory, if a cell phone is repeatedly held near one side of the head, then brain tumors are more likely to develop on that side of the head. To investigate this, a group of patients were studied who had used cell phones for at least six months prior to developing brain tumors. The patients were asked whether they routinely held the cell phone to a certain ear and, if so, which ear. The 88 responses (from those who preferred one side over the other) are shown in the following table.<sup>5</sup> Do the data provide sufficient

evidence to conclude that use of cellular telephones leads to an increase in brain tumors on that side of the head? Use a chi-square test against a directional alternative following the five steps (a–e) in Problem 10.2.4. Let  $\alpha = 0.05$ .

Brain tumor side	PHONE HOLDING SIDE	
	LEFT	RIGHT
Left	14	28
Right	19	27
Total	33	55

**10.2.7** Phenytoin is a standard anticonvulsant drug which unfortunately has many toxic side effects. A study was undertaken to compare phenytoin with valproate, another drug in the treatment of epilepsy. Patients were randomly allocated to receive either phenytoin or valproate for 12 months. Of 20 patients receiving valproate, 6 were free of seizures for the 12 months while 6 of 17 patients receiving phenytoin were seizure-free.<sup>6</sup>

- Use a chi-square test to compare the seizure-free response rates for the two drugs. Let  $H_A$  be non-directional and  $\alpha = 0.10$  following the five steps (a–e) in Problem 10.2.4.
- Do your conclusions in part (a) provide evidence that valproate and phenytoin are equally effective in preventing seizures? Discuss.

**10.2.8** Estrus synchronization products are used to bring cows into heat at a predictable time so that they can be reliably impregnated by artificial insemination. In a study of two estrus synchronization products, 42 mature cows (aged 4 to 8 years) were randomly allocated to receive either product A or product B, and then all cows were bred by artificial insemination. The table shows how many of the inseminations resulted in pregnancy.<sup>7</sup> Use a chi-square test to compare the effectiveness of the two products in producing pregnancy following the five steps (a–e) in Problem 10.2.4. Use a nondirectional alternative and let  $\alpha = 0.05$ .

	TREATMENT	
	PRODUCT A	PRODUCT B
Total number of cows	21	21
Number of cows pregnant	8	15

**10.2.9** Experimental studies of cancer often use strains of animals that have a naturally high incidence of tumors. In one such experiment, tumor-prone mice were kept in a sterile environment; one group of mice was maintained entirely germ free, while another group was exposed to the intestinal bacterium *Escherichia coli*.

The accompanying table shows the incidence of liver tumors.<sup>8</sup>

TREATMENT	TOTAL NUMBER OF MICE	MICE WITH LIVER TUMORS	
		NUMBER	PERCENT
Germ free	49	19	39%
<i>E. coli</i>	13	8	62%

- (a) How strong is the evidence that tumor incidence is higher in mice exposed to *E. coli*? Use a chi-square test against a directional alternative following the five steps (a–e) in Problem 10.2.4. Let  $\alpha = 0.05$ .
- (b) How would the result of part (a) change if the percentages (39% and 62%) of mice with tumors were the same, but the sample sizes were (i) doubled (98 and 26)? (ii) tripled (147 and 39)? [Hint: Part (b) requires almost no calculation.]

**10.2.10** In a randomized clinical trial to determine the most effective timing of administration of chemotherapeutic drugs to lung cancer patients, 16 patients were given four drugs simultaneously and 11 patients were given the same drugs sequentially. Objective response to the treatment (defined as shrinkage of the tumor by at least 50%) was observed in 11 of the patients treated simultaneously and in 3 of the patients treated sequentially.<sup>9</sup> Do the data provide evidence as to which timing is superior? Use a chi-square test against a nondirectional alternative following the five steps (a–e) in Problem 10.2.4. Let  $\alpha = 0.05$ .

**10.2.11** Physicians conducted an experiment to investigate the effectiveness of external hip protectors in preventing hip fractures in elderly people. They randomly assigned some people to get hip protectors and others to be the control group. They recorded the number of hip fractures in each group.<sup>10</sup> Do the data in the following table provide sufficient evidence to conclude that hip protectors reduce the likelihood of fracture? Use a chi-square test against a directional alternative following the five steps (a–e) in Problem 10.2.4. Let  $\alpha = 0.01$ .

Response	Hip fracture	TREATMENT	
		HIP PROTECTOR	CONTROL
	Hip fracture	13	67
	No hip fracture	640	1081
	Total	653	1148

**10.2.12** A sample of 276 healthy adult volunteers were asked about the variety of social networks that they were in (e.g., relationships with parents, close neighbors, workmates, etc.). They were then given nasal drops containing a rhinovirus and were quarantined for five days. Of the

123 subjects who were in five or fewer types of social relationships 57 (46.3%) developed colds. Of 153 who were in at least six types of social relationships 52 (34.0%) developed colds.<sup>11</sup> Thus, the data suggest that having more types of social relationships helps one develop resistance to the common cold. Determine whether this difference is statistically significant. That is, use a chi-square test to test the null hypothesis that the probability of getting a cold does not depend on the number of social relationships a person is in following the five steps (a–e) in Problem 10.2.4. Use a nondirectional alternative and let  $\alpha = 0.05$ .

**10.2.13** The drug ancrod was tested in a double-blind clinical trial in which subjects who had strokes were randomly assigned to get either ancrod or a placebo. One response variable in the study was whether or not a subject experienced intracranial hemorrhaging.<sup>12</sup> The data are provided in the following table. Use a chi-square test to determine whether the difference in hemorrhaging rates is statistically significant following the five steps (a–e) in Problem 10.2.4. Use a nondirectional alternative and let  $\alpha = 0.05$ .

Hemorrhage?	TREATMENT		
		ANCROD	PLACEBO
Yes		13	5
No		235	247
Total		248	252

**10.2.14** Do women respond to men's solicitations more readily during the fertile phase of their menstrual cycles? In a study of this question each of two hundred 18- to 25-year-old women who were walking alone in a city were approached by an attractive 20-year-old man who solicited the woman's telephone number. Previous research suggested that during the fertile phase of her menstrual cycle a woman would be more receptive to this kind of request than at other times. Of 60 women who were in the fertile phase of their cycles 13 gave out their phone numbers and 47 refused. The corresponding numbers for the 140 women not in the fertile phase of their cycles were 11 and 129.<sup>13</sup> The data are summarized in the following table. Use a chi-square test to determine whether the difference in success rates provides significant evidence in favor of an appropriate directional alternative, using  $\alpha = 0.02$  following the five steps (a–e) in Problem 10.2.4.

Success?	PHASE		
		FERTILE	NOT
Yes		13	11
No		47	129
Total		60	140

## 10.3 Independence and Association in the $2 \times 2$ Contingency Table

The  $2 \times 2$  contingency table is deceptively simple. In this section we explore further the relationships that it can express.

### Two Contexts for Contingency Tables

A  $2 \times 2$  contingency table can arise in two contexts, namely:

1. Two independent samples with a dichotomous observed variable
2. One sample with two dichotomous observed variables

The first context is illustrated by the migraine data of Example 10.1.1, which can be viewed as two independent samples—the real surgery group and the sham surgery group—of sizes  $n_1 = 49$  and  $n_2 = 26$ . The observed variable is success (or failure) of the surgery. Any study involving a dichotomous observed variable and completely randomized allocation to two treatments can be viewed this way. The second context is illustrated by the HIV data of Example 10.1.2, which can be viewed as a single sample of  $n = 120$  students, observed with respect to two dichotomous variables—sex (male or female) and HIV test status (whether or not the student had been tested for HIV).

The two contexts—two samples with one variable or one sample with two variables—are not always sharply differentiated. For instance, the HIV data of Example 10.1.2 could have been collected in two samples—61 women and 59 men—observed with respect to one dichotomous variable (HIV test status).

The arithmetic of the chi-square test is the same in both contexts, but the statement and interpretation of hypotheses and conclusions can be very different.

### Independence and Association

In many contingency tables, the columns of the table play a different role than the rows. For instance, in the migraine data of Example 10.1.1, the columns represent treatments and the rows represent responses. Also, in Example 10.1.2 it seems more natural to define the columnwise conditional probabilities  $\Pr\{\text{HIV test} | F\}$  and  $\Pr\{\text{HIV test} | M\}$  rather than the rowwise conditional probabilities  $\Pr\{F | \text{HIV test}\}$  and  $\Pr\{M | \text{HIV test}\}$ .

On the other hand, in some cases it is natural to think of the rows and the columns of the contingency table as playing interchangeable roles. In such a case, conditional probabilities may be calculated either rowwise or columnwise, and the null hypothesis for the chi-square test may be expressed either rowwise or columnwise. The following is an example.

---

#### Example 10.3.1

**Hair Color and Eye Color** To study the relationship between hair color and eye color in a German population, an anthropologist observed a sample of 6,800 men, with the results shown in Table 10.3.1.<sup>14</sup>

**Table 10.3.1** Hair color and eye color

		Hair color		Total
Eye Color	Dark	Dark	Light	
Eye Color	Dark	726	131	857
Total	3,129	2,814	5,943	6,800

The data of Table 10.3.1 would be naturally viewed as a single sample of size  $n = 6,800$  with two dichotomous observed variables—hair color and eye color. To describe the data, let us denote dark and light eyes by DE and LE, and dark and light hair by DH and LH. We may calculate estimated columnwise conditional probabilities as follows:

$$\hat{\Pr}\{DE|DH\} = \frac{726}{3855} \approx 0.19$$

$$\hat{\Pr}\{DE|LH\} = \frac{131}{2945} \approx 0.04$$

A natural way to analyze the data is to compare these values: 0.19 versus 0.04. On the other hand, it is just as natural to calculate and compare estimated rowwise conditional probabilities:

$$\hat{\Pr}\{DH|DE\} = \frac{726}{857} \approx 0.85$$

$$\hat{\Pr}\{DH|LE\} = \frac{3129}{5943} \approx 0.53$$

Corresponding to these two views of the contingency table, the null hypothesis for the chi-square test can be stated columnwise as

$$H_0: \Pr\{DE|DH\} = \Pr\{DE|LH\}$$

or rowwise as

$$H_0: \Pr\{DH|DE\} = \Pr\{DH|LE\}$$

As we shall see, these two hypotheses are equivalent—that is, any population that satisfies one of them also satisfies the other. ■

When a data set is viewed as a single sample with two observed variables, the relationship expressed by  $H_0$  is called **statistical independence** of the row variable and the column variable. Variables that are not independent are called **dependent** or **associated**. Thus, the chi-square test is sometimes called a “test of independence” or a “test for association.”

---

**Example 10.3.2**

**Hair Color and Eye Color** The null hypothesis of Example 10.3.1 can be stated verbally as

$$H_0: \text{Eye color is independent of hair color}$$

or

$$H_0: \text{Hair color is independent of eye color}$$

or, more symmetrically,

$$H_0: \text{Hair color and eye color are independent}$$



The null hypothesis of independence can be stated generically as follows. Two groups,  $G_1$  and  $G_2$ , are to be compared with respect to the probability of a characteristic  $C$ . The null hypothesis is

$$H_0: \Pr\{C|G_1\} = \Pr\{C|G_2\}$$

Note that each of the two statements of  $H_0$  in Example 10.3.1 is of this form.

To further clarify the meaning of the null hypothesis of independence, in the following example we examine a data set that agrees *exactly* with  $H_0$ .

**Example  
10.3.3**

**Plant Height and Disease Resistance** Consider a (fictitious) species of plant that can be categorized as short (S) or tall (T) and as resistant (R) or nonresistant (NR) to a certain disease. Consider the following null hypothesis:

$H_0$ : Plant height and disease resistance are independent

Each of the following is a valid statement of  $H_0$ :

1.  $H_0: \Pr\{R|S\} = \Pr\{R|T\}$
2.  $H_0: \Pr\{NR|S\} = \Pr\{NR|T\}$
3.  $H_0: \Pr\{S|R\} = \Pr\{S|NR\}$
4.  $H_0: \Pr\{T|R\} = \Pr\{T|NR\}$

The following is not a statement of  $H_0$ :

5.  $H_0: \Pr\{R|S\} = \Pr\{NR|S\}$

Note the difference between statements 5 and 1. Statement 1 compares two groups (short and tall plants) with respect to disease resistance, whereas statement 5 is a statement about the distribution of disease resistance in only *one* group (short plants); statement 5 merely asserts that half (50%) of short plants are resistant and half are nonresistant.

Suppose, now, that we choose a random sample of 100 plants from the population and we obtain the data in Table 10.3.2.

**Table 10.3.2** Plant height and disease resistance

		Height		Total
		S	T	
Resistance	R	12	18	30
	NR	28	42	70
	Total	40	60	100

The data in Table 10.3.2 agree exactly with  $H_0$ ; this agreement can be checked in four different ways, corresponding to the four different symbolic statements of  $H_0$ :

1.  $\hat{\Pr}\{R|S\} = \hat{\Pr}\{R|T\}$

$$\frac{12}{40} = 0.30 = \frac{18}{60}$$

2.  $\hat{\Pr}\{NR|S\} = \hat{\Pr}\{NR|T\}$

$$\frac{28}{40} = 0.70 = \frac{42}{60}$$

3.  $\hat{\Pr}\{S|R\} = \hat{\Pr}\{S|NR\}$

$$\frac{12}{30} = 0.40 = \frac{28}{70}$$

$$4. \hat{P}r\{T|R\} = \hat{P}r\{T|NR\}$$

$$\frac{18}{30} = 0.60 = \frac{42}{70}$$

Note that the data in Table 10.3.2 do *not* agree with statement 5:

$$\hat{P}r\{R|S\} = \frac{12}{40} = 0.30 \text{ and } \hat{P}r\{NR|S\} = \frac{28}{40} = 0.70$$

$$0.30 \neq 0.70$$

■

## Facts about Rows and Columns

The data in Table 10.3.2 display independence whether viewed rowwise or columnwise. This is no accident, as the following fact shows.

**Fact 10.3.1** The columns of a  $2 \times 2$  table are proportional if and only if the rows are proportional. Specifically, suppose that  $a, b, c$ , and  $d$  are any positive numbers, arranged as in Table 10.3.3.

**Table 10.3.3** A general  $2 \times 2$  contingency table

		Total
		$a + b$
Total	$a$	$a + b$
	$c$	$c + d$
Total	$a + c$	$b + d$

Then

$$\frac{a}{c} = \frac{b}{d} \text{ if and only if } \frac{a}{b} = \frac{c}{d}$$

Another way to express this is

$$\frac{a}{a+c} = \frac{b}{b+d} \text{ if and only if } \frac{a}{a+b} = \frac{c}{c+d}$$

You can easily show that Fact 10.3.1 is true; just use simple algebra. Because of Fact 10.3.1, the relationship of independence in a  $2 \times 2$  contingency table is the same whether the table is viewed rowwise or columnwise. Note also that the expected frequencies, and therefore the value of  $\chi^2_s$ , would remain the same if the rows and columns of the contingency table were interchanged. The following fact shows that the *direction* of dependence is also the same whether viewed rowwise or columnwise.

**Fact 10.3.2** Suppose that  $a, b, c$ , and  $d$  are any positive numbers, arranged as in Table 10.3.3. Then

$$\frac{a}{a+c} > \frac{b}{b+d} \text{ if and only if } \frac{a}{a+b} > \frac{c}{c+d}$$

Also

$$\frac{a}{a+c} < \frac{b}{b+d} \text{ if and only if } \frac{a}{a+b} < \frac{c}{c+d}$$

**Note:** For more discussion of conditional probability and independence, see optional Section 3.3.

## Verbal Description of Association

Ideas of logical implication are expressed in everyday English in subtle ways. The following excerpt is from *Alice in Wonderland*, by Lewis Carroll:

“... you should say what you mean,” the March Hare went on.

“I do,” Alice hastily replied; “at least—at least I mean what I say—that’s the same thing, you know.”

“Not the same thing a bit!” said the Hatter. “Why, you might just as well say that ‘I see what I eat’ is the same thing as ‘I eat what I see’!”

... “You might just as well say,” added the Dormouse..., “That ‘I breathe when I sleep’ is the same thing as ‘I sleep when I breathe’!”

“It *is* the same thing with you,” said the Hatter...

We also use ordinary language to express ideas of probability, conditional probability, and association. For instance, consider the following four statements:

Color-blindness is more common among males than among females.

Maleness is more common among color-blind people than femaleness.

Most color-blind people are male.

Most males are color-blind.

The first three statements are all true; they are actually just different ways of saying the same thing. However, the last statement is false.<sup>15</sup>

In interpreting contingency tables, it is often necessary to describe probabilistic relationships in words. This can be quite a challenge. If you become fluent in such description, then you can always “say what you mean” and “mean what you say.” The following two examples illustrate some of the issues.

---

### Example 10.3.4

**Plant Height and Disease Resistance** For the plant height and disease resistance study of Example 10.3.3, we considered the null hypothesis

$H_0$ : Height and resistance are independent.

This hypothesis could also be expressed verbally in various other ways, such as

$H_0$ : Short and tall plants are equally likely to be resistant.

$H_0$ : Resistant and nonresistant plants are equally likely to be tall.

$H_0$ : Resistance is equally common among short and tall plants.




---

### Example 10.3.5

**Hair Color and Eye Color** Let us consider the interpretation of Table 10.3.1. The chi-square statistic is  $\chi_s^2 = 314$ ; from Table 9 we see that the  $P$ -value is tiny, so the null hypothesis of independence is overwhelmingly rejected. We might state our

conclusion in various ways. For instance, suppose we focus on the incidence of dark eyes. From the data we found that

$$\hat{P}r\{DE|DH\} > \hat{P}r\{DE|LH\}$$

that is,

$$\frac{726}{3855} = 0.19 > \frac{131}{2945} = 0.04$$

A natural conclusion from this comparison would be

Conclusion 1: There is sufficient evidence to conclude that dark-haired men have a greater tendency to be dark-eyed than do light-haired men.

This statement is carefully phrased, because the statement

“Dark-haired men have a greater tendency to be dark-eyed.”

is ambiguous by itself; it could mean

“Dark-haired men have a greater tendency to be dark-eyed than do light-haired men.”

or

“Dark-haired men have a greater tendency to be dark-eyed than to be light-eyed.”

The first of these statements says that

$$\hat{P}r\{DE|DH\} > \hat{P}r\{DE|LH\}$$

whereas the second says that

$$\hat{P}r\{DE|DH\} > \hat{P}r\{LE|DH\}$$

The second statement asserts that more than half of dark-haired men have dark eyes. Note that the data do not support this assertion; of the 3,855 dark-haired men, only 19% have dark eyes.

Conclusion 1 is only one of several possible wordings of the conclusion from the contingency table analysis. For instance, one might focus on dark hair and find

Conclusion 2: There is sufficient evidence to conclude that dark-eyed men have a greater tendency to be dark-haired than do light-eyed men.

A more symmetrical phrasing would be

Conclusion 3: There is sufficient evidence to conclude that dark hair is associated with dark eyes.

However, the phrasing in conclusion 3 is easily misinterpreted; it may suggest something like

“There is sufficient evidence to conclude that most dark-haired men are dark-eyed.”

which is not a correct interpretation. ■

We emphasize once again the principle that we stated in Section 10.2: *The calculation and comparison of appropriate conditional probabilities or  $\hat{P}r$ 's is an essential part of the chi-square test.* Example 10.3.5 provides ample illustration of this point.

## Exercises 10.3.1–10.3.12

**10.3.1** Consider a fictitious population of mice. Each animal's coat is either black (B) or grey (G) in color and is either wavy (W) or smooth (S) in texture. Express each of the following relationships in terms of probabilities or conditional probabilities relating to the population of animals.

- Smooth coats are more common among black mice than among grey mice.
- Smooth coats are more common among black mice than wavy coats are.
- Smooth coats are more often black than are wavy coats.
- Smooth coats are more often black than grey.
- Smooth coats are more common than wavy coats.

**10.3.2** Consider a fictitious population of mice in which each animal's coat is either black (B) or grey (G) in color and is either wavy (W) or smooth (S) in texture (as in Exercise 10.3.1). Suppose a random sample of mice is selected from the population and the coat color and texture are observed; consider the accompanying partially complete contingency table for the data.

		HEIGHT		
		B	G	
Texture	W		50	
	S			
Total	60	150		

- Invent fictitious data sets that agree with the table and for which
  - $\hat{P}\{W|B\} > \hat{P}\{W|G\}$ ; (ii)  $\hat{P}\{W|B\} = \hat{P}\{W|G\}$
 In each case, verify your answer by calculating the estimated conditional probabilities.
- For each of the two data sets you invented in part (a), calculate  $\hat{P}\{B|W\}$  and  $\hat{P}\{B|S\}$ .
- Which of the data sets of part (a) has  $\hat{P}\{B|W\} > \hat{P}\{B|S\}$ ? Can you invent a data set for which  $\hat{P}\{W|B\} > \hat{P}\{W|G\}$  but  $\hat{P}\{B|W\} < \hat{P}\{B|S\}$ ?

If so, do it. If not, explain why not.

**10.3.3** Men with prostate cancer were randomly assigned to undergo surgery ( $n = 347$ ) or “watchful waiting” (no surgery,  $n = 348$ ). Over the next several years there were 83 deaths in the first group and 106 deaths in the second group. The results are given in the table.<sup>16</sup>

		TREATMENT		
		SURGERY	WW	TOTAL
Survival	Died	83	106	189
	Alive	264	242	506
Total		347	348	695

- Let D and A represent died and alive, respectively, and let S and WW represent surgery and watchful waiting. Estimate  $\Pr\{D|S\}$  and  $\Pr\{D|WW\}$ .

- The value of the contingency-table chi-square statistic for these data is  $\chi^2_s = 3.75$ . Test for a relationship between the treatment and survival. Use a nondirectional alternative and let  $\alpha = 0.05$ .

**10.3.4** In a study of behavioral asymmetries, 2,391 women were asked which hand they preferred to use (for instance, to write) and which foot they preferred to use (for instance, to kick a ball). The results are reported in the table.<sup>17</sup>

PREFERRED HAND	PREFERRED FOOT	NUMBER OF WOMEN
Right	Right	2,012
Right	Left	142
Left	Right	121
Left	Left	116
Total		2,391

- Estimate the conditional probability that a woman is right-footed, given that she is right-handed.
- Estimate the conditional probability that a woman is right-footed, given that she is left-handed.
- Suppose we want to test the null hypothesis that hand preference and foot preference are independent. Calculate the chi-square statistic for this hypothesis.
- Suppose we want to test the null hypothesis that right-handed women are equally likely to be right-footed or left-footed. Calculate the chi-square statistic for this hypothesis.

**10.3.5** Consider a study to investigate a certain suspected disease-causing agent. One thousand people are to be chosen at random from the population; each individual is to be classified as diseased or not diseased and as exposed or not exposed to the agent. The results are to be cast in the following contingency table:

		EXPOSURE	
		YES	NO
Disease	Yes		
	No		

Let EY and EN denote exposure and nonexposure and let DY and DN denote presence and absence of the disease. Express each of the following statements in terms of conditional probabilities. (Note that “a majority” means “more than half.”)

- (a) The disease is more common among exposed than among nonexposed people.
- (b) Exposure is more common among diseased people than among nondiseased people.
- (c) Exposure is more common among diseased people than is nonexposure.
- (d) A majority of diseased people are exposed.
- (e) A majority of exposed people are diseased.
- (f) Exposed people are more likely to be diseased than are nonexposed people.
- (g) Exposed people are more likely to be diseased than to be nondiseased.

**10.3.6** Refer to Exercise 10.3.5. Which of the statements express the assertion that occurrence of the disease is associated with exposure to the agent? (There may be more than one.)

**10.3.7** Refer to Exercise 10.3.5. Invent fictitious data sets as specified, and verify your answer by calculating appropriate estimated conditional probabilities. (Your data need not be statistically significant.)

- (a) Invent a data set for which

$$\hat{P}\{DY|EY\} > \hat{P}\{DY|EN\} \text{ but}$$

$$\hat{P}\{EY|DY\} < \hat{P}\{EN|DY\}$$

or explain why it is not possible.

- (b) Invent a data set that agrees with statement (a) of Exercise 10.3.5 but with neither (d) nor (e); or, explain why it is not possible.
- (c) Invent a data set for which

$$\hat{P}\{DY|EY\} > \hat{P}\{DY|EN\} \text{ but}$$

$$\hat{P}\{EY|DY\} < \hat{P}\{EY|DN\}$$

or explain why it is not possible.

**10.3.8** An ecologist studied the spatial distribution of tree species in a wooded area. From a total area of 21 acres, he randomly selected 144 quadrats (plots), each 38 feet square, and noted the presence or absence of maples and hickories in each quadrat. The results are shown in the table.<sup>18</sup>

		MAPLES	
		PRESENT	ABSENT
HICKORIES	PRESENT	26	63
Absent		29	26

The value of the chi-square statistic for this contingency table is  $\chi^2_s = 7.96$ . Test the null hypothesis that the two species are distributed independently of each other. Use a nondirectional alternative and let  $\alpha = 0.01$ . In stating your conclusion, indicate whether the data suggest attraction between the species or repulsion. Support your interpretation with estimated conditional probabilities from the data.

**10.3.9** Refer to Exercise 10.3.8. Suppose the data for fictitious tree species, A and B, were as presented in the accompanying table. The value of the chi-square statistic for this contingency table is  $\chi^2_s = 9.07$ . As in Exercise 10.3.9, test the null hypothesis of independence and interpret your conclusion in terms of attraction or repulsion between the species.

Species B	SPECIES A	
	PRESENT	ABSENT
Present	30	10
Absent	49	55

**10.3.10** A randomized experiment was conducted in which patients with coronary artery disease either had angioplasty or bypass surgery. The accompanying table shows the incidence of angina (chest pain) among the patients five years after treatment.<sup>19</sup>

Angina?	TREATMENT		
	ANGIOPLASTY	BYPASS	TOTAL
Yes	111	74	185
No	402	441	843
Total	513	515	1,028

Let A represent angioplasty and B represent bypass.

- (a) Calculate  $\hat{P}\{Yes|A\}$  and  $\hat{P}\{Yes|B\}$ .
- (b) Calculate  $\hat{P}\{A|Yes\}$  and  $\hat{P}\{A|No\}$ .

**10.3.11** Refer to Exercise 10.3.10. Invent a fictitious data set on coronary treatment and angina for 1,000 patients, for which  $\hat{P}\{Yes|A\}$  is twice as great as  $\hat{P}\{Yes|B\}$ , but nevertheless the majority of patients who have angina also had bypass surgery (as opposed to angioplasty).

**10.3.12** Suppose pairs of fraternal twins are examined and the handedness of each twin is determined; assume that all the twins are brother-sister pairs. Suppose data are collected for 1,000 twin pairs, with the results shown in the following table.<sup>20</sup> State whether each of the following statements is true or false.

- (a) Most of the brothers have the same handedness as their sisters.
- (b) Most of the sisters have the same handedness as their brothers.
- (c) Most of the twin pairs are either both right-handed or both left-handed.
- (d) Handedness of twin sister is independent of handedness of twin brother.
- (e) Most left-handed sisters have right-handed brothers.

Brother	SISTER		TOTAL
	LEFT	RIGHT	
Left	15	85	100
Right	135	765	900
Total	150	850	1,000

## 10.4 Fisher's Exact Test (Optional)

In this optional section we consider an alternative to the chi-square test for  $2 \times 2$  contingency tables. This procedure, known as **Fisher's exact test**, is particularly appropriate when dealing with small samples. Example 10.4.1 presents a situation in which Fisher's exact test can be used.

**Example  
10.4.1**

**ECMO** Extracorporeal membrane oxygenation (ECMO) is a potentially life-saving procedure that is used to treat newborn babies who suffer from severe respiratory failure. An experiment was conducted in which 29 babies were treated with ECMO and 10 babies were treated with conventional medical therapy (CMT). The data are shown in Table 10.4.1.<sup>21</sup>

		Treatment			
		CMT	ECMO	Total	
Outcome	Die	4	1	5	
	Live	6	28	34	
	Total	10	29	39	

The data in Table 10.4.1 show that 34 of the 39 babies survived, but 5 of them died. The death rate was 40% for those given CMT and was 3.4% for those given ECMO. However, the sample sizes here are quite small. Is it possible that the difference in death rates happened simply by chance?

The null hypothesis of interest is that outcome (live or die) is independent of treatment (CMT or ECMO). If the null hypothesis is true, then we can think of the data in the following way: The two column headings of “CMT” and “ECMO” are arbitrary labels. Five of the babies would have died no matter which treatment group they were in; 4 of these babies ended up in the CMT group by chance.

The alternative hypothesis asserts that probability of death depends on treatment group. This means that there is a real difference between CMT and ECMO survival rates, which accounts for the sample percentages being different.

Thus, a question of interest is this: “If the null hypothesis is true, how likely is it to get a table of data like Table 10.4.1?” In conducting Fisher's exact test we find the probability that the observed table, Table 10.4.1, would arise by chance, given that the marginal totals—5 deaths and 34 survivors, 10 given CMT and 29 given ECMO—are fixed. To make this more concrete, suppose the null hypothesis is true and another experiment is conducted, with 10 babies given CMT and 29 given ECMO. Further, suppose that 5 of these 39 babies are going to die, no matter which group they are in. That is, there are 5 babies who are so seriously ill that neither treatment would be able to save them. What is the probability that 4 of them will be assigned to the CMT group?

To find this probability, we need to determine the following:

1. The number of ways of assigning exactly 4 of the 5 babies who are fated to die to the CMT group
2. The number of ways of assigning exactly 6 of the 34 babies who are going to survive to the CMT group
3. The number of ways of assigning 10 of the 39 babies to the CMT group

The product of (1) and (2), divided by (3), gives the probability in question. ■

## Combinations

In Section 3.6 we presented the binomial distribution formula. Part of that formula is the quantity  ${}_nC_j$  (which in Section 3.6 we called a binomial coefficient). The quantity  ${}_nC_j$  is the number of ways in which  $j$  objects can be chosen out of a set of  $n$  objects. For instance, the number of ways that a group of 4 babies can be chosen out of 5 babies is  ${}_5C_4$ . The numerical value of  ${}_nC_j$  is given by formula 10.4.1:

$${}_nC_j = \frac{n!}{j!(n-j)!} \quad (10.4.1)$$

where  $n!$  (“ $n$  factorial”) is defined for any positive integer as

$$n! = n(n-1)(n-2)\dots(2)(1)$$

and  $0! = 1$ .

For example, if  $j = 1$  then we have  ${}_nC_1 = \frac{n!}{1!(n-1)!} = n$ , which makes sense:

There are  $n$  ways to choose 1 object from a set of  $n$  objects. If  $j = n$  then we have  ${}_nC_n = \frac{n!}{n!0!} = 1$ , since there is only one way to choose all  $n$  objects from a set of size  $n$ .

### Example 10.4.2

**ECMO** We can apply formula 10.4.1 as follows.

1. The number of ways of assigning 4 babies to the CMT group from among the 5 who are fated to die is  ${}_5C_4 = \frac{5!}{4!1!} = 5$ .
2. The number of ways of assigning 6 babies to the CMT group from among the 34 who are going to survive is  ${}_{34}C_6 = \frac{34!}{6!28!} = 1,344,904$ .
3. The number of ways of assigning 10 babies to the CMT group from among the 39 total babies is  ${}_{39}C_{10} = \frac{39!}{10!29!} = 635,745,396$ .\*

Thus, the probability of getting the same data as those in Table 10.4.1, given that the marginal totals are fixed, is  $\frac{{}_5C_4 \times {}_{34}C_6}{{}_{39}C_{10}} = \frac{5 \times 1344904}{635745396} = 0.01058$ . ■

When conducting Fisher’s exact test of a null hypothesis against a directional alternative, we need to find the probabilities of all tables of data (having the same margins as the observed table) that provide evidence as strongly against  $H_0$ , in the direction predicted by  $H_A$ , as the observed table.

### Example 10.4.3

**ECMO** Prior to this experiment described in Example 10.4.1, there was evidence that suggested that ECMO is better than CMT. Hence, a directional alternative hypothesis is appropriate:

$$H_A: \Pr\{\text{death} | \text{ECMO}\} < \Pr\{\text{death} | \text{CMT}\}$$

The data in the observed table, Table 10.4.1, support  $H_A$ . There is one other possible table, shown as Table 10.4.2, that has the same margins as Table 10.4.1 but which is even more extreme in supporting  $H_A$ . Given that 5 of 39 babies died and

---

\*It is evident from this example that a computer or a graphing calculator is a very handy tool when conducting Fisher’s exact test. This is a statistical procedure that is almost never carried out without the use of technology.

**Table 10.4.2** A more extreme table that could have resulted from the ECMO experiment

		Treatment		Total
		CMT	ECMO	
Outcome	Die	5	0	5
	Live	5	29	34
Total	10	29		39

that 10 babies were assigned to CMT, the most extreme possible result supporting the alternative hypothesis (that ECMO is better than CMT) is the table in which none of the ECMO babies die and all 5 deaths occur in the CMT group.

The probability of Table 10.4.2 occurring, if  $H_0$  is true, is  $\frac{5C_5 \times 34C_5}{39C_{10}} = \frac{1 \times 278256}{635745396} = 0.00044$ . The  $P$ -value is the probability of obtaining data at least as

extreme as those observed, if  $H_0$  is true. In this case, the  $P$ -value is the probability of obtaining either the data in Table 10.4.1 or in Table 10.4.2, if  $H_0$  is true. Thus,  $P$ -value =  $0.01058 + 0.00044 = 0.01102$ . This  $P$ -value is quite small, so the experiment provided strong evidence that  $H_0$  is false and that ECMO really is better than CMT. ■

### Comparison to the Chi-Square Test

The chi-square test presented in Section 10.2 is often used for analyzing  $2 \times 2$  contingency tables. One advantage of the chi-square test is that it can be extended to  $2 \times 3$  tables and other tables of larger dimension, as will be shown in Section 10.6. The  $P$ -value for the chi-square test is based on the chi-square distribution, as the name implies. It can be shown that as the sample size becomes large, this distribution provides a good approximation to the theoretical sampling distribution of the chi-square test statistic  $\chi_s^2$ . If the sample size is small, however, then the approximation can be poor and the  $P$ -value from the chi-square test can be misleading.

Fisher's exact test is called an "exact" test because the  $P$ -value is determined exactly, using calculations such as those shown in Example 10.4.2, rather than being based on an asymptotic approximation. Example 10.4.4 shows how the exact test and the chi-square test compare for the ECMO data.

---

**Example  
10.4.4**

**ECMO** Conducting a chi-square test on the ECMO experiment data in Table 10.4.1 gives a test statistic of

$$\begin{aligned}\chi_s^2 &= \frac{(4 - 1.28)^2}{1.28} + \frac{(1 - 3.72)^2}{3.72} + \frac{(6 - 8.72)^2}{8.72} + \frac{(28 - 25.28)^2}{25.28} \\ &= 8.89\end{aligned}$$

The  $P$ -value (using a directional alternative) is 0.0014. This is quite a bit smaller than the  $P$ -value found with the exact test of 0.01102. ■

### Nondirectional Alternatives and the Exact Test

Typically, the difference between a directional and a nondirectional test is that the  $P$ -value for the nondirectional test is twice the  $P$ -value for the directional test (assuming that the data deviate from  $H_0$  in the direction specified by  $H_A$ ).

For Fisher's exact test this is not true. The  $P$ -value when  $H_A$  is nondirectional is not found by simply doubling the  $P$ -value from the directional test. Rather, a generally accepted procedure is to find the probabilities of all tables that are as likely or less likely than the observed table. These probabilities are added together to get the  $P$ -value for the nondirectional test.\* Example 10.4.5 illustrates this idea.

**Example  
10.4.5**

**Flu Shots** A random sample of college students found that 13 of them had gotten a flu shot at the beginning of the winter and 28 had not. Of the 13 who had a flu shot, 3 got the flu during the winter. Of the 28 who did not get a flu shot, 15 got the flu.<sup>22</sup> These data are shown in Table 10.4.3. Consider the null hypothesis that the probability of getting the flu is the same whether or not one gets a flu shot. The probability of the data in Table 10.4.3, given that the margins are fixed, is

$$\frac{18C_3 \times 23C_{10}}{41C_{13}} = 0.05298.$$

Table	Probability
15      3 13      10	0.05298
16      2 12      11	0.01174
17      1 11      12	0.00138
18      0 10      13	0.00006

**Figure 10.4.1**

Table 10.4.3 Flu shot data			
	No shot	Flu shot	Total
Flu?	Yes	15      3	18
	No	13      10	23
Total	28	13	41

A natural directional alternative would be that getting a flu shot reduces one's chance of getting the flu. Figure 10.4.1 shows the obtained data (from Table 10.4.3) along with tables of possible outcomes that more strongly support  $H_A$ . The probability of each table is given in Figure 10.4.1, as well.

The  $P$ -value for the directional test is the sum of the probabilities of these tables:  $P$ -value = 0.05298 + 0.01174 + 0.00138 + 0.00006 = 0.06616.

A nondirectional alternative states that the probability of getting the flu depends on whether or not one gets a flu shot but does not state whether a flu shot increases or decreases the probability. (Some people might get the flu *because* of the shot, so it is plausible that the overall flu rate is higher among people who get the shot than among those who don't—although public health officials certainly hope otherwise!)

Figure 10.4.2 shows tables of possible outcomes for which the flu rate is higher among those who got the shot than among those who didn't. The probability of each table is given, as well. The first five tables all have probabilities less than 0.05298, which is the probability of the observed data in Table 10.4.3, but the probability of the sixth table is greater than 0.05298. Thus, the contribution to the  $P$ -value from this set of tables is the sum of the first five probabilities: 0.00000 + 0.00002 + 0.00046 + 0.00440 + 0.02443 = 0.02931. Adding this to the  $P$ -value for the directional test of 0.06616 gives the  $P$ -value for the nondirectional test:  $P$ -value = 0.06616 + 0.02931 = 0.09547.

As this example shows, the calculation of a  $P$ -value for Fisher's exact test is quite cumbersome, particularly when the alternative is nondirectional. It is highly recommended that statistics software be used to carry out the test. ■

Table	Probability
5      13 23      0	0.00000
6      12 22      1	0.00002
7      11 21      2	0.00046
8      10 20      3	0.00440
9      9 19      4	0.02443
10     8 18     5	0.08356

**Figure 10.4.2**

\*There is not universal agreement on this process. The  $P$ -value can be taken to be the sum of the probabilities of all "extreme" tables, but there are several ways to define "extreme." One alternative to the method presented here is to order tables according to the values of  $\chi^2_s$  and to count a table as extreme if it has a value of  $\chi^2_s$  that is at least as large as the  $\chi^2_s$  found from the observed table. Another approach is to order the tables according to  $|p_1 - p_2|$ . These methods will sometimes lead to a different  $P$ -value than the  $P$ -value being presented here.

## Exercises 10.4.1–10.4.8

**10.4.1** Consider conducting Fisher's exact test with the following fictitious table of data. Let the null hypothesis be that treatment and response are independent and let the alternative be the directional hypothesis that treatment B is better than treatment A. List the tables of possible outcomes that more strongly support  $H_A$ .

		TREATMENT		TOTAL
		A	B	
Outcome	Die	4	2	6
	Live	10	14	24
	Total	14	16	30

**10.4.2** Repeat Exercise 10.4.1 with the following table of data.

		TREATMENT		TOTAL
		A	B	
Outcome	Die	5	3	8
	Live	12	13	25
	Total	17	16	33

**10.4.3** In a randomized, double-blind clinical trial, 156 subjects were given an antidepressant medication to help them stop smoking; a second group of 153 subjects were given a placebo. Insomnia was more common in the anti-depressant group than in the placebo group; Fisher's exact test of the insomnia data gave a  $P$ -value of 0.008.<sup>23</sup> Interpret this  $P$ -value in the context of the clinical trial.

**10.4.4 (Computer exercise)** A random sample of 99 students in a Conservatory of Music found that 9 of the 48 women sampled had “perfect pitch” (the ability to identify, without error, the pitch of a musical note), but only 1 of the 51 men sampled had perfect pitch.<sup>24</sup>

Conduct Fisher's exact test of the null hypothesis that having perfect pitch is independent of sex. Use a directional alternative and let  $\alpha = 0.05$ . Do you reject  $H_0$ ? Why or why not?

**10.4.5** Consider the data from Exercise 10.4.4. Conduct a chi-square test and compare the results of the chi-square test to the results of Fisher's exact test.

**10.4.6 (Computer exercise)** The growth factor pleiotrophin is associated with cancer progression in humans. In an attempt to monitor the growth of tumors, doctors measured serum pleiotrophin levels in patients with pancreatic cancer and in a control group of patients. They found that only 2 of 28 control patients had serum levels more than two standard deviations above the control group mean, whereas 20 of 41 cancer patients had serum levels this high.<sup>25</sup> Use Fisher's exact test to determine whether a discrepancy this large (2 of 28 versus 20 of 41) is likely to happen by chance. Use a directional alternative and let  $\alpha = 0.05$ .

**10.4.7 (Computer exercise)** A group of 225 men with benign prostatic hyperplasia were randomly assigned to take saw palmetto extract or a placebo in a double-blind trial. One year into the experiment 45 of the 112 men in the saw palmetto group (40%) thought they were taking saw palmetto, compared to 52 of the 113 men in the placebo group (46%).<sup>26</sup> Is this difference consistent with chance variation? Conduct Fisher's exact test using a nondirectional alternative.

**10.4.8 (Computer exercise)** An experiment involving subjects with schizophrenia compared “personal therapy” to “family therapy.” Only 2 out of 23 subjects assigned to the personal therapy group suffered psychotic relapses in the first year of the study, compared to 8 of the 24 subjects assigned to the family therapy group.<sup>27</sup> Is this sufficient evidence to conclude, at the 0.05 level of significance, that the two types of therapies are not equally effective? Conduct Fisher's exact test using a nondirectional alternative.

## 10.5 The $r \times k$ Contingency Table

The ideas of Sections 10.2 and 10.3 extend readily to contingency tables that are larger than  $2 \times 2$ . We now consider a contingency table with  $r$  rows and  $k$  columns, which is termed an  **$r \times k$  contingency table**. Here is an example.

### Example 10.5.1

**Plover Nesting** Wildlife ecologists monitored the breeding habitats of mountain plovers for three years and made note of where the plovers nested. They found 66 nests on agricultural fields (AF), 67 nests in shortgrass prairie dog habitat (PD), and

20 nests on other grassland (G). The nesting choices varied across the years for these 153 sampled plover broods; Table 10.5.1 shows the data.<sup>28</sup>

<b>Table 10.5.1</b> Plover nest locations across three years				
Location	Year			Total
	2004	2005	2006	
<b>Agricultural field (AF)</b>	21	19	26	<b>66</b>
<b>Prairie dog habitat (PD)</b>	17	38	12	<b>67</b>
<b>Grassland (G)</b>	5	6	9	<b>20</b>
<b>Total</b>	<b>43</b>	<b>63</b>	<b>47</b>	<b>153</b>

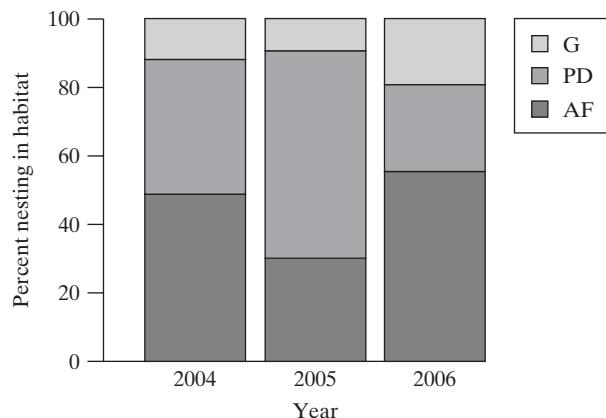
To compare the distributions in the three locations, we can calculate the columnwise percentages, as displayed in Table 10.5.2. (For instance, in the 2004 sample  $\frac{21}{43}$ , or 48.8%, of the nests were on agricultural fields.) Inspection of Table 10.5.2 shows some clear differences among the three percentage distributions (columns), with prairie dog habitat being much more common in 2005 than in the other years.

<b>Table 10.5.2</b> Percentage distributions of plover nests by year			
Location	Year		
	2004	2005	2006
<b>Agricultural field (AF)</b>	48.8	30.2	55.3
<b>Prairie dog habitat (PD)</b>	39.5	60.3	25.5
<b>Grassland (G)</b>	11.6	9.5	19.1
<b>Total</b>	<b>99.9*</b>	<b>100.0</b>	<b>99.9*</b>

\*The sums of the 2004 and 2006 percentages differ from 100% due to rounding.

Figure 10.5.1 is a bar chart of the data, which gives a visual impression of the distributions.

**Figure 10.5.1** Stacked relative frequency (percentage) chart of plover nest locations by year



## The Chi-Square Test for the $r \times k$ Table

The goal of statistical analysis of an  $r \times k$  contingency table is to investigate the relationship between the row variable and the column variable. Such an investigation can begin with an inspection of the columnwise or rowwise percentages, as in Table 10.5.2. One route to further analysis is to ask whether the discrepancies in percentages are too large to be explained as sampling error. This question can be answered by a chi-square test. The chi-square statistic is calculated from the familiar formula

$$\sum_{\text{all cells}} \frac{(o_i - e_i)^2}{e_i}$$

where the sum is over all  $I = r \times k$  cells of the contingency table, and the expected frequencies ( $e$ 's) are calculated as

$$e = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Grand total}}$$

This method of calculating the  $e$ 's can be justified by a simple extension of the rationale given in Section 10.2. Critical values for the chi-square test are obtained from Table 9 with

$$df = (r - 1)(k - 1)$$

The following example illustrates the chi-square test.

---

**Example  
10.5.2**

**Plover Nesting** Let us apply the chi-square test to the plover nesting data of Example 10.5.1. The null hypothesis is

$H_0$ : The population distributions of nest locations are the same in the three years.

This hypothesis can be stated symbolically in conditional probability notation as follows:

$$H_0: \left\{ \begin{array}{l} \Pr \{AF|2004\} = \Pr \{AF|2005\} = \Pr \{AF|2006\} \\ \Pr \{PD|2004\} = \Pr \{PD|2005\} = \Pr \{PD|2006\} \\ \Pr \{G|2004\} = \Pr \{G|2005\} = \Pr \{G|2006\} \end{array} \right\}$$

Note that the percentages in Table 10.5.2 are the estimated conditional probabilities; that is,

$$\hat{\Pr}\{AF|2004\} = 0.488$$

$$\hat{\Pr}\{AF|2005\} = 0.302$$

and so on. We test  $H_0$  against the nondirectional alternative hypothesis

$H_A$ : The population distributions of nest locations are not the same in all three years.

Table 10.5.3 shows the observed and expected frequencies.

**Table 10.5.3** Observed and expected frequencies of plover nests

Location	Year			Total
	2004	2005	2006	
<b>Agricultural field (AF)</b>	21 (18.55)	19 (21.18)	26 (20.27)	<b>66</b>
<b>Prairie dog habitat (PD)</b>	17 (18.83)	38 (27.59)	12 (20.58)	<b>67</b>
<b>Grassland (G)</b>	5 (5.62)	6 (8.24)	9 (6.14)	<b>20</b>
<b>Total</b>	<b>43</b>	<b>63</b>	<b>47</b>	<b>153</b>

From Table 10.5.3, we can calculate the test statistic as

$$\begin{aligned}\chi^2_s &= \frac{(21 - 18.55)^2}{18.55} + \frac{(19 - 21.18)^2}{21.18} + \dots + \frac{(9 - 6.14)^2}{6.14} \\ &= 14.09\end{aligned}$$

For these data,  $r = 3$  and  $k = 3$ , so that

$$df = (3 - 1)(3 - 1) = 4$$

From Table 9 with  $df = 4$ , we find that  $\chi^2_{4,0.01} = 13.28$  and  $\chi^2_{4,0.001} = 18.47$ , and so we have  $0.001 < P\text{-value} < 0.01$ . Thus, the chi-square test shows that there is significant evidence that the nesting location preferences differed across the three years. ■

Note that  $H_0$  in Example 10.5.2 is a compound null hypothesis in the sense defined in Section 9.4—that is,  $H_0$  contains more than one independent assertion. This will always be true for contingency tables larger than  $2 \times 2$ , and consequently for such tables the alternative hypothesis for the chi-square test will always be nondirectional and the conclusion, if  $H_0$  is rejected, will be nondirectional. Thus, the chi-square test will often not represent a complete analysis of an  $r \times k$  contingency table.

## Two Contexts for $r \times k$ Contingency Tables

We noted in Section 10.3 that a  $2 \times 2$  contingency table can arise in two different contexts. Similarly, an  $r \times k$  contingency table can arise in the following two contexts:

1.  $k$  independent samples; a categorical observed variable with  $r$  categories
2. One sample; two categorical observed variables—one with  $k$  categories and one with  $r$  categories

As with the  $2 \times 2$  table, the calculation of the chi-square statistic is the same for both contexts, but the statement of hypotheses and conclusions can differ. The following example illustrates the second context.

### Example 10.5.3

**Hair Color and Eye Color** Table 10.5.4 shows the relationship between hair color and eye color for 6,800 German men.<sup>29</sup> (This is the same study as in Example 10.3.2.)

Let us use a chi square test to test the hypothesis

$$H_0: \text{Hair color and eye color are independent.}$$

**Table 10.5.4** Hair color and eye color

		Hair color			
		Brown	Black	Fair	Red
Eye	Brown	438	288	115	16
	Grey or Green	1,387	746	946	53
Color	Blue	807	189	1,768	47

For the data of Table 10.5.4, one can calculate  $\chi_s^2 = 1,074$ . The degrees of freedom for the test are  $df = (3 - 1)(4 - 1) = 6$ . From Table 9 we find  $\chi_{6,0.0001}^2 = 27.86$ . Thus,  $H_0$  is overwhelmingly rejected and we conclude that there is extremely strong evidence that hair color and eye color are associated. ■

## Exercises 10.5.1–10.5.8

**10.5.1** Patients with painful knee osteoarthritis were randomly assigned in a clinical trial to one of five treatments: glucosamine, chondroitin, both, placebo, or Celebrex, the standard therapy. One outcome recorded was whether or not each patient experienced substantial improvement in pain or in ability to function. The data are given in the following table.<sup>30</sup>

TREATMENT	SUCCESSFUL OUTCOME		
	SAMPLE SIZE	NUMBER	PERCENT
Glucosamine	317	192	60.6
Chondroitin	318	202	63.5
Both	317	208	65.6
Placebo	313	178	56.9
Celebrex	318	214	67.3

- (a) Use a chi-square test to compare the success rates at  $\alpha = 0.05$ . (The value of the chi-square statistic is  $\chi_s^2 = 9.29$ .)  
(b) Verify the value of  $\chi_s^2$  given in part (a).

**10.5.2** For a study of free-living populations of the fruitfly *Drosophila subobscura*, researchers placed baited traps in two woodland sites and one open-ground area. The numbers of male and female flies trapped in a single day are given in the table.<sup>31</sup>

	WOODLAND SITE I	WOODLAND SITE II	OPEN GROUND
Males	89	34	74
Females	31	20	136
Total	120	54	210

- (a) Use a chi-square test to compare the sex ratios at the three sites. Let  $\alpha = 0.05$ .

- (b) Construct a table that displays the data in a more readable format, such as the one in Exercise 10.5.1.

**10.5.3** In a classic study of peptic ulcer, blood types were determined for 1,655 ulcer patients. The accompanying table shows the data for these patients and for an independently chosen group of 10,000 healthy controls from the same city.<sup>32</sup>

BLOOD TYPE	ULCER PATIENTS	CONTROLS
O	911	4,578
A	579	4,219
B	124	890
AB	41	313
Total	1,655	10,000

- (a) The value of the chi-square statistic for this contingency table is  $\chi_s^2 = 49.0$ . Carry out the chi-square test at  $\alpha = 0.01$ .

- (b) Construct a table showing the percentage distributions of blood type for patients and for controls.

- (c) Verify the value of  $\chi_s^2$  given in part (a).

**10.5.4** The two claws of the lobster (*Homarus americanus*) are identical in the juvenile stages. By adulthood, however, the two claws normally have differentiated into a stout claw called a “crusher” and a slender claw called a “cutter.” In a study of the differentiation process, 26 juvenile animals were reared in smooth plastic trays and 18 were reared in trays containing oyster chips (which they could use to exercise their claws). Another 23 animals were reared in trays containing only one oyster chip. The claw configurations of all the animals as adults are summarized in the table.<sup>33</sup>

TREATMENT	CLAW CONFIGURATION		
	RIGHT CRUSHER, LEFT	RIGHT CUTTER, LEFT	RIGHT CUTTER, LEFT
	CUTTER	CRUSHER	CUTTER
Oyster chips	8	9	1
Smooth plastic	2	4	20
One oyster chip	7	9	7

- (a) The value of the contingency-table chi-square statistic for these data is  $\chi_s^2 = 24.35$ . Carry out the chi-square test at  $\alpha = 0.01$ .
- (b) Verify the value of  $\chi_s^2$  given in part (a).
- (c) Construct a table showing the percentage distribution of claw configurations for each of the three treatments.
- (d) Interpret the table from part (c): In what way is claw configuration related to treatment? (For example, if you wanted a lobster with two cutter claws, which treatment would you choose and why?)

**10.5.5** A randomized, double-blind, placebo-controlled experiment was conducted in which patients with Alzheimer's disease were given either extract of Ginkgo biloba (EGb) or a placebo for one year. The change in each patient's Alzheimer's Disease Assessment Scale—Cognitive subscale (ADAS-Cog) score was measured. The results are given in the table.<sup>34</sup> (Note: If the ADAS-Cog went down, then the patient improved.)

CHANGE IN ADAS-Cog SCORE					
	-4	-2	-1	+2	+4
OR	TO	TO	TO		OR
BETTER	-3	+1	+3		WORSE
EGb	22	18	12	7	16
Placebo	10	11	19	11	24

- (a) Use a chi-square test to compare the prevalence rates at  $\alpha = 0.05$ . (The value of the chi-square statistic is  $\chi_s^2 = 10.26$ .)
- (b) Verify the value of  $\chi_s^2$  given in part (a).

**10.5.6** Marine biologists have noticed that the color of the outermost growth band on a clam tends to be related to the time of the year in which the clam dies. A biologist conducted a small investigation of whether this is true for the species *Protothaca staminea*. She collected a sample of 78 clam shells from this species and cross-classified them according to (1) month when the clam died and (2) color of the outermost growth band. The data are shown in the following table.<sup>35</sup>

COLOR			
	CLEAR	DARK	UNREADABLE
February	9	26	9
March	6	25	3
Total	15	51	12

Use a chi-square test to compare the color distributions for the two months. Let  $\alpha = 0.10$ .

**10.5.7** A group of patients with a binge-eating disorder were randomly assigned to take either the experimental drug fluvoxamine or a placebo in a nine-week long double-blind clinical trial. At the end of the trial the condition of each patient was classified into one of four categories: no response, moderate response, marked response, or remission. The following table shows a cross classification of the data.<sup>36</sup> Is there statistically significant evidence, at the 0.10 level, to conclude that there is an association between treatment group (fluvoxamine versus placebo) and condition?

	NO RESPONSE	MODERATE RESPONSE	MARKED RESPONSE	REMISION	TOTAL
Fluvoxamine	15	7	3	15	40
Placebo	22	7	3	11	43
Total	37	14	6	26	

**10.5.8** Patients with coronary artery disease were randomly assigned to either receive angioplasty plus medical therapy ( $n = 1149$ ) or medical therapy alone ( $n = 1138$ ) in a clinical trial. Over the next several years 85 angioplasty and 95 medical therapy patients died, with cause

of death categorized as cardiac, other, or unknown. The following table shows a cross classification of the data.<sup>37</sup> Is there statistically significant evidence, at the 0.10 level, to conclude that there is an association between treatment group (angioplasty versus medical therapy) and outcome?

	CARDIAC DEATH	OTHER DEATH	UNKNOWN CAUSE OF DEATH	ALIVE	TOTAL
Angioplasty	23	45	17	1,064	1,149
Medical therapy	25	51	19	1,043	1,138
Total	48	96	36	2,107	2,287

## 10.6 Applicability of Methods

In this section we discuss guidelines for deciding when to use a chi-square test.

### Conditions for Validity

A chi-square test is valid under the following conditions:

1. *Design conditions* For the contingency-table chi-square test, it must be appropriate to view the data in one of the following ways:
  - (a) As two or more independent random samples, observed with respect to a categorical variable; or
  - (b) As one random sample, observed with respect to two categorical variables.
 For either type of chi-square test, the observations within a sample must be independent of each other.
2. *Sample size conditions* The sample size must be large enough. The critical values given in Table 9 are only approximately correct for determining the  $P$ -value associated with  $\chi^2$ . As a rule of thumb, the approximation is considered adequate if each expected frequency ( $e$ ) is at least equal to 5.\* (If the expected frequencies are small and the data form a  $2 \times 2$  contingency table, then Fisher's exact test might be appropriate—see optional Section 10.4.)
3. *Form of  $H_0$*  A generic form of the null hypothesis for the contingency-table chi-square test may be stated as follows:
 
$$H_0: \text{The row variable and the column variable are independent.}$$
4. *Scope of inference* As with other statistical tests, if the data arise from an experiment with random assignment of treatments, as in Example 10.1.1, then we can draw a causal inference; if the experimental units were drawn at random from a population, then we can extend the causal inference to that population. However, if the data arise from an observational study, as in Example 10.1.2, then a small  $P$ -value only allows us to infer that the observed association is not due to chance, but we cannot rule out other explanations.

\*For an  $r \times k$  table with more than 2 rows and columns, the approximation is adequate if the average expected frequency is at least 5, even if some of the cell counts are smaller.

## Verification of Design Conditions

To verify the design conditions, we need to identify a population from which the data may be viewed as a random sample. If the data consist of several samples [situation 1(a)], then the samples are required to be independent of each other. Failure to observe this restriction may result in a loss of power. If the design includes any pairing or matching of experimental units, then the samples would not be independent. A method of analysis for dependent samples is described in Section 10.8.

As always, bias in the sampling procedure must be ruled out. Moreover, chi-square methods are not appropriate when complex random sampling schemes such as cluster sampling or stratified random sampling are used. Finally, there must be no dependency or hierarchical structure in the design. Failure to observe this restriction can result in a vastly inflated chance of Type I error (which is usually much more serious than a loss of power). The following examples show the relevance of checking for dependency in the observations.

### Example 10.6.1

**Food Choice by Insect Larvae** In a behavioral study of the clover root curculio *Sitona hispidulus*, 20 larvae were released into each of six petri dishes. Each dish contained nodulated and nonnodulated alfalfa roots, arranged in a symmetric pattern. (This experiment was more fully described in Example 1.1.5.) After 24 hours the location of each larva was noted, with the results shown in Table 10.6.1.<sup>38</sup>

**Table 10.6.1** Food choice by *Sitona* larvae

Dish	Number of larvae		
	Nodulated Roots	Nonnodulated Roots	Other (died, lost, etc.)
1	5	3	12
2	9	1	10
3	6	3	11
4	7	1	12
5	5	1	14
6	14	3	3
Total	46	12	62

Suppose the following analysis is proposed. A total of 58 larvae made a choice; the observed frequencies of choosing nodulated and nonnodulated roots were 46 and 12, and the corresponding expected frequencies (assuming random choice) would be 29 and 29; these data yield  $\chi^2_s = 19.93$ , from which (using a directional alternative) we find from Table 9 that  $P\text{-value} < 0.00005$ . The validity of this proposed analysis is highly doubtful because it depends on the assumption that all the observations in a given dish are independent of each other; this assumption would certainly be false if (as is biologically plausible) the larvae tend to follow each other in their search for food.

How, then, should the data be analyzed? One approach is to make the reasonable assumption that the observations in one dish are independent of those in another dish. Under this assumption one could use a paired analysis on the six dishes ( $n_d = 6$ ); a paired  $t$  test yields  $P\text{-value} \approx 0.005$  and a sign test yields  $P\text{-value} \approx 0.02$ . Note that the questionable assumption of independence within dishes led to a  $P\text{-value}$  that was much too small. ■

**Example  
10.6.2**

**Pollination of Flowers** A study was conducted to determine the adaptive significance of flower color in the scarlet gilia (*Ipomopsis aggregata*). Six red-flowered plants and six white-flowered plants were chosen for observation in field conditions; hummingbirds were permitted to visit the flowers, but the other major pollinator, a moth, was excluded by covering the plants at night. Table 10.6.2 shows, for each plant, the total number of flowers at the end of the season and the number that had set fruit.<sup>39</sup>

**Table 10.6.2** Fruit set in scarlet gilia flowers

Red-flowered plants			White-flowered plants		
Number of flowers	Number setting fruit	Percent setting fruit	Number of flowers	Number setting fruit	Percent setting fruit
140	26	19	125	21	17
116	11	9	134	17	13
34	0	0	273	81	30
79	9	11	146	38	26
185	28	15	103	17	17
<u>106</u>	<u>11</u>	<u>10</u>	<u>82</u>	<u>24</u>	<u>29</u>
Sum	660	85	863	198	

The question of interest is whether the percentage of fruit set is different for red-flowered than for white-flowered plants. Suppose this question is approached by regarding the individual flower as the observational unit; then the data could be cast in the contingency table format of Table 10.6.3.

**Table 10.6.3** Fruit set in scarlet gilia flowers

		Flower color	
		Red	White
Fruit set	Yes	85	198
	No	575	665
Total		660	863
Percent setting fruit		13	23

Table 10.6.3 yields  $\chi^2_s = 25.0$ , for which Table 9 gives  $P\text{-value} < 0.0001$ . However, this analysis is not correct, because the observations on flowers on the same plant are not independent of each other; they are dependent because the pollinator (the hummingbird) tends to visit flowers in groups, and perhaps also because the flowers on the same plant are physiologically and genetically related. The chi-square test is invalidated by the hierarchical structure in the data.

A better approach would be to treat the entire plant as the observational unit. For instance, one could take the “Percentage Setting Fruit” column of Table 10.6.2 as the basic observations; applying a  $t$  test to the values yields  $t_s = 2.88$  (with  $0.01 < P\text{-value} < 0.02$ ), and applying a Wilcoxon-Mann-Whitney test yields  $U_s = 32$  (with  $0.02 < P\text{-value} < 0.05$ ). Thus, the  $P\text{-value}$  from the inappropriate chi-square analysis is much too small. ■

## Power Considerations

In many studies the chi-square test is valid but is not as powerful as a more appropriate test. Specifically, consider a situation in which the rows or the columns (or both) of the contingency table correspond to a *rankable* categorical variable with more than two categories. The following is an example.

### Example 10.6.3

**Physiotherapy** A randomized clinical trial was conducted to determine whether the addition of Saturday physiotherapy sessions (the “treatment”) to the usual Monday–Friday sessions (the “control”) would benefit patients undergoing rehabilitation in a hospital. One outcome measure was the destination of a patient upon being discharged, with the categories being home, low-level residential care (LLRC), high-level residential care (HLRC), or acute hospital transfer (AHT). The results are shown in Table 10.6.4.<sup>40</sup>

<b>Table 10.6.4</b> Discharge destination for physiotherapy patients			
		Group	
		Treatment	Control
Discharge destination	Home	107	103
	LLRC	10	15
	HLRC	6	1
	AHT	7	13
Total		130	132

A contingency-table chi-square test would be valid to compare treatment and control, but the test would lack power because it does not use the information contained in the *ordering* of the discharge destination categories (home is preferred to LLRC, which is preferred to HLRC, which is preferred to AHT). A related weakness of the chi-square test is that, even if  $H_0$  is rejected, the test does not yield a directional conclusion such as “the treatment leads to better discharge destinations than does the control.” ■

Methods are available to analyze contingency tables with rankable row and/or column variables; such methods, however, are beyond the scope of this book.

## Exercises 10.6.1–10.6.3

**10.6.1** Refer to the chemotherapy data of Exercise 10.2.10. Are the sample sizes large enough for the approximate validity of the chi-square test?

**10.6.2** In a study of prenatal influences on susceptibility to seizures in mice, pregnant females were randomly allocated to a control group or a “handled” group. Handled

mice were given sham injections three times during gestation, while control mice were not touched. The offspring were tested for their susceptibility to seizures induced by a loud noise. The investigators noted that the response varied considerably from litter to litter. The accompanying table summarizes the results.<sup>41</sup>

TREATMENT	NUMBER OF LITTERS	NUMBER OF MICE	RESPONSE TO LOUD NOISE		
			NO RESPONSE	WILD RUNNING	SEIZURE
Handled	19	104	23	10	71
Control	20	120	47	13	60

If these data are analyzed as a  $2 \times 3$  contingency table, the chi-square statistic is  $\chi_s^2 = 8.45$  and Table 9 gives  $0.01 < P\text{-value} < 0.02$ . Is this an appropriate analysis for this experiment? Explain. (*Hint:* Does the design meet the conditions for validity of the chi-square test?)

**10.6.3** In control of diabetes it is important to know how blood glucose levels change after eating various foods. Ten volunteers participated in a study to compare the effects of two foods—a sugar and a starch. A blood specimen was drawn before each volunteer consumed a measured amount of food; then additional blood specimens were drawn at 11 times during the next four hours. Each volunteer repeated the entire test on another occasion with the other food. Of particular concern were blood glucose levels that dropped below the initial level; the accompanying table shows the number of such values.<sup>42</sup>

FOOD	NO. OF VALUES LESS THAN INITIAL VALUE	TOTAL NUMBER OF OBSERVATIONS
Sugar	26	110
Starch	14	110

Suppose we analyze the given data as a contingency table. The test statistic would be

$$\begin{aligned}\chi_s^2 &= \frac{(26 - 20)^2}{20} + \frac{(14 - 20)^2}{20} \\ &\quad + \frac{(84 - 90)^2}{90} + \frac{(96 - 90)^2}{90} = 4.40\end{aligned}$$

At  $\alpha = 0.05$  we would reject  $H_0$  and find that there is sufficient evidence to conclude that blood glucose values below the initial value occur more often after ingestion of sugar than after ingestion of starch. This analysis contains two flaws. What are they? (*Hint:* Are the conditions for validity of the test satisfied?)

## 10.7 Confidence Interval for Difference between Probabilities

The chi-square test for a  $2 \times 2$  contingency table answers only a limited question: Do the estimated probabilities—call them  $\hat{p}_1$  and  $\hat{p}_2$ —differ enough to conclude that the true probabilities—call them  $p_1$  and  $p_2$ —are not equal? A complementary mode of analysis is to construct a confidence interval for the magnitude of the difference,  $(p_1 - p_2)$ .

When we discussed constructing a confidence interval for a single proportion,  $p$ , in Section 9.2, we defined an estimate  $\tilde{p}$ , based on the idea of “adding 2 successes and 2 failures to the data.” Making this adjustment to the data resulted in a confidence interval procedure that has good coverage properties. Likewise, when constructing a confidence interval for the difference in two proportions, we will define new estimates that are based on the idea of adding 1 observation to each cell of the table (so that a *total* of 2 successes and 2 failures are added to the data).

Consider a  $2 \times 2$  contingency table that can be viewed as a comparison of two samples, of sizes  $n_1$  and  $n_2$ , with respect to a dichotomous response variable. Let the  $2 \times 2$  table be given as

Sample 1	Sample 2
$y_1$	$y_2$
$n_1 - y_1$	$n_2 - y_2$
$n_1$	$n_2$

We define

$$\tilde{p}_1 = \frac{y_1 + 1}{n_1 + 2}$$

and

$$\tilde{p}_2 = \frac{y_2 + 1}{n_2 + 2}$$

We will use the difference in the new values,  $(\tilde{p}_1 - \tilde{p}_2)$ , to construct a confidence interval for  $(p_1 - p_2)$ . Like all quantities calculated from samples, the quantity  $(\tilde{P}_1 - \tilde{P}_2)$  is subject to sampling error. The magnitude of the sampling error can be expressed by the standard error of  $(\tilde{P}_1 - \tilde{P}_2)$ , which is calculated from the following formula:

$$\text{SE}(\tilde{P}_1 - \tilde{P}_2) = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}$$

Note that  $\text{SE}(\tilde{P}_1 - \tilde{P}_2)$  is analogous to  $\text{SE}(\bar{Y}_1 - \bar{Y}_2)$  as described in Section 6.6.

An approximate confidence interval can be based on  $\text{SE}(\tilde{P}_1 - \tilde{P}_2)$ ; for instance, a 95% confidence interval is

$$(\tilde{p}_1 - \tilde{p}_2) \pm (1.96) \text{SE}(\tilde{P}_1 - \tilde{P}_2)$$

Confidence intervals constructed this way have good coverage properties (i.e., approximately 95% of all 95% confidence intervals cover the true difference  $p_1 - p_2$ ) for almost any sample sizes  $n_1$  and  $n_2$ .<sup>43</sup> The following example illustrates the construction of the confidence interval.\*

**Example  
10.7.1**

**Migraine Headache** For the migraine headache data of Example 10.1.1, the sample sizes are  $n_1 = 49$  and  $n_2 = 26$ , and the estimated probabilities of substantial reduction in migraines are

$$\tilde{p}_1 = \frac{42}{51} = 0.824$$

$$\tilde{p}_2 = \frac{16}{28} = 0.571$$

The difference between these is

$$\begin{aligned}\tilde{p}_1 - \tilde{p}_2 &= 0.824 - 0.571 \\ &= 0.253 \\ &\approx 0.25\end{aligned}$$

Thus, we estimate that the real surgery increases the probability of substantial reduction in migraines by 0.25, compared to the sham surgery. To set confidence limits on this estimate, we calculate the standard error as

$$\begin{aligned}\text{SE}(\tilde{P}_1 - \tilde{P}_2) &= \sqrt{\frac{0.824(0.176)}{51} + \frac{0.571(0.429)}{28}} \\ &= 0.1077\end{aligned}$$

---

\*In Section 9.3 we presented a general version of the “add 2 successes and 2 failures” idea, in which the formula for  $\tilde{p}$  depends on the confidence level (95%, 90%, etc.). When constructing a confidence interval for a difference in two proportions, the coverage properties of the interval are best when 1 is added to each cell in the  $2 \times 2$  table, no matter what confidence level is being used.<sup>44</sup>

The 95% confidence interval is

$$\begin{aligned} & 0.253 \pm (1.96)(0.1077) \\ & 0.253 \pm 0.211 \\ & 0.042 < p_1 - p_2 < 0.464 \end{aligned}$$

We are 95% confident that the probability of substantial reduction in migraines is between 0.042 and 0.464 higher with the real surgery than with the sham surgery. ■

**Relationship to Test** The chi-square test for a  $2 \times 2$  contingency table (Section 10.2) is approximately, but not exactly, equivalent to checking whether a confidence interval for  $(p_1 - p_2)$  includes zero. [Recall from Section 7.3 that there is an exact equivalence between a *t* test and a confidence interval for  $(\mu_1 - \mu_2)$ .]

## Exercises 10.7.1–10.7.6

**10.7.1** Elderly patients who had suffered hip fractures were randomly assigned to receive either a placebo ( $n = 1,062$ ) or zolendronic acid ( $n = 1,065$ ) in a double-blind clinical trial. During the trial 139 placebo patients and 92 zolendronic acid patients had new fractures.<sup>45</sup> Let  $p_1$  and  $p_2$  represent the probabilities of fracture on placebo and zolendronic acid, respectively. Construct a 95% confidence interval for  $(p_1 - p_2)$ .

**10.7.2** Refer to the liver tumor data of Exercise 10.2.9.

- (a) Construct a 95% confidence interval for  $(\Pr\{\text{liver tumor} | \text{germ-free}\} - \Pr\{\text{liver tumor} | E. coli\})$ .
- (b) Interpret the confidence interval from part (a). That is, explain what the interval tells you about tumor probabilities.

**10.7.3** For women who are pregnant with twins, complete bed rest in late pregnancy is commonly prescribed in order to reduce the risk of premature delivery. To test the value of this practice, 212 women with twin pregnancies were randomly allocated to a bed-rest group or a control group. The accompanying table shows the incidence of preterm delivery (less than 37 weeks of gestation).<sup>46</sup>

	BED REST	CONTROLS
No. of preterm deliveries	32	20
No. of women	105	107

Construct a 95% confidence interval for  $(\Pr\{\text{preterm} | \text{bed rest}\} - \Pr\{\text{preterm} | \text{control}\})$ . Does the confidence interval suggest that bed rest is beneficial?

**10.7.4** Refer to Exercise 10.7.3. The numbers of infants with low birthweight (2,500 gm or less) born to the women are shown in the table.

	BED REST	CONTROLS
No. of low-birthweight babies	76	92
Total no. of babies	210	214

Let  $p_1$  and  $p_2$  represent the probabilities of a low-birthweight baby in the two conditions. Explain why the above information is not sufficient to construct a confidence interval for  $(p_1 - p_2)$ .

**10.7.5** Refer to the blood type data of Exercise 10.5.3. Let  $p_1$  and  $p_2$  represent the probabilities of type O blood in the patient population and the control population, respectively.

- (a) Construct a 95% confidence interval for  $(p_1 - p_2)$ .
- (b) Interpret the confidence interval from part (a). That is, explain what the interval tells you about the difference in probabilities of type O blood.

**10.7.6** In an experiment to treat patients with “generalized anxiety disorder,” the drug hydroxyzine was given to 71 patients and 30 of them improved. A group of 70 patients were given a placebo and 20 of them improved.<sup>47</sup> Let  $p_1$  and  $p_2$  represent the probabilities of improvement using hydroxyzine and the placebo, respectively. Construct a 95% confidence interval for  $(p_1 - p_2)$ .

## 10.8 Paired Data and $2 \times 2$ Tables (Optional)

In Chapter 8 we considered paired data when the response variable is continuous. In this section we consider the analysis of paired categorical data.

**Example  
10.8.1**

**HIV Transmission to Children** A study was conducted to determine a woman's risk of transmitting HIV to her unborn child. A sample of 114 HIV-infected women who gave birth to two children found that HIV infection occurred in 19 of the 114 older siblings and in 20 of the 114 younger siblings.<sup>48</sup> These data are shown in Table 10.8.1.

<b>Table 10.8.1</b> HIV infection data			
		Older sibling	Younger sibling
HIV?	Yes	19	20
	No	95	94
	Total	114	114

At first glance, it might appear that a regular chi-square test could be used to test the null hypothesis that the probability of HIV infection is the same for older siblings as for younger siblings. However, as we stated in Section 10.6, for the chi-square test to be valid the two samples—of 114 older siblings and of 114 younger siblings—must be independent of each other. In this case the samples are clearly dependent. Indeed, these are paired data, with a family generating the pair (older sibling, younger sibling).

Table 10.8.2 presents the data in a different format. This format helps focus attention on the relevant part of the data.\*

<b>Table 10.8.2</b> HIV infection data shown by pairs		
		Younger sibling HIV?
		Yes      No
Older sibling	Yes	2      17
HIV?	No	18      77

From Table 10.8.2 we can see that there are 79 pairs in which both siblings have the same HIV status: Two are “yes/yes” pairs and 77 are “no/no” pairs. These 79 pairs, which are called **concordant pairs**, do not help us determine whether HIV infection is more likely for younger siblings than for older siblings. The remaining 35 pairs—17 “yes/no” pairs and 18 “no/yes” pairs—do provide information on the relative likelihood of HIV infection for older and younger siblings. These pairs are called **discordant pairs**; we will focus on these 35 pairs in our analysis.

---

\*Note that Table 10.8.2 cannot be derived from Table 10.8.1.

If the chance of HIV infection is the same for older siblings as it is for younger siblings, then the two kinds of pairs—“yes/no” and “no/yes”—are equally likely. Thus, the null hypothesis

$$H_0: \text{the probability of HIV infection is the same for older siblings as it is for younger siblings}$$

is equivalent to

$$H_0: \text{among discordant pairs, } \Pr(\text{"yes/no"}) = \Pr(\text{"no/yes"}) = \frac{1}{2} \quad \blacksquare$$

## McNemar's Test

The hypothesis that discordant pairs are equally likely to be “yes/no” or “no/yes” can be tested with the chi-square goodness-of-fit test developed in Section 9.4. This application of the goodness-of-fit test is known as **McNemar's test** and has a particularly simple form.\* Let  $n_{11}$  denote the number of “yes/yes” pairs,  $n_{12}$  the number of “yes/no” pairs,  $n_{21}$  the number of “no/yes” pairs, and  $n_{22}$  the number of “no/no” pairs, as shown in Table 10.8.3. If  $H_0$  is true, the expected number of “yes/no” pairs is  $\frac{n_{12} + n_{21}}{2}$ , as is the expected number of “no/yes” pairs. Thus,

the test statistic is

$$\chi_s^2 = \frac{\left(n_{12} - \frac{(n_{12} + n_{21})}{2}\right)^2}{\frac{(n_{12} + n_{21})}{2}} + \frac{\left(n_{21} - \frac{(n_{12} + n_{21})}{2}\right)^2}{\frac{(n_{12} + n_{21})}{2}}$$

which simplifies to

$$\chi_s^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

The distribution of  $\chi_s^2$  under the null hypothesis is approximately a  $\chi^2$  distribution with 1 degree of freedom.

**Table 10.8.3** A general table of paired proportion data

	Yes	No
Yes	$n_{11}$	$n_{12}$
No	$n_{21}$	$n_{22}$

\*The null hypothesis tested by McNemar's test can also be tested by using the binomial distribution. The null hypothesis states that among discordant pairs,  $\Pr(\text{"yes/no"}) = \Pr(\text{"no/yes"}) = 1/2$ . Thus, under the null hypothesis, the number of “yes/no” pairs has a binomial distribution with  $n$  = the number of discordant pairs and  $p = 0.5$ .

**Example  
10.8.2**

**HIV Transmission to Children** For the data given in Example 10.8.1,  $n_{12} = 17$  and  $n_{21} = 18$ . Thus,

$$\chi_s^2 = \frac{(17 - 18)^2}{17 + 18} = 0.0286$$

From Table 9 we see that the  $P$ -value is greater than 0.20. (Using a computer gives  $P$ -value = 0.87.) The data are very much consistent with the null hypothesis that the probability of HIV infection is the same for older siblings as it is for younger siblings. ■

### Exercises 10.8.1–10.8.4

**10.8.1** As part of a study of risk factors for stroke, 155 women who had experienced a hemorrhagic stroke (cases) were interviewed. For each case, a control was chosen who had not experienced a stroke; the control was matched to the case by neighborhood of residence, age, and race. Each woman was asked whether she used oral contraceptives. The data for the 155 pairs are displayed in the table. “Yes” and “No” refer to use of oral contraceptives.<sup>49</sup>

		CASE	
		NO	YES
Control	No	107	30
	Yes	13	5

To test for association between oral contraceptive use and stroke, consider only the 43 discordant pairs (pairs who answered differently) and test the hypothesis that a discordant pair is equally likely to be “yes/no” or “no/yes.” Use McNemar’s test to test the hypothesis that having a stroke is independent of use of oral contraceptives against a nondirectional alternative at  $\alpha = 0.05$ .

**10.8.2** Example 10.8.1 referred to a sample of HIV-infected women who gave birth to two children. One of the outcomes that was studied was whether the gestational age of the child was less than 38 weeks; this information was recorded for 106 of the families. The data for this variable are shown in the following table. Analyze these data using McNemar’s test. Use a nondirectional alternative and let  $\alpha = 0.10$ .

		YOUNGER SIBLING < 38 WEEKS?	
		YES	NO
Older sibling	Yes	26	5
	< 38 weeks?	21	54

**10.8.3** A study of 85 patients with Hodgkin’s disease found that 41 had had their tonsils removed. Each patient was matched with a sibling of the same sex. Only 33 of the siblings had undergone tonsillectomy. The data are shown in the following table.<sup>50</sup> Use McNemar’s test to test the hypothesis that “yes/no” and “no/yes” pairs are equally likely. Previous research had suggested that having a tonsillectomy is associated with an increased risk of Hodgkin’s disease; thus, use a directional alternative. Let  $\alpha = 0.05$ .

		SIBLING TONSILLECTOMY?	
		YES	NO
Hodgkin’s patient	Yes	26	15
	No	7	37

**10.8.4** In a study of the mating behavior of *Gryllus campestris*, pairs of female crickets were placed in a plexiglass arena with a single male cricket. There were 54 cases in which the females fought; these resulted in 42 cases in which the winning female copulated with the male, 8 cases in which the losing female copulated with the male, and 4 cases that ended with no copulation. The data are summarized in the following table.<sup>51</sup> Use McNemar’s test to test the hypothesis that winners and losers are equally likely to copulate. Use an appropriate directional alternative and let  $\alpha = 0.05$ .

		WINNERS	
		YES	NO
Losers	COPULATE	0	8
	No	42	4

## 10.9 Relative Risk and the Odds Ratio (Optional)

It is quite common to test the null hypothesis that two population proportions,  $p_1$  and  $p_2$ , are equal. A chi-square test, based on a  $2 \times 2$  table, is often used for this purpose. A confidence interval for  $(p_1 - p_2)$  provides information about the magnitude of the difference between  $p_1$  and  $p_2$ . In this section we consider two other measures of dependence: the relative risk and the odds ratio.

### Relative Risk

Sometimes researchers prefer to compare probabilities in terms of their *ratio*, rather than their difference. When the outcome event is deleterious (such as having a heart attack or getting cancer) the ratio of probabilities is called the **relative risk**, or the risk ratio. The relative risk is defined as  $p_1/p_2$ . This measure is widely used in studies of human health. The following is an example.

**Example  
10.9.1**

**Smoking and Lung Cancer** The health histories of 11,900 middle-aged men were tracked over many years. During the study 126 of the men developed lung cancer, including 89 men who were smokers and 37 men who were former smokers. Table 10.9.1 shows the data.<sup>52</sup>

		Smoking history	
		Smoker	Former smoker
Lung cancer?	Yes	89	37
	No	6,063	5,711
	Total	6,152	5,748

The probabilities of primary interest are the columnwise conditional probabilities:

$$p_1 = \Pr\{\text{lung cancer} \mid \text{smoker}\}$$

$$p_2 = \Pr\{\text{lung cancer} \mid \text{former smoker}\}$$

The estimates of these from the data are

$$\hat{p}_1 = \frac{89}{6152} = 0.01447 \approx 0.014$$

$$\hat{p}_2 = \frac{37}{5748} = 0.00644 \approx 0.006$$

The estimated relative risk is

$$\frac{\hat{p}_1}{\hat{p}_2} = \frac{0.01447}{0.00644} = 2.247 \approx 2.2$$

Thus, we estimate that the risk (i.e., the conditional probability) of developing lung cancer is about 2.2 times as great for smokers as for former smokers. (Of course, because this is an observational study, we would not be justified in concluding that smoking *causes* lung cancer.) ■

## The Odds Ratio

Another way to compare two probabilities is in terms of **odds**. The odds of an event  $E$  is defined to be the ratio of the probability that  $E$  occurs to the probability that  $E$  does not occur:

$$\text{odds of } E = \frac{\Pr\{E\}}{1 - \Pr\{E\}}$$

For instance, if the probability of an event is  $1/4$ , then the odds of the event are  $\frac{1/4}{3/4} = 1/3$  or 1:3. As another example, if the probability of an event is  $1/2$ , then the odds of the event are  $\frac{1/2}{1/2} = 1$  or 1:1.

The **odds ratio** is simply the ratio of odds under two conditions. Specifically, suppose that  $p_1$  and  $p_2$  are the conditional probabilities of an event under two different conditions. Then the odds ratio, which we will denote by  $\theta$  ("theta"), is defined as follows:

$$\theta = \frac{\frac{p_1}{1 - p_1}}{\frac{p_2}{1 - p_2}}$$

If the estimated probabilities  $\hat{p}_1$  and  $\hat{p}_2$  are calculated from a  $2 \times 2$  contingency table, the corresponding estimated odds ratio, denoted  $\hat{\theta}$ , is calculated as

$$\hat{\theta} = \frac{\frac{\hat{p}_1}{1 - \hat{p}_1}}{\frac{\hat{p}_2}{1 - \hat{p}_2}}$$

We illustrate with an example.

---

**Example 10.9.2**

**Smoking and Lung Cancer** From the data of Example 10.9.1, we estimate the odds of developing lung cancer as follows:

$$\hat{\text{odds}} = \frac{0.01447}{1 - 0.01447} = 0.01468 \text{ among smokers}$$

$$\hat{\text{odds}} = \frac{0.00644}{1 - 0.00644} = 0.00648 \text{ among former smokers}$$

The estimated odds ratio is

$$\hat{\theta} = \frac{0.01468}{0.00648} = 2.265 \approx 2.3$$

Thus, we estimate that the odds of developing lung cancer are about 2.3 times as great for smokers as for former smokers. ■

## Odds Ratio and Relative Risk

The odds ratio measures association in an unfamiliar way; the relative risk is a more natural measure. Fortunately, in many applications the two measures are approximately equal. In general the relationship between the odds ratio and the relative risk is given by

$$\text{odds ratio} = \text{relative risk} \times \frac{1 - p_2}{1 - p_1}$$

Notice that if  $p_1$  and  $p_2$  are small, then the relative risk is approximately equal to the odds ratio. We illustrate with the smoking and lung cancer data.

---

**Example  
10.9.3**

**Smoking and Lung Cancer** For the data in Table 10.9.1 we found that the estimated relative risk of lung cancer is

$$\text{estimated relative risk} = 2.247$$

and the estimated odds ratio is

$$\hat{\theta} = 2.265$$

These are approximately equal because the outcome of interest (developing lung cancer) is rare, so that  $\hat{p}_1$  and  $\hat{p}_2$  are small. ■

## Advantage of the Odds Ratio

Both the relative risk  $p_1/p_2$  and the difference  $(p_1 - p_2)$  are easier to interpret than the odds ratio. Why, then, is the odds ratio used at all? One important advantage of the odds ratio is that, in certain kinds of studies, the odds ratio can be estimated even though  $p_1$  and  $p_2$  *cannot* be estimated. To explain this property, we must first discuss the question of estimability of conditional probabilities in contingency tables.

In a  $2 \times 2$  contingency table, the conditional probabilities can be defined by rows or by columns. Whether these probabilities can be estimated from the observed data depends on the study design. The following example illustrates this point.

---

**Example  
10.9.4**

**Smoking and Lung Cancer** In studying the relationship between smoking and lung cancer, the conditional probabilities of primary interest are

$$p_1 = \Pr\{\text{lung cancer} \mid \text{smoker}\}$$

and

$$p_2 = \Pr\{\text{lung cancer} \mid \text{former smoker}\}$$

These are columnwise probabilities in a table like Table 10.9.1. One could, however, also consider the following rowwise conditional probabilities:

$$p_1^* = \Pr\{\text{smoker} \mid \text{lung cancer}\}$$

and

$$p_2^* = \Pr\{\text{smoker} \mid \text{no lung cancer}\}$$

(Of course,  $p_1^*$  and  $p_2^*$  are not particularly meaningful biologically.) From the study described in Example 10.9.1, that is, a single sample of size  $n = 11,900$  observed with respect to smoking status and lung cancer, one can estimate not only  $p_1$  and  $p_2$  but also  $p_1^*$  and  $p_2^*$ . However, there are other important study designs that do not provide enough information to estimate all these conditional probabilities. For example, suppose that a study is conducted by choosing a group of 500 smokers and a group of 500 former smokers and then observing how many of them develop lung cancer. This kind of study is called a prospective study

or **cohort study**. Such a study might produce the fictitious but realistic data of Table 10.9.2.

**Table 10.9.2** Fictitious data for cohort study of smoking and lung cancer mortality

		Smoking history	
		Smoker	Former smoker
Lung cancer?	Yes	7	3
	No	473	497
Total		500	500

The data of Table 10.9.2 can be viewed as two independent samples. From the data we can estimate the conditional probabilities of lung cancer in the two populations (smokers and former smokers):

$$\hat{p}_1 = \frac{7}{500} = 0.014 \quad \hat{p}_2 = \frac{3}{500} = 0.006$$

By contrast, the rowwise probabilities  $p_1^*$  and  $p_2^*$  cannot be estimated from Table 10.9.2. Because the relative numbers of smokers and former smokers were predetermined by the design of the study ( $n_1 = 500$  and  $n_2 = 500$ ), the data contain no information about the prevalence of smoking, and therefore no information about the population values of

$$\Pr\{\text{smoker} \mid \text{lung cancer}\} \text{ and } \Pr\{\text{smoker} \mid \text{no lung cancer}\}$$

Table 10.9.2 was generated by fixing the column totals and observing the row variable. Consider now the reverse sort of design. Suppose we choose 500 men who died from lung cancer and 500 men who did not die from lung cancer and we then determine the smoking histories of the men. This design is called a **case-control design**. Such a design might generate the fictitious but realistic data of Table 10.9.3.

**Table 10.9.3** Fictitious data for cohort study of smoking and lung cancer mortality

		Smoking history		Total
		Smoker	Former smoker	
Lung cancer?	Yes	273	227	500
	No	173	327	500

From Table 10.9.3 we can estimate the rowwise conditional probabilities

$$\hat{p}_1^* = \frac{273}{500} = 0.546 \approx 0.55$$

$$\hat{p}_2^* = \frac{173}{500} = 0.346 \approx 0.35$$

However, from the data in Table 10.9.3 we cannot estimate the columnwise conditional probabilities  $p_1$  and  $p_2$ : Because the row totals were predetermined by design, the data contain no information about  $\Pr\{\text{lung cancer} \mid \text{smoker}\}$  and  $\Pr\{\text{lung cancer} \mid \text{former smoker}\}$ .

The preceding example shows that, depending on the design, a study may not permit estimation of both columnwise probabilities  $p_1$  and  $p_2$  and rowwise probabilities  $p_1^*$  and  $p_2^*$ . Fortunately, the odds ratio is the same whether it is determined columnwise or rowwise. Specifically,

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{\frac{p_1^*}{1-p_1^*}}{\frac{p_2^*}{1-p_2^*}}$$

Because of this relationship, the odds ratio  $\theta$  can be estimated by estimating  $p_1$  and  $p_2$  or by estimating  $p_1^*$  and  $p_2^*$ . This fact has important applications, especially for case-control studies, as illustrated by the following example.

**Example  
10.9.5**

**Smoking and Lung Cancer** To characterize the relationship between smoking and lung cancer mortality, the columnwise probabilities  $p_1$  and  $p_2$  are more biologically meaningful than the rowwise probabilities  $p_1^*$  and  $p_2^*$ . If we investigate the relationship using a case-control design, neither  $p_1$  nor  $p_2$  can be estimated from the data. (See Example 10.9.4.) However, the odds ratio *can* be estimated from the data. For instance, from Table 10.9.3 we obtain

$$\begin{aligned}\hat{\theta} &= \frac{\frac{\hat{p}_1^*}{1-\hat{p}_1^*}}{\frac{\hat{p}_2^*}{1-\hat{p}_2^*}} \\ &= \frac{\frac{0.546}{1-0.546}}{\frac{0.346}{1-0.346}} = 2.265 \approx 2.27\end{aligned}$$

We can interpret this odds ratio as follows: We know that the outcome event—developing lung cancer—is rare, and so we know that the odds ratio is approximately equal to the relative risk,  $p_1/p_2$ . We therefore estimate that the risk of lung cancer is about 2.3 times as great for smokers as for former smokers. ■

There is an easier way to compute the odds ratio for a  $2 \times 2$  contingency table. For a general  $2 \times 2$  table, let  $n_{11}$  denote the number of observations in the first row and the first column. Likewise, let  $n_{12}$  be the number of observations in the first row and second column, and so on. The general  $2 \times 2$  table then has the form

$n_{11}$	$n_{12}$
$n_{21}$	$n_{22}$

The estimated odds ratio from the table is

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

**Example  
10.9.6**

**Smoking and Lung Cancer** From the data in Table 10.9.1, we can calculate the estimated odds ratio as

$$\hat{\theta} = \frac{89 \times 5,711}{37 \times 6,063} = 2.265 \approx 2.27$$

The case-control design is often the most efficient design for investigating rare outcome events, such as rare diseases. Although Table 10.9.3 was constructed assuming that the two samples, cases and controls, were chosen independently, a more common design is to incorporate matching of cases and controls with respect to potential confounding factors (for example, age). As we have seen, by taking advantage of the odds ratio, one can estimate the relative risk from a case-control study of a rare event even though one cannot estimate the risks  $p_1$  and  $p_2$  separately.

If the odds ratio (or the relative risk) is equal to 1.0, then the odds (or the risk) are the same for both of the groups being compared. In the smoking and lung cancer data of Table 10.9.1 the calculated odds ratio was *greater* than 1.0, indicating that the odds of lung cancer are greater for smokers than for former smokers. Notice that we could have focused attention on the odds of not getting lung cancer. In this case, the odds ratio would be *less* than 1.0, as shown in Example 10.9.7.

**Example  
10.9.7**

**Smoking and Lung Cancer** Suppose we rearrange the data in Table 10.9.1 by putting lung cancer in the second row and not getting lung cancer in the first row:

		Smoking History	
		Smoker	Former Smoker
Lung cancer?	No	6,063	5,711
	Yes	89	37
	Total	6,152	5,748

In this case the odds ratio is the odds of not getting lung cancer for a smoker divided by the odds of not getting lung cancer for a former smoker. We can calculate the estimated odds ratio as

$$\hat{\theta} = \frac{6,063 \times 37}{5,711 \times 89} = 0.44$$

This is the reciprocal of the odds ratio calculated in Example 10.9.6:  $\frac{1}{2.27} = 0.44$ .

The fact that the odds ratio is less than 1.0 means that the event (being free of lung cancer) is less likely for smokers than for former smokers.

## Confidence Interval for the Odds Ratio

In Chapter 6 we discussed confidence intervals for proportions, which are of the form  $\tilde{p} \pm z_{\alpha/2} \text{SE}_{\tilde{p}}$ , where  $\tilde{p} = \frac{y+2}{n+4}$ . In particular, a 95% confidence interval for  $p$  is given by  $\tilde{p} \pm z_{0.025} \text{SE}_{\tilde{p}}$ . Such confidence intervals are based on the fact that for large samples the sampling distribution of  $\tilde{p}$  is approximately normal (according to the Central Limit Theorem).

In a similar way, we can construct a confidence interval for an odds ratio. One problem is that the sampling distribution of  $\hat{\theta}$  is not normal. However, if we take the natural logarithm of  $\hat{\theta}$ , then we have a distribution that is approximately normal. Hence, we construct a confidence interval for  $\theta$  by first finding a confidence interval for  $\ln(\theta)$  and then transforming the endpoints back to the original scale.

In order to construct a confidence interval for  $\ln(\theta)$ , we need the standard error of  $\ln(\hat{\theta})$ . The formula for the standard error of  $\ln(\hat{\theta})$  is given in the box.

### Standard Error of $\ln(\hat{\theta})$

$$\text{SE}_{\ln(\hat{\theta})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

A 95% confidence interval for  $\ln(\theta)$  is given by  $\ln(\hat{\theta}) \pm (1.96)\text{SE}_{\ln(\hat{\theta})}$ . We then exponentiate the two endpoints of the interval to get a 95% confidence interval for  $\theta$ . Intervals with other confidence coefficients are constructed analogously; for instance, for a 90% confidence interval one would use  $z_{0.05}$  (1.645) instead of  $z_{0.025}$  (1.960). The process for finding a confidence interval for  $\theta$  is summarized in the following box.\*

### Confidence Interval for $\theta$

To construct a 95% confidence interval for  $\theta$ ,

1. Calculate  $\ln(\hat{\theta})$ .
2. Construct a confidence interval for  $\ln(\theta)$  using the formula  $\ln(\hat{\theta}) \pm (1.96)\text{SE}_{\ln(\hat{\theta})}$ .
3. Exponentiate the endpoints to get a confidence interval for  $\theta$ .

This process is illustrated in the following examples.

#### Example 10.9.8

**Smoking and Lung Cancer** From the data in Table 10.9.1, the estimated odds ratio is

$$\hat{\theta} = \frac{89 \times 5711}{37 \times 6063} = 2.27$$

Thus,  $\ln(\hat{\theta}) = \ln(2.27) = 0.820$ .

The standard error is given by  $\text{SE}_{\ln(\hat{\theta})} = \sqrt{\frac{1}{89} + \frac{1}{37} + \frac{1}{6063} + \frac{1}{5711}} = 0.1965$ .

A 95% confidence interval for  $\ln(\theta)$  is  $0.820 \pm (1.96)(0.1965)$  or  $0.820 \pm 0.385$ . This interval is  $(0.435, 1.205)$ .

To get a 95% confidence interval for  $\theta$ , we evaluate  $e^{0.435} = 1.54$  and  $e^{1.205} = 3.24$ . Thus, we are 95% confident that the population value of the odds ratio is between 1.54 and 3.24. ■

---

\*A confidence interval for the relative risk can be found in a suitably modified manner for those situations in which the relative risk can be estimated from the data.

**Example  
10.9.9**

**Heart Attacks and Aspirin** During the Physician's Health Study, 11,037 physicians were randomly assigned to take 325 mg of aspirin every other day; 104 of them had heart attacks during the study. Another 11,034 physicians were randomly assigned to take a placebo; 189 of them had heart attacks. These data are shown in Table 10.9.4.<sup>53</sup> The odds ratio for comparing the heart attack rate on aspirin to the heart attack rate on placebo is

$$\hat{\theta} = \frac{189 \times 10,933}{104 \times 10,845} = 1.832$$

Thus,  $\ln(\hat{\theta}) = \ln(1.832) = 0.605$ .

The standard error is

$$SE_{\ln(\hat{\theta})} = \sqrt{\frac{1}{189} + \frac{1}{104} + \frac{1}{10,845} + \frac{1}{10,933}} = 0.123.$$

A 95% confidence interval for  $\ln(\theta)$  is  $0.605 \pm (1.96)(0.123)$  or  $0.605 \pm 0.241$ . This interval is (0.364, 0.846).

To get a 95% confidence interval for  $\theta$ , we evaluate  $e^{0.364} = 1.44$  and  $e^{0.846} = 2.33$ . Thus, we are 95% confident that the population value of the odds ratio is between 1.44 and 2.33. Because heart attacks are relatively rare in this data set, the relative risk is nearly equal to the odds ratio. Thus, we can say that we are 95% confident that the probability of a heart attack is about 1.44 to 2.33 times greater when taking the placebo than when taking aspirin. ■

**Table 10.9.4** Heart attacks on placebo and on aspirin

	Placebo	Aspirin
Heart attack	189	104
No heart attack	10,845	10,933
Total	11,034	11,037

## Exercises 10.9.1–10.9.8

- 10.9.1** For each of the following tables, calculate (i) the relative risk and (ii) the odds ratio.

(a)

25	23
492	614

(b)

12	8
93	84

- 10.9.2** For each of the following tables, calculate (i) the relative risk and (ii) the odds ratio.

(a)

14	16
322	412

(b)

15	7
338	82

- 10.9.3** Hip dysplasia is a hip socket abnormality that affects many large breed dogs. A review of medical records of dogs seen at 27 veterinary medical teaching hospitals found that hip dysplasia was more common in Golden Retrievers than in Border Collies; the data are shown in the following table.<sup>54</sup> Calculate the relative risk of hip dysplasia for Golden Retrievers compared to Border Collies.

	GOLDEN RETRIEVER	BORDER COLLIE
Hip dysplasia? Yes	3,995	221
Hip dysplasia? No	42,946	5,007
Total	46,941	5,228

**10.9.4** Consider the data from Exercise 10.9.3.

- Calculate the sample value of the odds ratio.
- Construct a 95% confidence interval for the population value of the odds ratio.
- Interpret the confidence interval from part (b) in the context of this setting.

**10.9.5** As part of the National Health Interview Survey, occupational injury data were collected on thousands of American workers. The table below summarizes part of these data.<sup>55</sup>

		SELF-EMPLOYED	EMPLOYED BY OTHERS
Injured?	Yes	210	4,391
	No	33,724	421,502
Total		33,934	425,893

- Calculate the sample value of the odds ratio.
- According to the odds ratio, are self-employed workers more likely, or less likely, to be injured than persons who work for others?
- Construct a 95% confidence interval for the population value of the odds ratio.
- Interpret the confidence interval from part (b) in the context of this setting.

**10.9.6** Many over-the-counter decongestants and appetite suppressants contain the ingredient phenylpropanolamine. A study was conducted to investigate whether this ingredient is associated with strokes. The study found that 6 of 702 stroke victims had used an appetite suppressant containing phenylpropanolamine, compared to only 1 of 1,376 subjects in a control group. The following table summarizes these data.<sup>56</sup>

		STROKE	NO STROKE
Appetite suppressant?	Yes	6	1
	No	696	1,375
Total		702	1,376

- Calculate the sample value of the odds ratio.

- Construct a 95% confidence interval for the population value of the odds ratio.

- Upon hearing of these data, some scientists called the study “inconclusive” because the numbers of users of appetite suppressants containing phenylpropanolamine (7 total: 6 in one group and 1 in the other) are so small. What is your response to these scientists?

**10.9.7** Two treatments, heparin and enoxaparin, were compared in a double-blind, randomized clinical trial of patients with coronary artery disease. The subjects can be classified as having a positive or negative response to treatment; the data are given in the following table.<sup>57</sup>

		HEPARIN	ENOXAPARIN
Outcome	Negative	309	266
	Positive	1,255	1,341
Total		1,564	1,607

- Calculate the sample value of the odds ratio.

- Construct a 95% confidence interval for the population value of the odds ratio.

- Interpret the confidence interval from part (b) in the context of this setting.

**10.9.8** Consider the data from Exercise 10.7.1. Given that there were 139 hip fractures in 1,062 placebo patients and 92 hip fractures in 1,065 zoledronic acid patients, construct a 95% confidence interval for the population value of the odds ratio.<sup>45</sup>

## 10.10 Summary of Chi-Square Test

The chi-square test is often applied to contingency tables; it is summarized here.

### Summary of Chi-Square Test for a Contingency Table

*Null hypothesis:*

$H_0$ : Row variable and column variable are independent

*Calculation of expected frequencies:*

$$e_i = \frac{(Row\ Total) \times (Column\ Total)}{Grand\ Total}$$

*Test statistic:*

$$\chi_s^2 = \sum_{\text{all cells}} \frac{(o_i - e_i)^2}{e_i}$$

*Null distribution (approximate):*

$$\chi^2 \text{ distribution with } df = (r - 1)(k - 1)$$

where  $r$  is the number of rows and  $k$  is the number of columns in the contingency table. This approximation is adequate if  $e_i \geq 5$  for every cell. If  $r$  and  $k$  are large, the condition that  $e_i \geq 5$  is less critical and the  $\chi^2$  approximation is adequate if the average expected frequency is at least 5, even if some of the cell counts are smaller.

The observations must be independent of one another. If paired data are collected for a  $2 \times 2$  table, then McNemar's test is appropriate (Section 10.8).

## Supplementary Exercises 10.S.1–10.S.19

(Note: Exercises preceded by an asterisk refer to optional sections.)

**10.S.1** In the Women's Health Initiative Dietary Modification Trial women were randomly assigned to an intervention or a control group. The intervention included counseling sessions designed to reduce fat intake and to increase consumption of fruits and vegetables. Over six years data were collected on coronary heart disease (CHD); results are shown in the table.<sup>58</sup> Do the data provide evidence that the intervention makes a difference? The value of the chi-square statistic for this contingency table is  $\chi_s^2 = 0.69$ . Carry out the chi-square test against a nondirectional alternative at  $\alpha = 0.10$ .

		GROUP		
		INTERVENTION	CONTROL	
CHD?	Yes	1,000	1,549	2,549
	No	18,541	27,745	46,286
Total	19,541	29,294	48,835	

**10.S.2** Use the data from Exercise 10.S.1 to construct a 95% confidence interval for  $(\Pr\{\text{CHD} | \text{intervention}\} - \Pr\{\text{CHD} | \text{control}\})$ .

**10.S.3** As part of a study of environmental influences on sex determination in the fish *Menidia*, eggs from a single mating were divided into two groups and raised in either

a warm or a cold environment. It was found that 73 of 141 offspring in the warm environment and 107 of 169 offspring in the cold environment were females.<sup>59</sup> In each of the following chi-square tests, use a nondirectional alternative and let  $\alpha = 0.05$ .

- Test the hypothesis that the population sex ratio is 1:1 in the warm environment.
- Test the hypothesis that the population sex ratio is 1:1 in the cold environment.
- Test the hypothesis that the population sex ratio is the same in the warm as in the cold environment.
- Define the population to which the conclusions reached in parts (a)–(c) apply. (Is it the entire genus *Menidia*?)

**10.S.4** The cilia are hairlike structures that line the nose and help to protect the respiratory tract from dust and foreign particles. A medical team obtained specimens of nasal tissue from nursery school children who had viral upper respiratory infections, and also from healthy children in the same classroom. The tissue was sectioned and the cilia were examined with a microscope for specific defects, with the results shown in the accompanying table.<sup>60</sup> The data show that the percentage of defective cilia was much higher in the tissue from infected children (15.7% versus 3.1%). Would it be valid to apply a chi-square test to compare these percentages? If so, do it. If not, explain why not.

		CILIA WITH DEFECTS			
		NUMBER OF CHILDREN	TOTAL NUMBER OF CILIA COUNTED	NUMBER	PERCENT
	Control	7	556	17	3.1
	Respiratory infection	22	1,493	235	15.7

**10.S.5** A group of mountain climbers participated in a trial to investigate the usefulness of the drug acetazolamide in preventing altitude sickness. The climbers were randomly assigned to receive either drug or placebo during an ascent of Mt. Rainier. The experiment was supposed to be double-blind, but the question arose whether some of the climbers might have received clues (perhaps from the presence or absence of side effects or from a perceived therapeutic effect or lack of it) as to which treatment they were receiving. To investigate this possibility, the climbers were asked (after the trial was over) to guess which treatment they had received.<sup>61</sup> The results can be cast in the following contingency table, for which  $\chi^2_s = 5.07$ :

		TREATMENT RECEIVED	
		DRUG	PLACEBO
Guess	Correct	20	12
	Incorrect	11	21

Alternatively the same results can be rearranged in the following contingency table, for which  $\chi^2_s = 0.01$ :

		TREATMENT RECEIVED	
		DRUG	PLACEBO
Guess	Drug	20	21
	Placebo	11	12

Consider the null hypothesis

$H_0$ : The blinding was perfect (the climbers received no clues).

Carry out the chi-square test of  $H_0$  against the alternative that the climbers did receive clues. Let  $\alpha = 0.05$ . (You must decide which contingency table is relevant to this question.) (*Hint:* To clarify the issue for yourself, try inventing a fictitious data set in which most of the climbers *have* received strong clues, so that we would expect a large value of  $\chi^2_s$ ; then arrange your fictitious data in each of the two contingency table formats and note which table would yield a larger value of  $\chi^2_s$ .)

**\*10.S.6** Desert lizards (*Dipsosaurus dorsalis*) regulate their body temperature by basking in the sun or moving into the shade, as required. Normally the lizards will maintain a daytime temperature of about  $38^\circ\text{C}$ . When they are sick, however, they maintain a temperature about  $2^\circ$  to  $4^\circ$  higher—that is, a “fever.” In an experiment to see whether this fever might be beneficial, lizards were given a bacterial infection; then 36 of the animals were prevented from developing a fever by keeping them in a  $38^\circ$  enclosure, while 12 animals were kept

at a temperature of  $40^\circ$ . The following table describes the mortality after 24 hours.<sup>62</sup> How strongly do these results support the hypothesis that fever has survival value? Use Fisher’s exact test against a directional alternative. Let  $\alpha = 0.05$ .

	38°	40°
Died	18	2
Survived	18	10
Total	36	12

**10.S.7** Consider the data from Exercise 10.S.6. Analyze these data with a chi-square test. Let  $\alpha = 0.05$ .

**10.S.8** In a randomized clinical trial, 154 women with breast cancer were assigned to receive chemotherapy. Another 164 women were assigned to receive chemotherapy combined with radiation therapy. Survival data after 15 years are given in the following table.<sup>63</sup> Use these data to conduct a test of the null hypothesis that type of treatment does not affect survival rate. Let  $\alpha = 0.05$ .

	CHEMOTHERAPY ONLY	CHEMOTHERAPY AND RADIATION THERAPY
Died	78	66
Survived	76	98
Total	154	164

**\*10.S.9** Refer to the data in Exercise 10.S.8.

- Calculate the sample odds ratio.
- Find a 95% confidence interval for the population value of the odds ratio.

**10.S.10** Two drugs, zidovudine and didanosine, were tested for their effectiveness in preventing progression of HIV disease in children. In a double-blind clinical trial, 276 children with HIV were given zidovudine, 281 were given didanosine, and 274 were given zidovudine plus didanosine. The following table shows the survival data for the three groups.<sup>64</sup> Use these data to conduct a test of the null hypothesis that survival and treatment are independent. Let  $\alpha = 0.10$ .

	ZIDOVUDINE	DIDANOSINE	ZIDOVUDINE AND DIDANOSINE
Died	17	7	10
Survived	259	274	264
Total	276	281	274

**10.S.11** The blood types of malaria patients at a clinic in India were compared with those obtained in a sample of visitors to a nearby hospital. The data are shown in the following table.<sup>65</sup> Use these data to conduct a test of the null hypothesis that blood type is independent of contracting malaria. Let  $\alpha = 0.05$ .

	A	B	O	AB	TOTAL
Malaria cases	138	199	106	33	476
Controls	229	535	428	96	1,300

**10.S.12** The habitat selection behavior of the fruitfly *Drosophila subobscura* was studied by capturing flies from two different habitat sites. The flies were marked with colored fluorescent dust to indicate the site of capture and then released at a point midway between the original sites. On the following two days, flies were recaptured at the two sites. The results are summarized in the table.<sup>66</sup> The value of the chi-square statistic for this contingency table is  $\chi^2_s = 10.44$ . Test the null hypothesis of independence against the alternative that the flies preferentially tend to return to their site of capture. Let  $\alpha = 0.01$ .

		SITE OF RECAPTURE		
		I	II	
Site of original capture	I	78	56	
	II	33	58	

**10.S.13** In the garden pea *Pisum sativum*, seed color can be yellow (Y) or green (G), and seed shape can be round (R) or wrinkled (W). Consider the following three hypotheses describing a population of plants:

$$H_0^{(1)}: \Pr\{Y\} = \frac{3}{4}$$

$$H_0^{(2)}: \Pr\{R\} = \frac{3}{4}$$

$$H_0^{(3)}: \Pr\{R|Y\} = \Pr\{R|G\}$$

The first hypothesis asserts that yellow and green plants occur in a 3:1 ratio; the second hypothesis asserts that round and wrinkled plants occur in a 3:1 ratio, and the third hypothesis asserts that color and shape are independent. (In fact, for a population of plants produced by a certain cross—the dihybrid cross—all three hypotheses are known to be true.)

Suppose a random sample of 1,600 plants is to be observed, with the data to be arranged in the following contingency table:

		COLOR		
		y	G	
Shape	R			
	W			
				1,600

Invent fictitious data sets as specified, and verify each answer by calculating the estimated conditional probabilities. (*Hint:* In each case, begin with the marginal frequencies.)

- (a) A data set that agrees perfectly with  $H_0^{(1)}$ ,  $H_0^{(2)}$ , and  $H_0^{(3)}$ .
- (b) A data set that agrees perfectly with  $H_0^{(1)}$  and  $H_0^{(2)}$  but not with  $H_0^{(3)}$ .
- (c) A data set that agrees perfectly with  $H_0^{(3)}$  but not with  $H_0^{(1)}$  or  $H_0^{(2)}$ .

**\*10.S.14** A study of 36,080 persons who had heart attacks found that men were more likely to survive than were women. The following table shows some of the data collected in the study.<sup>67</sup>

		MEN	WOMEN
Survived at least 24 hours?	Yes	25,339	8,914
	No	1,141	686
Total		26,480	9,600

- (a) Calculate the odds ratio for comparing survival of men to survival of women.
- (b) Calculate a 95% confidence interval for the population value of the odds ratio.
- (c) Does the odds ratio give a good approximation to the relative risk for these data? Why or why not?

**\*10.S.15** In the study described in Exercise 10.9.6, one of the variables measured was whether the subjects had used any products containing phenypropanolamine. The odds ratio was calculated to be 1.49, with stroke victims more likely than the control subjects to have used a product containing phenypropanolamine.<sup>56</sup> A 95% confidence interval for the population value of the odds ratio is (0.84, 2.64). Interpret this confidence interval in the context of this setting.

**10.S.16 (Computer exercise)** In a study of the effects of smoking cigarettes during pregnancy, researchers examined the placenta from each of 58 women after childbirth. They noted the presence or absence (P or A) of a particular placental abnormality—atrophied villi. In addition, each woman was categorized as a nonsmoker (N), moderate smoker (M), or heavy smoker (H). The following table shows, for each woman, an ID number (#) and the results for smoking (S) and atrophied villi (V).<sup>68</sup>

#	S	V	#	S	V	#	S	V	#	S	V
1	N	A	16	H	P	31	M	A	46	M	A
2	M	A	17	H	P	32	M	A	47	H	P
3	N	A	18	N	A	33	N	A	48	H	P
4	M	A	19	M	P	34	N	A	49	H	A
5	M	A	20	N	P	35	N	A	50	N	P
6	M	P	21	M	A	36	H	P	51	N	A
7	H	P	22	H	A	37	N	A	52	M	P
8	N	A	23	M	P	38	H	P	53	M	A
9	N	A	24	N	A	39	H	P	54	H	P
10	M	P	25	N	P	40	N	A	55	H	A
11	N	A	26	N	A	41	M	A	56	M	P
12	N	P	27	N	A	42	N	A	57	H	P
13	H	P	28	M	P	43	H	A	58	H	P
14	M	A	29	N	A	44	M	A			
15	M	P	30	N	A	45	M	P			

- (a) Test for a relationship between smoking status and atrophied villi. Use a chi-square test at  $\alpha = 0.05$ .
- (b) Prepare a table that shows the total number of women in each smoking category, and the number and percentage in each category who had atrophied villi.
- (c) What pattern appears in the table of part (b) that is not used by the test of part (a)?

\***10.S.17** Researchers studied the cellular telephone records of 699 persons who had automobile accidents. They determined that 170 of the 699 had made a cellular telephone call during the 10-minute period prior to their accident; this period is called the hazard interval. There were 37 persons who had made a call during a corresponding 10-minute period on the day before their accident; this period is called the control interval. Finally, there were 13 who made calls both during the hazard interval and the control interval.<sup>69</sup> Do these data indicate that use of a cellular telephone is associated with an increase in accident rate? Analyze these data using McNemar's test. Use a directional alternative and let  $\alpha = 0.01$ .

Call during control interval?	CALL DURING CONTROL INTERVAL?	
	YES	NO
Call during hazard interval?	Yes	13      157
	No	24      505

**10.S.18** Prior to an influenza season subjects were randomly assigned to receive either a flu vaccine or a placebo. During that season there were 28 cases of the flu among 813 vaccine recipients and 35 cases of the flu among the 325 subjects who were given the placebo.<sup>70</sup> Do these data indicate that the vaccine was effective? Conduct an appropriate test using a directional alternative with  $\alpha = 0.05$ .

\***10.S.19** Refer to the data in Exercise 10.S.18.

- (a) Calculate the sample odds ratio.
- (b) Find a 95% confidence interval for the population value of the odds ratio.

**10.S.20** Consider Exercise 9.S.18. The romantic partners of the 36 men discussed in Exercise 9.S.18 were also tested, in the same manner as the men (i.e., they were blindfolded and asked to identify their partner by touching the backs of the hands of three men, one of whom was their partner). Among the women, 25 were successful and 11 were not. Are these data significant evidence for the hypothesis that men and women differ in their ability to identify their partners?

Conduct a test, using  $\alpha = 0.05$ ; use a nondirectional alternative.

# COMPARING THE MEANS OF MANY INDEPENDENT SAMPLES

## Objectives

In this chapter we study analysis of variance (ANOVA). We will

- discuss when and why an analysis of variance may be conducted.
- develop the intuition behind the ANOVA model.
- demonstrate how ANOVA calculations are carried out.
- describe and examine the conditions under which ANOVA is valid.
- see how blocking is used and how to conduct randomized blocks ANOVA.
- describe interactions and main effects in factorial ANOVA models.
- construct contrasts and other linear combinations of means.
- introduce and compare several methods for dealing with multiple comparisons.

### 11.1 Introduction

In Chapter 7 we considered the comparison of two independent samples with respect to a quantitative variable  $Y$ . The classical techniques for comparing the two sample means  $\bar{Y}_1$  and  $\bar{Y}_2$  are the test and the confidence interval based on Student's  $t$  distribution. In the present chapter we consider the comparison of the means of  $I$  independent samples, where  $I$  may be greater than 2. The following example illustrates an experiment with  $I = 5$ .

**Example**  
**11.1.1**

**Sweet Corn** When growing sweet corn, can organic methods be used successfully to control harmful insects and limit their effect on the corn? In a study of this question researchers compared the weights of ears of corn under five conditions in an experiment in which sweet corn was grown using organic methods. In one plot of corn a beneficial soil nematode was introduced. In a second plot a parasitic wasp was used. A third plot was treated with both the nematode and the wasp. In a fourth plot a bacterium was used. Finally, a fifth plot of corn acted as a control; no special treatment was applied here. Thus, the treatments were

- Treatment 1: Nematodes
- Treatment 2: Wasps
- Treatment 3: Nematodes and wasps
- Treatment 4: Bacteria
- Treatment 5: Control

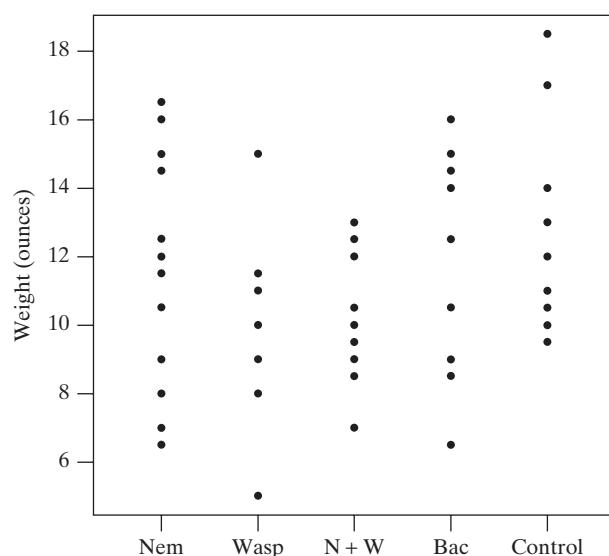
Ears of corn were randomly sampled from each plot and weighed. The results are given in Table 11.1.1 and plotted in Figure 11.1.1.<sup>1</sup> Note that in addition to the differences between the treatment means, there is also considerable variation within each treatment group.

We will discuss the classical method of analyzing data from  $I$  independent samples. The method is called an **analysis of variance**, or **ANOVA**. In applying analysis of variance, the data are regarded as random samples from  $I$  populations. We will denote the means of these populations as  $\mu_1, \mu_2, \dots, \mu_I$  and the standard deviations as  $\sigma_1, \sigma_2, \dots, \sigma_I$ .

**Table 11.1.1** Weights (ounces) of ears of sweet corn

	Treatment				
	1	2	3	4	5
16.5	11.0	8.5	16.0	13.0	
15.0	15.0	13.0	14.5	10.5	
11.5	9.0	12.0	15.0	11.0	
12.0	9.0	10.0	9.0	10.0	
12.5	11.5	12.5	10.5	14.0	
9.0	11.0	8.5	14.0	12.0	
16.0	9.0	9.5	12.5	11.0	
6.5	10.0	7.0	9.0	9.5	
8.0	9.0	10.5	9.0	18.5	
14.5	8.0	10.5	9.0	17.0	
7.0	8.0	13.0	6.5	10.0	
10.5	5.0	9.0	8.5	11.0	
Mean	11.5	9.6	10.3	11.1	12.3
SD	3.5	2.4	2.0	3.1	2.9
$n$	12	12	12	12	12

**Figure 11.1.1** Weights of ears of corn receiving five different treatments



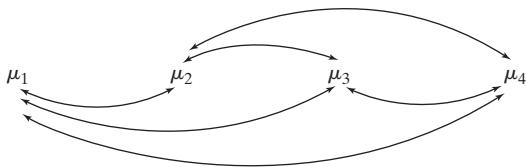
## Why Not Repeated $t$ Tests?

It is natural to wonder why the comparison of the means of  $I$  samples requires any new methods. For instance, why not just use a two-sample  $t$  test on each pair of samples? There are three reasons why this is not a good idea.

1. *The problem of multiple comparisons* The most serious difficulty with a naive “repeated  $t$  tests” procedure concerns Type I error: The probability of false rejection of a null hypothesis may be much higher than it appears to be. For instance, suppose  $I = 4$  and consider the null hypothesis that all four population means are equal ( $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ ) versus the alternative hypothesis that the four means are not all equal.\* Among four means there are six possible pairs to compare. The pairings are displayed in Figure 11.1.2. The six resulting hypotheses are

$$\begin{array}{lll} H_0: \mu_1 = \mu_2 & H_0: \mu_1 = \mu_3 & H_0: \mu_1 = \mu_4 \\ H_0: \mu_2 = \mu_3 & H_0: \mu_2 = \mu_4 & H_0: \mu_3 = \mu_4 \end{array}$$

**Figure 11.1.2** Comparing four population means requires six comparisons



Let’s consider the risk of a Type I error for testing our primary null hypothesis that all four means are equal by conducting six separate  $t$  tests. If *any* of the six  $t$  tests finds a significant difference between a pair of means, we would reject our primary null hypothesis that all four means are equal. A Type I error would occur if *any* of the six  $t$  tests found a significant difference between a pair of means when in fact all four means are equal. Thus, using  $\alpha = 0.05$  for each of the individual  $t$  tests carries an overall risk of a Type I error that is greater than 5%.

Our intuition might suggest that the risk of an overall Type I error in the preceding example should be  $6 \times 0.05 = 0.3 = 30\%$  (in each of six tests we had a 5% chance of wrongly finding evidence for a difference), but this is not the case. The computation of this overall Type I error rate is more complex. Table 11.1.2 displays the overall risk of Type I error,<sup>†</sup> that is,

Overall Type I error risk = Probability that at least one of the  $t$  tests will reject its null hypothesis, when in fact  $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_I$ .

**Table 11.1.2** Overall risk of Type I error in using repeated  $t$  tests at  $\alpha = 0.05$

$I$	Overall risk
2	0.05
3	0.12
4	0.20
6	0.37
8	0.51
10	0.63

\*In Section 11.2 we will elaborate more on the form of this alternative hypothesis.

<sup>†</sup>Table 11.1.2 was computed assuming that the sample sizes are large and equal and that the population distributions are normal with equal standard deviations.

If  $I = 2$ , then the overall risk is 0.05, as it should be, but with larger  $I$  the risk increases rapidly; for  $I = 6$  it is 0.37. It is clear from Table 11.1.2 that the researcher who uses repeated  $t$  tests is highly vulnerable to Type I error unless  $I$  is quite small.

The difficulties illustrated by Table 11.1.2 are due to **multiple comparisons**—that is, many comparisons on the same set of data. These difficulties can be reduced when the comparison of several groups is approached through ANOVA.

2. *Estimation of the standard deviation.* The ANOVA technique combines information on variability from all the samples simultaneously. This global sharing of information can yield improved precision in the analysis.
3. *Structure in the groups.* In many studies the logical structure of the treatments or groups to be compared may inspire questions that cannot be answered by simple pairwise comparisons. For example, we may wish to study the effects of two experimental factors simultaneously. ANOVA can be used to analyze data in such settings (see Sections 11.6, 11.7, and 11.8).

## A Graphical Perspective on ANOVA

When data are analyzed by analysis of variance, the usual first step is to test the following global null hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_I$$

which asserts that all the population means are equal. A statistical test of  $H_0$  will be described in Section 11.4. However, we will first consider analysis of variance from a graphical perspective.

Consider the dotplots shown in Figure 11.1.3(a). These dotplots were generated in a setting in which  $H_0$  is true. The sample means, which are shown as lines on the graph, differ from one another only as a result of chance error. For the data shown in Figure 11.1.3(b),  $H_0$  is false. The sample means are quite different—there is substantial variability between the group means, which provides evidence that the corresponding population means ( $\mu_1, \mu_2, \mu_3$ , and  $\mu_4$ ) are not all equal. In this particular case, it appears that  $\mu_1$  and  $\mu_2$  differ from  $\mu_3$  and  $\mu_4$ .

**Figure 11.1.3** (a)  $H_0$  true, (b)  $H_0$  false, with small SDs for the groups

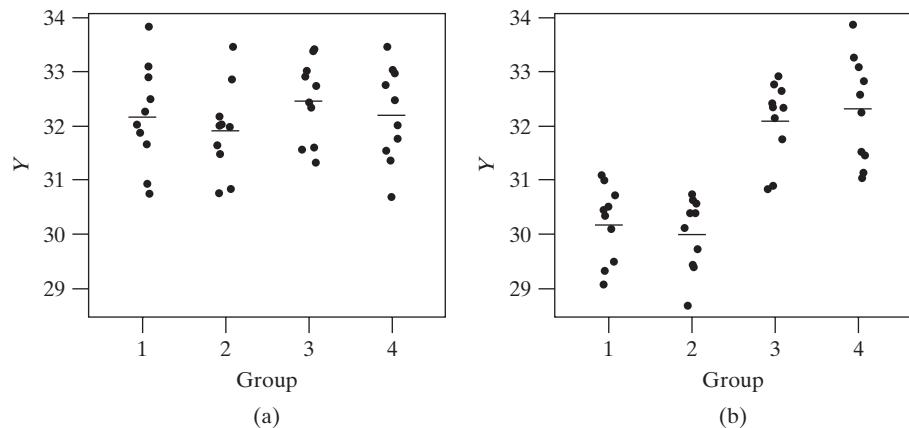
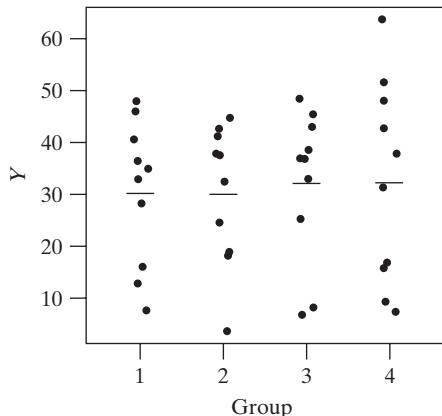


Figure 11.1.4 shows a situation that is less clear. In fact,  $H_0$  is false here—the means in Figure 11.1.4 are identical to those in Figure 11.1.3(b). However, the individual group standard deviations are quite large, which makes it hard to tell that the population means differ.\*

\*Note the change in scale on the vertical axis in Figure 11.1.4.

**Figure 11.1.4**  $H_0$  false, with large SDs for the groups



We need to know how much inherent variability there is in the data before we can judge whether a difference in sample means is fairly small and attributable to chance or whether it is too large to be due to chance alone. As Figures 11.1.3 and 11.1.4 illustrate, in order to find compelling evidence for a difference in population means, not only must there be (1) variation among the group means, but it must be large *relative* to (2) the inherent variability in the groups. It is through comparing the relative magnitudes of these two kinds of *variability*—this “analysis of variance”—that we are able to make an inference about *means*.

### A Look Ahead

If the global null hypothesis that  $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_I$  is rejected, then the data provide sufficient evidence to conclude that at least *some* of the  $\mu$ 's are unequal; the researcher would usually proceed to detailed comparisons to determine the *pattern* of differences among the  $\mu$ 's. If there is a lack of evidence against the global null hypothesis, then the researcher might choose to construct one or more confidence intervals to characterize the lack of significant differences among the  $\mu$ 's.

All the statistical procedures of this chapter—the test of the global null hypothesis and various methods of making detailed comparisons among the means—depend on the same basic calculations. These calculations are presented in Section 11.2.

## 11.2 The Basic One-Way Analysis of Variance

The ANOVA model presented in Section 11.1 that compares the means of three or more groups is called a **one-way ANOVA**. The term “one-way” refers to the fact that there is one variable that defines the groups or treatments (e.g., in the sweet corn example the treatments were based on the type of harmful insect/bacteria). Later in this chapter we will examine other ANOVA models such as the randomized complete block ANOVA (Section 11.6) and the two-way ANOVA model (Section 11.7), which consider the impact of having more than one variable defining the groups or how treatments are assigned to experimental units.

In this section we present the basic one-way ANOVA calculations that are used to describe the data and to facilitate further analysis. In the previous section we noted that if the between-group mean variability is large relative to within-group

variability, we will take this as evidence against the null hypothesis that the population means are all equal. Hence, the analysis of variance of  $I$  samples, or groups, begins with the calculation of quantities that describe the variability of the data *between* the groups and *within* the groups.\* (For clarity, in this chapter we will often refer to the samples as “groups” of observations.)

## Notation

To describe several groups of quantitative observations, we will use two subscripts: one to keep track of group membership and the other to keep track of observations within the groups. Thus, we will denote observation  $j$  in group  $i$  as

$$y_{ij} = \text{observation } j \text{ in group } i$$

Thus, the first observation in the first group is  $y_{11}$ , the second observation in the first group is  $y_{12}$ , the third observation in the second group is  $y_{23}$ , and so on.

We will also use the following notation:

$$I = \text{number of groups}$$

$$n_i = \text{number of observations in group } i$$

$$\bar{y}_i = \text{mean for group } i$$

$$s_i = \text{standard deviation for group } i$$

The total number of observations is

$$n_* = \sum_{i=1}^I n_i$$

Finally, the **grand mean**—the mean of all the observations—is

$$\bar{\bar{y}} = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} y_{ij}}{n_*}$$

Equivalently we can express  $\bar{\bar{y}}$  as a weighted average of the group means

$$\bar{\bar{y}} = \frac{\sum_{i=1}^I n_i \bar{y}_i}{\sum_i n_i} = \frac{\sum_{i=1}^I n_i \bar{y}_i}{n_*}$$

The following example illustrates this notation.

### Example 11.2.1

**Weight Gain of Lambs** Table 11.2.1 shows the weight gains (in two weeks) of young lambs on three different diets. (These data are fictitious, but are realistic in all respects except for the fact that the group means are whole numbers.)<sup>2</sup>

The total number of observations is

$$n_* = 3 + 5 + 4 = 12$$

---

\*Grammatically speaking, the word *among* should be used rather than *between* when referring to three or more groups; however, we will use “between” because it more clearly suggests that the groups are being compared against each other.

<b>Table 11.2.1</b> Weight gains of lambs (lb)*			
	Diet 1	Diet 2	Diet 3
	8	9	15
	16	16	10
	9	21	17
		11	6
		18	
$n_i$	3	5	4
Sum = $\sum_{j=1}^{n_i} y_{ij}$	33	75	48
Mean = $\bar{y}_i$	11.000	15.000	12.000
SD = $s_i$	4.359	4.950	4.967

\*Extra digits are reported for accuracy of subsequent calculations.

and the total of all the observations is

$$\sum_{i=1}^I \sum_{j=1}^{n_i} y_{ij} = 33 + 75 + 48 = 156 \text{ or, equivalently } 3 \times 11 + 5 \times 15 + 4 \times 12 = 156$$

The grand mean is

$$\bar{y} = \frac{156}{12} = 13 \text{ lb}$$

If the sample sizes ( $n_i$ 's) are all equal, then the grand mean  $\bar{y}$  is just the ordinary average (i.e., mean) of the group means (the  $\bar{y}_i$ 's); but if the sample sizes are unequal, this is not the case. For instance, in Example 11.2.1 note that

$$\frac{11 + 15 + 12}{3} \neq 13$$

## Measuring Variation within Groups

A combined measure of variation within the  $I$  groups is the pooled standard deviation  $s_{\text{pooled}}$ , often simply denoted as just  $s$ , which is computed as follows.\*

### Pooled Standard Deviation

$$s_{\text{pooled}} = s = \sqrt{\frac{\sum_{i=1}^I (n_i - 1)s_i^2}{\sum_{i=1}^I (n_i - 1)}} = \sqrt{\frac{\sum_{i=1}^I (n_i - 1)s_i^2}{n_{\bullet} - I}}$$

\*There is no ambiguity in this notation since  $s_i$  (i.e.,  $s$  with a subscript) denotes an individual group sample standard deviation.

We call  $s_{\text{pooled}}^2 = s^2$  the pooled variance\*

$$s_{\text{pooled}}^2 = s^2 = \frac{\sum_{i=1}^I (n_i - 1)s_i^2}{\sum_{i=1}^I (n_i - 1)}$$

Examining the formula we can see that the pooled variance is a weighted average of the group sample variances, and thus the pooled standard deviation can be very loosely interpreted as a weighted average of the group standard deviations.

The following example illustrates the computation of the pooled standard deviation,  $s$ .

### Example 11.2.2

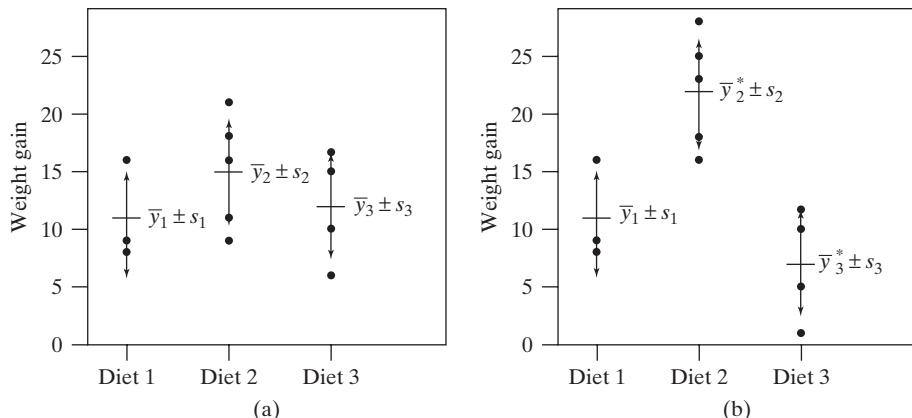
**Weight Gain of Lambs** Table 11.2.1 shows the group sample sizes and standard deviations for the lamb weight-gain data. The pooled variance and standard deviation are calculated as

$$s^2 = \frac{(3 - 1)4.359^2 + (5 - 1)4.950^2 + (4 - 1)4.967^2}{12 - 3} = \frac{210.025}{9} = 23.336$$

$$s = \sqrt{23.336} = 4.831$$

Observe that the pooled standard deviation, 4.831 lb, is a sensible representative value for the three group standard deviations, 4.359, 4.950, and 4.967 lb. If we assume that the population standard deviation of weight gains is the same for all three diets, then we would estimate this common value to be 4.83 lb. This estimate depends only on the variability within the groups and not on their mean values. Figure 11.2.1(a) displays the data from Table 11.2.1 while Figure 11.2.1(b) displays a modified version of the data for which 7 has been added to each Diet 2 observation and 5 has been subtracted from each Diet 3 observation. We see that while the group means are different for these two data sets, the pooled standard deviation—the inherent variability in each group—is the same.

**Figure 11.2.1** Examining within-group standard deviations. Plot (a) displays the weight gain data from Table 11.2.1 with  $s = 4.831$ . Plot (b) displays modified data with the same individual group standard deviations, and thus the same pooled standard deviation  $s = 4.831$



### ANOVA Notation

While our preceding formulas use familiar notation and terms, we will find it convenient to decompose the pooled variance into parts and subsequently define new terms to be used in the context of analysis of variance.

\*Recall from Chapter 2 that the variance is simply the standard deviation squared.

The numerator of the pooled variance is known as the **sum of squares within groups**, **SS(within)**, while the denominator is known as the **degrees of freedom within groups**, **df(within)**. The formulas for these are displayed in the following box.\*

### Sum of Squares and df within Groups

$$\text{SS}(\text{within}) = \sum_{i=1}^I (n_i - 1)s_i^2$$

$$\text{df}(\text{within}) = n_* - I$$

Their ratio is defined as the **mean square within groups**, or **MS(within)**. Note that **MS(within)** is just another name for the pooled variance.

### Mean Square within Groups

$$\text{MS}(\text{within}) = \frac{\text{SS}(\text{within})}{\text{df}(\text{within})}$$

Hence, the quantity **MS(within)** measures the variability within the groups.<sup>†</sup>

The following example illustrates the calculation of **SS(within)**, **df(within)**, and **MS(within)**.

#### Example 11.2.3

**Weight Gain of Lambs** In Example 11.2.2 when computing the pooled variance, we found

$$s^2 = \frac{(3 - 1)4.359^2 + (5 - 1)4.950^2 + (4 - 1)4.967^2}{12 - 3} = \frac{210.025}{9} = 23.336$$

Thus, **SS(within)** = 210.025, **df(within)** = 9, and **MS(within)** = 23.336. ■

### Variation between Groups

For two groups, the difference between the groups is simply described by  $(\bar{y}_1 - \bar{y}_2)$ . How can we describe between-group variability for more than two groups? One naive idea is to simply compute the sample variance of the group means. The **mean square between groups**, or **MS(between)** is motivated by this idea. In fact, were it not for the  $n_i$  in the numerator of the following expression (to adjust for the sample sizes of the groups), the **MS(between)** would indeed be the sample variance of the group means.

### Mean Square between Groups

$$\text{MS}(\text{between}) = \frac{\sum_{i=1}^I n_i (\bar{y}_i - \bar{y})^2}{I - 1}$$

\*A popular but less intuitive formula for **SS(within)** is given by  $\text{SS}(\text{within}) = \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ .

<sup>†</sup>If there were only one group, with  $n$  observations, then **df(within)** would be  $n - 1$  and the **SS(within)** would be  $(n - 1)s^2$ . **MS(within)** would then simply be  $\frac{(n - 1)s^2}{(n - 1)} = s^2$ , the sample variance.

As with the measures used for the within-group variation,  $MS(\text{within})$ , it is convenient to define the numerator of  $MS(\text{between})$  as the **sum of squares between groups** or **SS(between)** and the denominator as the **degrees of freedom between groups** or **df(between)** so that

$$MS(\text{between}) = \frac{SS(\text{between})}{df(\text{between})}$$

where  $SS(\text{between})$  and  $df(\text{between})$  are explicitly defined as follows.

### Sum of Squares and df between Groups

$$SS(\text{between}) = \sum_{i=1}^I n_i (\bar{y}_i - \bar{y})^2$$

$$df(\text{between}) = I - 1$$

The following example illustrates these definitions.

#### Example 11.2.4

**Weight Gain of Lambs** For the data of Example 11.2.1, the quantities that enter  $SS(\text{between})$  are shown in Table 11.2.2.

Table 11.2.2 Calculation of $SS(\text{between})$ for lamb weight gains			
	Diet 1	Diet 2	Diet 3
Mean: $\bar{y}_i$	11	15	12
$n_i$	3	5	4
Grand mean $\bar{y} = 13$			

From Table 11.2.2 we calculate

$$SS(\text{between}) = 3(11 - 13)^2 + 5(15 - 13)^2 + 4(12 - 13)^2 = 36$$

Since  $I = 3$ , we have

$$df(\text{between}) = 3 - 1 = 2$$

so that

$$MS(\text{between}) = \frac{36}{2} = 18$$

The  $SS(\text{between})$  and  $MS(\text{between})$  measure the variability between the samples means of the groups. This variability is shown graphically in Figure 11.2.2. ■

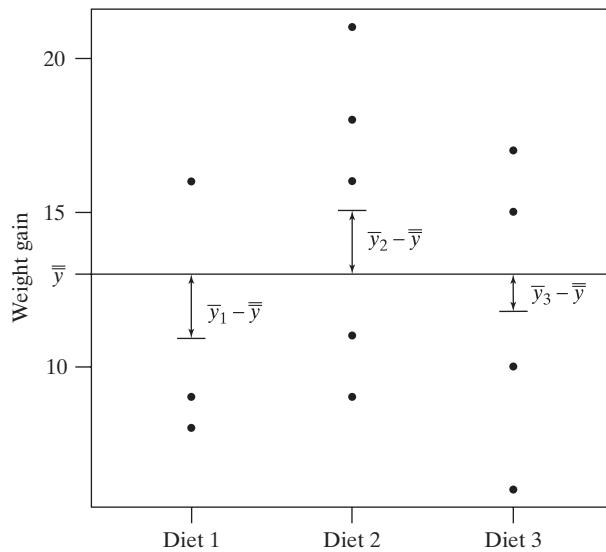
### A Fundamental Relationship of ANOVA

The name *analysis of variance* derives from a fundamental relationship involving  $SS(\text{between})$  and  $SS(\text{within})$ . Consider an individual observation  $y_{ij}$ . It is obviously true that

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

This equation expresses the deviation of an observation from the grand mean as the sum of two parts: a within-group deviation ( $y_{ij} - \bar{y}_i$ ) and a between-group deviation

**Figure 11.2.2** Measuring the differences between group means



$(\bar{y}_i - \bar{y})$ . It is also true (but not at all obvious) that the analogous relationship holds for the corresponding sums of squares; that is,

$$\sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^I \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \quad (11.2.1)$$

which, by rewriting each of the sums on the right-hand side can be expressed as

$$\begin{aligned} \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 &= \sum_{i=1}^I (n_i - 1)s_i^2 + \sum_{i=1}^I n_i(\bar{y}_i - \bar{y})^2 \\ &= \text{SS(within)} + \text{SS(between)} \end{aligned}$$

The quantity on the left-hand side of formula (11.2.1) is called the **total sum of squares**, or **SS(total)**:

#### Definition of Total Sum of Squares

$$\text{SS(total)} = \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

Note that SS(total) measures variability among all  $n$  observations in the  $I$  groups. The relationship [formula (11.2.1)] can be written as

#### Relationship between Sums of Squares

$$\text{SS(total)} = \text{SS(between)} + \text{SS(within)}$$

The preceding fundamental relationship shows how the total variation in the data set can be analyzed, or broken down, into two interpretable components: between-sample variation and within-sample variation. This partition is an analysis of variance.

The **total degrees of freedom**, or **df(total)**, is defined as follows:

Total df

$$\text{df}(\text{total}) = n_{\bullet} - 1$$

With this definition, the degrees of freedom add, just as the sums of squares do; that is,

$$\text{df}(\text{total}) = \text{df}(\text{within}) + \text{df}(\text{between})$$

$$n_{\bullet} - 1 = (n_{\bullet} - I) + (I - 1)$$

Notice that, if we were to consider all  $n_{\bullet}$  observations as a single sample, then the SS for that sample (that is, the numerator of the variance) would be SS(total) and the associated df (that is, the denominator of the variance) would be df(total). Consequently,  $\sqrt{\frac{\text{SS}(\text{total})}{\text{df}(\text{total})}}$  is the standard deviation of the entire data set when group membership is ignored.

The following example illustrates the fundamental relationships between the sums of squares and degrees of freedom.

**Example  
11.2.5**

**Weight Gain of Lambs** For the data of Table 11.2.1, we found  $\bar{y} = 13$ ; we calculate SS(total) as

$$\begin{aligned}\text{SS}(\text{total}) &= \sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \\ &= [(8 - 13)^2 + (16 - 13)^2 + (9 - 13)^2] \\ &\quad + [(9 - 13)^2 + (16 - 13)^2 + (21 - 13)^2 + (11 - 13)^2 + (18 - 13)^2] \\ &\quad + [(15 - 13)^2 + (10 - 13)^2 + (17 - 13)^2 + (6 - 13)^2] \\ &= 246\end{aligned}$$

For these data, we found that  $\text{SS}(\text{between}) = 36$  and  $\text{SS}(\text{within}) = 210$ . We verify that

$$246 = 36 + 210$$

Also, we found that  $\text{df}(\text{within}) = 9$  and  $\text{df}(\text{between}) = 2$ . We verify that

$$\text{df}(\text{total}) = 12 - 1 = 11 = 9 + 2$$

## The ANOVA Table

When working with the ANOVA quantities, it is customary to arrange them in a table. The following example shows a typical format for the ANOVA table.

**Example  
11.2.6**

**Weight Gain of Lambs** Table 11.2.3 shows the ANOVA for the lamb weight-gain data. Notice that the ANOVA table clearly shows the additivity of the sums of squares and the degrees of freedom.

**Comment on terminology.** While the terms “between-groups” and “within-groups” are not technical terms, they are useful in describing and understanding the ANOVA model. Computer software and other texts commonly refer to these sources of variability as **treatment** (between groups) and **error** (within groups).

Source	df	SS	MS
Between diets	2	36	18.00
Within diets	9	210	23.33
Total	11	246	

## Summary of Formulas

For convenient reference, we display in the box the definitional formulas for the basic ANOVA quantities.

### ANOVA Quantities with Formulas

Source	df	SS (Sum of Squares)	MS (Mean Square)
Between groups	$I - 1$	$\sum_{i=1}^I n_i (\bar{y}_i - \bar{y})^2$	SS/df
Within groups	$n_{\bullet} - I$	$\sum_{i=1}^I (n_i - 1) s_i^2$	SS/df
Total	$n_{\bullet} - 1$	$\sum_{i=1}^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	

## Exercises 11.2.1–11.2.7

**11.2.1** The accompanying table shows fictitious data for three samples.

SAMPLE			
	1	2	3
48	40	39	
39	48	30	
42	44	32	
43		35	
Mean	43.00	44.00	34.00
SD	3.74	4.00	3.92

- (a) Compute SS(between) and SS(within).
- (b) Compute SS(total), and verify the relationship between SS(between), SS(within), and SS(total).
- (c) Compute MS(between), MS(within), and  $s_{\text{pooled}}$ .

**11.2.2** Proceed as in Exercise 11.2.1 for the following data:

SAMPLE			
	1	2	3
23	18	20	
29	12	16	
25	15	17	
23		23	
		19	
Mean	25.00	15.00	19.00
SD	2.83	3.00	3.16

**11.2.3** For the following data, SS(within) = 116 and SS(total) = 338.769.

SAMPLE		
1	2	3
31	30	39
34	26	45
39	35	39
32	29	37
	30	

- (a) Find SS(between).  
 (b) Compute MS(between), MS(within), and  $s_{\text{pooled}}$ .

**11.2.4** The following ANOVA table is only partially completed.

SOURCE	DF	SS	MS
Between groups	3	45	
Within groups	12	337	
Total		472	

- (a) Complete the table.  
 (b) How many groups were there in the study?  
 (c) How many total observations were there in the study?

**11.2.5** The following ANOVA table is only partially completed.

SOURCE	DF	SS	MS
Between groups	4		
Within groups		964	
Total	53	1123	

- (a) Complete the table.  
 (b) How many groups were there in the study?  
 (c) How many total observations were there in the study?

**11.2.6** The following ANOVA table is only partially completed.

SOURCE	DF	SS	MS
Between groups		258	
Within groups	26		
Total	29	898	

- (a) Complete the table.  
 (b) How many groups were there in the study?  
 (c) How many total observations were there in the study?

**11.2.7** Invent examples of data with

- (a) SS(between) = 0 and SS(within) > 0  
 (b) SS(between) > 0 and SS(within) = 0  
 (c) For each example, use three samples, each of size 5.

## 11.3 The Analysis of Variance Model

In Section 11.2 we introduced the notation  $y_{ij}$  for the  $j$ th observation in group  $i$ . We think of  $y_{ij}$  as a random observation from group  $i$ , where the population mean of group  $i$  is  $\mu_i$ . We use analysis of variance to investigate the null hypothesis that  $\mu_1 = \mu_2 = \dots = \mu_I$ . It can be helpful to think of ANOVA in terms of the following model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

In this model,  $\mu$  represents the grand population mean—the population mean when all the groups are combined. If the null hypothesis is true, then  $\mu$  is the common population mean. If the null hypothesis is false, then at least some of the  $\mu_i$ 's differ from the grand population mean of  $\mu$ .

The term  $\tau_i$  represents the effect of group  $i$ —that is, the difference between the population mean for group  $i$ ,  $\mu_i$ , and the grand population mean,  $\mu$ . ( $\tau$  is the Greek letter “tau.”) Thus,

$$\tau_i = \mu_i - \mu$$

The null hypothesis

$$H_0: \mu_1 = \mu_2 = \dots = \mu_I$$

is equivalent to

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_I = 0$$

If  $H_0$  is false, then at least some of the groups differ from the others. If  $\tau_i$  is positive, then observations from group  $i$  tend to be greater than the overall average; if  $\tau_i$  is negative, then data from group  $i$  tend to be less than the overall average.

The term  $e_{ij}$  in the model represents random error associated with observation  $j$  in group  $i$ . Thus, the model

$$y_{ij} = \mu + \tau_i + e_{ij}$$

can be stated in words as

observation = overall average + group effect + random error

We estimate the overall average,  $\mu$ , with the grand mean of the data:

$$\hat{\mu} = \bar{y}$$

Likewise, we estimate the population average for group  $i$  with the sample average for group  $i$ :

$$\hat{\mu}_i = \bar{y}_i$$

Since the group effect is

$$\tau_i = \mu_i - \mu$$

we estimate  $\tau_i$  as

$$\hat{\tau}_i = \bar{y}_i - \bar{y}$$

Finally, we estimate the random error,  $e_{ij}$ , for observation  $y_{ij}$  as

$$\hat{e}_{ij} = y_{ij} - \bar{y}_i$$

Putting these estimates together, we have

$$y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

or

$$y_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{e}_{ij}$$

**Note.** Some authors use the terminology SS(error) for what we have called SS(within). This is due to the fact that the within-groups component  $y_{ij} - \bar{y}_i$  estimates the random error term in the ANOVA model.

---

### Example 11.3.1

**Weight Gain of Lambs** For the data of Example 11.2.1, the estimate of the grand population mean is  $\hat{\mu} = 13$ . The estimated group effects are

$$\hat{\tau}_1 = \bar{y}_1 - \bar{y} = 11 - 13 = -2$$

$$\hat{\tau}_2 = 15 - 13 = 2$$

and

$$\hat{\tau}_3 = 12 - 13 = -1$$

Thus, we estimate that Diet 2 increases weight gain by 2 lb on average (when compared to the average of the three diets), Diet 1 decreases weight gain by an average of 2 lb, and Diet 3 decreases weight gain by 1 lb, on average. ■

When we conduct an analysis of variance, we are comparing the sizes of the sample group effects, the  $\hat{\tau}_i$ 's, to the sizes of the random errors in the data, the  $\hat{\varepsilon}_{ij}$ 's. We can see that

$$\text{SS(between)} = \sum_{i=1}^I n_i \hat{\tau}_i^2$$

and

$$\text{SS(within)} = \sum_{i=1}^I \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$$

## 11.4 The Global F Test

The global null hypothesis is

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

We consider testing  $H_0$  against the nondirectional (or omnidirectional) alternative hypothesis

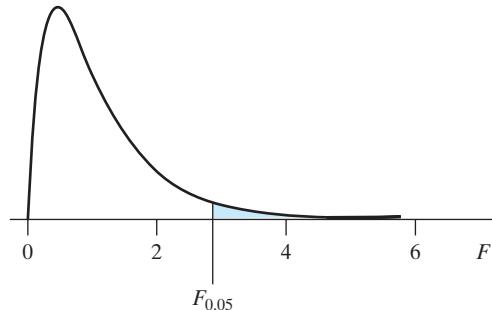
$$H_A: \text{The } \mu_i \text{'s are not all equal}$$

Note that  $H_0$  is compound (unless  $I = 2$ ), and so rejection of  $H_0$  does not specify which  $\mu_i$ 's are different. If we reject  $H_0$ , then we conduct a further analysis to make detailed comparisons among the  $\mu_i$ 's. Testing the global null hypothesis may be likened to looking at a microscope slide through a low-power lens to see if there is anything on it; if we find something, we switch to a greater magnification to examine its fine structure.

## The F Distributions

The **F distributions**, named after the statistician and geneticist R. A. Fisher, are probability distributions that are used in many kinds of statistical analysis. The form of an F distribution depends on two parameters: the **numerator degrees of freedom** and the **denominator degrees of freedom**. Figure 11.4.1 shows an F distribution with numerator df = 4 and denominator df = 20. Critical values for the F distribution are given in Table 10 at the end of this book. Note that Table 10 occupies 10 pages, each page having a different value of the numerator df. As a specific example, for numerator df = 4 and denominator df = 20, we find in Table 10 that  $F(4, 20)_{0.05} = 2.87$ ; this value is shown in Figure 11.4.1.

**Figure 11.4.1** The F distribution with numerator df = 4 and denominator df = 20



## The *F* Test

The ***F* test** is a classical test of the global null hypothesis. The test statistic, the ***F* statistic**, is calculated as follows:

$$F_s = \frac{\text{MS(between)}}{\text{MS(within)}}$$

From the definitions of the mean squares (Section 11.2), it is clear that  $F_s$  will be large if the discrepancies among the group means ( $\bar{Y}_i$ 's) are large relative to the variability within the groups. Thus, large values of  $F_s$  tend to provide evidence against  $H_0$ —evidence for a difference among the group means.

To carry out the *F* test of the global null hypothesis, critical values are obtained from an *F* distribution (Table 10) with

$$\text{Numerator df} = \text{df(between)}$$

and

$$\text{Denominator df} = \text{df(within)}$$

It can be shown that (when suitable conditions for validity are met) the null distribution of  $F_s$  is an *F* distribution with df as given above.

The following example illustrates the global *F* test.

---

### Example 11.4.1

**Weight Gain of Lambs** For the lamb feeding experiment of Example 11.2.1, the global null hypothesis and alternative can be stated verbally as

$H_0$ : Mean weight gain is the same on all three diets.

$H_A$ : Mean weight gain is not the same on all three diets.

or symbolically as

$H_0: \mu_1 = \mu_2 = \mu_3$

$H_A$ : The  $\mu_i$ 's are not all equal

We saw in Figure 11.2.2 that the three sample means do not differ much when compared to the variability within the groups, which is not very strong evidence against  $H_0$ . Let us confirm this visual impression by carrying out the *F* test at  $\alpha = 0.05$ . From the ANOVA table (Table 11.2.3) we find

$$F_s = \frac{18.00}{23.33} = 0.77$$

The degrees of freedom can also be read from the ANOVA table as

$$\text{Numerator df} = 2$$

$$\text{Denominator df} = 9$$

From Table 10 we find  $F(2, 9)_{0.20} = 1.93$ , so that  $P > 0.20$ . Thus, there is a lack of significant evidence against  $H_0$ ; there is insufficient evidence to conclude that there is any difference among the diets with respect to population mean weight gain. The observed differences in the mean gains in the samples can readily be attributed to chance variation. Because this study was an experiment (as opposed to

an observational study), we can even make a slightly stronger summary of the results: There is insufficient evidence to conclude that among these three diets, diet *affects* weight gain.

## Relationship between F Test and t Test

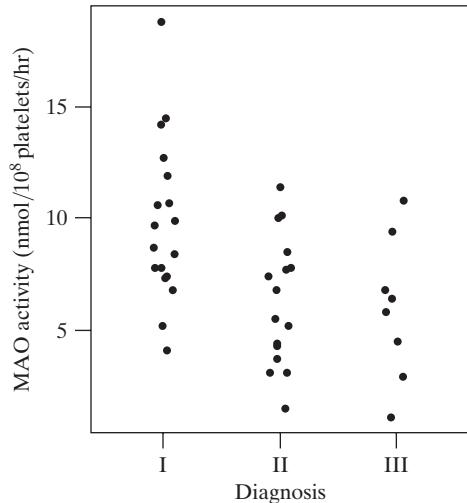
Suppose only two groups are to be compared ( $I = 2$ ). Then one could test  $H_0: \mu_1 = \mu_2$  against  $H_A: \mu_1 \neq \mu_2$  using either the *F* test or the *t* test. The *t* test from Chapter 7 can be modified slightly by replacing each sample standard deviation by  $s_{\text{pooled}}$ , as defined in Section 11.2, before calculating the standard error of  $(\bar{Y}_1 - \bar{Y}_2)$ . It can be shown that the *F* test and this “pooled” *t* test are actually equivalent procedures. The relationship between the test statistics is  $t_s^2 = F$ ; that is, the value of the *F* statistic for any set of data is necessarily equal to the square of the value of the (pooled) *t* statistic. The corresponding relationship between the critical values is  $t_{0.025}^2 = F_{0.05}$ ,  $t_{0.005}^2 = F_{0.01}$ , and so on. For example, suppose  $n_1 = 10$  and  $n_2 = 7$ . Then the appropriate *t* distribution has  $\text{df} = n_1 + n_2 - 2 = 15$ , and  $t_{15,0.025} = 2.131$ , whereas the *F* distribution has numerator  $\text{df} = I - 1 = 1$  and denominator  $\text{df} = n_{\cdot} - I = 15$ , so that  $F(1,15)_{0.05} = 4.54$ ; note that  $(2.131)^2 = 4.54$ . Because of the equivalence of the tests, the application of the *F* test to compare the means of two samples will always give exactly the same *P*-value as the pooled *t* test applied to the same data.

### Exercises 11.4.1–11.4.7

**11.4.1** Monoamine oxidase (MAO) is an enzyme that is thought to play a role in the regulation of behavior. To see whether different categories of schizophrenic patients have different levels of MAO activity, researchers collected blood specimens from 42 patients and measured the MAO activity in the platelets. The results are summarized in the accompanying table. (Values are expressed as nmol benzylaldehyde product/ $10^8$  platelets/hour.)<sup>3</sup> Calculations based on the raw data yielded  $\text{SS}(\text{between}) = 136.12$  and  $\text{SS}(\text{within}) = 418.25$ .

DIAGNOSIS	MAO ACTIVITY		
	MEAN	SD	NO. OF PATIENTS
Chronic undifferentiated schizophrenic	9.81	3.62	18
Undifferentiated with paranoid features	6.28	2.88	16
Paranoid schizophrenic	5.97	3.19	8

- (a) Dotplots of these data follow. Based on this graphical display, does it appear that the null hypothesis is true? Why or why not?
- (b) Construct the ANOVA table and test the global null hypothesis at  $\alpha = 0.05$ .
- (c) Calculate the pooled standard deviation,  $s_{\text{pooled}}$ .



**11.4.2** It is thought that stress may increase susceptibility to illness through suppression of the immune system. In an experiment to investigate this theory, 48 rats were randomly allocated to four treatment groups: no stress, mild stress, moderate stress, and high stress. The stress conditions involved various amounts of restraint and electric shock. The concentration of lymphocytes (cells/ml  $\times 10^{-6}$ ) in the peripheral blood was measured for each rat with the results given in the accompanying table.<sup>4</sup> Calculations based on the raw data yielded  $\text{SS}(\text{between}) = 89.036$  and  $\text{SS}(\text{within}) = 340.24$ .

	NO STRESS	MILD STRESS	MODERATE STRESS	HIGH STRESS
$\bar{y}$	6.64	4.84	3.98	2.92
$s$	2.77	2.42	3.91	1.45
$n$	12	12	12	12

- (a) Construct the ANOVA table and test the global null hypothesis at  $\alpha = 0.05$ .  
(b) Calculate the pooled standard deviation,  $s_{\text{pooled}}$ .

**11.4.3** Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress. An exercise physiologist measured the resting (unstressed) blood concentration of HBE in three groups of men: 15 who had just entered a physical fitness program, 11 who had been jogging regularly for some time, and 10 sedentary people. The HBE levels (pg/ml) are shown in the following table.<sup>5</sup> Calculations based on the raw data yielded SS(between) = 240.69 and SS(within) = 6,887.6.

	FITNESS PROGRAM		
	ENTRANTS	JOGGERS	SEDENTARY
Mean	38.7	35.7	42.5
SD	16.1	13.4	12.8
$n$	15	11	10

- (a) State the appropriate null hypothesis in words, in the context of this setting.  
(b) State the null hypothesis in symbols.  
(c) Construct the ANOVA table and test the null hypothesis. Let  $\alpha = 0.05$ .  
(d) Calculate the pooled standard deviation,  $s_{\text{pooled}}$ .

**11.4.4** An experiment was conducted in which the antiviral medication zanamivir was given to patients who had the flu. The length of time until the alleviation of major flu symptoms was measured for three groups: 85 patients who were given inhaled zanamivir, 88 patients who were given inhaled and intranasal zanamivir, and 89 patients who were given a placebo. Summary statistics are given in the following table.<sup>6</sup> The ANOVA SS(between) is 53.67 and the SS(within) is 2034.52.

	INHALED ZANAMIVIR	INHALED AND INTRANASAL ZANAMIVIR	PLACEBO
Mean	5.4	5.3	6.3
SD	2.7	2.8	2.9
$n$	85	88	89

- (a) State the appropriate null hypothesis in words, in the context of this setting.  
(b) State the null hypothesis in symbols.

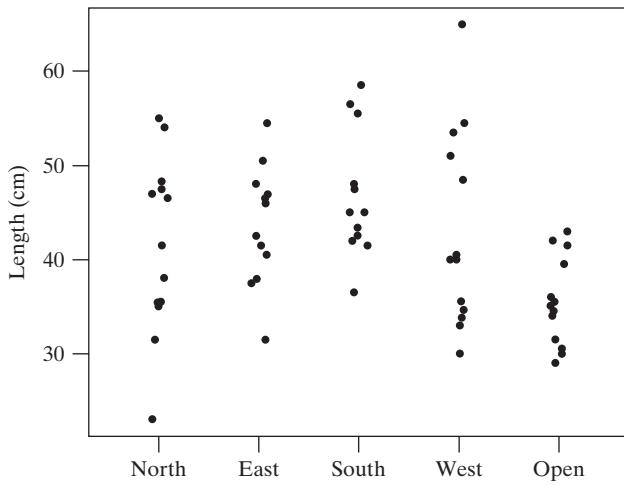
- (c) Construct the ANOVA table and test the null hypothesis. Let  $\alpha = 0.05$ .

- (d) Calculate the pooled standard deviation,  $s_{\text{pooled}}$ .

**11.4.5** A researcher collected daffodils from four sides of a building and from an open area nearby. She wondered whether the average stem length of a daffodil depends on the side of the building on which it is growing. Summary statistics are given in the following table.<sup>7</sup> The ANOVA SS(between) is 871.408 and the SS(within) is 3588.54.

	NORTH	EAST	SOUTH	WEST	OPEN
Mean	41.4	43.8	46.5	43.2	35.5
SD	9.3	6.1	6.6	10.4	4.7
$n$	13	13	13	13	13

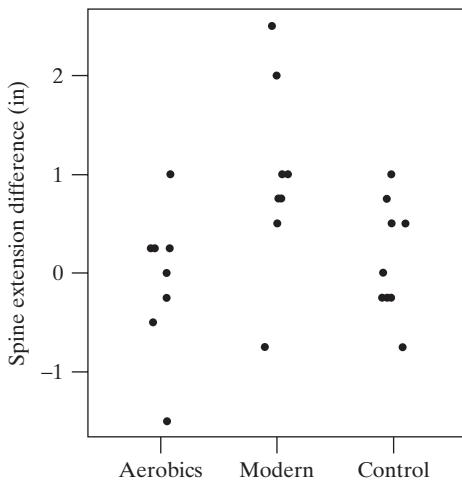
- (a) Dotplots of these data follow. Based on the dotplots, does it appear that the null hypothesis is true? Why or why not?  
(b) State the null hypothesis in symbols.  
(c) Construct the ANOVA table and test the null hypothesis. Let  $\alpha = 0.10$ .



**11.4.6** A researcher studied the flexibility of 10 women in an aerobic exercise class, 10 women in a modern dance class, and a control group of 9 women. One measurement she made on each woman was spinal extension, which is a measure of how far the woman could bend her back. Measurements were made before and after a 16-week training period. The change in spinal extension was recorded for each woman. Summary statistics are given in the following table.<sup>8</sup> The ANOVA SS(between) is 7.04 and the SS(within) is 15.08.

	AEROBICS	MODERN DANCE	CONTROL
Mean	-0.18	0.98	0.13
SD	0.80	0.86	0.57
$n$	10	10	9

- (a) Dotplots of these data were shown below. Based on the dotplots, does it appear that the null hypothesis is true? Why or why not?
- (b) State the null hypothesis in symbols.
- (c) Construct the ANOVA table and test the null hypothesis. Let  $\alpha = 0.01$ .



**11.4.7** The following computer output is for an analysis of variance in which yields (bu/acre) of different varieties of oats were compared.<sup>9</sup>

SOURCE	DF	SUMS OF SQUARES	MEAN SQUARE	F RATIO	PROB
Group	2	76.8950	38.4475	0.40245	0.6801
Error	9	859.808	95.5342		
Total	11	936.703			

- (a) How many varieties (groups) were in the experiment?  
 (b) State the conclusion of the ANOVA.  
 (c) What is the pooled standard deviation,  $s_{\text{pooled}}$ ?

## 11.5 Applicability of Methods

Like all other methods of statistical inference, the calculations and interpretations of ANOVA are based on certain conditions.

### Standard Conditions

The ANOVA techniques described in this chapter, including the global  $F$  test, are valid if the following conditions hold.

1. *Design conditions*
  - (a) It must be reasonable to regard the groups of observations as random samples from their respective populations.
  - (b) The  $I$  samples must be independent of each other.
2. *Population conditions* The  $I$  population distributions must be (approximately) normal with equal standard deviations:

$$\sigma_1 = \sigma_2 = \dots = \sigma_I$$

These conditions are extensions of the conditions given in Chapter 7 for the independent-samples  $t$  test with the added condition that the standard deviations be equal. The condition of normal populations with equal standard deviations is less crucial if the sample sizes ( $n_i$ ) are large and approximately equal.

### Verification of Conditions

The design conditions may be verified as for the independent-samples  $t$  test. To check condition 1(a), one looks for biases or hierarchical structure in the collection of the data. A completely randomized design assures independence of the samples

[condition 1(b)]. If units have been allocated to treatment groups in a nonrandom manner (e.g., by a randomized blocks design to be discussed in Section 11.6), or if observations on the same experimental unit appear in different samples (e.g., for  $I = 2$ , paired data as seen in Chapter 9), then the samples are not independent.

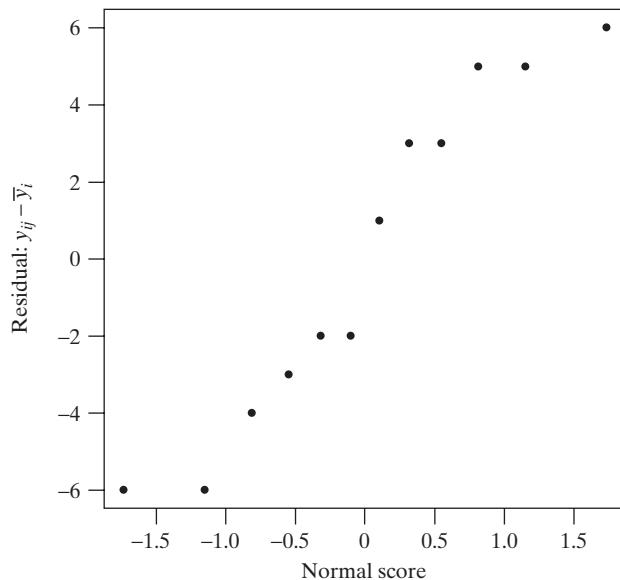
As with the independent-samples  $t$  test, the population conditions can be roughly checked from the data. To check normality, a separate histogram or normal probability plot can be made for each sample. Another option is to make a single histogram or normal probability plot of the deviations ( $y_{ij} - \bar{y}_i$ ) from all the samples combined. In the context of analysis of variance we call these deviations from the group means **residuals**. Thus, a residual measures how far a data value falls from its respective group mean.

Equality of the population SDs is checked by comparing the sample SDs; one useful trick is to plot the SDs against the means ( $\bar{y}_i$ 's) to check for a trend. Another approach is to make a plot of the residuals ( $y_{ij} - \bar{y}_i$ ) against the means ( $\bar{y}_i$ 's). As a rule of thumb, we would like the largest sample SD divided by the smallest sample SD to be less than 2 or so. If this ratio is much larger than 2, then we cannot be confident in the  $P$ -value from the ANOVA, particularly if the sample sizes are small and unequal. In particular, if the sample sizes are unequal and the sample SD from a small sample is quite a bit larger than the other SDs, then the  $P$ -value can be quite inaccurate.

### Example 11.5.1

**Weight Gain of Lambs** Consider the lamb feeding experiment of Example 11.2.2. Figure 11.2.1 (in Section 11.2) shows that the variability within groups is nearly equal across the three diets: The three sample SDs are 4.36, 4.95, and 4.97. Figure 11.5.1 is a normal probability plot of the 12 residuals ( $y_{ij} - \bar{y}_i$ ) (3 from Diet 1, 5 from Diet 2, and 4 from Diet 3). This plot is close to linear, which provides no evidence to cast doubt on the normality condition. ■

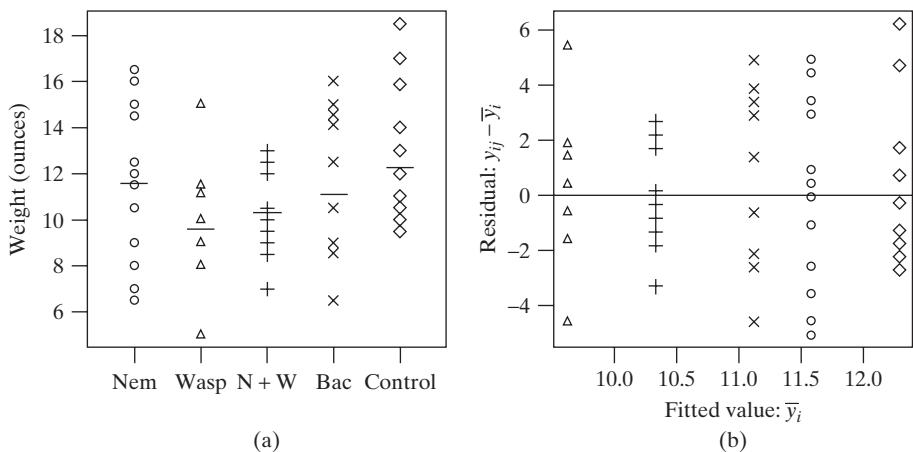
**Figure 11.5.1** Normal probability plot of residuals ( $y_{ij} - \bar{y}_i$ ) in weight-gain data



### Example 11.5.2

**Sweet Corn** Consider the sweet corn data of Example 11.1.1. Figure 11.5.2(a) shows the data with each group receiving its own plotting symbol. Using those same plotting symbols for each group, Figure 11.5.2(b) displays the residuals ( $y_{ij} - \bar{y}_i$ ) plotted

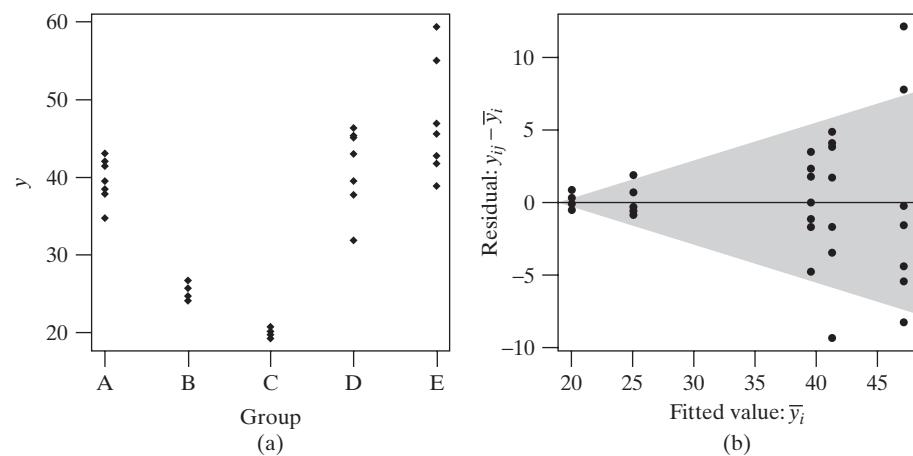
**Figure 11.5.2** Plot of residuals versus sample mean for the sweet corn data



against each group's mean ( $\bar{y}_i$ ) (also known as a **fitted value** in the context of analysis of variance). This second graph shows that the variability (as measured visually by the vertical spread) does not appreciably change as the mean changes (which is good—if the variability increased as the mean increased, then condition 2 would be violated). ■

While one could look at a basic plot of the data, as in Figure 11.5.2(a), to visually inspect that the SDs are similar across all groups, plotting the data as in Figure 11.5.2(b) provides some visual advantages. First, by examining the residuals (Figure 11.5.2(b)) and not the raw data (Figure 11.5.2(a)), one can scan the graph from left to right allowing the eyes to more clearly compare the variability among the groups without being distracted by the changing means. Furthermore, a common violation of the equal SD requirement is that the group SDs grow with the means. To illustrate this violation, consider the fictitious data graphed in Figure 11.5.3(a) consisting of five treatment groups and seven observations per group. Clearly the variability is not the same in all five groups. The plot of the residuals versus means in Figure 11.5.3(b) exposes this problem more clearly and shows that the SD (represented by vertical spread) increases with the mean. We often describe this as *funnel* or *horn* shape in the residuals.

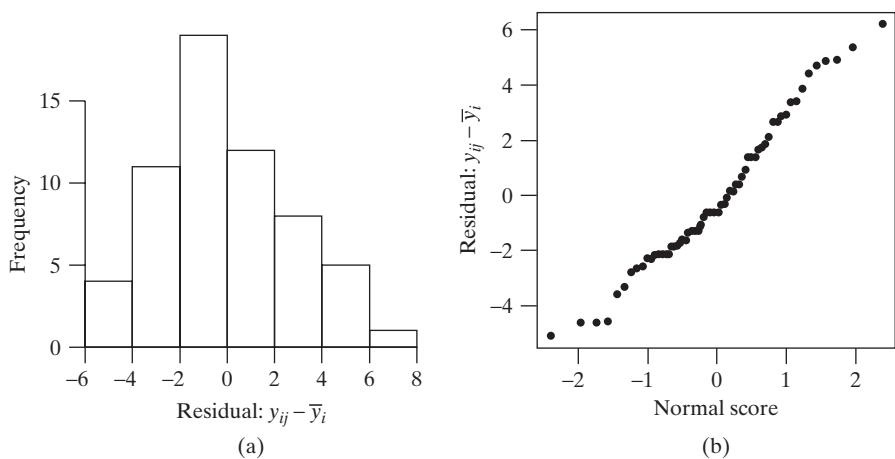
**Figure 11.5.3** Plot of residuals versus sample means for a fictitious data set for which the standard deviation increases with the mean



**Example  
11.5.3**

**Sweet Corn** Again considering the sweet corn data of Example 11.2.1, we examine the normality of the groups through examination of the residuals. Figure 11.5.4 contains a histogram and a normal probability plot of the 60 residuals ( $y_{ij} - \bar{y}_i$ ). The bell-shaped nature of plot (a) and linearity of plot (b) cast little doubt upon the normality condition. ■

**Figure 11.5.4** Histogram and normal probability plot of deviations ( $y_{ij} - \bar{y}_i$ ) in sweet corn data



## Further Analysis

In addition to their relevance to the  $F$  test, the standard conditions underlie many classical methods for further analysis of the data.

If the  $I$  populations have the same SD, then a pooled estimate of that SD from the data is

$$s_{\text{pooled}} = \sqrt{\text{MS}(\text{within})}$$

from the ANOVA. This pooled standard deviation  $s_{\text{pooled}}$  is a better estimate than any individual sample SD because  $s_{\text{pooled}}$  is based on more observations.

A simple way to see the advantage of  $s_{\text{pooled}}$  is to consider the standard error of an individual sample mean, which can be calculated as

$$\text{SE}_{\bar{Y}} = \frac{s_{\text{pooled}}}{\sqrt{n}}$$

where  $n$  is the size of the individual sample. The df associated with this standard error is  $\text{df}(\text{within})$ , which is the sum of the degrees of freedom of all the samples. By contrast, if the individual SD were used in calculating  $\text{SE}_{\bar{Y}}$ , it would have only  $(n - 1)$  df. When the SE is used for inference, larger df yield smaller critical values (see Table 4), which in turn lead to improved power and narrower confidence intervals.

In optional Sections 11.7 and 11.8 we will consider methods for detailed analysis of the group means  $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_I$ . Like the  $F$  test, these methods were designed for independent samples from normal populations with equal standard deviations. The methods use standard errors based on the pooled standard deviation estimate  $s_{\text{pooled}}$ .

## Exercises 11.5.1–11.5.2

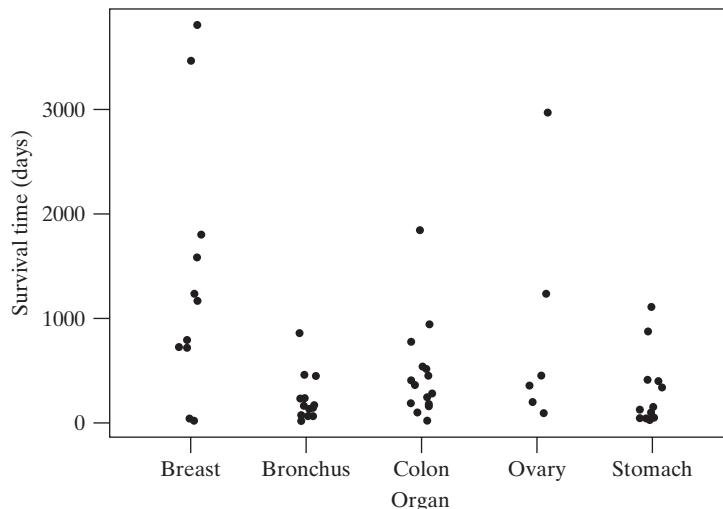
**11.5.1** Refer to the lymphocyte data of Exercise 11.4.2. The global  $F$  test is based on certain conditions concerning the population distributions.

- State the conditions.
- Which features of the data suggest that the conditions may be doubtful in this case?

**11.5.2** Patients with advanced cancers of the stomach, bronchus, colon, ovary, or breast were treated with ascor-

bate. The purpose of the study was to determine if the survival times differ with respect to the organ affected by the cancer. The variable of interest is survival time (in days).<sup>10</sup> Here are parallel dotplots of the raw data.

An ANOVA was done after a square root transformation was applied to the raw data. There were two (related) reasons that the data were transformed. What were those two reasons?



## 11.6 One-Way Randomized Blocks Design

The completely randomized design makes no distinctions among the experimental units. Often an experiment can be improved by a more refined approach, one that takes advantage of known patterns of variability in the experimental units.

In a **randomized blocks design**, we first group the experimental units into sets, or **blocks**, of relatively similar units and then we randomly allocate treatments within each block. Here is an example.

### Example 11.6.1

**Alfalfa and Acid Rain** Researchers were interested in the effect that acid has on the growth rate of alfalfa plants. They created three treatment groups in an experiment: low acid, high acid, and control. The response variable in their experiment was the height of the alfalfa plants in a Styrofoam cup after five days of growth.\* They had 5 cups for each of the 3 treatments, for a total of 15 observations. However, the cups were arranged near a window and they wanted to account for the effect of differing amounts of sunlight. Thus, they created 5 blocks—each block was a fixed

\*More precisely, the response variable was the average height of plants within a cup, so that the observational unit was a cup, rather than individual plants.

**Figure 11.6.1** Design of the alfalfa experiment

	Block 1	Block 2	Block 3	Block 4	Block 5
Window	high	control	control	control	high
	control	low	high	low	low
	low	high	low	high	control

Organization of blocks for alfalfa experiment

distance away from the window (block 1 being the closest through block 5, the farthest). Within each block the three treatments were randomly assigned, as shown in Figure 11.6.1.<sup>11</sup> ■

Example 11.6.1 is an illustration of a randomized blocks design. To carry out a randomized blocks design, the experimenter creates or identifies suitable blocks of experimental units and then randomly assigns treatments within each block in such a way that each treatment appears in each block.\* In Example 11.6.1, the rows of cups at each of the five distances from the window serve as blocks. In general, we create blocks in order to reduce or eliminate variability caused by extraneous variables, so that the precision of the experiment is increased. We want the experimental units within a block to be homogenous; we want the extraneous variability to occur *between* the blocks. Here are more examples of randomized blocks designs in biological experiments.

**Example  
11.6.2**

**Blocking by Litter** How does experience affect the anatomy of the brain? In a typical experiment to study this question, young rats are placed in one of three environments for 80 days:

*T<sub>1</sub>: Standard environment.* The rat is housed with a single companion in a standard lab cage.

*T<sub>2</sub>: Enriched environment.* The rat is housed with several companions in a large cage, furnished with various playthings.

*T<sub>3</sub>: Impoverished environment.* The rat lives alone in a standard lab cage.

At the end of the 80-day experience, various anatomical measurements are made on the rats' brains.

Suppose a researcher plans to conduct the above experiment using 30 rats. To minimize variation in response, all 30 animals will be male, of the same age and strain. To reduce variation even further, the researcher can take advantage of the similarity of animals from the same litter. In this approach, the researcher would obtain three male rats from each of 10 litters. The three littermates from each litter would be assigned at random: one to *T<sub>1</sub>*, one to *T<sub>2</sub>*, and one to *T<sub>3</sub>*.<sup>12</sup> ■

Another way to visualize the experimental design is in tabular form, as shown in Table 11.6.1. Each "Y" in the table represents an observation on one rat. Using the layout of Table 11.6.1, the experimenter can compare the responses of rats that received *different* treatments but are in the *same* litter. Such comparisons are not affected by any difference (genetic and other) that may exist between one litter and another.

---

\*Strictly speaking, the design we discuss is termed a *randomized complete blocks design* because every treatment appears in every block. In an *incomplete blocks design*, each block contains some, but not necessarily all, of the treatments.

**Table 11.6.1** Format for rat brain data

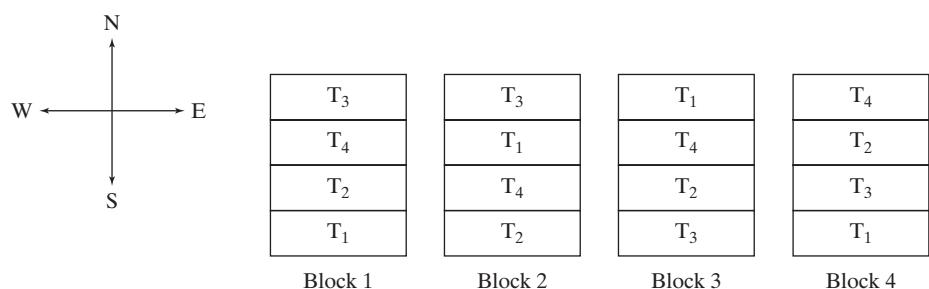
	Treatment		
	$T_1$	$T_2$	$T_3$
Litter 1	Y	Y	Y
Litter 2	Y	Y	Y
Litter 3	Y	Y	Y
.	.	.	.
.	.	.	.
.	.	.	.
Litter 10	Y	Y	Y

**Example 11.6.3**

**Within-Subject Blocking (Pairing)** A dermatologist is planning a study to compare two medicated lotions for their effectiveness in treating acne. Twenty patients are to participate in the study. Each patient will use lotion A on one side of his or her face and lotion B on the other; the dermatologist will observe the improvement on each side during a three-month period. For each patient, the side of the face to receive lotion A is randomly selected; the other side receives lotion B. The bottles of medication have coded labels so that neither the patient nor the physician knows which bottle contains A and which contains B—that is, in addition to blocking, the experiment also makes use of blinding.<sup>13</sup> This example, with blocks of size 2, is an example of pairing: The left side of the face is paired with the right side of the face. We have considered the analysis of paired data in Chapter 8. ■

**Example 11.6.4**

**Blocking in an Agricultural Field Study** When comparing several varieties of grain, an agronomist will generally plant many field plots of each variety and measure the yield of each plot. Differences in yields may reflect not only genuine differences among the varieties, but also differences among the plots in soil fertility, pH, water-holding capacity, and so on. Consequently, the spatial arrangement of the plots in the field is important. An efficient way to use the available field area is to divide the field into large regions—the blocks—and to subdivide each block into several plots. Within each block the various varieties of grain are then randomly allocated to the plots, with a separate randomization done for each block. For instance, suppose we want to test four varieties of barley. Then each block would contain four plots. The resulting randomized allocation might look like Figure 11.6.2, which is a schematic map of the field. The “treatments”  $T_1$ ,  $T_2$ ,  $T_3$ , and  $T_4$  are the four varieties of barley. ■

**Figure 11.6.2** Layout of an agricultural randomized blocks design

## Creating the Blocks

As the preceding examples show, blocking is a way of *organizing* the inherent variation that exists among experimental units. Ideally, the blocking should be arranged so as to increase the information available from the experiment. To achieve this goal, *the experimenter should try to create blocks that are as homogeneous within themselves as possible, so that the inherent variation between experimental units becomes, as far as possible, variation between blocks rather than within blocks.* This principle was illustrated in the preceding examples (e.g., in Example 11.6.2, where blocking by litter exploits the fact that littermates are more similar to each other than to nonlittermates). The following is another illustration.

### Example 11.6.5

**Agricultural Field Study** For the barley experiment of Example 11.6.4, how would agronomists determine the best arrangement or layout of blocks in a field? They would design the blocks to take advantage of any prior knowledge they may have of fertility patterns in the field. For instance, if they know that an east–west fertility gradient exists in the field (perhaps the field slopes from east to west, with the result that the west end has a thicker layer of good soil or receives better irrigation), then they might choose blocks as in Figure 11.6.2; the layout maximizes soil differences between the blocks and minimizes differences between plots within each block. (But even if a field appears to be uniform, blocking is usually used in agronomic experiments, because plots closer together in the field are generally more similar than plots farther apart.) ■

To add solidity to this example, let us look at a set of data from a randomized blocks experiment on barley. Each entry in Table 11.6.2 shows the yield (bushels of barley per acre) of a plot 3.5 ft wide by 80 ft long.<sup>14</sup>

**Table 11.6.2** Yield (lb) of barley

	Block 1	Block 2	Block 3	Block 4	Variety mean
Variety 1	93.5	66.6	50.5	42.4	63.3
Variety 2	102.9	53.2	47.4	43.8	61.8
Variety 3	67.0	54.7	50.0	40.1	53.0
Variety 4	86.3	61.3	50.7	46.4	61.2
Block Mean	87.4	59.0	49.7	43.2	

It appears from Table 11.6.2 that the yield potential of the blocks varies greatly; the data indicate a definite fertility gradient from block 1 to block 4. Because of the blocked design, comparison of the varieties is relatively unaffected by the fertility gradient. Of course, there also appears to be substantial variation within blocks. [You might find it an interesting exercise to peruse the data and ask yourself whether the observed differences between varieties are large enough to conclude that, for example, variety 1 is superior (in mean yield) to variety 3; use your intuition rather than a formal statistical analysis. The truth is revealed in Note 14.]

## The Randomization Procedure

Once the blocks have been created, the blocked allocation of experimental units is straightforward: It is as if a mini-experiment is conducted within each block. Randomization is carried out for each block separately, as illustrated in the following example.

**Example  
11.6.6**

**Agricultural Field Study** Consider the agricultural field experiment of Example 11.6.4. In block 1, let us label the plots 1, 2, 3, 4, from north to south (see Figure 11.6.2); we will allocate one plot to each variety. The allocation proceeds as for the completely randomized design, by choosing plots at random from the four, and assigning the first plot chosen to  $T_1$ , the second to  $T_2$ , and so on. For instance, using a computer to randomly permute the numbers 1 through 4 (or even shuffled cards numbered 1 through 4) we might obtain the sequence 4, 3, 1, 2 which would lead to the following treatment allocation.

Block 1

$T_1$ : Plot 4

$T_2$ : Plot 3

$T_3$ : Plot 1

$T_4$ : Plot 2

This is in fact the assignment shown in Figure 11.6.2 for block 1. We can then repeat this procedure for blocks 2, 3, and so on. ■

### Analyzing Data from a Randomized Block Experiment

In the same way we cannot use a two-sample  $t$  test when data are paired, when an experiment has been blocked, we no longer can use our ANOVA methods of Section 11.4. Instead, we will use a **randomized blocks ANOVA** model. We will illustrate the concepts as we reconsider the alfalfa and acid rain experiment of Example 11.6.1 in which the researchers blocked the experiment based on rows of cups placed parallel to a window so that each block has roughly the same light exposure. The data are given in Table 11.6.3 and are graphed in Figure 11.6.3.

**Table 11.6.3** Alfalfa plant height after five days (cm)

	High acid	Low acid	Control	Block mean
Block 1	1.30	1.78	2.67	1.917
Block 2	1.15	1.25	2.25	1.550
Block 3	0.50	1.27	1.46	1.077
Block 4	0.30	0.55	1.66	0.837
Block 5	1.30	0.80	0.80	0.967
Treatment mean = $\bar{y}_i$	0.910	1.130	1.768	
$n$	5	5	5	

Our usual ANOVA null hypothesis for comparing  $I$  populations or treatments is

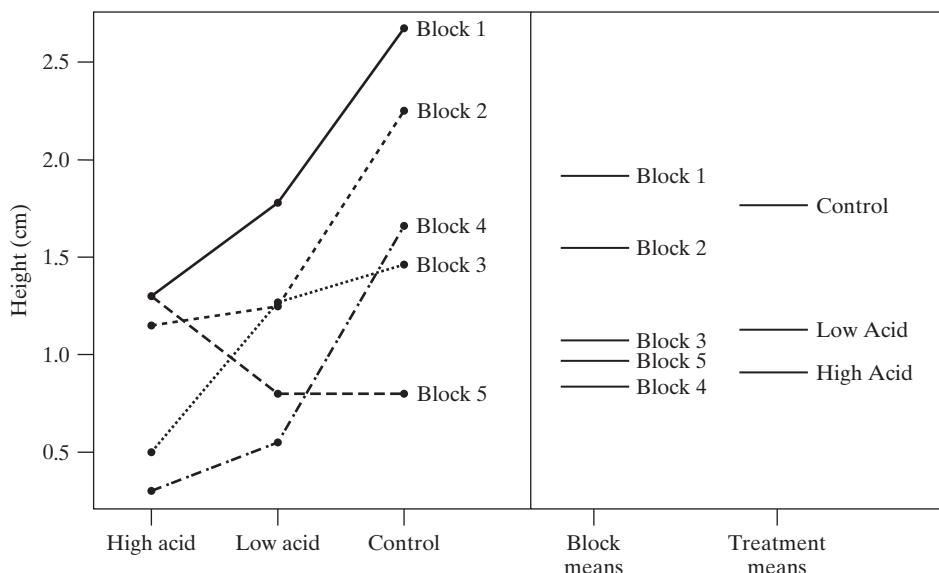
$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

**Example  
11.6.7**

**Alfalfa and Acid Rain** The null hypothesis for the alfalfa growth experiment is that acid has no effect on five-day growth. (We can make a strong causal claim like this because this was an experiment.) More directly, the null hypothesis is that the mean five-day growth is the same for all three treatments (high acid, low acid, and control).

$$H_0: \mu_1 = \mu_2 = \mu_3$$

**Figure 11.6.3** Dotplots of the alfalfa growth data with a summary of block and treatment means



This hypothesis can be tested with an analysis of variance  $F$  test, but first we want to remove the variability in the data that is due to differences between the blocks. To do this, we extend the ANOVA model presented in Section 11.3 to the following model:

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$$

In this model  $y_{ijk}$  is the  $k$ th observation when treatment  $i$  is applied in block  $j$ . (In Example 11.6.1 there is only one observation for each treatment in each block, but in general there might be more than one.) Here, as before,  $\mu$  represents the grand population mean and the term  $\tau_i$  represents the effect of group  $i$  (that is, treatment  $i$ ). The new term in the model is  $\beta_j$ , which represents the effect of the  $j$ th block.

## Visualizing the Block Effects

To visualize how blocking affects our ANOVA, we can think of our model in a slightly different way:

$$(y_{ijk} - \tau_i) = \mu + \beta_j + \varepsilon_{ijk}$$

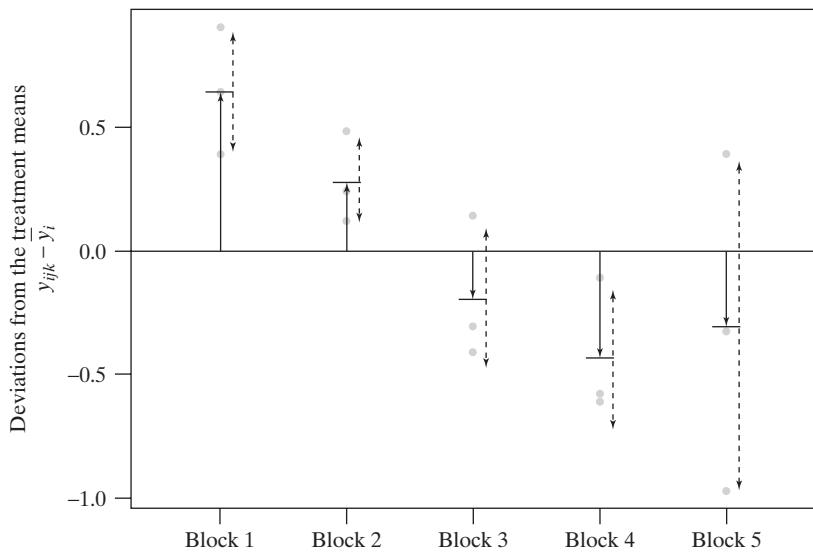
The left-hand side of the equation describes the data after treatment effects have been removed. With our data we estimate this left-hand side as

$$y_{ijk} - \hat{\tau}_i = y_{ijk} - \bar{y}_{i\bullet}$$

That is, within each treatment group, the treatment mean is subtracted from each data value.\* We've seen this before—in the context of a one-way ANOVA (Section 11.2) we called these deviations or residuals. Figure 11.6.4 is a plot of the deviations from the treatment means for the alfalfa data broken down by block. We can see that there is still a lot of structure in the data: The mean deviations in blocks 1 and 2 are greater than zero while blocks 3, 4, and 5 are below zero (corresponding to above average growth near the window and below average growth farther from the

\*Here we write  $\bar{y}_{i\bullet}$  rather than  $\bar{y}_i$  to distinguish the treatment means from the block means  $\bar{y}_{j\bullet}$ .

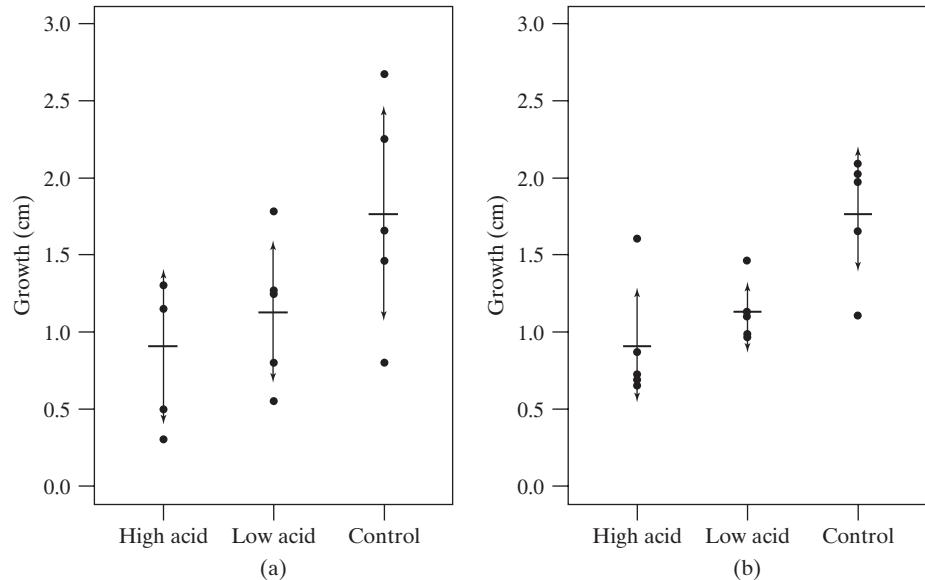
**Figure 11.6.4** Deviations from the treatment means for the alfalfa growth data by blocks



window). The fact that these mean deviations are not all zero is a consequence of the variability due to the blocks. Soon we shall describe how to measure the variability of these mean deviations for the blocks through the **mean squares for blocks**, or **MS(blocks)**.

To visualize how accounting for this block-to-block variation improves our ability to detect treatment effects, consider the alfalfa and acid rain data graphed in Figure 11.6.5. Figure 11.6.5(a) displays the growth data for each treatment group and simply ignores the blocks entirely while Figure 11.6.5(b) displays the growth data after adjusting for the estimated block effects.\* While the variability among the

**Figure 11.6.5** Visualizing the effect of blocking when comparing mean growth under the three acid treatments in the alfalfa experiment. Plot (a) displays the raw growth data while (b) displays the growth data after adjusting for the estimated block effects. Treatment means are indicated by horizontal lines and within-group standard deviations by arrows



\*To account for the blocking, the adjusted growth data on the y-axis for each treatment group is computed as  $y_{ijk} - \bar{y}_{\bullet j}$ .

treatment means is unchanged between the plots, we observe that the variability within the treatment groups is much smaller after accounting for the blocks and thus the differences among the treatments are more pronounced.

## The One-Way Randomized Complete Block *F* Test

Recall that the ANOVA *F* test is a ratio that compares the variability among the treatment means to the within-group variability. As seen in Figure 11.6.5, accounting for the blocks has reduced the within-group variability and will thus increase the *F* statistic value. We now briefly discuss the computations involved in computing the ANOVA table for the randomized complete block *F* test.

In Section 11.2 for a one-way ANOVA, we discussed how the total sum of squares, SS(total), is broken down into SS(between), which measures variability attributed to differences among the treatment means, and SS(within), which measures unexplained random variation in the data. For a randomized blocks experiment, we write SS(treatments) rather than SS(between) to describe the variability between treatment means to be clear that we're measuring variability between treatments and not blocks. For a randomized blocks experiment we also split the one-way ANOVA SS(within) into two parts: SS(blocks), which measures variability among the block means, and SS(within), which measures the remaining unexplained variation in the data. Thus, we have

$$\begin{array}{ll} \text{One-way ANOVA:} & \text{SS(total)} = \text{SS(within)} + \text{SS(between)} \\ & \swarrow \qquad \qquad \searrow \\ \text{One-way ANOVA with blocks: } & \text{SS(total)} = \text{SS(within)} + \text{SS(blocks)} + \text{SS(treatments)} \end{array}$$

Usually we are not interested in testing a hypothesis about the blocks, but nonetheless we want to take into consideration the effect that blocking has on the response variable. Refining the one-way ANOVA by calculating SS(blocks) accomplishes this goal and furthermore, if blocks are chosen wisely, can lead to more powerful tests.

Computing the sums of squares is typically left to a computer and rarely performed by hand. Nonetheless, the formulas are worth noting as they mathematically reveal how the blocks are being accounted for.

The **mean squares between blocks** is calculated in a manner similar to our computation of MS(between) from the one-way ANOVA of Section 11.2. Roughly speaking, we compute a sort of weighted variance of the block means in which we weight the differences between a block mean and the overall mean by the block sample size. If we define the average of the observations in block  $j$  to be  $\bar{y}_{\bullet j}$  and we let  $m_j$  denote the number of observations in block  $j$ , then the mean squares due to blocks is defined as follows:

### Mean Squares between Blocks

$$\text{MS(blocks)} = \frac{\sum_{j=1}^J m_j (\bar{y}_{\bullet j} - \bar{y})^2}{J - 1}$$

Analogous to our formulas in Section 11.2 we define SS(blocks) and df(blocks) as the numerator and denominator of MS(blocks) as follows:

### Sum of Squares and df between Blocks

$$\begin{aligned} \text{SS(blocks)} &= \sum_{j=1}^J m_j (\bar{y}_{\cdot j} - \bar{\bar{y}})^2 \\ \text{df(blocks)} &= J - 1 \end{aligned}$$

As noted previously, the blocking reduces MS(within). To compute MS(within) for the randomized complete block experiment we compute

$$\text{SS(within)} = \text{SS(total)} - \text{SS(treatment)} - \text{SS(blocks)}$$

where SS(treatment) and SS(total) are computed as in Section 11.2. As sums of squares are always nonnegative values, the preceding formula shows directly how the blocks reduce the within-group variability.

Similarly, to compute df(within) for the randomized complete block experiment, we have

$$\begin{aligned} \text{df(within)} &= \text{df(total)} - \text{df(treatment)} - \text{df(blocks)} \\ &= (n_{\cdot} - 1) - (I - 1) - (J - 1) \\ &= n_{\cdot} - I - J + 1 \end{aligned}$$

#### **Example 11.6.8**

**Alfalfa and Acid Rain** For the alfalfa growth data in Table 11.6.2, the total of all the observations is  $1.30 + 1.15 + \dots + 0.80 = 19.04$  and the grand mean is

$$\bar{\bar{y}} = \frac{19.04}{15} = 1.269$$

We calculate

$$\text{SS(treatments)} = 5(0.910 - 1.269)^2 + 5(1.130 - 1.269)^2 + 5(1.768 - 1.269)^2 = 1.986$$

Since  $I = 3$ , we have

$$\text{df(treatments)} = 3 - 1 = 2$$

so that

$$\text{MS(treatments)} = \frac{1.986}{2} = 0.993$$

We calculate

$$\begin{aligned} \text{SS(blocks)} &= 3(1.917 - 1.269)^2 + 3(1.550 - 1.269)^2 \\ &\quad + 3(1.077 - 1.269)^2 + 3(1.837 - 1.269)^2 \\ &\quad + 3(1.967 - 1.269)^2 \\ &= 2.441 \end{aligned}$$

Since  $J = 5$ , we have

$$\text{df(blocks)} = 5 - 1 = 4$$

and

$$\text{MS(blocks)} = \frac{2.441}{4} = 0.610$$

The total sum of squares is found as  $(1.30 - 1.269)^2 + \dots + (0.80 - 1.269)^2 = 5.879$ .

By subtraction, we compute SS(within):

$$\begin{aligned} \text{SS}(\text{within}) &= \text{SS}(\text{total}) - \text{SS}(\text{treatments}) - \text{SS}(\text{blocks}) \\ &= 5.879 - 1.986 - 2.441 = 1.452 \end{aligned}$$

Similarly, we compute df(within) as

$$\text{df}(\text{within}) = \text{df}(\text{total}) + \text{df}(\text{treatments}) + \text{df}(\text{blocks})$$

which in this case gives us  $14 - 2 - 4 = 8$ .

$$\text{Thus, } \text{MS}(\text{within}) = \frac{1.452}{8} = 0.182.$$

■

The sums of squares, degrees of freedom, and resulting mean squares are collected in an expanded ANOVA table, which includes a line for the effect of the blocks.

To test the null hypothesis, we calculate

$$F_s = \frac{\text{MS}(\text{treatments})}{\text{MS}(\text{within})}$$

and reject  $H_0$  if the  $P$ -value is too small.

### Example 11.6.9

**Alfalfa and Acid Rain** For the alfalfa growth data of Example 11.6.1, the ANOVA summary is given in Table 11.6.4. The  $F$  statistic is  $0.993/0.182 = 5.47$ , with degrees of freedom 2 for the numerator and 8 for the denominator. From Table 10 we bracket the  $P$ -value as  $0.02 < P\text{-value} < 0.05$ . (Using a computer gives  $P\text{-value} = 0.0318$ .) The  $P$ -value is small, indicating that the differences between the three sample means are greater than would be expected by chance alone. There is significant evidence that acid affects the growth of alfalfa plants. (It is worth noting that if we ignore the blocks and conduct an erroneous one-way ANOVA, we would find  $P\text{-value} = 0.0842$ , which would not provide significant evidence for an acid effect at  $\alpha = 0.05$ ). ■

**Table 11.6.4** ANOVA table for alfalfa experiment

Source	df	SS	MS	$F$ ratio
Between treatments	2	1.986	0.993	5.47
Between blocks	4	2.441	0.610	
Within groups	8	1.452	0.182	
Total	14	4.278		

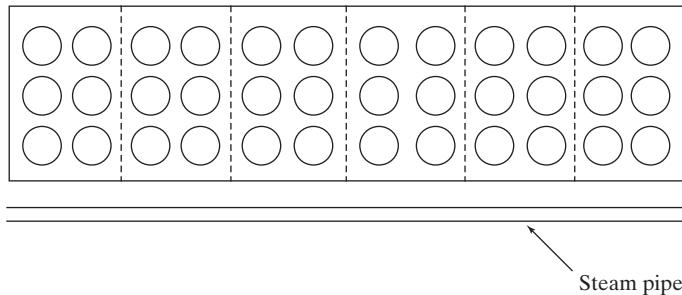
### Exercises 11.6.1–11.6.10

(Note: In several of these exercises you are asked to prepare a randomized allocation. For this purpose you can use either Table 1, random digits from your calculator, or a computer.)

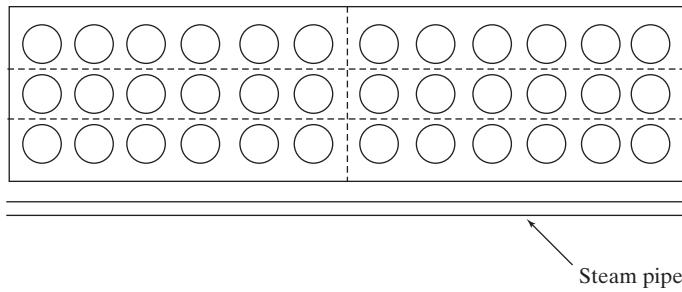
**11.6.1** In an experiment to compare six different fertilizers for tomatoes, 36 individually potted seedlings are to

be used, 6 to receive each fertilizer. The tomato plants will be grown in a greenhouse, and the total yield of tomatoes will be observed for each plant. The experimenter has decided to use a randomized blocks design: The pots are to be arranged in six blocks of 6 plants each on the greenhouse bench. Two possible arrangements of the blocks are shown in the accompanying figure.

Arrangement I:



Arrangement II:



One factor that affects tomato yield is temperature, which cannot be held exactly constant throughout the greenhouse. In fact, a temperature gradient across the bench is likely. Heat for the greenhouse is provided by a steam pipe that runs lengthwise under one edge of the bench, and so the side of the bench near the steam pipe is likely to be warmer.

- Which arrangement of blocks (I or II) is better? Why?
- Prepare a randomized allocation of treatments to the pots within each block. (Refer to Example 11.6.4 as a guide; assume that the assignments of seedlings to pots and of pots to positions within the block have already been made.)

**11.6.2** An experiment on vitamin supplements is to be conducted on young piglets, using litters as blocks in a randomized blocks design. There will be five treatments: four types of supplement and a control. Thus, five piglets from each litter will be used. The experiment will include five litters. Prepare a randomized blocks allocation of piglets to treatments. (Refer to Example 11.6.4 as a guide.)

**11.6.3** Refer to the vitamin experiment of Exercise 11.6.2. Suppose a colleague of the experimenter proposes an alternative design: All pigs in a given litter are to receive the same treatment, with the five litters being randomly allocated to the five treatments. He points out that

his proposal would save labor and greatly simplify the record keeping. If you were the experimenter, how would you reply to this proposal?

**11.6.4** In a pharmacological experiment on eating behavior in rats, 18 rats are to be randomly allocated to three treatment groups:  $T_1$ ,  $T_2$ , and  $T_3$ . While under observation, the animals will be kept in individual cages in a rack. The rack has three tiers with six cages per tier. In spite of efforts to keep the lighting uniform, the lighting conditions vary somewhat from one tier to another (the bottom tier is darkest), and the experimenter is concerned about this because lighting is thought to influence eating behavior in rats. The following three plans are proposed for allocating the rats to positions in the rack (to be done after the allocation of rats to treatment groups):

Plan I. Randomly allocate the 18 rats to the 18 positions in the rack.

Plan II. Put all  $T_1$  rats on the first tier, all  $T_2$  rats on the second, and all  $T_3$  rats on the third tier.

Plan III. On each tier, put two  $T_1$  rats, two  $T_2$  rats, and two  $T_3$  rats.

Put these three plans in order, from best to worst. Explain your reasoning.

**11.6.5** An experimenter is planning an agricultural field experiment to compare the yields of 25 varieties of corn.

She will use a randomized blocks design with six blocks; thus, there will be 150 plots, and the yield of each plot must be measured. The experimenter realizes that the time required to harvest and weigh all the plots is so long that rain might interrupt the operation. If rain should intervene, there could be a yield difference between the harvests before and after the rain. The experimenter is considering the following plans.

**Plan I.** Harvest all plots of variety 1 first, all of variety 2 next, and so on.

**Plan II.** Harvest all plots of block 1 first, all of block 2 next, and so on.

Which plan is better? Why?

**11.6.6** For an experiment to compare two methods of artificial insemination in cattle, the following cows are available:

Heifers (14–15 months old): 8 animals

Young cows (2–3 years old): 8 animals

Mature cows (4–8 years old): 10 animals

The animals are to be randomly allocated to the two treatment groups, using the three age groups as blocks. Prepare a suitable allocation, randomly dividing each stratum into two equal groups.

**11.6.7** True or false (and say why): The primary reason for using a randomized blocks design in an experiment is to reduce bias.

**11.6.8** In an experiment to understand the impact of fish grazing on invertebrate populations in streams, researchers established nine observation channels in three streams (three channels per stream). Each of the three channels within a stream received one of three treatments: No fish were added, *Galaxias* fish were added, or *Trout* fish were added. (The channels were constructed with mesh to prevent fish from entering or leaving.) Twelve days after establishing the channels, the number of *Deleatidium* mayfly nymphs present in a specified region in the center of the channel were counted. The number of nymphs for each treatment in each creek follows.<sup>15</sup>

		CREEK		
		A	B	C
Treatment	No Fish	11	8	7
	Galaxias	9	4	4
	Trout	6	4	0

- (a) Identify the blocking, treatment (i.e., the explanatory variable of interest), and response variables in this study.

- (b) In the context of this problem, explain to someone who has never taken a statistics course how blocking may help better identify treatment differences should they exist.

### 11.6.9 (Continuation of 11.6.8)

- (a) The accompanying table is an (improper) ANOVA table for the data in Exercise 11.6.8. This analysis does not account for the blocking that was performed in the experiment. Based on this analysis, is there evidence that fish affect the number of mayfly nymphs present in the channels? Use  $\alpha = 0.05$ .

	DF	SUM SQ	MEAN SQ	F VALUE
Between groups	2	42.889	21.444	2.924
Within groups	6	44.000	7.333	
Total	8	86.889		

- (b) The proper ANOVA table for the data, which accounts for blocking, follows. Based on this proper analysis, is there evidence that fish affect the number of mayfly nymphs present in the channels? Use  $\alpha = 0.05$ .

	DF	SUM SQ	MEAN SQ	F VALUE
Between groups	2	42.889	21.444	16.783
Between blocks	2	38.889	19.444	15.217
Within groups	4	5.111	1.278	
Total	8	86.889		

- (c) Compute and compare  $s_{\text{pooled}}$  using the ANOVA table from parts (a) and (b). Why is one estimate larger than the other? What is  $s_{\text{pooled}}$  measuring in part (a)? In part (b)?

**11.6.10** Consider the experiment described in Exercise 11.6.8. In addition to measuring the number of mayfly nymphs at the end of 12 days, stones of the same size were removed from each channel and the algal ash free dry mass ( $\text{mg/cm}^2$ ) was measured for each of nine stones. These data produced  $\text{SS}(\text{blocks}) = 0.889$ ,  $\text{SS}(\text{within}) = 0.444$ , and  $\text{SS}(\text{total}) = 2.889$ .

- (a) Construct an ANOVA table similar to Table 11.6.4 to summarize these data.
- (b) Is there evidence that the presence or type of fish is associated with the mean algal ash free dry mass in the channels? Use  $\alpha = 0.05$ .
- (c) Can a causal conclusion be drawn from the analysis performed in part (b) based on these data? If so, what causal conclusion can be made? If not, explain why no causal conclusion is appropriate.

## 11.7 Two-Way ANOVA

### Factorial ANOVA

In a typical analysis of variance application there is a single explanatory variable or **factor** under study. For example, in the weight gain setting of Example 11.2.1, the factor is “type of diet,” which takes on three **levels**: diet 1, diet 2, and diet 3. However, some analysis of variance settings involves the simultaneous study of two or more factors. The following is an example.

---

**Example**  
**11.7.1**

**Growth of Soybeans** A plant physiologist investigated the effect of mechanical stress on the growth of soybean plants. Individually potted seedlings were randomly allocated to four treatment groups of 13 seedlings each. Seedlings in two groups were stressed by shaking for 20 minutes twice daily, while two control groups were not stressed. Thus, the first factor in the experiment was presence or absence of stress, with two levels: control or stress. Also, plants were grown in either low or moderate light. Thus, the second factor was amount of light, with two levels: low light or moderate light. This experiment is an example of a  $2 \times 2$  factorial experiment; it includes four treatments:

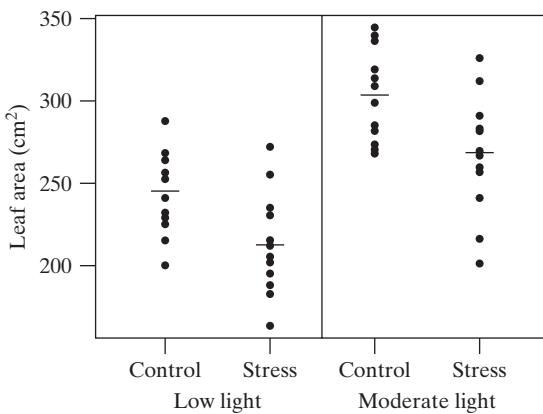
- Treatment 1: Control, low light
- Treatment 2: Stress, low light
- Treatment 3: Control, moderate light
- Treatment 4: Stress, moderate light

After 16 days of growth, the plants were harvested, and the total leaf area ( $\text{cm}^2$ ) of each plant was measured. The results are given in Table 11.7.1 and plotted in Figure 11.7.1.<sup>16</sup>

**Table 11.7.1** Leaf area ( $\text{cm}^2$ ) of soybean plants

	Treatment			
	Control, low light	Stress, low light	Control, moderate light	Stress, moderate light
264	235	314	283	
200	188	320	312	
225	195	310	291	
268	205	340	259	
215	212	299	216	
241	214	268	201	
232	182	345	267	
256	215	271	326	
229	272	285	241	
288	163	309	291	
253	230	337	269	
288	255	282	282	
230	202	273	257	
Mean	245.3	212.9	304.1	268.8
SD	27.0	29.7	26.9	35.2
<i>n</i>	13	13	13	13

**Figure 11.7.1** Leaf area of soybean plants receiving four different treatments. Group means indicated by (-)



There is evidence in Figure 11.7.1 that stress reduces leaf area. This is true under low light and under moderate light. Likewise, moderate light increases leaf area, whether or not the seedlings are stressed. ■

A model for this setting is

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$$

where  $y_{ijk}$  is the  $k$ th observation of level  $i$  of the first factor and level  $j$  of the second factor. The term  $\tau_i$  represents the effect of level  $i$  of the first factor (stress condition in Example 11.7.1) and now the term  $\beta_j$  represents the effect of level  $j$  of the second factor (light condition in Example 11.7.1).

When studying two factors within a single experiment it helps to organize the sample means in a table that reflects the structure of the experiment and to present the means in a graph that features this structure.

### Example 11.7.2

**Growth of Soybeans** Table 11.7.2 summarizes the data of Example 11.7.1. For example, when the first factor is at its first level (control) and the second factor is at its first level (low light), the sample mean is  $\bar{y}_{11} = 245.3$ . The format of this table permits us easily to consider the two factors—stress condition and light condition—separately and together. The last column shows the effect of light at each stress level. The numbers in this column confirm the visual impression of Figure 11.7.1: Moderate light increases average leaf area by roughly the same amount when the seedlings are stressed as it does when they are not stressed. Likewise, the last row ( $-32.4$  versus  $-35.3$ ) shows that the effect of stress is roughly the same at each level of light. ■

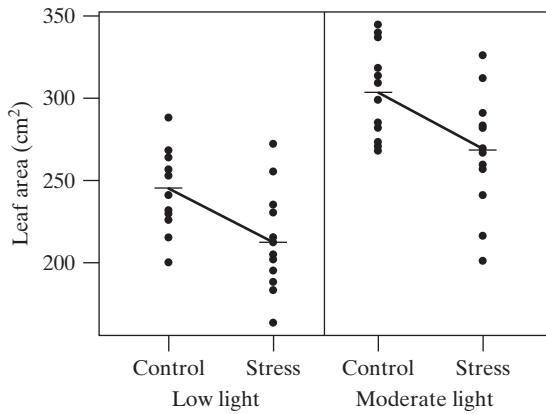
**Table 11.7.2** Mean leaf areas for soybean experiment

		Light condition		
		Low light	Moderate light	Difference
Shaking condition	Control	245.3	304.1	58.8
	Stress	212.9	268.8	55.9
	Difference	-32.4	-35.3	

If the joint influence of two factors is equal to the sum of their separate influences, the two factors are said to be **additive** in their effects. For instance, consider the soybean experiment of Example 11.7.1. If stress reduces mean leaf area by the same

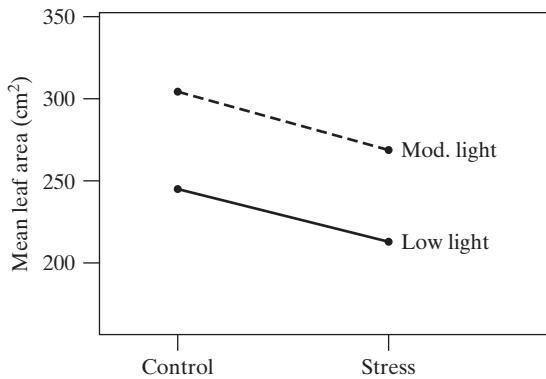
amount in either light condition, then the effect of stress (a negative effect in this case) is *added* to the effect of light. To visualize this additivity of effects, consider Figure 11.7.2, which shows the data with the four treatment means. The solid lines connecting treatment means are almost parallel because the data display a pattern of nearly perfect additivity.\*

**Figure 11.7.2** Data and treatment means for soybean experiment



When the effects of factors are additive we say that there is no **interaction** between the factors. A graph that displays only the treatment means is often called an interaction graph. Figure 11.7.3, which is a summary version of Figure 11.7.2, is an interaction graph highlighting the effect of stress on mean leaf area for the two light conditions. Analogous graphs can be made to draw the focus to comparing the effect of light on mean leaf area for the two stress conditions.

**Figure 11.7.3** Interaction graph for soybean experiment



Sometimes the effect that one factor has on a response variable depends on the level of a second factor. When this happens we say that the two factors interact in their effect on the response. The following is an example.

---

\*The difference between the mean leaf area for stress under low light (212.9) and the mean leaf area for control under low light of (245.3) is called the **simple effect** of shaking under low light. Thus, the simple effect of shaking under low light is  $212.9 - 245.3 = -32.4$ . Likewise, the simple effect of shaking under moderate light is  $268.8 - 304.1 = -35.3$ . A **main effect** is an average of simple effects. For example, the main effect of light is  $(-32.4 + -35.3)/2 = -33.85$ . The main effect of light is  $(58.8 + 55.9)/2 = 57.35$ .

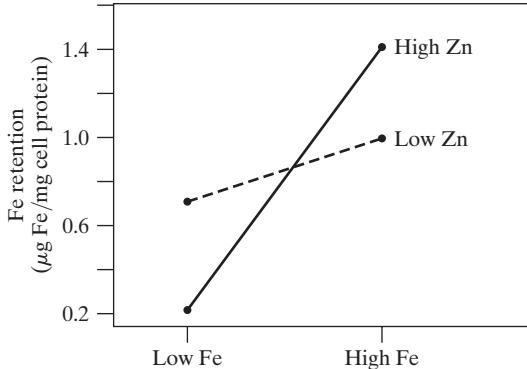
**Example  
11.7.3**

**Iron Supplements in Milk-Based Fruit Beverages** Iron and zinc fortification of milk-based fruit drinks are common practice. To better understand the effects of drink fortification on the cellular retention of iron, researchers conducted an experiment by fortifying milk-based fruit drinks with low and high levels of iron (Fe) and zinc (Zn). The drinks were digested in a simulated gastrointestinal tract and cellular iron retention was measured ( $\mu\text{g Fe/mg cell protein}$ ). Table 11.7.3 summarizes the data, which included eight observations for each combination of Fe and Zn supplementation levels.<sup>17</sup> Figure 11.7.4 is an interaction graph showing the four means. Note that when the Zn supplementation level is low, the effect of the Fe supplementation on cellular retention is much smaller than when the Zn supplementation level is high (i.e., the slopes of the two lines differ—the lines are not parallel). Thus, the effect of Fe supplementation on mean cellular retention depends on the amount of Zn supplementation used. We say that Fe and Zn interact in their effects on cellular retention. ■

**Table 11.7.3** Mean iron retention ( $\mu\text{g Fe/mg cell protein}$ ) for drink supplement experiment

		Zn Level		Difference	
Fe Level	Lo	Hi			
	0.707	0.215	−0.492		
	0.994	1.412	0.418		
Difference		0.287	1.197		

**Figure 11.7.4** Interaction graph for drink supplementation experiment



When we suspect that two factors interact in an ANOVA setting, we can extend our model by adding an interaction term:

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

Here the term  $\gamma_{ij}$  is the effect of the interaction between level  $i$  of the first factor and level  $j$  of the second factor. As before, if there are  $n$  total observations, then  $\text{df}(\text{total}) = n - 1$ . If there are  $I$  levels of the first factor, then it has  $I - 1$  degrees of freedom. Likewise, if there are  $J$  levels of the second factor, then it has  $J - 1$  degrees of freedom. There are  $(I - 1) \times (J - 1)$  interaction degrees of freedom. With  $I$  levels of the first factor and  $J$  levels of the second factor there are  $IJ$  treatment combinations. Thus,  $\text{df}(\text{within}) = n - IJ$ .\*

\*This is analogous to the definition of  $\text{df}(\text{within}) = n - I$  for one-way ANOVA from Section 11.2. In each setting  $\text{df}(\text{within}) = \text{total number of observations} - \text{number of treatments}$ .

A null hypothesis of interest is that all interaction terms are zero:

$$H_0: \gamma_{11} = \gamma_{12} = \cdots = \gamma_{IJ} = 0$$

To test this null hypothesis we calculate

$$F_s = \frac{\text{MS(interaction)}}{\text{MS(within)}}$$

and reject  $H_0$  if the  $P$ -value is too small.

**Example  
11.7.4**

**Iron Supplements in Milk-Based Fruit Beverages** Table 11.7.4 shows the analysis of variance results for the drink supplement experiment of Example 11.7.3. This table includes a line for the interaction term.\* There were eight observations at each combination of Fe and Zn supplementation level; thus  $n_\cdot = 32$  and  $\text{df(total)} = 31$ . In this example  $I = J = 2$ , so  $\text{df(Fe levels)} = \text{df(Zn levels)} = \text{df(interaction)} = 1$ . We can find  $\text{df(within)}$  by subtraction:  $\text{df(within)} = 31 - 1 - 1 - 1 = 28$ . (This agrees with the formula  $\text{df(within)} = n_\cdot - IJ = 32 - 2 \times 2$ .)

To test whether Fe and Zn supplementation levels interact we use the  $F$  ratio  $1.6555/0.0019 = 871.3$ , which has degrees of freedom 1 for the numerator and 28 for the denominator. From Table 10 we bracket the  $P$ -value as  $P\text{-value} < 0.0001$ . The  $P$ -value is extremely small, indicating that the interaction pattern seen in Figure 11.7.4 is more pronounced than would be expected by chance alone. Thus, we reject  $H_0$ . ■

<b>Table 11.7.4</b> ANOVA table for drink supplement experiment				
Source	df	SS	MS	$F$ ratio
Between Fe levels	1	4.4023	4.4023	2317.0
Between Zn levels	1	0.0109	0.0109	5.736
Interaction	1	1.6555	1.6555	871.3
Within groups	28	0.0523	0.0019	
Total	31	6.1210		

The concept of interaction occurs throughout biology. The terms “synergism” and “antagonism” describe interactions between biological agents. The term “epistasis” describes interaction between genes at two loci.

When interactions are present, as in Example 11.7.3, the main effects of factors don’t have their usual interpretations. Regarding Example 11.7.3, it is difficult to state the independent effect of Fe because the nature and magnitude of the effect depends on the particular level of Zn supplementation. Because of this, we usually test for the presence of interactions first. If interactions are present, as in the drink supplementation example, then we often stop the analysis at this stage. If no evidence for an interaction effect is found (that is, if we do not reject  $H_0$ ), then we proceed to testing the main effects of the individual factors. The following example illustrates this process.

---

\*The ANOVA formulas that are used to calculate the sum of squares due to interaction are rather messy and aren’t presented here. In particular, it matters whether or not the design is “balanced.” The drink supplementation experiment is balanced in that there are eight observations in each of the four combinations of factor levels shown in Table 11.7.3. However, unbalanced designs, which lead to complicated calculations and analyses, are possible. We rely here on computer software to calculate the necessary sums of squares.

**Example  
11.7.5**

**Growth of Soybeans** Table 11.7.5 is an analysis of variance table for the soybean growth data of Example 11.7.1. The null hypothesis

$$H_0: \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$$

is tested with the  $F$  ratio

$$F_s = \frac{\text{MS(interaction)}}{\text{MS(within)}} = \frac{26.3}{895.34} = 0.029$$

Looking in Table 10 with degrees of freedom 1 and 12, we see that the  $P$ -value is greater than 0.20; thus there is no significant evidence for an interaction and we do not reject  $H_0$ .

Since there is no evidence of interactions, we test the main effect of stress level. Here the  $F$  ratio is

$$F_s = \frac{\text{MS(between stress levels)}}{\text{MS(within)}} = \frac{14858.5}{895.34} = 16.6$$

This is highly significant (i.e., the  $P$ -value is very small) and we reject  $H_0$ .

Likewise, the test for the main effect of light levels has an  $F$  ratio of

$$F_s = \frac{\text{MS(between light levels)}}{\text{MS(within)}} = \frac{42751.6}{895.34} = 47.75$$

Again, this is highly significant and we reject  $H_0$ . ■

**Table 11.7.5** ANOVA table for soybean growth experiment

Source	df	SS	MS	F ratio
Between stress levels	1	14858.5	14858.5	16.60
Between light levels	1	42751.6	42751.6	47.75
Interaction	1	26.3	26.3	0.029
Within groups	48	42976.3	895.34	
Total	51	100612.7		

Interaction graphs can be used when there are more than two levels for a factor, as in the next example.

**Example  
11.7.6**

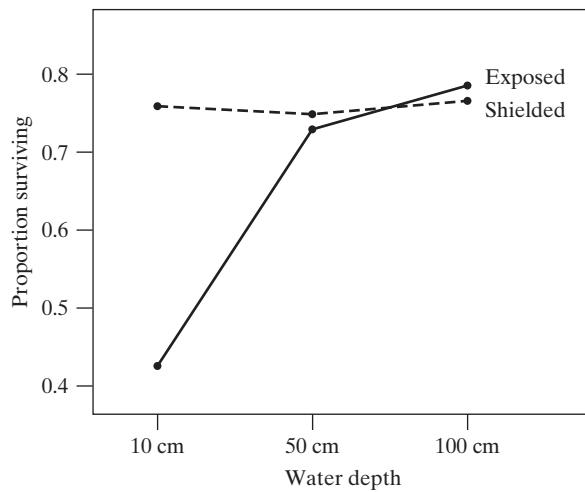
**Toads** Researchers studied the effect that exposure to ultraviolet-B radiation has on the survival of embryos of the western toad *Bufo boreas*. They conducted an experiment in which several *B. boreas* embryos were placed at one of three water depths—10 cm, 50 cm, or 100 cm—and one of two radiation settings—exposed to UV-B radiation or shielded. The response variable was the percentage of embryos surviving to hatching. Table 11.7.6 summarizes the data, which

**Table 11.7.6** Percent embryos surviving for toads experiment

		UV-B		Difference
		Exposed	Shielded	
Water depth	10 cm	0.425	0.759	0.334
	50 cm	0.729	0.748	0.019
	100 cm	0.785	0.766	-0.019

included four observations at each combination of depth and UV-B exposure. Figure 11.7.5 is an interaction graph showing the six means. The presence of interactions here is readily apparent. Table 11.7.7 summarizes the analysis of variance.<sup>18</sup>

**Figure 11.7.5** Interaction graph for toad experiment



**Table 11.7.7** ANOVA table for toad experiment

Source	df	SS	MS	F ratio
Between water depths	2	0.150676	0.075338	13.92
Between UV-B levels	1	0.074371	0.074371	13.74
Interaction	2	0.150185	0.075093	13.88
Within groups	18	0.097401	0.005411	
Total	23	0.472633		

The topic of interactions is also discussed in Section 11.8.

## Exercises 11.7.1–11.7.6

- 11.7.1** A plant physiologist investigated the effect of flooding on root metabolism in two tree species: flood-tolerant river birch and the intolerant European birch. Four seedlings of each species were flooded for one day and four were used as controls. The concentration of adenosine triphosphate (ATP) in the roots of each plant was measured. The data (nmol ATP per mg tissue) are shown in the table.<sup>19</sup>

	RIVER BIRCH		EUROPEAN BIRCH	
	FLOODED	CONTROL	FLOODED	CONTROL
	1.45	1.70	0.21	1.34
	1.19	2.04	0.58	0.99
	1.05	1.49	0.11	1.17
	1.07	1.91	0.27	1.30
Mean	1.19	1.785	0.2925	1.20

Prepare an interaction graph (like Figure 11.7.3).

**11.7.2** Consider the data from Exercise 11.7.1. For these data,  $SS(\text{species of birch}) = 2.19781$ ,  $SS(\text{flooding}) = 2.25751$ ,  $SS(\text{interaction}) = 0.097656$ , and  $SS(\text{within}) = 0.47438$ .

- Construct the ANOVA table.
- Carry out an  $F$  test for interactions; use  $\alpha = 0.05$ .
- Test the null hypothesis that species has no effect on ATP concentration. Use  $\alpha = 0.01$ .
- Assuming that each of the four populations has the same standard deviation, use the data to calculate an estimate of that standard deviation.

**11.7.3** A completely randomized double-blind clinical trial was conducted to compare two drugs, ticrynafen (T) and hydrochlorothiazide (H), for effectiveness in treatment of high blood pressure. Each drug was given at either a low or a high dosage level for six weeks. The accompanying table shows the results for the drop (baseline minus final value) in systolic blood pressure (mm Hg).<sup>20</sup>

TICRYNAFEN (T)		HYDROCHLOROTHIAZIDE (H)	
LOW DOSE	HIGH DOSE	LOW DOSE	HIGH DOSE
Mean	13.9	17.1	15.8
No. of patients	53	57	55

Prepare an interaction graph (like Figure 11.7.3).

**11.7.4** Consider the data from Exercise 11.7.3. The difference in response between T and H appears to be larger for the low dose than for the high dose.

- Carry out an  $F$  test for interactions to assess whether this pattern can be ascribed to chance variation.

Let  $\alpha = 0.10$ . For these data  $SS(\text{interaction}) = 31.33$  and  $SS(\text{within}) = 30648.81$ .

- Based on your results in part (a), is it sensible to examine and interpret the main effects of drug and of dose?

**11.7.5** Consider the data from Exercise 11.7.3. For these data,  $SS(\text{drug}) = 69.22$ ,  $SS(\text{dose}) = 330.00$ ,  $SS(\text{interaction}) = 31.33$ , and  $SS(\text{within}) = 30648.81$ .

- Construct the ANOVA table.
- Carry out a test of the null hypothesis that the effects of the two drugs (T and H) are equal. Let  $\alpha = 0.05$ .

**11.7.6** In a study of lettuce growth, 36 seedlings were randomly allocated to receive either high or low light and to be grown in either a standard nutrient solution or one containing extra nitrogen. After 16 days of growth, the lettuce plants were harvested and the dry weight of the leaves was determined for each plant. The accompanying table shows the mean leaf dry weight (gm) of the 9 plants in each treatment group.<sup>21</sup>

	NUTRIENT SOLUTION	
	STANDARD	EXTRA NITROGEN
Low light	2.16	3.09
High light	3.26	4.48

For these data,  $SS(\text{nutrient solution}) = 10.4006$ ,  $SS(\text{light}) = 13.95023$ ,  $SS(\text{interaction}) = 0.18923$ , and  $SS(\text{within}) = 11.1392$ .

- Construct the ANOVA table.
- Carry out an  $F$  test for interactions; use  $\alpha = 0.05$ .
- Test the null hypothesis that nutrient solution has no effect on weight. Use  $\alpha = 0.01$ .

## 11.8 Linear Combinations of Means (Optional)

In many studies, interesting questions can be addressed by considering linear combinations of the group means. A **linear combination**  $L$  is a quantity of the form

$$L = m_1\bar{y}_1 + m_2\bar{y}_2 + \cdots + m_I\bar{y}_I$$

where the  $m$ 's are multipliers of the  $\bar{y}$ 's.

### Linear Combinations for Adjustment

One use of linear combinations is to “adjust” for an extraneous variable, as illustrated by the following example.

#### Example 11.8.1

**Forced Vital Capacity** One measure of lung function is forced vital capacity (FVC), which is the maximal amount of air a person can expire in one breath. In a public health survey, researchers measured FVC in a large sample of people. The results for male ex-smokers, stratified by age, are shown in Table 11.8.1.<sup>22</sup>

<b>Table 11.8.1</b> FVC in male ex-smokers			
Age (years)	FVC (liters)		
	n	Mean	SD
25–34	83	5.29	0.76
35–44	102	5.05	0.77
45–54	126	4.51	0.74
55–64	97	4.24	0.80
65–74	73	3.58	0.82
25–74	481	4.56	

Suppose it is desired to calculate a summary value for FVC in male ex-smokers. One possibility would be simply to calculate the grand mean of the 481 observed values, which is 4.56 liters. But the grand mean has a serious drawback: It cannot be meaningfully compared with other populations that may have different age distributions. For instance, suppose we were to compare ex-smokers with nonsmokers; the observed difference in FVC would be distorted because ex-smokers as a group are (not surprisingly) older than nonsmokers. A summary measure that does not have this disadvantage is the “age-adjusted” mean, which is an estimate of the mean FVC value in a reference population with a specified age distribution. To illustrate, we will use the reference distribution in Table 11.8.2, which is (approximately) the distribution for the entire U.S. population.<sup>23</sup>

<b>Table 11.8.2</b> Age distribution in reference population	
Age	Relative frequency
25–34	0.23
35–44	0.22
45–54	0.24
55–64	0.22
65–74	0.09

The “age-adjusted” mean FVC value is the following linear combination:

$$L = 0.23\bar{y}_1 + 0.22\bar{y}_2 + 0.24\bar{y}_3 + 0.22\bar{y}_4 + 0.09\bar{y}_5$$

Note that the multipliers ( $m$ 's) are the relative frequencies in the reference population. From Table 11.8.1, the value of  $L$  is

$$\begin{aligned} L &= (0.23)(5.29) + (0.22)(5.05) + (0.24)(4.51) + (0.22)(4.24) + (0.09)(3.58) \\ &= 4.67 \text{ liters} \end{aligned}$$

This value is an estimate of the mean FVC in an idealized population of people who are biologically like male ex-smokers, but whose age distribution is that of the reference population. ■

## Contrasts

A linear combination whose multipliers ( $m$ 's) add to zero is called a **contrast**. The following example shows how contrasts can be used to describe the results of an experiment.

**Example  
11.8.2**

**Growth of Soybeans** Table 11.8.3 shows the treatment means and sample sizes for the soybean growth experiment of Example 11.6.8. We can use contrasts to describe the effects of stress in the two temperature conditions.

Table 11.8.3 Soybean growth data		
Treatment	Mean leaf area ( $\text{cm}^2$ )	$n$
1. Control, low light	245.3	13
2. Stress, low light	212.9	13
3. Control, moderate light	304.1	13
4. Stress, moderate light	268.8	13

- (a) First, note that an ordinary pairwise difference is a contrast. For instance, to measure the effect of stress in low light we can consider the contrast

$$L = \bar{y}_1 - \bar{y}_2 = 245.3 - 212.9 = 32.4$$

For this contrast, the multipliers are  $m_1 = 1, m_2 = -1, m_3 = 0, m_4 = 0$ ; note that they add to zero.

- (b) To measure the effect of stress in moderate light we can consider the contrast

$$L = \bar{y}_3 - \bar{y}_4 = 304.1 - 268.8 = 35.3$$

For this contrast, the multipliers are  $m_1 = 0, m_2 = 0, m_3 = 1, m_4 = -1$ .

- (c) To measure the overall effect of stress, we can average the contrasts in parts (a) and (b) to obtain the contrast

$$\begin{aligned} L &= \frac{1}{2}(\bar{y}_1 - \bar{y}_2) + \frac{1}{2}(\bar{y}_3 - \bar{y}_4) \\ &= \frac{1}{2}(32.4) + \frac{1}{2}(35.3) = 33.85 \end{aligned}$$

For this contrast, the multipliers are  $m_1 = \frac{1}{2}, m_2 = -\frac{1}{2}, m_3 = \frac{1}{2}, m_4 = -\frac{1}{2}$ . ■

## Standard Error of a Linear Combination

Each linear combination  $L$  is an estimate, based on the  $\bar{y}$ 's, of the corresponding linear combination of the population means ( $\mu$ 's). As a basis for statistical inference, we need to consider the standard error of a linear combination, which is calculated as follows.

### Standard Error of $L$

The standard error of the linear combination

$$L = m_1\bar{y}_1 + m_2\bar{y}_2 + \cdots + m_I\bar{y}_I$$

is

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\sum_{i=1}^I \frac{m_i^2}{n_i}}$$

where  $s_{\text{pooled}} = \sqrt{\text{MS}(\text{within})}$  from the ANOVA.

The SE can be written explicitly as

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\left( \frac{m_1^2}{n_1} + \frac{m_2^2}{n_2} + \dots + \frac{m_I^2}{n_I} \right)}$$

If all the sample sizes ( $n_i$ ) are equal, the SE can be written as

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\frac{(m_1^2 + m_2^2 + \dots + m_I^2)}{n}} = s_{\text{pooled}} \sqrt{\frac{1}{n} \sum_{i=1}^I m_i^2}$$

The following two examples illustrate the application of the standard error formula.

### Example 11.8.3

**Forced Vital Capacity** For the linear combination  $L$  defined in Example 11.8.1, we find that

$$\begin{aligned} \sum_{i=1}^I \frac{m_i^2}{n_i} &= \frac{0.23^2}{83} + \frac{0.22^2}{102} + \frac{0.24^2}{126} + \frac{0.22^2}{97} + \frac{0.09^2}{73} \\ &= 0.0021789 \end{aligned}$$

The ANOVA for these data yields  $s_{\text{pooled}} = \sqrt{0.59989} = 0.77453$ . Thus, the standard error of  $L$  is

$$\text{SE}_L = 0.77453 \sqrt{0.0021789} = 0.0362$$

### Example 11.8.4

**Growth of Soybeans** For the linear combination  $L$  defined in Example 11.8.2(a), we find that

$$\sum_{i=1}^I m_i^2 = (1)^2 + (-1)^2 + (0)^2 + (0)^2 = 2$$

so that

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\frac{2}{13}}$$

## Confidence Intervals

Linear combinations of means can be used for testing hypotheses and for constructing confidence intervals. Critical values are obtained from Student's  $t$  distribution with

$$\text{df} = \text{df}(\text{within})$$

from the ANOVA.\* Confidence intervals are constructed using the familiar Student's  $t$  format. For instance, a 95% confidence interval is

$$L \pm t_{0.025} \text{SE}_L$$

The following example illustrates the construction of the confidence interval.

### Example 11.8.5

**Growth of Soybeans** Consider the contrast defined in Example 11.8.2(c):

$$L = \frac{1}{2} (\bar{y}_1 - \bar{y}_2) + \frac{1}{2} (\bar{y}_3 - \bar{y}_4)$$

---

\*This method of determining critical values does not take account of multiple comparisons. See Section 11.9.

This contrast is an estimate of the quantity

$$\lambda = \frac{1}{2}(\mu_1 - \mu_2) + \frac{1}{2}(\mu_3 - \mu_4)$$

which can be described as the true (population) effect of stress, averaged over the light conditions. Let us construct a 95% confidence interval for this true difference.

We found in Example 11.8.2 that the value of  $L$  is

$$L = 33.85$$

To calculate  $\text{SE}_L$ , we first calculate

$$\sum_{i=1}^I \frac{m_i^2}{n_i} = \frac{\left(\frac{1}{2}\right)^2}{13} + \frac{\left(-\frac{1}{2}\right)^2}{13} + \frac{\left(\frac{1}{2}\right)^2}{13} + \frac{\left(-\frac{1}{2}\right)^2}{13} = \frac{1}{13}$$

From the ANOVA, which is shown in Table 11.8.4, we find that  $s_{\text{pooled}} = \sqrt{895.34} = 29.922$ ; thus,

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\sum_{i=1}^I \frac{m_i^2}{n_i}} = 29.922 \sqrt{\frac{1}{13}} = 8.299$$

**Table 11.8.4** ANOVA table for soybean growth experiment

Source	df	SS	MS	F ratio
Between stress depths	1	14858.5	14858.5	16.60
Between light levels	1	42751.6	42751.6	47.75
Interaction	1	26.3	26.3	0.029
Within groups	48	42976.3	895.34	
Total	51	100613		

From Table 4 with  $df = 40 \approx 48$ , we find  $t_{40,0.025} = 2.021$ . The confidence interval is

$$33.85 \pm (2.021)(8.299)$$

$$33.85 \pm 16.77$$

or  $(17.1, 50.6)$ .

We are 95% confident that the effect of stress, averaged over the light conditions, is to reduce the leaf area by an amount whose mean value is between  $17.1 \text{ cm}^2$  and  $50.6 \text{ cm}^2$ . ■

## t Tests

To test the null hypothesis that the population value of a contrast is zero, the test statistic is calculated as

$$t_s = \frac{L}{\text{SE}_L}$$

and the  $t$  test is carried out in the usual way. The  $t$  test will be illustrated in Example 11.8.6.

## Contrasts to Assess Interaction

Sometimes an investigator wishes to study the separate and joint effects of two or more factors on a response variable  $Y$ . In Section 11.7 the concept of interaction between two factors was introduced. Linear contrasts provide another way to study such interactions. The following is an example.

### Example 11.8.6

**Growth of Soybeans** In the soybean growth experiment (Example 11.6.8 and Example 11.8.2), the two factors of interest are stress condition and light level. Table 11.8.5 shows the treatment means, arranged in a new format that permits us easily to consider the factors separately and together.

**Table 11.8.5** Mean leaf areas for soybean experiment

		Light condition		
		Low light	Moderate light	Difference
Shaking condition	Control	245.3 (1)	304.1 (3)	58.8
	Stress	212.9 (2)	268.8 (4)	55.9
	Difference	-32.4	-35.3	

At each light level, the mean effect of stress can be measured by a contrast:

$$\text{Effect of stress in low light: } \bar{y}_2 - \bar{y}_1 = 212.9 - 245.3 = -32.4$$

$$\text{Effect of stress in moderate light: } \bar{y}_4 - \bar{y}_3 = 268.8 - 304.1 = -35.3$$

Now consider the question: Is the reduction in leaf area due to stress the same in both light conditions? One way to address this question is to compare  $(\bar{y}_2 - \bar{y}_1)$  versus  $(\bar{y}_4 - \bar{y}_3)$ ; the difference between these two values is a contrast:

$$\begin{aligned} L &= (\bar{y}_2 - \bar{y}_1) - (\bar{y}_4 - \bar{y}_3) \\ &= -32.4 - (-35.3) = 2.9 \end{aligned}$$

This contrast  $L$  can be used as the basis for a confidence interval or a test of hypothesis. We illustrate the test. The null hypothesis is

$$H_0: (\mu_2 - \mu_1) = (\mu_4 - \mu_3)$$

or, in words,

$H_0$ : The effect of stress is the same in the two light conditions.

For the preceding  $L$ ,  $\sum_{i=1}^I \frac{m_i^2}{n_i} = \frac{4}{13}$ , and the standard error is

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\sum_{i=1}^I \frac{m_i^2}{n_i}} = s_{\text{pooled}} \sqrt{\left( \frac{4}{13} \right)} = 29.922 \sqrt{\frac{4}{13}} = 16.6$$

The test statistic is

$$t_s = \frac{2.9}{16.6} = 0.2$$

From Table 4 with  $df = 40$  we find  $t_{40,0.20} = 1.303$ . The data provide virtually no evidence that the effect of stress is different in the two light conditions. This is consistent with the  $F$  test for interactions conducted in Example 11.7.5. ■

The statistical definition of interaction introduced in Section 11.7 and viewed through the lens of contrasts here is rather specialized. It is defined in terms of the observed variable rather than in terms of a biological mechanism. Further, interaction as measured by a contrast is defined by *differences* between means. In some applications the biologist might feel that ratios of means are more meaningful or relevant than differences. The following example shows that the two points of view can lead to different answers.

**Example  
11.8.7**

**Chromosomal Aberrations** A research team investigated the separate and joint effects in mice of exposure to high temperature ( $35^{\circ}\text{C}$ ) and injection with the cancer drug cyclophosphamide (CTX). A completely randomized design was used, with eight mice in each treatment group. For each animal, the researchers measured the incidence of a certain chromosomal aberration in the bone marrow; the result is expressed as the number of abnormal cells per 1,000 cells. The treatment means are shown in Table 11.8.6.<sup>24</sup>

		Injection	
		CTX	None
Temperature	Room	23.5	2.7
	High	75.4	20.9

Is the observed effect of CTX greater at room temperature or at high temperature? The answer depends on whether “effect” is measured absolutely or relatively.

Measured as a difference, the effect of CTX is

$$\text{Room temperature: } 23.5 - 2.7 = 20.8$$

$$\text{High temperature: } 75.4 - 20.9 = 54.5$$

Thus, the absolute effect of CTX is greater at the high temperature. However, this relationship is reversed if we express the effect of CTX as a ratio rather than as a difference:

$$\text{Room temperature: } \frac{23.5}{2.7} = 8.70$$

$$\text{High temperature: } \frac{75.4}{20.9} = 3.61$$

At room temperature CTX produces almost a ninefold increase in chromosomal aberrations, whereas at high temperature the increase is less than fourfold; thus, in relative terms, the effect of CTX is much greater at room temperature. ■

If the phenomenon under study is thought to be multiplicative rather than additive, so that relative rather than absolute change is of primary interest, then ordinary contrasts should not be used. One simple approach in this situation is to use a logarithmic transformation—that is, to compute  $Y' = \log(Y)$ , and then analyze  $Y'$  using contrasts. The motivation for this approach is that relations of constant *relative* magnitude in the  $Y$  scale become relations of constant *absolute* magnitude in the  $Y'$  scale.

## Exercises 11.8.1–11.8.10

**11.8.1** Refer to the FVC data of Example 11.8.1.

- Verify that the grand mean of all 481 FVC values is 4.56.
- Taking into account the age distribution among the 481 subjects and the age distribution in the U.S. population, explain intuitively why the grand mean (4.56 liters) is smaller than the age-adjusted mean (4.67 liters).

**11.8.2** To see if there is any relationship between blood pressure and childbearing, researchers examined data from a large health survey. The following table shows the data on systolic blood pressure (mm Hg) for random samples from two populations of women: women who had borne no children and women who had borne five or more children. The pooled standard deviation from all eight groups was  $s_{\text{pooled}} = 18 \text{ mm Hg}$ .<sup>25</sup>

AGE	NO CHILDREN		FIVE OR MORE CHILDREN	
	MEAN BLOOD PRESSURE	NO. OF WOMEN	MEAN BLOOD PRESSURE	NO. OF WOMEN
18–24	113	230	114	7
25–34	118	110	116	82
35–44	125	105	124	127
45–54	134	123	138	124
18–54	121	568	127	340

Carry out age adjustment, as directed, using the following reference distribution, which is the approximate distribution for U.S. women:<sup>26</sup>

AGE	RELATIVE FREQUENCY
18–24	0.17
25–34	0.29
35–44	0.31
45–54	0.23

- Calculate the age-adjusted mean blood pressure for women with no children.
- Calculate the age-adjusted mean blood pressure for women with five or more children.
- Calculate the difference between the values obtained in parts (a) and (b). Explain intuitively why the result is smaller than the unadjusted difference of  $127 - 121 = 6 \text{ mg Hg}$ .
- Calculate the standard error of the value calculated in part (a).
- Calculate the standard error of the value calculated in part (c).

**11.8.3** Refer to the ATP data of Exercise 11.7.1. The sample means and standard deviations are as follows:

	RIVER BIRCH		EUROPEAN BIRCH	
	FLOODED	CONTROL	FLOODED	CONTROL
$\bar{y}$	1.19	1.78	0.29	1.20
$s$	0.18	0.24	0.20	0.16

Define linear combinations (that is, specify the multipliers) to measure each of the following:

- The effect of flooding in river birch
- The effect of flooding in European birch
- The difference between river birch and European birch with respect to the effect of flooding (that is, the interaction between flooding and species)

**11.8.4** (*Continuation of Exercise 11.8.3*)

- Use a  $t$  test to investigate whether flooding has the same effect in river birch and in European birch. Use a nondirectional alternative and let  $\alpha = 0.05$ . (The pooled standard deviation is  $s_{\text{pooled}} = 0.199$ .)
- If the sample sizes were  $n = 10$  rather than  $n = 4$  for each group, but the means, standard deviations, and  $s_{\text{pooled}}$  remained the same, how would the result of part (a) change?

**11.8.5** (*Continuation of Exercise 11.8.4*)

Consider the null hypothesis that flooding has no effect on ATP level in river birch. This hypothesis could be tested in two ways: as a contrast (using the method of Section 11.8), or with a two-sample  $t$  test (as in Exercise 7.2.11). Answer the following questions; do not actually carry out the tests.

- In what way or ways do the two test procedures differ?
- In what way or ways do the conditions for validity of the two procedures differ?
- One of the two procedures requires more conditions for its validity, but if the conditions are met, then this procedure has certain advantages over the other one. What are these advantages?

**11.8.6** Consider the data from Exercise 11.7.3 in which the drugs ticrynafen (T) and hydrochlorothiazide (H) were compared. The data are summarized in the following table. The pooled standard deviation is  $s_{\text{pooled}} = 11.83 \text{ mm Hg}$ .

	TICRYNAFEN (T)		HYDROCHLOROTHIAZIDE (H)	
	LOW DOSE	HIGH DOSE	LOW DOSE	HIGH DOSE
Mean	13.9	17.1	15.8	17.5
No. of patients	53	57	55	58

If the two drugs have equal effects on blood pressure, then T might be preferable because it has fewer side effects.

- Construct a 95% confidence interval for the difference between the drugs (with respect to mean blood pressure reduction), averaged over the two dosage levels.
- Interpret the confidence interval from part (a) in the context of this setting.

**11.8.7** Consider the lettuce growth experiment described in Exercise 11.7.6. The accompanying table shows the mean leaf dry weight (gm) of the nine plants in each treatment group. MS(within) from the ANOVA was 0.3481.

	NUTRIENT SOLUTION		AREA (MM <sup>2</sup> )
	STANDARD	EXTRA NITROGEN	
Low light	2.16	3.09	
High light	3.26	4.48	

Construct a 95% confidence interval for the effect of extra nitrogen, averaged over the two light conditions.

- 11.8.8** Refer to the MAO data of Exercise 11.4.1.
- Define a contrast to compare the MAO activity for schizophrenics without paranoid features versus the average of the two types with paranoid features.
  - Calculate the value of the contrast in part (a) and its standard error.
  - Apply a *t* test to the contrast in part (a). Let  $H_A$  be nondirectional and  $\alpha = 0.05$ .

**11.8.9** Are the brains of left-handed people anatomically different? To investigate this question, a neuroscientist conducted postmortem brain examinations in 42 people. Each person had been evaluated before death for hand

preference and categorized as consistently right-handed (CRH) or mixed-handed (MH). The table shows the results on the area of the anterior half of the corpus callosum (the structure that links the left and right hemispheres of the brain).<sup>27</sup> The MS(within) from the ANOVA was 2,498.

GROUP	AREA (MM <sup>2</sup> )		
	MEAN	SD	n
1. Males: MH	423	48	5
2. Males: CRH	367	49	7
3. Females: MH	377	63	10
4. Females: CRH	345	43	20

- The difference between MH and CRH is 56 mm<sup>2</sup> for males and 32 mm<sup>2</sup> for females. Is this sufficient evidence to conclude that the corresponding population difference is greater for males than for females? Test an appropriate hypothesis. (Use a nondirectional alternative and let  $\alpha = 0.10$ .)
- As an overall measure of the difference between MH and CRH, one can consider the quantity  $0.5(\mu_1 - \mu_2) + 0.5(\mu_3 - \mu_4)$ . Construct a 95% confidence interval for this quantity. (This is a sex-adjusted comparison of MH and CRH, where the reference population is 50% male and 50% female.)

**11.8.10** Consider the daffodil data of Exercise 11.4.5.

- Define a contrast to compare the stem length for daffodils from the open area versus the average of the north, south, east and west sides of the building.
- Calculate the value of the contrast in part (a) and its standard error.
- Apply a *t* test to the contrast in part (a). Let  $H_A$  be nondirectional and  $\alpha = 0.05$ .

## 11.9 Multiple Comparisons (Optional)

After conducting a global *F* test, we may find that there is significant evidence for a difference among the population means  $\mu_1, \mu_2, \dots, \mu_I$ . In this situation, we are often interested in a detailed analysis of the sample means  $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_I$  considering all pairwise comparisons. That is, we wish to test all possible pairwise hypotheses:

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 = \mu_3$$

$$H_0: \mu_2 = \mu_3$$

and so on.

We saw in Section 11.1 that using repeated  $t$  tests leads to an increased overall risk of Type I error (e.g., finding evidence for a difference in population means when, in fact, there is no difference). In fact, it was this increased risk of Type I error that motivated the global  $F$  test in the first place. In this section we describe three multiple comparison methods to control the overall risk of Type I error: Bonferroni's method, Fisher's Least Significant Difference, and Tukey's Honest Significant Difference. First, however, we must examine the different types of Type I error that arise when considering multiple comparisons.

## Experimentwise versus Comparisonwise Error

Consider a study involving the comparison of four population means:  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$ , and  $\mu_4$ . As noted in Section 11.1, there are six possible comparisons:

$$H_0: \mu_1 = \mu_2 \quad H_0: \mu_1 = \mu_3 \quad H_0: \mu_1 = \mu_4 \quad H_0: \mu_2 = \mu_3 \quad H_0: \mu_2 = \mu_4 \quad H_0: \mu_3 = \mu_4$$

When considering these six comparisons we can speak of the chance of a Type I error for a particular comparison, say  $H_0: \mu_1 = \mu_2$ , called the **comparisonwise Type I error rate** ( $\alpha_{cw}$ ), or we can speak of the chance of making a Type I error among *any* of the six comparisons, called the **experimentwise Type I error rate** ( $\alpha_{ew}$ ).\* For example, Table 11.1.2 displays the experimentwise Type I error rates for comparing different numbers of groups when the comparisonwise Type I error rate is  $\alpha_{cw} = 0.05$ .

While the relationship between  $\alpha_{cw}$  and  $\alpha_{ew}$  may be complex, it is always true that

$$\alpha_{ew} \leq k \times \alpha_{cw}$$

where  $k$  is the number of comparisons. Thus, if six independent comparisons were made at the  $\alpha_{cw} = 0.05$  level, the experimentwise Type I error rate ( $\alpha_{ew}$ ) is at most  $6 \times 0.05 = 0.30$ .

## Fisher's Least Significant Difference

In optional Section 11.8 we described a procedure for estimating linear contrasts. Fisher's Least Significant Difference (LSD) uses this procedure to produce all pairwise confidence intervals for differences of population means using  $\alpha_{cw} = \alpha$ , the Type I error rate used in the ANOVA. Intervals that do not contain zero provide evidence for a significant difference between the compared population means.

An example of the procedure follows.

---

### Example 11.9.1

**Oysters and Seagrass** In a study to investigate the effect of oyster density on seagrass biomass, researchers introduced oysters to thirty 1-m<sup>2</sup> plots of healthy seagrass. At the beginning of the study the seagrass was clipped short in all plots. Next, 10 randomly chosen plots received a high density of oysters; 10, an intermediate density; and 10, a low density. As a control, an additional 10 randomly chosen clipped 1-m<sup>2</sup> plots received no oysters. After two weeks, the belowground

---

\*Although the term *experimentwise* contains the word experiment, this terminology pertains to both experiments and observational studies.

seagrass biomass was measured in each plot ( $\text{g/m}^2$ ). Data from some plots are missing. A summary of the data (Table 11.9.1) as well as the ANOVA table (Table 11.9.2) follow.<sup>28</sup>

**Table 11.9.1** Belowground seagrass biomass ( $\text{g/m}^2$ )

	Oyster density			
	None (1)	Low (2)	Intermediate (3)	High (4)
Mean	34.81	33.13	28.33	15.00
SD	13.44	17.36	17.11	10.97
<i>n</i>	9	10	8	10

**Table 11.9.2** ANOVA summary of belowground seagrass biomass ( $\text{g/m}^2$ )

	df	Sum of squares	Mean squares	F	P-value
Between	3	2365.5	788.51	3.5688	0.0243
Within	33	7291.1	220.94		
Total	36	9656.6			

The *P*-value for the ANOVA is 0.0243, indicating that there is significant evidence of a difference among the biomass means under these experimental conditions. Having evidence for a difference we proceed with comparisons.

Recall that for any linear contrast  $L = m_1\bar{y}_1 + m_2\bar{y}_2 + \cdots + m_I\bar{y}_I$ ,

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\sum_{i=1}^I \frac{m_i^2}{n_i}}$$

where

$$s_{\text{pooled}} = \sqrt{\text{MS}(\text{within})}$$

Thus, to compare the no oyster condition (1) to the low oyster density condition (2) we define  $D_{12} = \bar{Y}_1 - \bar{Y}_2$  so that as a linear contrast we have

$$\begin{aligned} d_{12} &= 1\bar{y}_1 + (-1)\bar{y}_2 + 0\bar{y}_3 + 0\bar{y}_4 \\ &= (1)(34.81) + (-1)(33.13) + (0)(28.33) + (0)(15.00) \\ &= 34.81 - 33.13 = 1.68 \end{aligned}$$

and, since  $s_{\text{pooled}} = \sqrt{220.94} = 14.86$ , we have

$$\begin{aligned} \text{SE}_{D_{12}} &= 14.86 \times \sqrt{\frac{1^2}{9} + \frac{(-1)^2}{10} + \frac{0^2}{8} + \frac{0^2}{10}} \\ &= 14.86 \times \sqrt{\frac{1}{9} + \frac{1}{10}} \\ &= 6.82 \end{aligned}$$

A 95% confidence interval for the population mean difference in belowground biomass for the no oyster condition compared to the low oyster density condition,  $\mu_1 - \mu_2$ , is given by

$$\begin{aligned} d_{12} \pm t_{33,0.025} \times \text{SE}_{D_{12}} &= 1.68 \pm 2.0345 \times 6.82 \\ &= 1.68 \pm 13.89 \\ &= (-12.21, 15.57) \end{aligned}$$

We are 95% confident that the mean belowground biomass when there are no oysters is between 12.21 g/m<sup>2</sup> lower to 15.57 g/m<sup>2</sup> higher than when there is a low density of oysters. Since this interval contains zero, there is no evidence that the mean belowground biomass differs for these two conditions.

Repeating this process for the remaining five comparisons produces the intermediate computations and final intervals summarized in Table 11.9.3.

**Table 11.9.3** Intermediate computations and 95% Fisher's LSD intervals comparing belowground biomass under different oyster density conditions\*

Comparison	$d_{ab} = \bar{y}_a - \bar{y}_b$	$\sqrt{(1/n_a) + (1/n_b)}$	$\text{SE}_{D_{ab}} = s_{\text{pooled}} \times \sqrt{(1/n_a) + (1/n_b)}$	$t_{33,0.025} \times \text{SE}_{D_{ab}}$
None–low	1.68	0.459	6.828	13.891
None–intermediate	6.48	0.486	7.221	14.690
<i>None–high</i>	<i>19.81</i>	<i>0.459</i>	<i>6.828</i>	<i>13.891</i>
Low–intermediate	4.80	0.474	7.049	14.341
<i>Low–high</i>	<i>18.13</i>	<i>0.447</i>	<i>6.646</i>	<i>13.520</i>
Intermediate–high	13.33	0.474	7.049	14.341
Comparison	Lower 95%	Upper 95%		
None–low	-12.2	15.6		
None–intermediate	-8.2	21.2		
<i>None–high</i>	<i>5.9</i>	<i>33.7</i>		
Low–intermediate	-9.5	19.1		
<i>Low–high</i>	<i>4.6</i>	<i>31.7</i>		
Intermediate–high	-1.0	27.7		

\*Intervals not containing zero (i.e., there is a statistically significant difference between the group means) are in italics. Note that an interval will not contain zero whenever  $|D_{ab}| > t \times \text{SE}_{D_{ab}}$ . (The value of  $t_{33,0.025} = 2.0345$  was determined using a computer. Using Table 4 we would obtain very similar results using the value listed for 30 degrees of freedom,  $t_{30,0.025} = 2.042$ .)

From Table 11.9.3 we observe that the only comparisons showing significant differences in mean biomass are the no- to high-oyster density and low- to high-oyster densities. ■

A general formula for computing a  $100(1 - \alpha)\%$  Fisher LSD interval for  $(\mu_a - \mu_b)$  is given in the following box.

100(1 –  $\alpha$ )% Fisher LSD Interval for  $(\mu_a - \mu_b)$

$$d_{ab} \pm t_{df,\alpha/2} \times \text{SE}_{D_{ab}}$$

where

$$d_{ab} = \bar{y}_a - \bar{y}_b$$

$$\text{SE}_{D_{ab}} = s_{\text{pooled}} \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}$$

$$s_{\text{pooled}} = \sqrt{\text{MS}(\text{within})}$$

and

$$df = df(\text{within})$$

**How does Fisher's LSD control the experimentwise Type I error rate?** One should use Fisher's LSD comparisons only after rejecting the ANOVA global null hypothesis that all population means are equal:  $H_0: \mu_1 = \mu_2 = \dots = \mu_I$ . The ANOVA global  $F$  test acts as a screening procedure for the multiple comparisons and thus offers control over  $\alpha_{ew}$ .

## Displaying Results

The presentation of all six Fisher LSD intervals for the seagrass example in Table 11.9.3 is a useful working summary but is not suitable for effective communication of results. To organize the results for presentation in a simple table we take the following steps.

**Step 1** *Array of group labels.* Arrange the group labels in increasing order of their means.

**Step 2** *Systematic comparison of means, underlining nonsignificant differences.*

- (a) Begin by examining the interval comparing the largest and smallest means. If the interval contains zero, the difference in means is not statistically significant and a line is drawn under the array of group labels to “connect” the groups with the largest and smallest means. If the interval does not contain zero, proceed to the next step.
- (b) Ignore the group with the smallest mean and compare the remaining subarray of  $I - 1$  means. As in step (2a), if the interval contains zero, the difference in means is not statistically significant and a line is drawn under the array of group labels being compared to “connect” the groups. Next consider the other subarray of  $I - 1$  means—the means that remain if the group with the largest mean is ignored. Again, underline this subarray if the interval contains zero.
- (c) Repeat step (2b) by successively comparing all subarrays of size  $I - 2$ ,  $I - 3$ , and so on, until an interval is produced that contains zero or no more comparisons are possible.

**Important Notes:** During this procedure, never make a comparison within any subarray that has already been underlined; these group means are automatically declared not statistically significantly different. Also, when underlining, use a separate line for each step; never join a line to one that has already been drawn.

**Step 3** *Translate the underlines to a tabular summary.* Create a summary table of the data using superscript letters to indicate which groups are not statistically significantly different.

**Example**  
**11.9.2**

**Oysters and Seagrass** In this example we will follow the preceding procedure to display the oyster and seagrass Fisher's LSD comparisons displayed in Table 11.9.3.

**Step 1** We first arrange the labels in order of the means (shown in Table 11.9.1).

High   Intermediate   Low   None

**Step 2** We compare the groups with the smallest (high oyster density) and largest (no oysters) means:  $\mu_{\text{None}} - \mu_{\text{High}} = (5.9, 33.7)$ . This interval does not contain zero, so these means are significantly different and no underline is made.\* We now proceed to the next set (step 2b), the comparisons of subarrays of three means. First, we compare Intermediate to None:

$$\mu_{\text{None}} - \mu_{\text{Intermediate}} = (-8.2, 21.2)$$

This interval contains zero, so an underline is drawn as shown.

High   Intermediate   Low   None

This underline indicates that these three groups do not have significantly different means. We now compare the next subarray of three means, High to Low:  $\mu_{\text{High}} - \mu_{\text{Low}} = (4.6, 31.7)$ . This interval does not contain zero, so no underlines are drawn. There is evidence for a difference in mean belowground biomass between the high and low oyster-density conditions.

Having compared all subarrays of three means, we continue with subarrays of two means. The only subarray of two means not already connected with an underline is the High–Intermediate comparison. This interval  $\mu_{\text{Intermediate}} - \mu_{\text{High}} = (-1.0, 27.7)$  contains zero, so an underline is drawn as shown.

High   Intermediate   Low   None

**Step 3** Communicating these results, we give each line a letter and display these letters as superscripts in our table of group means as shown below and in Table 11.9.4. A graphical display is also possible and is displayed in Figure 11.9.1.

High   Intermediate   Low   None  
a   \_\_\_\_\_  
b   \_\_\_\_\_

\*Intuitively, this interval should not contain zero since we have rejected the global  $F$  test null hypothesis, though there are some instances where the results of our multiple comparison procedure and global  $F$  test may not agree.

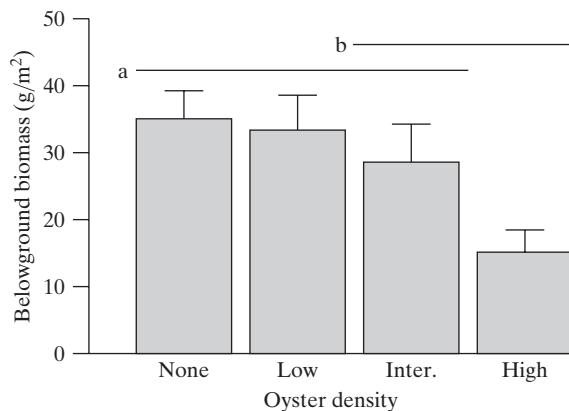
**Table 11.9.4** Belowground seagrass biomass ( $\text{g/m}^2$ ) for different levels of oyster density

	Oyster density			
	None	Low	Intermediate	High
Mean	34.8 <sup>a</sup>	33.1 <sup>a</sup>	28.3 <sup>a,b</sup>	15.0 <sup>b</sup>
SD	13.4	17.4	17.1	11.0
n	9	10	8	10

\*Groups sharing a common superscript have means that are not statistically significantly different based on Fisher's LSD comparisons with  $\alpha_{cw} = 0.05$ .

**Figure 11.9.1**

Belowground seagrass biomass ( $\text{g/m}^2$ ) for different levels of oyster density. Bars display means plus one standard error. Groups sharing a common overbar are not statistically significantly different based on Fisher's LSD comparisons with  $\alpha_{cw} = 0.05$



## The Bonferroni Method

The **Bonferroni method** is based on a very simple and general relationship: The probability that at least one of several events will occur cannot exceed the sum of the individual probabilities. For instance, suppose we conduct six tests of hypotheses, each at  $\alpha_{cw} = 0.01$ . Then the overall risk of Type I error  $\alpha_{ew}$ —that is, the chance of rejecting at least one of the six hypotheses when in fact all of them are true—cannot exceed

$$0.01 + 0.01 + 0.01 + 0.01 + 0.01 + 0.01 = (6)(0.01) = 0.06$$

Turning this logic around, suppose an investigator plans to conduct six tests of hypotheses and wants the overall risk of Type I error not to exceed  $\alpha_{ew} = 0.05$ . A conservative approach is to conduct each of the separate tests at the significance level  $\alpha_{cw} = 0.05/6 = 0.0083$ ; this is called a **Bonferroni adjustment**.

Note that the Bonferroni technique is very broadly applicable. The separate tests may relate to different response variables, different subsets, and so on; some may be *t* tests, some chi-square tests, and so on.

The Bonferroni approach can be used by a person reading a research report, if the author has included explicit *P*-values. For instance, if the report contains six *P*-values and the reader desires overall 5%-level protection against Type I error, then the reader will not regard a *P*-value as sufficient evidence of an effect unless it is smaller than  $\alpha_{cw} = 0.0083$ .

A Bonferroni adjustment can also be made for confidence intervals. For instance, suppose we wish to construct six confidence intervals and desire an overall probability of 95% that *all* the intervals contain their respective parameters ( $\alpha_{ew} = 0.05$ ). Then this can be accomplished by constructing each interval at confidence level 99.17% (because  $0.05/6 = 0.0083$  and  $1 - 0.0083 = 0.9917$ ).

In general, to construct  $k$  Bonferroni-adjusted confidence intervals with an overall probability of  $100(1 - \alpha_{ew})\%$  that *all* the intervals contain their respective parameters, we construct each interval at confidence level  $100(1 - \alpha_{cw})\%$  where  $\alpha_{cw} = \alpha_{ew}/k$ . The mechanics of the computations are identical to those used for Fisher's LSD except the value of the  $t$  multiplier is modified:  $t_{df,\alpha_{cw}/2}$ . Note that the application of this idea requires unusual critical values, so that standard tables are not sufficient. Table 11 (at the end of this book) provides Bonferroni multipliers for confidence intervals that are based on a  $t$  distribution. Software can also be used to produce appropriate multipliers. Example 11.9.3 illustrates this idea.

**Example  
11.9.3**

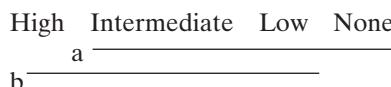
**Oysters and Seagrass** To compute the Bonferroni adjusted experimentwise 95% ( $\alpha_{ew} = 0.05$ ) confidence intervals for our oyster and seagrass example, we first recall that a total of six comparisons are required so that  $\alpha_{cw} = 0.05/6 = 0.0083$  and  $t_{30,0.0083/2} = 2.825$  [because not all values of df are listed in Table 12, we use  $df = 30$ , the closest value to  $df(\text{within}) = 33$ ]. Table 11.9.5 summarizes the collection of intervals in a manner similar to the Fisher LSD intervals in Table 11.9.3.

**Table 11.9.5** Intermediate computations and experimentwise 95% (99.17% comparisonwise) Bonferroni intervals comparing belowground biomass under different oyster density conditions

Comparison	$d_{ab} = \bar{y}_a - \bar{y}_b$	$SE_{D_{ab}}$	$t_{30,0.025/6} \times SE_{D_{ab}}$	Lower 99.17%	Upper 99.17%
None–low	1.68	6.828	13.891	-17.6	21.0
None–intermediate	6.48	7.221	14.690	-13.9	26.9
<i>None–high</i>	<i>19.81</i>	<i>6.828</i>	<i>13.891</i>	<i>0.5</i>	<i>39.1</i>
Low–intermediate	4.80	7.049	14.341	-15.1	24.7
Low–high	18.13	6.646	13.520	-0.6	36.9
Intermediate–high	13.33	7.049	14.341	-6.6	33.2

\*Intervals not containing zero (i.e., where there is a statistically significant difference between the group means) are in italics. Note the first two columns ( $d_{ab}$  and  $SE_{D_{ab}}$ ) are identical to those presented in Table 11.9.3.

Using the method of underlining to visualize the comparisons, we have



The underlines indicate that the only significant difference in mean belowground seagrass biomass is between the high oyster density and no oyster conditions. A summary of the results is presented in Table 11.9.6. ■

**Table 11.9.6** Belowground seagrass biomass ( $\text{g/m}^2$ ) for different levels of oyster density

	Oyster density			
	None	Low	Intermediate	High
Mean	34.8 <sup>a</sup>	33.1 <sup>a,b</sup>	28.3 <sup>a,b</sup>	15.0 <sup>b</sup>
SD	13.4	17.4	17.1	11.0
<i>n</i>	9	10	8	10

\*Groups sharing a common superscript have means that are not statistically significantly different based on Bonferroni comparisons with  $\alpha_{ew} = 0.05$ .

Note that the Fisher LSD intervals and the Bonferroni intervals are not identical (the Bonferroni are wider due to the smaller value of  $\alpha_{cw}$ ). Additionally, the conclusions differ as well. The Fisher LSD intervals indicate that there is evidence that the low and high oyster density conditions have different population means, while the Bonferroni intervals do not indicate a difference. This is because the Bonferroni intervals are less powerful and thus more conservative than the Fisher intervals. Unlike the Fisher intervals, the Bonferroni intervals are guaranteed to have  $\alpha_{ew}$  less than or equal to the desired experimentwise Type I error rate.

Unfortunately, the Bonferroni intervals are often overly conservative so that the actual value of  $\alpha_{ew}$  is much less than the desired experimentwise Type I error rate, and thus too much power is sacrificed for Type I error protection. A more complex procedure that (when sample sizes are equal) is able to achieve the desired experimentwise error exactly (and thus achieve higher power than Bonferroni) is Tukey's Honest Significant Difference.

## Tukey's Honest Significant Difference

Tukey's Honest Significant Difference (HSD) is very similar to the Fisher's LSD and Bonferroni adjusted intervals, but rather than using  $t$  multipliers in the confidence interval formulas, related values from a distribution known as the Studentized range distribution are used. Most computer packages will display all Tukey HSD pairwise intervals for any desired experimentwise Type I error rate,  $\alpha_{ew}$ . As an example, Figure 11.9.2 displays the Tukey output from the statistical software package R using our oyster and seagrass data. Note that in addition to the intervals, most software also provides an “adjusted”  $P$ -value. Even though multiple comparisons are being made, if these “adjusted”  $P$ -values are compared to  $\alpha_{ew}$ , an overall experimentwise Type I error rate of  $\alpha_{ew}$  will still be maintained.

**Figure 11.9.2** R software output presenting experimentwise 95% Tukey HSD intervals for the oyster and seagrass example

	diff	lwr	upr	p-adj
int-high	13.33	-5.74	32.40	0.2515
low-high	18.13	0.15	36.11	0.0475
no-high	19.81	1.34	38.28	0.0318
low-int	4.80	-14.27	23.87	0.9037
zero-int	6.48	-13.06	26.02	0.8063
zero-low	1.68	-16.79	20.15	0.9947

The intervals in Figure 11.9.2 show that the conclusions drawn from the Tukey HSD intervals match those from the Fisher LSD intervals: the high and low oyster density as well as the high and no oyster density means differ significantly. The endpoints of the experimentwise 95% Tukey HSD intervals are, however, different from both the Fisher LSD and Bonferroni intervals.

## Conditions for Validity

All three multiple comparison procedures as described require the same standard ANOVA conditions given in Section 11.5. In addition, the validity conditions for Fisher's LSD intervals also require that the procedure not be used unless the global null hypothesis of all means being equal is rejected. In contrast, Tukey's HSD and Bonferroni intervals do not require that the global  $F$  test be performed a priori (though the computation of  $s_{\text{pooled}}$  is still needed). To exactly achieve the desired experimentwise Type I error rate, Tukey's HSD requires that all samples be the same

size. If the sample sizes are unequal, the actual error rate will be somewhat less than the nominal rate resulting in a loss of power.

An advantage of the Bonferroni method is that it is widely applicable and can easily be generalized to situations beyond ANOVA. One such example appears in the exercises.

## Exercises 11.9.1–11.9.8

**11.9.1** A botanist used a completely randomized design to allocate 45 individually potted eggplant plants to five different soil treatments. The observed variable was the total plant dry weight without roots (gm) after 31 days of growth. The treatment means were as shown in the following table.<sup>29</sup> The MS(within) was 0.2246. Use Fisher's LSD intervals to compare all pairs of means at  $\alpha_{cw} = 0.05$ . Present your results in a summary table similar to Table 11.9.4. (*Hint:* Take note that all sample sizes are equal; thus the calculated margin of error need only be calculated once for all comparisons. There is a total of 10 comparisons possible).

TREATMENT	A	B	C	D	E
Mean	4.37	4.76	3.70	5.41	5.38
n	9	9	9	9	9

**11.9.2** Repeat Exercise 11.9.1, but use Bonferroni intervals with  $\alpha_{ew} = 0.05$ .

**11.9.3** In a study of the dietary treatment of anemia in cattle, researchers randomly divided 144 cows into four treatment groups. Group A was a control group, and groups B, C, and D received different regimens of dietary supplementation with selenium. After a year of treatment, blood samples were drawn and assayed for selenium. The accompanying table shows the mean selenium concentrations ( $\mu\text{g}/\text{dl}$ ).<sup>30</sup> The MS(within) from the ANOVA was 2.071.

GROUP	MEAN	n
A	0.8	36
B	5.4	36
C	6.2	36
D	5.0	36

(a) Compute three Bonferroni-adjusted intervals comparing diets B, C, and D to the control (diet A) using  $\alpha_{ew} = 0.05$ . (*Note:* This is an example of a situation for which the Bonferroni comparisons may be preferred over the Tukey HSD comparisons since not all comparisons are considered—we are only interested in comparing the control to each of the other three treatments.)

(b) In the context of the problem, interpret the Bonferroni interval computed in part (a) that compares the control (group A) to the group that is most different from it.

**11.9.4** Consider the experiment and data in Exercise 11.9.3. The experimentwise 95% Tukey HSD intervals are displayed using the statistical software package R.

	diff	lwr	upr
B-A	4.6	3.72	5.48
C-A	5.4	4.52	6.28
D-A	4.2	3.32	5.08
C-B	0.8	-0.08	1.68
D-B	-0.4	-1.28	0.48
D-C	-1.2	-2.08	-0.32

(a) Using the preceding output to support your answer, is there evidence that each of the groups/diets B, C, and D, differs from the control, A?

(b) According to the preceding Tukey HSD intervals and summary of the data in Exercise 11.9.3, diet C yields the greatest mean selenium concentration and is significantly higher than the control. If the goal of the researchers is to find a diet that maximizes selenium concentration, is diet C the clear choice? That is, should we rule out diet B, diet D, or both? Refer to the Tukey HSD intervals to justify your answer.

**11.9.5** Ten treatments were compared for their effect on the liver in mice. There were 13 animals in each treatment group. The ANOVA gave  $\text{MS}(\text{within}) = 0.5842$ . The mean liver weights are given in the table.<sup>31</sup>

TREATMENT	MEAN LIVER WEIGHT (GM)		TREATMENT	MEAN LIVER WEIGHT (GM)
	1	2		
1	2.59		6	2.84
2	2.28		7	2.29
3	2.34		8	2.45
4	2.07		9	2.76
5	2.40		10	2.37

- (a) Use Fisher LSD intervals to compare all pairs of means with  $\alpha_{cw} = 0.05$  and summarize the results in a table similar to Table 11.9.4. [Time Saving Hints: First note that the sample sizes are equal; hence the same margin of error ( $t \times \text{SE}_{D_{ab}}$ ) can be used for all comparisons. Furthermore, since a summary table is desired, the actual intervals need not be computed: Simply check if  $|d_{ab}| > t \times \text{SE}_{D_{ab}}$ . If it is, then the computed interval would not contain zero, so the difference is significant. Finally, note that not all possible comparisons (there are 45) need to be checked: when using the method of underlining to summarize results, once a subarray of groups has been underlined all comparisons within the subarray are considered nonsignificant.]
- (b) If Bonferroni's method is used with  $\alpha_{ew} = 0.05$  instead of Fisher's LSD in part (a), are any pairs of means significantly different?

**11.9.6** Consider the data from Example 11.2.1 on the weight gain of lambs. The MS(within) from the ANOVA for these data was 23.333. The sample mean of diet 2 was 15 and of diet 1 was 11.

- (a) Use the Bonferroni method to construct a 95% confidence interval for the difference in population means of these two diets (assuming that intervals will also be computed for the other two possible comparisons as well).
- (b) Suppose that the comparison in part (a) was the *only* comparison of interest (i.e., one comparison rather than three). How would the interval in part (a) change? Would it be wider, narrower, or stay the same? Explain.

**11.9.7** As mentioned in this section, the Bonferroni procedure can be used in a variety of circumstances. Consider the plover nesting example from Section 10.5, which compares plover nest locations across three years. The percentage distribution appears in the following table.

LOCATION	YEAR		
	2004	2005	2006
Agricultural field (AF)	48.8	30.2	55.3
Prairie dog habitat (PD)	39.5	60.3	25.5
Grassland (G)	11.6	9.5	19.1
Total	99.9*	100.0	99.9*

\*The sums of the 2004 and 2006 percentages differ from 100% due to rounding.

The  $P$ -value for the chi-square test of these data was found to be 0.007, indicating a significant difference in the distribution of nesting locations across the three years with  $\alpha = 0.10$ . Considering reduced tables and using chi-square tests to compare nesting distributions for pairs of years, we obtain the following  $P$ -values:

YEARS COMPARED	P-VALUE
2004 to 2005	0.100
2004 to 2006	0.307
2005 to 2006	0.001

Using a Bonferroni adjustment to achieve  $\alpha_{ew} = 0.10$ , for which pair(s) of years is there evidence of a significant difference in nesting location distributions? Indicate the value of  $\alpha_{cw}$  used.

**11.9.8** Exercise 10.5.1 presented the following problem: Patients with painful knee osteoarthritis were randomly assigned in a clinical trial to one of five treatments: glucosamine, chondroitin, both, placebo, or Celebrex, the standard therapy. One outcome recorded was whether or not each patient experienced substantial improvement in pain or in ability to function. The data are given in the following table.

TREATMENT	SUCCESSFUL OUTCOME		
	SAMPLE SIZE	NUMBER	PERCENT
Glucosamine	317	192	60.6
Chondroitin	318	202	63.5
Both	317	208	65.6
Placebo	313	178	56.9
Celebrex	318	214	67.3

- (a) Suppose we wished to compare only the success rates of each of the treatments to the control (placebo) using four separate  $2 \times 2$  chi-square tests. The  $P$ -values for these comparisons follow. Using a Bonferroni adjustment with  $\alpha_{ew} = 0.05$ , which treatments perform significantly different from the placebo? Indicate the value of  $\alpha_{cw}$  used.

TREATMENTS COMPARED TO PLACEBO	P-VALUE
Glucosamine	0.346
Chondroitin	0.088
Both	0.024
Celebrex	0.007

- (b) The  $P$ -value of the chi-square test that considers the entire  $5 \times 2$  table is 0.054, which provides insufficient evidence to demonstrate any difference among the success rates of the five treatments using  $\alpha = 0.05$ . Explain why this result does not contradict the results of part (a). [Hint: How many comparisons are being considered by this chi-square test as compared to the number of comparisons in part (a)? To achieve  $\alpha_{ew} = 0.05$  using a Bonferroni adjustment, how large would  $\alpha_{cw}$  need to be? How large was it in part (a)? How does conducting many tests with a Bonferroni adjustment affect the power of each test?]

## 11.10 Perspective

In Chapter 11 we have introduced some statistical issues that arise when analyzing data from more than two samples and we have considered some classical methods of analysis. In this section we review these issues and briefly mention some alternative methods of analysis.

### Advantages of Global Approach

Let us recapitulate the advantages of analyzing  $I$  independent samples by a global approach rather than by viewing each pairwise comparison separately.

1. *Multiple comparisons* In Section 11.1 we saw that the use of repeated  $t$  tests can greatly inflate the overall risk of Type I error. Some control of Type I error can be gained by the simple device of beginning the data analysis with a global  $F$  test. For more stringent control of Type I error, other multiple comparison methods are available (e.g., Bonferroni and Tukey HSD) and are described in optional Section 11.9. (Note that the problem of multiple comparisons is not confined to an ANOVA setting.)
2. *Use of structure in the treatments or groups* Analysis of suitable combinations of group means can be very useful in interpreting data. Many of the relevant techniques are beyond the scope of this book. The discussion in optional Sections 11.7 and 11.8 gave a hint of the possibilities. In Chapter 12 we will discuss some ideas that are applicable when the treatments themselves are quantitative (for instance, doses).
3. *Use of a pooled SD* We have seen that pooling all of the within-sample variability into a single pooled SD leads to a better estimate of the common population SD and thus to a more precise analysis. This is particularly advantageous if the individual sample sizes ( $n$ 's) are small, in which case the individual SD estimates are quite imprecise. Of course, using a pooled SD is proper only if the population SDs are equal. It sometimes happens that one cannot take advantage of pooling the SDs because the assumption of equal population SDs is not tenable. One approach that can be helpful in this case is to analyze a transformed variable, such as  $\log(Y)$ ; the SDs may be more nearly equal in the transformed scale.

### Other Experimental Designs

The techniques of this chapter are valid only for independent samples. But the basic idea—partitioning variability within and between treatments into interpretable components—can be applied in many experimental designs. For instance, all the techniques discussed in this chapter can be adapted (by suitable modification of the SE calculation) to analysis of data from an experiment with more than two experimental factors or situations for which all or some experimental factors are numeric rather than categorical. These and related techniques belong to the large subject called *analysis of variance*, of which we have discussed only a small part.

### Nonparametric Approaches

There are  $k$ -sample analogs of the Wilcoxon-Mann-Whitney test and other nonparametric tests (e.g., the Kruskal-Wallis test). These tests have the advantage of not assuming underlying normal distributions. However, many of the advantages of the parametric techniques—such as the use of linear combinations—do not easily carry over to the nonparametric setting.

## Ranking and Selection

In some investigations the primary aim of the investigator is not to answer research questions about the populations but simply to *select* one or several “best” populations. For instance, suppose 10 populations (stocks) of laying hens are available and it is desired to select the one population with the highest egg-laying potential. The investigator will select a random sample of  $n$  chickens from each stock and will observe for each chicken  $Y$  = total number of eggs laid in 500 days.<sup>32</sup> One relevant question is: How large should  $n$  be so that the stock that is *actually* best (has the highest  $\mu$ ) is likely to also *appear* best (have the highest  $\bar{Y}$ )? This and similar questions are addressed by a branch of statistics called *ranking and selection theory*.

### Supplementary Exercises 11.S.1–11.S.19

(Note: Exercises preceded by an asterisk refer to optional sections.)

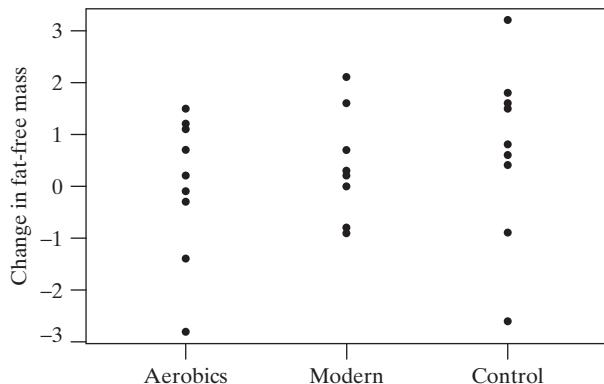
**11.S.1** Consider the research described in Exercise 11.4.6 in which 10 women in an aerobic exercise class, 10 women in a modern dance class, and a control group of 9 women were studied. One measurement made on each woman was change in fat-free mass over the course of the 16-week training period. Summary statistics are given in the following table.<sup>8</sup> The ANOVA SS(between) is 2.465 and the SS(within) is 50.133.

	AEROBICS	MODERN DANCE	CONTROL
Mean	0.00	0.44	0.71
SD	1.31	1.17	1.68
$n$	10	10	9

- (a) State in words, in the context of this problem, the null hypothesis that is tested by the analysis of variance.
- (b) Construct the ANOVA table and test the null hypothesis. Let  $\alpha = 0.05$ .

**11.S.2** Refer to Exercise 11.S.1. The  $F$  test is based on certain conditions concerning the population distributions.

- (a) State the conditions.
- (b) The following dotplots show the raw data. Based on these plots and on the information given in Exercise 11.S.1, does it appear that the  $F$  test conditions are met? Why or why not?



**11.S.3** In a study of the eye disease retinitis pigmentosa (RP), 211 patients were classified into four groups according to the pattern of inheritance of their disease. Visual acuity (spherical refractive error, in diopters) was measured for each eye, and the two values were then averaged to give one observation per person. The accompanying table shows the number of persons in each group and the group mean refractive error.<sup>33</sup> The ANOVA of the 211 observations yields SS(between) = 129.49 and SS(within) = 2,506.8. Construct the ANOVA table and carry out the  $F$  test at  $\alpha = 0.05$ .

GROUP	NUMBER OF PERSONS	MEAN REFRACTIVE ERROR
Autosomal dominant RP	27	+0.07
Autosomal recessive RP	20	-0.83
Sex-linked RP	18	-3.30
Isolate RP	146	-0.84
Total	211	

**11.S.4** (Continuation of Exercise 11.S.3) Another approach to the data analysis is to use the eye, rather than the person, as the observational unit. For the 211 persons there were 422 measurements of refractive error; the accompanying table summarizes these measurements. The ANOVA of the 422 observations yields SS(between) = 258.97 and SS(within) = 5,143.9.

GROUP	NUMBER OF EYES	MEAN REFRACTIVE ERROR
Autosomal dominant RP	54	+0.07
Autosomal recessive RP	40	-0.83
Sex-linked RP	36	-3.30
Isolate RP	292	-0.84
Total	422	

- (a) Construct the ANOVA table and bracket the  $P$ -value for the  $F$  test. Compare with the  $P$ -value obtained in Exercise 11.S.3. Which of the two  $P$ -values is of doubtful validity, and why?
- (b) The mean refractive error for the sex-linked RP patients was  $-3.30$ . Calculate the standard error of this mean two ways: (i) regarding the person as the observational unit and using  $s_{\text{pooled}}$  from the ANOVA of Exercise 11.S.3; (ii) regarding the eye as the observational unit and using  $s_{\text{pooled}}$  from the ANOVA of this exercise. Which of these standard errors is of doubtful validity, and why?

**\*11.S.5** In a study of the mutual effects of the air pollutants ozone and sulfur dioxide, Blue Lake snap beans were grown in open-top field chambers. Some chambers were fumigated repeatedly with sulfur dioxide. The air in some chambers was carbon filtered to remove ambient ozone. There were three chambers per treatment combination, allocated at random. After one month of treatment, total yield (kg) of bean pods was recorded for each chamber, with results shown in the accompanying table.<sup>34</sup> For these data,  $\text{SS}(\text{between}) = 1.3538$  and  $\text{SS}(\text{within}) = 0.27513$ . Complete the ANOVA table and carry out the  $F$  test at  $\alpha = 0.05$ .

OZONE ABSENT		OZONE PRESENT	
SULFUR DIOXIDE		SULFUR DIOXIDE	
ABSENT	PRESENT	ABSENT	PRESENT
1.52	1.49	1.15	0.65
1.85	1.55	1.30	0.76
1.39	1.21	1.57	0.69
Mean	1.587	1.417	1.340
SD	0.237	0.181	0.213
			0.056

Prepare an interaction graph (like Figure 11.7.3).

**\*11.S.6** Consider the data from Exercise 11.S.5. For these data,  $\text{SS}(\text{ozone}) = 0.696$ ,  $\text{SS}(\text{sulfur}) = 0.492$ ,  $\text{SS}(\text{interaction}) = 0.166$ , and  $\text{SS}(\text{within}) = 0.275$ .

- (a) Construct the ANOVA table.
- (b) Carry out an  $F$  test for interactions; use  $\alpha = 0.05$ .
- (c) Test the null hypothesis that ozone has no effect on yield. Use  $\alpha = 0.05$ .

**\*11.S.7** Refer to Exercise 11.S.5. Define contrasts to measure each effect specified, and calculate the value of each contrast.

- (a) The effect of sulfur dioxide in the absence of ozone
- (b) The effect of sulfur dioxide in the presence of ozone
- (c) The interaction between sulfur dioxide and ozone

**\*11.S.8** (*Continuation of Exercises 11.S.6 and 11.S.7*) For the snap-bean data, use a  $t$  test to test the null hypothesis of no interaction against the alternative that sulfur dioxide is more harmful in the presence of ozone than in its absence.

Let  $\alpha = 0.05$ . How does this compare with the  $F$  test of Exercise 11.S.6(b) (which has a nondirectional alternative)?

**\*11.S.9** (*Computer exercise*) Refer to the snap-bean data of Exercise 11.S.5. Apply a reciprocal transformation to the data. That is, for each yield value  $Y$ , calculate  $Y' = 1/Y$ .

- (a) Calculate the ANOVA table for  $Y'$  and carry out the  $F$  test.
- (b) It often happens that the SDs are more nearly equal for transformed data than for the original data. Is this true for the snap-bean data when a reciprocal transformation is used?
- (c) Make a normal probability plot of the residuals,  $(y'_{ij} - \bar{y}'_i)$ . Does this plot support the condition that the populations are normal?

**\*11.S.10** (*Computer exercise—continuation of Exercises 11.S.8 and 11.S.9*) Repeat the test in Exercise 11.S.7 using  $Y'$  instead of  $Y$ , and compare with the results of Exercise 11.S.7.

**11.S.11** Suppose a drug for treating high blood pressure is to be compared to a standard blood pressure drug in a study of humans.

- (a) Describe an experimental design for a study that makes use of blocking. Be careful to note which parts of the design involve randomness and which parts do not.
- (b) Can the experiment you described in part (a) involve blinding? If so, explain how blinding could be used.

**11.S.12** In a study of balloon angioplasty, patients with coronary artery disease were randomly assigned to one of four treatment groups: placebo, probucol (an experimental drug), multivitamins (a combination of beta carotene, vitamin E, and vitamin C), or probucol combined with multivitamins. Balloon angioplasty was performed on each of the patients. Later, “minimal luminal diameter” (a measurement of how well the angioplasty did in dilating the artery) was recorded for each of the patients. Summary statistics are given in the following table.<sup>35</sup>

	MULTI-PLACEBO	PROBUCOL	VITAMINS	PROBUCOL AND MULTIVITAMINS
<i>n</i>	62	58	54	56
Mean	1.43	1.79	1.40	1.54
SD	0.58	0.45	0.55	0.61

- (a) Complete the ANOVA table and bracket the  $P$ -value for the  $F$  test.

SOURCE	DF	SS	MS	F
Between treatments	—	5.4336	—	—
Within treatments	—	—	—	—
Total	229	73.9945	—	—

- (b) If  $\alpha = 0.01$ , do you reject the null hypothesis of equal population means? Why or why not?

**\*11.S.13** Refer to Exercise 11.S.12. Define contrasts to measure each effect specified, and calculate the value of each contrast.

- The effect of probucol in the absence of multivitamins
- The effect of probucol in the presence of multivitamins
- The interaction between probucol and multivitamins

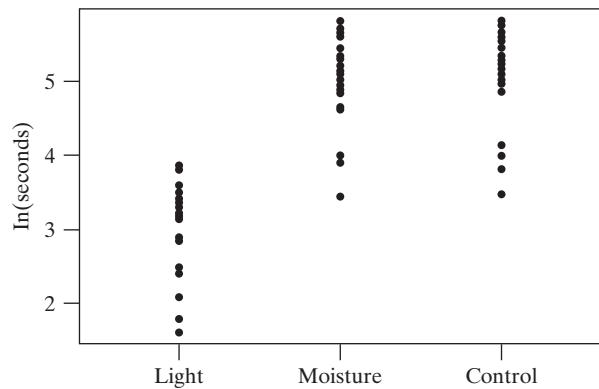
**\*11.S.14** Refer to Exercise 11.S.12. Construct a 95% confidence interval ( $\alpha_{cw} = 0.05$ ) for the effect of probucol in the absence of multivitamins. That is, construct a 95% confidence interval for  $\mu_{\text{probucol}} - \mu_{\text{placebo}}$ .

**\*11.S.15** Refer to Exercise 11.S.12. Assuming all possible comparisons of group means will be computed, use the Bonferroni method to construct a 95% confidence interval for the effect of probucol in the absence of multivitamins. That is, construct a Bonferroni-adjusted 95% ( $\alpha_{ew} = 0.05$ ) confidence interval for  $\mu_{\text{probucol}} - \mu_{\text{placebo}}$ .

**\*11.S.16** Three college students collected several pillbugs from a woodpile and used them in an experiment in which they measured the time, in seconds, that it took for a bug to move 6 inches within an apparatus they had created. There were three groups of bugs: one group was exposed to strong light, for one group the stimulus was moisture, and a third group served as a control. The data are shown in the following table.<sup>36</sup>

LIGHT	MOISTURE	CONTROL
23	170	229
12	182	126
29	286	140
12	103	260
5	330	330
47	55	310
18	49	45
30	31	248
8	132	280
45	150	140
36	165	160
27	206	192
29	200	159
33	270	62
24	298	180
17	100	32
11	162	54
25	126	149
6	229	201
34	140	173
Mean	23.6	169.2
SD	12.3	83.5
n	20	20

Clearly the SDs show that the variability is not constant between groups, so a transformation is needed. Taking the natural logarithm of each observation results in the following dotplots and summary statistics.



	LIGHT	MOISTURE	CONTROL
Mean	2.99	4.98	4.99
SD	0.65	0.62	0.66

For the transformed data, the ANOVA SS(between) is 53.1103 and the SS(within) is 23.5669.

- State the null hypothesis in symbols.
- Construct the ANOVA table and test the null hypothesis. Let  $\alpha = 0.05$ .
- Calculate the pooled standard deviation,  $s_{\text{pooled}}$ .

**\*11.S.17** Mountain climbers often experience several symptoms when they reach high altitudes during their climbs. Researchers studied the effects of exposure to high altitude on human skeletal muscle tissue. They set up a  $2 \times 2$  factorial experiment in which subjects trained for six weeks on a bicycle. The first factor was whether subjects trained under hypoxic conditions (corresponding to an altitude of 3,850m) or normal conditions. The second factor was whether subjects trained at a high level of energy expenditure or at a low level (25% less than the high level). There were either 7 or 8 subjects at each combination of factor levels. The accompanying table shows the results for the response variable “percentage change in vascular endothelial growth factor mRNA.”<sup>37</sup>

ENERGY	HYPOXIC		NORMAL	
	LOW LEVEL	HIGH LEVEL	LOW LEVEL	HIGH LEVEL
Mean	117.7	173.2	95.1	114.6
No. of patients	7	7	8	8

Prepare an interaction graph (like Figure 11.7.3).

**\*11.S.18** Consider the data from Exercise 11.S.17.

- (a) Complete the following ANOVA table.

SOURCE	DF	SS	MS	F RATIO
Between hypoxic and normal	1	12126.5	—	—
Between energy level	1	10035.7	—	—
Interaction within groups	1	—	—	—
Total	26	56076.0	—	—
	29	80738.7	—	—

- (b) Conduct a test for interactions. Use  $\alpha = 0.05$ .
- (c) Based on your conclusions in part (b), is it sensible to examine the main effects of condition and of energy level?
- (d) Test the null hypothesis that energy level has no effect on the response. Use  $\alpha = 0.05$ .
- (e) Test the null hypothesis that the effect on the response of hypoxic training is the same as the effect on the response of normal training. Use  $\alpha = 0.05$ .

**\*11.S.19** In a study to examine the utility of using ammonia gas to sanitize animal feeds, researchers inoculated corn silage with a strain of *Salmonella*. Next, two petri dishes of 5 g of contaminated feed were exposed to concentrated anhydrous ammonia gas and two control petri

dishes of 5 g of contaminated feed were not treated with the gas. This experiment was repeated twice, for a total of three trials, as only two petri dishes could be placed in the pressurized gas chamber at any given time. Twenty-four hours after inoculation and gassing, the number of bacterial colonies (colony forming units or cfu) on each dish were counted. Because the data were highly skewed, the log(cfu) was analyzed.<sup>38</sup>

- (a) Identify the blocking, treatment, and response variables in this problem.
- (b) Complete the following ANOVA table for this blocked analysis.

	DF	SS	MS	F RATIO
Between treatments	1	1.141	1.141	7.107
Between trials	2	3.611	—	—
Within groups	8	—	—	—
Total	11	6.036		

- (c) Using the complete table from part (b), is there evidence that the ammonia gas treatment affects the contamination level (i.e., mean log cfu)? Use  $\alpha = 0.05$ .
- (d) Do the preceding analysis and information allow you to infer that ammonia reduces contamination? If not, what other information would be necessary to make such a claim?

# LINEAR REGRESSION AND CORRELATION

## Objectives

In this chapter we study correlation and regression. We will

- demonstrate how the correlation coefficient is calculated and interpreted.
- show how least-squares regression models are fit to data.
- examine the relationship between the regression line, sample correlation, and the prediction of means.
- show how to test whether a regression relationship is statistically significant.
- extend regression ideas to multiple regression, analysis of covariance, and logistic regression.

## 12.1 Introduction

In this chapter we discuss some methods for analyzing the relationship between two quantitative variables,  $X$  and  $Y$ . **Linear regression** and **correlation analysis** are techniques based on fitting a straight line to the data.

### Examples

Data for regression and correlation analysis consist of pairs of observations  $(X, Y)$ . Here are two examples.

---

**Example  
12.1.1**

**Amphetamine and Food Consumption** Amphetamine is a drug that suppresses appetite. In a study of this effect, a pharmacologist randomly allocated 24 rats to three treatment groups to receive an injection of amphetamine at one of two dosage levels, or an injection of saline solution. She measured the amount of food consumed by each animal in the 3-hour period following injection. The results (gm of food consumed per kg body weight) are shown in Table 12.1.1.<sup>1</sup>

Figure 12.1.1 shows a **scatterplot** of

$Y$  = Food consumption

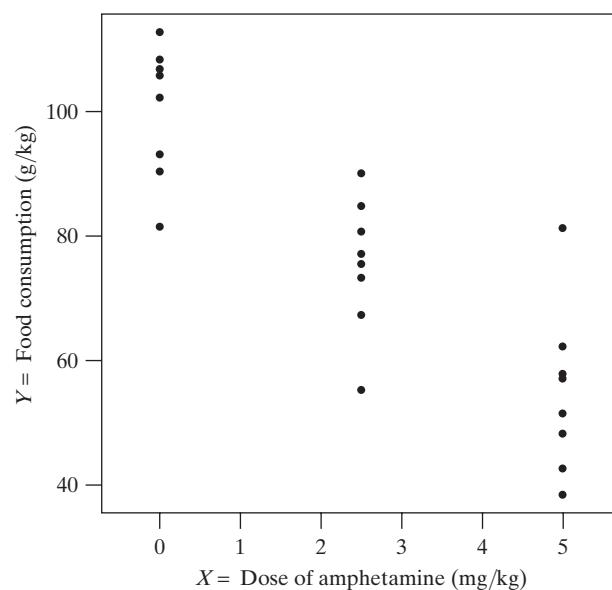
against

$X$  = Dose of amphetamine

The scatterplot suggests a definite dose-response relationship, with larger values of  $X$  tending to be associated with smaller values of  $Y$ .\*

<b>Table 12.1.1</b> Food consumption ( $Y$ ) of rats (gm/kg)			
	$X = \text{Dose of amphetamine (mg/kg)}$		
	0	2.5	5.0
112.6	73.3	38.5	
102.1	84.8	81.3	
90.2	67.3	57.1	
81.5	55.3	62.3	
105.6	80.7	51.5	
93.0	90.0	48.3	
106.6	75.5	42.7	
108.3	77.1	57.9	
Mean	100.0	75.5	55.0
SD	10.7	10.7	13.3
No. of animals	8	8	8

**Figure 12.1.1** Scatterplot of food consumption against dose of amphetamine



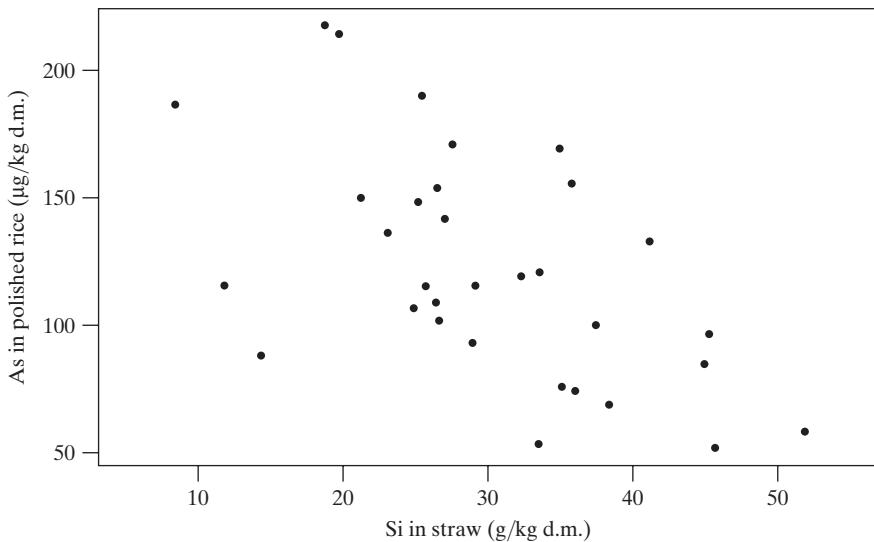
### Example 12.1.2

**Arsenic in Rice** Environmental pollutants may enter the food supply as contaminants leach into the soil from which the food is grown. It is hypothesized that naturally occurring silicon in rice plants may inhibit the absorption of some pollutants. In a study to investigate compounds that mitigate arsenic absorption in rice, researchers sampled 32 rice plants and measured the concentration of arsenic in the polished rice ( $\mu\text{g/kg}$  rice) as well as the concentration of silicon in the straw ( $\text{g/kg}$  straw) of each plant.<sup>2</sup> Figure 12.1.2 shows a scatterplot of

$Y = \text{rice arsenic concentration}$

\*In many dose-response relationships, the response depends linearly on  $\log(\text{dose})$  rather than on dose itself. We have chosen a linear portion of the dose-response curve to simplify the exposition.

**Figure 12.1.2** Scatterplot of rice arsenic concentration against straw silicon concentration



against

$X$  = straw silicon concentration

The scatterplot suggests that higher straw silicon concentrations ( $X$ ) tend to be associated with lower rice arsenic concentrations ( $Y$ ). ■

## 12.2 The Correlation Coefficient

Suppose we have a sample of  $n$  pairs for which each pair represents the measurements of two variables,  $X$  and  $Y$ . If a scatterplot of  $Y$  versus  $X$  shows a general linear trend, then it is natural to try to describe the strength of the linear association. In this section we will learn how to measure the strength of linear association using the **correlation coefficient**. The following example illustrates the kind of situation we wish to consider.

**Example  
12.2.1**

**Length and Weight of Snakes** In a study of a free-living population of the snake *Vipera beretis*, researchers caught and measured nine adult females.<sup>3</sup> Their body lengths and weights are shown in Table 12.2.1 and are displayed as a scatterplot in Figure 12.2.1. The number of observations is  $n = 9$ . ■

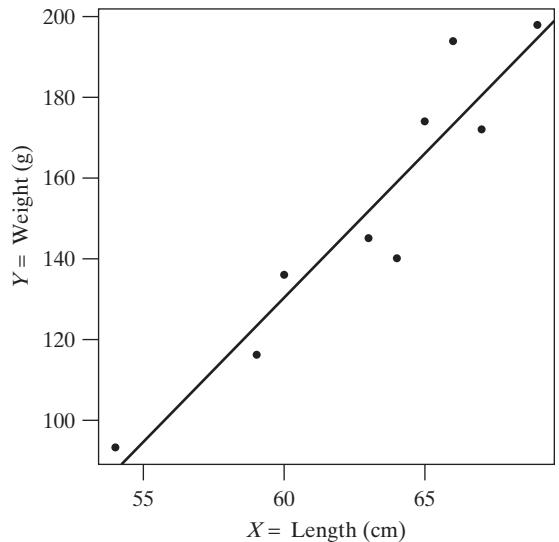
The scatterplot shown in Figure 12.2.1 shows a clear upward trend. We say that weight shows a **positive association** with length, indicating that greater lengths are associated with greater weights. Thus, snakes that are longer than the average length of  $\bar{x} = 63$  tend to be heavier than the average weight of  $\bar{y} = 152$ . The line superimposed on the plot is called the **least-squares line** or **fitted regression line** of  $Y$  on  $X$ . We will learn how to compute and interpret the regression line in Section 12.3.

### Measuring Strength of Linear Association

How strong is the linear relationship between snake length and weight? Are the data points tightly clustered around the regression line, or is the scatter loose? To answer these questions we will compute the **correlation coefficient**, a scale-invariant numeric measure of the strength of linear association between two quantitative variables. Being scale invariant means that the correlation coefficient is unaffected

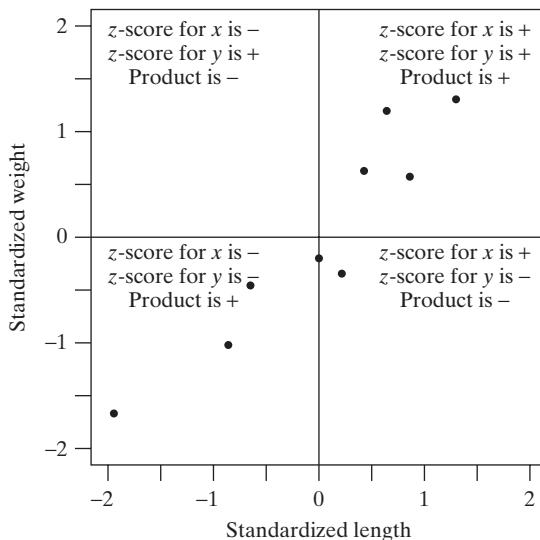
**Table 12.2.1**

Length $X$ (cm)	Weight $Y$ (g)
60	136
69	198
66	194
64	140
54	93
67	172
59	116
65	174
63	145
Mean	63
SD	4.6
	152
	35.3

**Figure 12.2.1** Body length and weight of nine snakes with fitted regression line

by any changes in measurement scales. That is, the correlation between length and weight will be the same whether measured in centimeters and grams or inches and pounds. To understand how the correlation coefficient works, consider again the snake length and weight example. Rather than plotting the original data, Figure 12.2.2 plots the standardized data ( $z$ -scores) displayed in Table 12.2.2; note that this plot looks identical to our original plot except now our scales are unit-less.

Dividing the plot into quadrants based on the sign of the standardized score, we see that most of these points fall into the upper-right and lower-left quadrants. Points falling in these quadrants will have standardized scores whose *products* are positive. Likewise, points falling in the upper-left and lower-right quadrants will have standardized score products that are negative. Computing the sum of these products provides a numeric measure of where our points fall (i.e., which quadrants are dominant). In our case, since there is a positive association between length and weight, most points fall in the positive product quadrants; thus, the sum of the

**Figure 12.2.2** Scatterplot of standardized weight versus standardized length

**Table 12.2.2** Standardized snake weights, lengths, and their products

Weight	Length	Standardized weight	Standardized length	Product of standardized values
X	Y	$z_x = \frac{x - \bar{x}}{s_x}$	$z_y = \frac{y - \bar{y}}{s_y}$	$z_x z_y$
60	136	-0.65...	-0.45...	0.29...
69	198	1.29...	1.30...	1.68...
66	194	0.65...	1.19...	0.77...
64	140	0.22...	-0.34...	-0.07...
54	93	-1.94...	-1.67...	3.24...
67	172	0.86...	0.57...	0.49...
59	116	-0.86...	-1.02...	0.88...
65	174	0.43...	0.62...	0.27...
63	145	0.00...	-0.20...	0.00...
Sum	567	1368	0.00	0.00
Mean	63.000	152.000	0.00	0.00
SD	4.637	35.338	1.00	1.00

Values in the table are truncated for ease of reading. Because the summary values will be used in subsequent calculations, they include more digits than one would typically report when following our rounding conventions.

products of standardized scores is positive. If a negative relationship were present, most of the points would fall in the negative quadrants and the sum would be negative. And, if there were no *linear* relationship, the points would fall in evenly in all four quadrants so that the positive and negative products would balance and their sum would be zero.

The correlation coefficient is based on this sum. It is computed as the average product of standardized scores (using  $n - 1$  rather than  $n$  to compute the average):\*

### The correlation coefficient, $r$

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

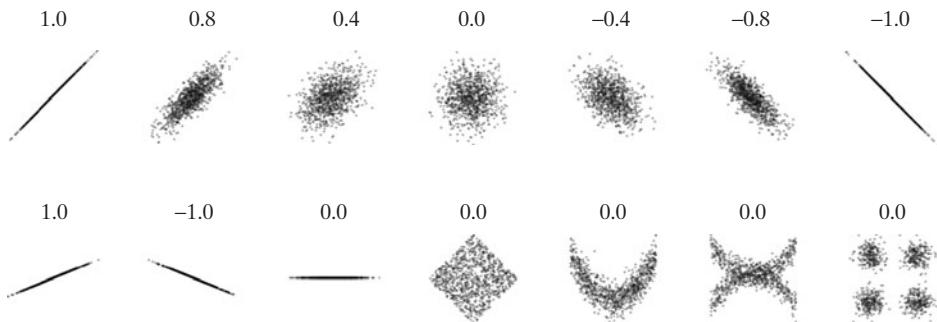
From this formula it is clear that  $X$  and  $Y$  enter  $r$  symmetrically; therefore, if we were to interchange the labels  $X$  and  $Y$  of our variables,  $r$  would remain unchanged. In fact, this is one of the advantages of the correlation coefficient as a summary statistic: In interpreting  $r$ , it is not necessary to know (or to decide) which variable is labeled  $X$  and which is labeled  $Y$ .

\*By substituting  $\sqrt{\sum_{i=1}^n (x - \bar{x})^2 / (n - 1)}$  for  $s_x$  and  $\sqrt{\sum_{i=1}^n (y - \bar{y})^2 / (n - 1)}$  for  $s_y$ , the equation for the correlation coefficient can be rewritten as  $r = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^n (x - \bar{x})^2 \sum_{i=1}^n (y - \bar{y})^2}}$ .

## Interpreting the Correlation Coefficient

Mathematically, the correlation coefficient is unit free and always between  $-1$  and  $1$ . The sign of the correlation indicates the sign of the relationship and matches the sign of the slope of the regression line: positive (increasing) or negative (decreasing). The closer the correlation is to  $-1$  or  $1$ , the stronger the linear relationship between  $X$  and  $Y$ . A correlation equal to  $-1$  or  $1$  indicates a perfect linear relationship between the two variables—a scatterplot of such data would display the data falling exactly on a straight line. Interestingly, a correlation of zero does not necessarily mean that there is no relationship between  $X$  and  $Y$ —it only means that there is no *linear* relationship between  $X$  and  $Y$ . The preceding computation of the correlation indicates that the sum of the products of standardized values will be zero whenever the positive and negative products balance; this can happen in many ways. Figure 12.2.3 displays several examples with a variety of correlation coefficient values.

**Figure 12.2.3** Scatterplots of data with a variety of sample correlation values



### Example 12.2.2

**Length and Weight of Snakes** In Table 12.2.2 we showed that for the snake data the sum of the products of the standardized scores is  $7.5494$ . Thus, the correlation coefficient for the lengths and weights of our sample of nine snakes is about  $0.94$ .

$$r = \frac{1}{9-1} \times 7.5494 \approx 0.94$$

In this example we may also refer to the value  $0.94$  as the **sample correlation**, since the lengths and weights of these nine snakes comprise a sample from a larger population. The sample correlation is an estimate of the **population correlation** (often denoted by the Greek letter “rho,”  $\rho$ )—in this case the correlation coefficient for the entire population of adult female *Vipera berus* snakes. In order to regard the sample correlation coefficient  $r$  as an estimate of a population parameter, it must be reasonable to assume that both the  $X$  and the  $Y$  values were selected at random, as in the following **bivariate random sampling model**:

#### Bivariate Random Sampling Model:

We regard each pair  $(x_i, y_i)$  as having been sampled at random from a population of  $(x, y)$  pairs.

In the bivariate random sampling model, the observed  $X$ 's are regarded as a random sample and the observed  $Y$ 's are also regarded as a random sample, so that the marginal statistics  $\bar{x}$ ,  $\bar{y}$ ,  $s_x$ , and  $s_y$  are estimates of corresponding population values  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x$ , and  $\sigma_y$ .

For many investigations the random sampling model is reasonable, but the additional assumption of a bivariate random sampling model is not. This is generally the case when the values of  $X$  are specified by the experimenter as in Example 12.1.1 where the researchers assigned rats to one of three dosages of amphetamine. This type of sampling model is called the random subsampling model and is defined in Section 12.4. In these cases the sample correlation coefficient is not an appropriate estimate of the population correlation.

## Inference Concerning Correlation

We have described how the correlation coefficient describes a data set within the bivariate random sampling model. Now we shall consider statistical inference based on  $r$  for data from this model. Consider the following example.

### Testing the Hypothesis $H_0: \rho = 0$

In some investigations it is not a foregone conclusion that there is any relationship between  $X$  and  $Y$ . It then may be relevant to consider the possibility that any apparent trend in the data is illusory and reflects only sampling variability. In this situation it is natural to formulate the null hypothesis

$H_0$ :  $X$  and  $Y$  are uncorrelated in the population.

or, equivalently

$H_0$ : There is no linear relationship between  $X$  and  $Y$ .

A  $t$  test of  $H_0$  is based on the test statistic

$$t_s = r \sqrt{\frac{n - 2}{1 - r^2}}$$

Critical values are obtained from Student's  $t$  distribution with

$$\text{df} = n - 2$$

The following example illustrates the application of this  $t$  test.

---

**Example  
12.2.3**

**Blood Pressure and Platelet Calcium** It is suspected that calcium in blood platelets may be related to blood pressure. As part of a study of this relationship, researchers recruited 38 subjects whose blood pressure was normal (that is, not abnormally elevated).<sup>4</sup> For each subject two measurements were made: pressure (average of systolic and diastolic measurements) and calcium concentration in the blood platelets. The data are shown in Figure 12.2.4. The sample size is  $n = 38$  and the sample correlation is  $r = 0.5832$ .

Is there evidence that blood pressure and platelet calcium are linearly related? We will test the null hypothesis

$$H_0: \rho = 0$$

against the nondirectional alternative

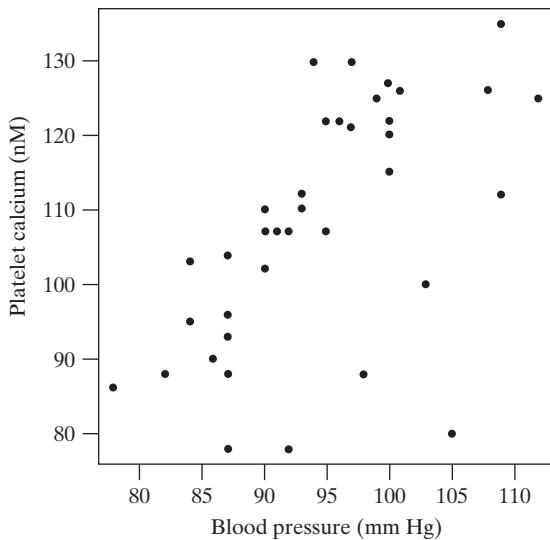
$$H_A: \rho \neq 0$$

These hypotheses are translations of the verbal hypotheses

$H_0$ : Platelet calcium is not linearly related to blood pressure.

$H_A$ : Platelet calcium is linearly related to blood pressure.

**Figure 12.2.4** Blood pressure and platelet calcium for 38 persons with normal blood pressure



Let us choose  $\alpha = 0.05$ . The test statistic is

$$t_s = 0.5832 \sqrt{\frac{38 - 2}{1 - 0.5832^2}} = 4.308$$

From Table 4 with  $df = n - 2 = 36 \approx 40$ , we find  $t_{40,0.0005} = 3.551$ . Thus, we find  $P\text{-value} < 0.0005 \times 2 = 0.001$  (since  $H_A$  is nondirectional) and we reject  $H_0$ . The data provide strong evidence that platelet calcium is linearly related to blood pressure ( $t_s = 4.308$ ,  $df = 36$ ,  $P\text{-value} < 0.001$ ). ■

**Why  $n - 2$ ?** The  $t$  statistic in the hypothesis test for the preceding population correlation coefficient has an associated  $df = n - 2$ . The origin of the  $n - 2$  is easy to explain. Any two points determine a straight line, yet such a small data set ( $n = 2$ ) provides no information about the inherent variability in the scatter of the points (or, equivalently, the strength of association between  $X$  and  $Y$ ). It is not until we observe a third point that we are able to begin estimating the strength of any relationship. As in our earlier contexts related to  $t$  distributions and  $F$ -distributions (Chapters 6, 7, 8, and 11), the degrees of freedom is the number of pieces of information provided by the data about the “noise” from which the investigator wants to extract the “signal.”

### Confidence Interval for $\rho$ (Optional)

If the sample size is large, it is possible to construct a confidence interval for  $\rho$ . The sampling distribution of the sample correlation coefficient,  $r$ , is skewed, so in order to construct the confidence interval we apply what is known as the Fisher transformation of  $r$ :

$$z_r = \frac{1}{2} \ln \left[ \frac{1 + r}{1 - r} \right]$$

where  $\ln$  is the natural logarithm (base  $e$ ). We can then construct a 95% confidence interval for  $\frac{1}{2} \ln \left[ \frac{1 + \rho}{1 - \rho} \right]$  as

$$z_r \pm 1.96 \frac{1}{\sqrt{n - 3}}$$

Finally, we can convert the limits of the confidence interval for  $\frac{1}{2} \ln \left[ \frac{1 + \rho}{1 - \rho} \right]$  into a confidence interval for  $\rho$  by solving for  $\rho$  in the equations given by

$$\frac{1}{2} \ln \left[ \frac{1 + \rho}{1 - \rho} \right] = z_r \pm 1.96 \frac{1}{\sqrt{n - 3}}$$

Intervals with other confidence levels are constructed analogously. For example, to construct a 90% confidence interval, replace 1.96 with 1.645. The construction of a confidence interval for a correlation coefficient is illustrated in Example 12.2.4.

**Example  
12.2.4**

**Blood Pressure and Platelet Calcium** For the data of Example 12.2.3 the sample size is  $n = 38$  and the sample correlation is  $r = 0.5832$ . The Fisher transformation of  $r$  gives

$$z_r = \frac{1}{2} \ln \left[ \frac{1 + 0.5832}{1 - 0.5832} \right] = \frac{1}{2} \ln \left[ \frac{1.5832}{0.4168} \right] = 0.6673$$

A 95% confidence interval for  $\frac{1}{2} \ln \left[ \frac{1 + \rho}{1 - \rho} \right]$  is

$$0.6673 \pm 1.96 \frac{1}{\sqrt{38 - 3}}$$

or  $0.6673 \pm 0.3313$ , which is  $(0.3360, 0.9986)$ .

Setting

$$\frac{1}{2} \ln \left[ \frac{1 + \rho}{1 - \rho} \right] = 0.3360 \text{ gives } \rho = \frac{e^{2(0.3360)} - 1}{e^{2(0.3360)} + 1} = 0.32$$

Setting

$$\frac{1}{2} \ln \left[ \frac{1 + \rho}{1 - \rho} \right] = 0.9986 \text{ gives } \rho = \frac{e^{2(0.9986)} - 1}{e^{2(0.9986)} + 1} = 0.76$$

We are 95% confident that the correlation between blood pressure and platelet calcium in the population is between 0.32 and 0.76. Thus, a 95% confidence interval for  $\rho$  is  $(0.32, 0.76)$ . ■

## Correlation and Causation

We have noted earlier that an observed association between two variables does not necessarily indicate any causal connection between them. It is important to remember this caution when interpreting correlation. The following example shows that even strongly correlated variables may be causally unrelated.

**Example  
12.2.5**

**Reproduction of an Alga** Akinetes are sporelike reproductive structures produced by the green alga *Pithophora oedogonia*. In a study of the life cycle of the alga, researchers counted akinetes in specimens of alga obtained from an Indiana lake on 26 occasions over a 17-month period. Low counts indicated germination of the akinetes. The researchers also recorded the water temperature and the photoperiod (hours of daylight) on each of the 26 occasions. The data showed a rather strong negative correlation between akinete counts and photoperiod; the correlation coefficient was  $r = -0.72$ . The researchers, however, recognized that this observed

correlation might not reflect a causal relationship. Longer days (increasing photoperiod) also tend to bring higher temperatures, and the akinetes might actually be responding to temperature rather than photoperiod. To resolve the question, the researchers conducted laboratory experiments in which temperature and photoperiod were varied independently; these experiments showed that temperature, not photoperiod, was the causal agent.<sup>5</sup>

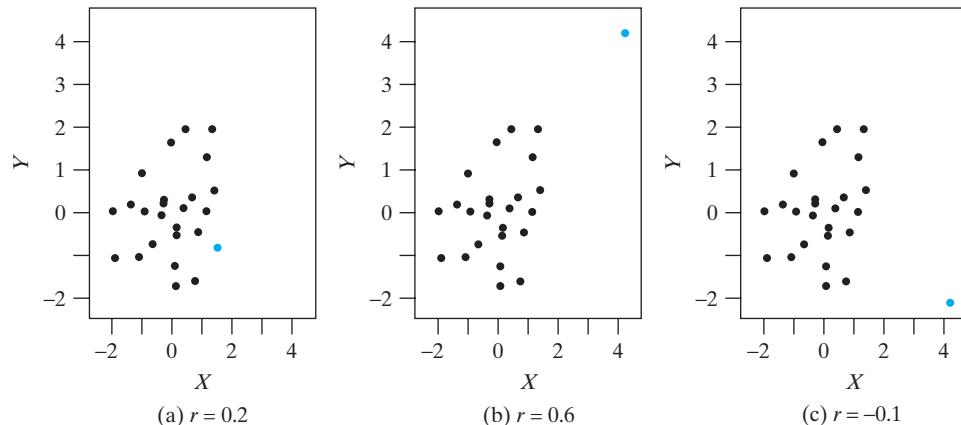
As Example 12.2.5 shows, one way to establish causality is to conduct a controlled experiment in which the putative causal factor is varied and all other factors are either held constant or controlled by randomization. When such an experiment is not possible, indirect approaches using statistical analysis can shed some light on potentially causal relationships. (One such approach will be illustrated in Example 12.8.3.)

### Cautionary Notes

To describe the results of testing a correlation coefficient, investigators often use the term *significant*, which can be misleading. For instance, a statement such as “A highly significant correlation was noted” is easily misunderstood. It is important to remember that statistical significance simply indicates rejection of a null hypothesis; it does not necessarily indicate a large or important effect. A “significant” correlation may in fact be quite a weak one; its “significance” means only that it cannot easily be explained away as a chance pattern. From the formula  $t_s = r\sqrt{\frac{n-2}{1-r^2}}$  we can see that for a fixed value of  $r$ ,  $t_s$  increases as  $n$  increases. Thus, if the sample size is large enough,  $t_s$  will be large enough for the correlation to be “significant” no matter how small  $r$  is. It is always wise to assess the practical significance of any result by considering a confidence interval for the population parameter of interest.

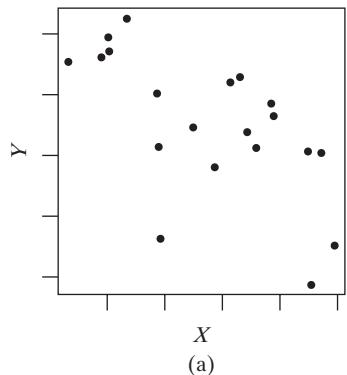
The correlation coefficient is highly sensitive to extreme points. For example, Figure 12.2.5(a) shows a scatterplot of 25 points with a correlation of  $r = 0.2$ ; one of the points has been plotted as a blue dot. Figure 12.2.5(b) shows the same points, except that the point plotted as a blue dot has been changed. The change of that single point causes the correlation coefficient to climb from 0.2 to 0.6. Figure 12.2.5(c) shows a third version of the data. In this case  $r = -0.1$ . These three graphs illustrate how a single point can greatly influence the size of the correlation coefficient. It is important to always plot the data before using  $r$  (or any other statistic) to summarize the data.

**Figure 12.2.5** The effect of outliers on the sample correlation coefficient

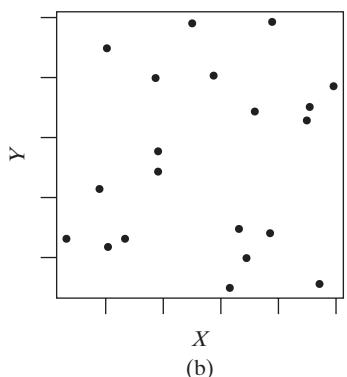


## Exercises 12.2.1–12.2.10

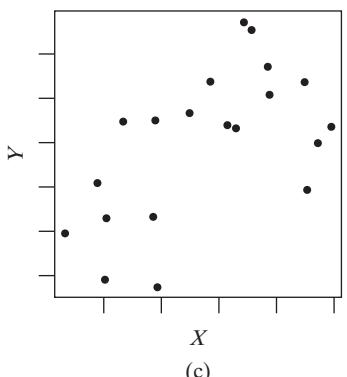
**12.2.1** Arrange the following plots in order of their correlations (from closest to  $-1$  to closest to  $+1$ ).



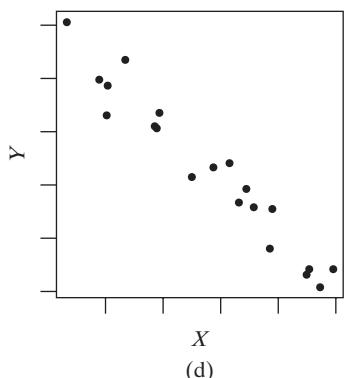
(a)



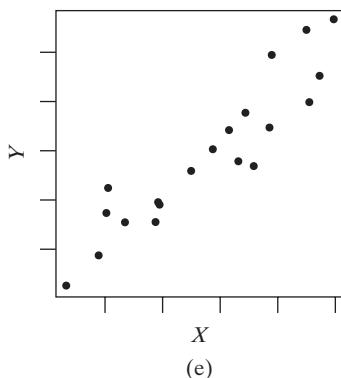
(b)



(c)



(d)



(e)

**12.2.2** Consider the following data.

	X	Y
6	6	
1	7	
3	3	
2	2	
5	14	
Mean	3.4	6.4
SD	2.1	4.7

- Plot the data. Does there appear to be a relationship between  $X$  and  $Y$ ? Is it linear or nonlinear? Weak or strong?
- Compute the sample correlation coefficient between  $X$  and  $Y$ .
- Is there significant evidence that  $X$  and  $Y$  are correlated? Conduct a test using  $\alpha = 0.05$ .

**12.2.3** In a study of natural variation in blood chemistry, blood specimens were obtained from 284 healthy people. The concentrations of urea and of uric acid were measured for each specimen, and the correlation between these two concentrations was found to be  $r = 0.2291$ . Test the hypothesis that the population correlation coefficient is zero against the alternative that it is positive.<sup>6</sup> Let  $\alpha = 0.05$ .

**12.2.4** Researchers measured the number of neurons in the CA1 region of the hippocampus in the brains of eight persons who had died of causes unrelated to brain function. They found that these data were negatively correlated with age. The sample value of  $r$  was  $-0.63$ .<sup>7</sup>

- Is this correlation coefficient significantly different from zero? Conduct a test using  $\alpha = 0.10$ .
- Suppose in part (a) you found that the correlation does significantly differ from zero. Does this provide evidence that aging is a cause for CA1 neuron loss? If not, what could be said? Briefly explain.

**12.2.5** Twenty plots, each  $10 \times 4$  meters, were randomly chosen in a large field of corn. For each plot, the plant density (number of plants in the plot) and the mean cob weight (gm of grain per cob) were observed. The results are given in the table.<sup>8</sup>

PLANT DENSITY X	COB WEIGHT Y	PLANT DENSITY X	COB WEIGHT Y
137	212	173	194
107	241	124	241
132	215	157	196
135	225	184	193
115	250	112	224
103	241	80	257
102	237	165	200
65	282	160	190
149	206	157	208
85	246	119	224

Preliminary calculations yield the following results:

$$\begin{aligned}\bar{x} &= 128.05 & \bar{y} &= 224.10 \\ s_x &= 32.61332 & s_y &= 24.95448 \\ r &= -0.94180\end{aligned}$$

- Is there significant evidence for a linear relationship between cob weight and plant density? Carry out an appropriate test using  $\alpha = 0.05$ .
- Is this study an observational study or an experiment?
- Farmers are interested in whether manipulating plant density can alter cob weight. Could these data be used to answer this question? If not, what could be said? Briefly explain.

**12.2.6** Laetisaric acid is a compound that holds promise for control of fungus diseases in crop plants. The accompanying data show the results of growing the fungus *Pythium ultimum* in various concentrations of laetisaric acid. Each growth value is the average of four radial measurements of a *P. ultimum* colony grown in a petri dish for 24 hours; there were two petri dishes at each concentration.<sup>9</sup>

- Is there significant evidence for a linear relationship between fungus growth and acid concentration? Carry out an appropriate test using  $\alpha = 0.05$ .
- Is this study an observational study or an experiment?
- It is suggested that acid could be used to retard fungus growth. Could these data be used to verify this claim? If not, what could be said? Briefly explain.

LAEATISARIC ACID CONCENTRATION X ( $\mu\text{G/ml}$ )	FUNGUS GROWTH Y (mm)
0	33.3
0	31.0
3	29.8
3	27.8
6	28.0
6	29.0
10	25.5
10	23.8
20	18.3
20	15.5
30	11.7
30	10.0
Mean	11.500
SD	10.884

$r = -0.98754$

**12.2.7** To investigate the dependence of energy expenditure on body build, researchers used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure for each man during conditions of quiet sedentary activity. The results are shown in the table.<sup>10</sup> (See also Exercise 12.5.5.)

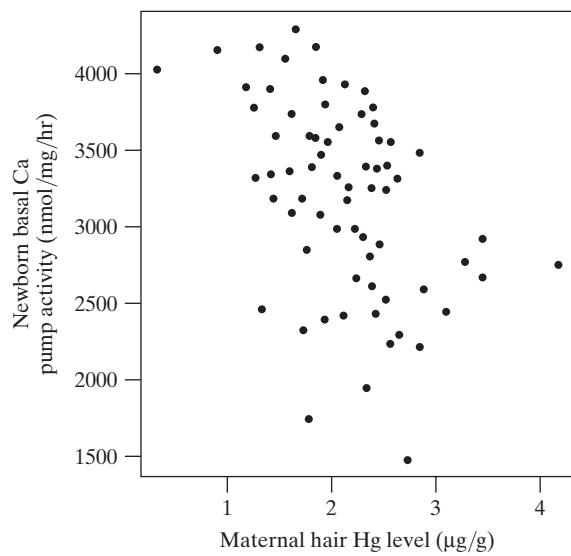
SUBJECT	FAT-FREE MASS X (kg)	ENERGY EXPENDITURE Y (kcal)
1	49.3	1,894
2	59.3	2,050
3	68.3	2,353
4	48.1	1,838
5	57.6	1,948
6	78.1	2,528
7	76.1	2,568
Mean	62.400	2,168.429
SD	12.095	308.254

$r = 0.98139$

- The correlation between energy expenditure and fat-free mass is very large (near 1). It is 0.98139, but the sample size is quite small, only 7. Is there enough evidence to claim the correlation is different from zero? Carry out an appropriate test using  $\alpha = 0.05$ .
- Is this study an observational study or an experiment?

- (c) Persons who exercise could increase their fat-free mass. Could these data be used to claim that their energy expenditure would also increase? If not, what could be said? Briefly explain.

**12.2.8** Cellular ability to regulate homeostasis is measured by basal Ca pump activity. Deregulation of calcium homeostasis can trigger serious effects of cell functioning. Can maternal mercury exposure measured by mercury deposits in hair ( $\mu\text{g/g}$ ) affect newborn's basal Ca pump activity ( $\text{nmol/mg/hr}$ )? The following data summaries and graph are from a human study involving a sample of 75 newborns and their mothers.<sup>11</sup>



$$\bar{x} = 2.11183 \quad \bar{y} = 3196.8196$$

$$s_x = 0.61166 \quad s_y = 611.34876$$

$$r = -0.45289$$

- (a) It is a good habit to always plot our data before analysis. Examining the preceding scatterplot, does

there seem to be a linear trend in the data? Is it increasing or decreasing? Is it weak or strong?

- (b) Examining the plot, we see there is a mother with a maternal hair level around  $4.2 \mu\text{g/g}$ . If her child's basal Ca pump activity were changed from about 2800 to about 2000 nmol/mg/hr, would the sample correlation increase or decrease?
- (c) Is there evidence that newborn basal Ca pump activity linearly decreases with maternal hair level? Carry out an appropriate test using  $\alpha = 0.05$ .
- (d) In part (c) you should have found that there is strong evidence for a linearly decreasing relationship between  $X$  and  $Y$ . Explain how the evidence can be so strong even though the graph displays substantial scatter and the sample correlation is not close to  $-1$ .
- (e) Based on your answer to part (c) and the design of this study, what can we say regarding the primary research question: Is there statistical evidence that maternal mercury exposure measured by mercury deposits in hair ( $\mu\text{g/g}$ ) *affects* newborn's basal Ca pump activity ( $\text{nmol/mg/hr}$ )?

**12.2.9** For each of the following examples, explain whether or not it is reasonable to treat the sample correlation coefficient,  $r$ , as an estimate of a population correlation coefficient  $\rho$ . Briefly justify your answer.

- (a) The blood chemistry data from Exercise 12.2.3.  
 (b) The CA1 neuron data from Exercise 12.2.4.  
 (c) The cob weight data from Exercise 12.2.5.  
 (d) The fungus growth data from Exercise 12.2.6.  
 (e) The basal Ca pump activity from Exercise 12.2.8.

**12.2.10** (optional) For each of the following data sets, compute a 95% confidence interval for the population correlation coefficient.

- (a) The blood chemistry data from Exercise 12.2.3.  
 (b) The cob weight data from Exercise 12.2.5.  
 (c) The energy expenditure data from Exercise 12.2.7.

## 12.3 The Fitted Regression Line

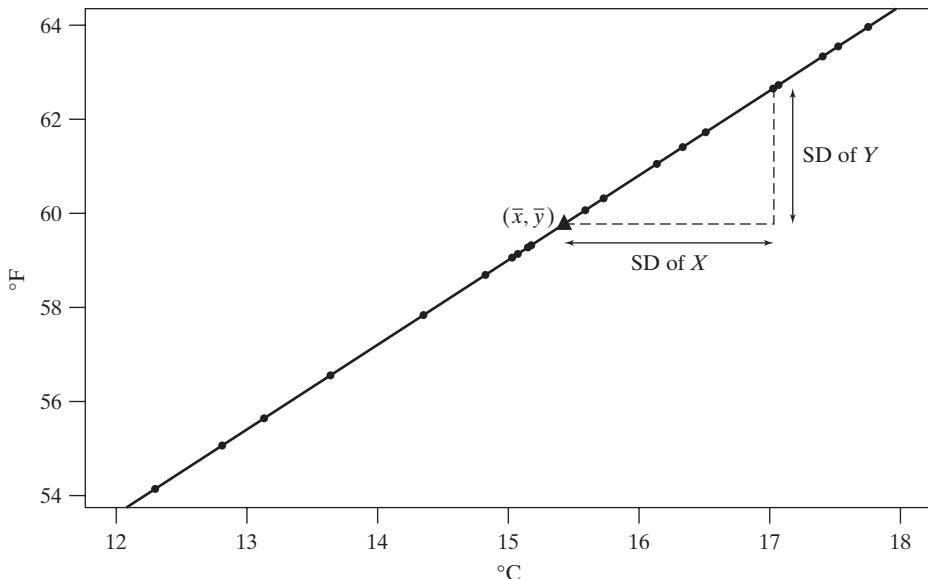
In Section 12.2 we learned how the correlation coefficient describes the strength of linear association between two numeric variables,  $X$  and  $Y$ . In this section we will learn how to find and interpret the line that best summarizes their linear relationship.

### Example 12.3.1

**Ocean Temperature** Consider a data set for which there is a perfect linear relationship between  $X$  and  $Y$  for example, temperature measured in  $X = \text{Celsius}$  and  $Y = \text{Fahrenheit}$ . Figure 12.3.1 displays 20 weekly ocean temperatures (in both  $^{\circ}\text{C}$  and  $^{\circ}\text{F}$ ) for a coastal California city along with a line that perfectly describes the relationship:<sup>\*</sup>  $y = 32 + \frac{9}{5}x$ . A summary of the data appears in Table 12.3.1.<sup>12</sup>

\*This equation is the Celsius to Fahrenheit conversion formula.

**Figure 12.3.1** Scatterplot of  $Y$  = ocean temperature in °F versus  $X$  = ocean temperature in °C. The mean value  $(\bar{x}, \bar{y})$  is denoted with a ▲



**Table 12.3.1** Summary of water temperature data

$X$ = temperature (°C)		$Y$ = temperature (°F)
Mean	15.43	59.77
SD	1.60	2.88

Because  $X$  and  $Y$  are measuring the same variable (temperature), it stands to reason that a water specimen that is 1 SD above average in °C ( $s_x = 1.60$ ) will also be 1 SD above average in °F ( $s_y = 2.88$ ). Combined, these values can describe the slope of the line that fits these data exactly:

$$\frac{\text{rise}}{\text{run}} = \frac{s_y}{s_x} = \frac{2.88}{1.60} = 1.80$$

In this example we also happen to know the equation of the line that describes the Celsius to Fahrenheit conversion. The slope of this line is  $9/5 = 1.80$ , the same value we found previously. ■

## The SD Line

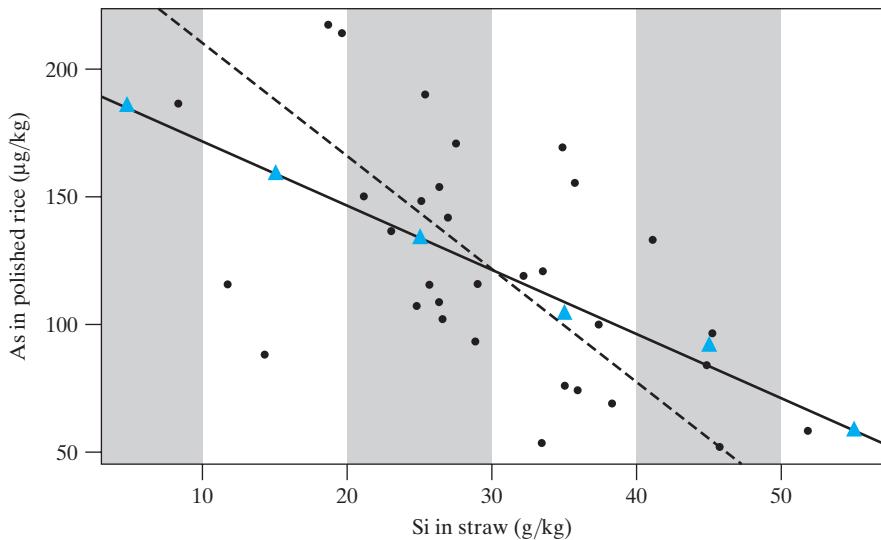
In perfect linear relationships (i.e., when  $r = \pm 1$ ) the line that fits the data exactly will have slope  $\pm s_y/s_x$  (the sign of the slope matches the sign of the correlation coefficient) and passes through the point  $(\bar{x}, \bar{y})$ . This line is sometimes referred to as the **SD line**. Our previous temperature example displays this property. But what about situations in which  $r$  is not exactly  $\pm 1$ , that is, when the relationship between  $X$  and  $Y$  is less than perfectly linear?

### Example 12.3.2

**Arsenic in Rice** In Section 12.1 we observed a scatterplot indicating that the amount of arsenic in rice and silicon in rice straw appears to be linearly related ( $r = -0.566$ ). Figure 12.3.2 displays a scatterplot of these data along with the SD line (dashed line). At first glance the SD line appears to be a good fit to these data; however, further investigation suggests otherwise. Suppose we wanted to estimate the mean arsenic concentration in rice for plants with straw silicon concentrations of

**Figure 12.3.2**

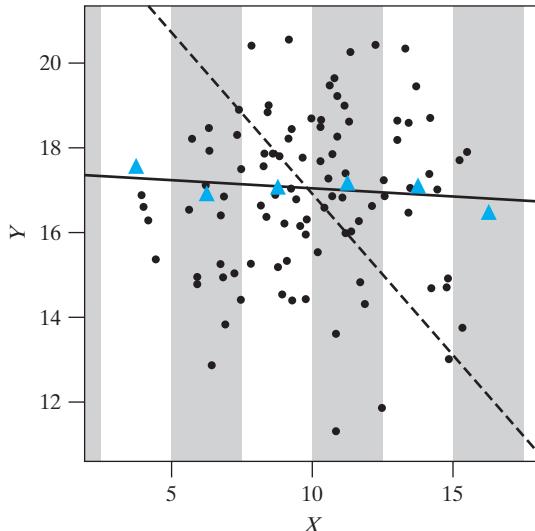
Concentrations of arsenic in rice versus silicon in straw for 32 rice plants. The dashed and solid lines are the SD and fitted regression lines, respectively. Each ▲ symbol indicates the mean rice arsenic concentration for a range of straw silicon concentrations specified by the shading.



15 g/kg. The SD line suggests an estimated mean rice arsenic concentration of approximately 190 µg/kg. Another way to estimate this value would be to simply use the mean rice arsenic concentration for plants in our sample that have straw silicon concentrations around 15 g/kg. The mean arsenic concentration in rice for straw silicon concentrations between 10 and 20 g/kg is 158.6 µg/kg (denoted by a ▲ on the graph), which is considerably less than the 190 µg/kg value given by the SD line. Similarly, for plants with straw silicon concentrations around 45 g/kg, the SD line indicates an arsenic level of about 55 µg/kg while the mean arsenic level for plants with silicon between 40 and 50 g/kg in our sample is 91.4 µg/kg, a much larger value. ■

The rice arsenic example shows that the SD line tends to overestimate the mean value of  $Y$  for below average  $X$  values and underestimate the mean value of  $Y$  for above average  $X$  values. Figure 12.3.3 shows an even more exaggerated example for a data set with a correlation even farther from  $\pm 1$ ; it is near zero ( $r = -0.05$ ). Recall that a correlation of zero indicates no linear relationship between  $X$  and  $Y$ . This lack of linear relationship is demonstrated by the fact that the

**Figure 12.3.3** Scatterplot, SD line (dashed), and fitted regression line (solid) for a sample of 100 data ( $x, y$ ) values with a correlation near zero. The ▲ symbols indicate the mean  $Y$  values for ranges of  $X$  values specified by the shading.



mean value of  $Y$  is about the same ( $\approx 17$ ) regardless of the value of  $X$  (most of the  $\blacktriangle$ 's in the plot are near 17).

If the SD line can be such a poor summary, why bother studying it? Because it is an ideal starting place based on a perfect linear relationship. With a perfect (positive) linear relationship, the SD line is the best fitting line and has a slope of  $s_y/s_x$ . Our examples illustrate that if the relationship is not perfect, the relationship between the mean  $Y$  values and  $X$  values has a flatter slope. Mathematically, it can be shown that the line that is best suited to predicting  $Y$  (in a certain sense)—the so called **least-squares** or **fitted regression** line—has a slope equal to  $r(s_y/s_x)$  and passes through the point  $(\bar{x}, \bar{y})$ . That is, for  $X$  values one standard deviation above average, the mean  $Y$  value will only be  $r$  standard deviations above average (assuming that  $r$  is positive; if  $r$  is negative, then for  $X$  values one standard deviation above average, the mean  $Y$  value will be  $r$  standard deviations below average).

### Example 12.3.3

**Arsenic in Rice** A summary and scatterplot of our rice arsenic data appear in Table 12.3.2 and Figure 12.3.4. In this example we estimate that plants with straw silicon concentrations that are  $s_x = 10.04$  g/kg above average (i.e., one standard deviation above average) will have rice arsenic concentrations that are  $25.19 \mu\text{g}/\text{kg}$  lower than average ( $r \times s_y = -0.566 \times 44.50 = -25.19$ ). Equivalently, the slope of the fitted regression line is

$$r(s_y/s_x) = -0.566 \times (44.50/10.04) = -2.51 (\mu\text{g As/kg rice})/(\text{g Si/kg straw})$$

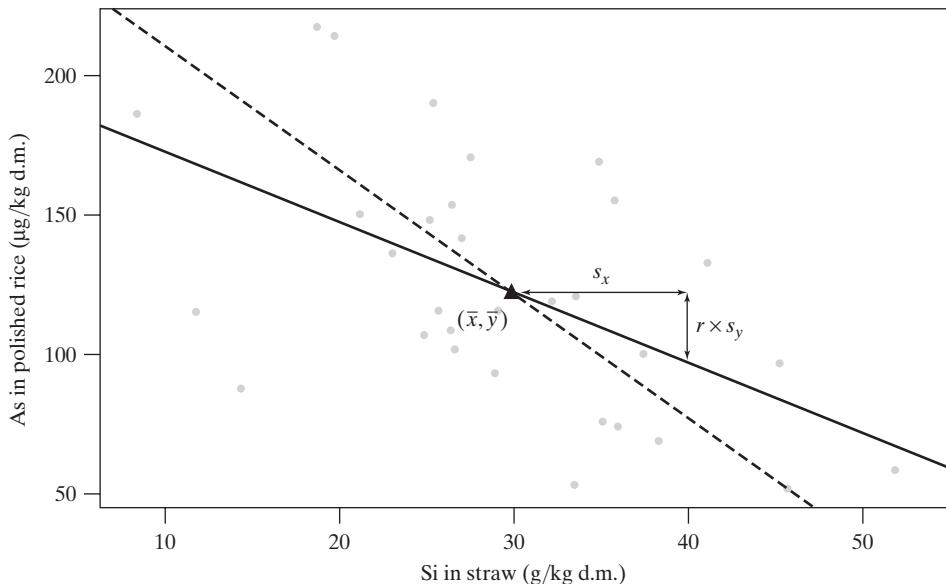
meaning that each additional 1-g/kg increase in straw silicon concentration is associated with a  $2.51-\mu\text{g}/\text{kg}$  decrease in the rice arsenic concentration, on average. ■

**Table 12.3.2** Summary of rice arsenic data

	$X = \text{Si in straw (g/kg)}$	$Y = \text{As in rice (\mu g/kg)}$
Mean	29.85	122.25
SD	10.04	44.50
	$r = -0.566$	

**Figure 12.3.4**

Concentrations of arsenic in rice versus silicon in straw for 32 rice plants with SD line (dashed) and fitted regression line (solid)



## Equation of the Fitted Regression Line

The equation of a straight line can be written as

$$Y = b_0 + b_1 X$$

where  $b_0$  is the  $y$ -intercept and  $b_1$  is the slope of the line. The slope  $b_1$  is the rate of change of  $Y$  with respect to  $X$ .

The fitted regression line of  $Y$  on  $X$  is written  $\hat{y} = b_0 + b_1 x$ . We write  $\hat{y}$  (read “ $Y$ -hat”) in place of  $Y$  to remind us that this line is providing only estimated or predicted  $Y$  values; unless the correlation is  $\pm 1$ , we don’t expect the data values to fall exactly on the line. The fitted regression line estimates the mean value of  $Y$  for any given value of  $X$ . We discuss this concept of the regression line as a *line of averages* in further detail below.

The slope and intercept of the least-squares\* regression line are calculated from the data as follows:

### Least-Squares Regression Line of $Y$ on $X$

$$\text{Slope: } b_1 = r \left( \frac{s_y}{s_x} \right)$$

$$\text{Intercept: } b_0 = \bar{y} - b_1 \bar{x}$$

Previously we saw the motivation for the formula for the slope,  $b_1$ . The formula for the intercept is also easy to motivate. We can rewrite the  $Y$ -intercept formula as

$$\bar{y} = b_0 + b_1 \bar{x}$$

which shows that *regression line passes through the joint mean*  $(\bar{x}, \bar{y})$  of our data.

We illustrate the use of these formulas by continuing our rice arsenic example.

#### **Example** 12.3.4

**Arsenic in Rice** Previously we found the slope of the regression line to be  $b_1 = r(s_y/s_x) = -2.51$  ( $\mu\text{g As/kg rice}/(\mu\text{g Si/kg straw})$ ). Using this value we find the  $Y$ -intercept,

$$b_0 = 122.25 - (-2.51) \times 29.85 = 197.17 \mu\text{g/kg}$$

Thus, our fitted regression line is  $\hat{y} = 197.17 - 2.51x$  as previously displayed in Figure 12.3.4. ■

Note that the  $Y$ -intercept, the point  $(0, b_0) = (0, 197.17)$ , does not appear on the scatterplot in Figure 12.3.4 as the  $X$ -scale limits do not extend to zero; they range from about 5 to 55 to produce a plot for which the data fill the picture nicely.

## Graph of Averages

If we have several observations of  $Y$  at a given level of  $X$ , we can estimate the population mean  $Y$  value for the given  $X$  value ( $\mu_{Y|X}$ ) by simply using the sample average of  $Y$ ,  $\bar{y}$ , for that given value of  $X$ ; we can denote this sample average as  $\bar{y}|X$ .†

\*There are other methods of finding fitted regression lines. In this text, we consider only the least-squares regression line, which aims to minimize the squared vertical distances between the data values and the fitted line.

†A more detailed exposition of these “conditional means” appears in Section 12.4.

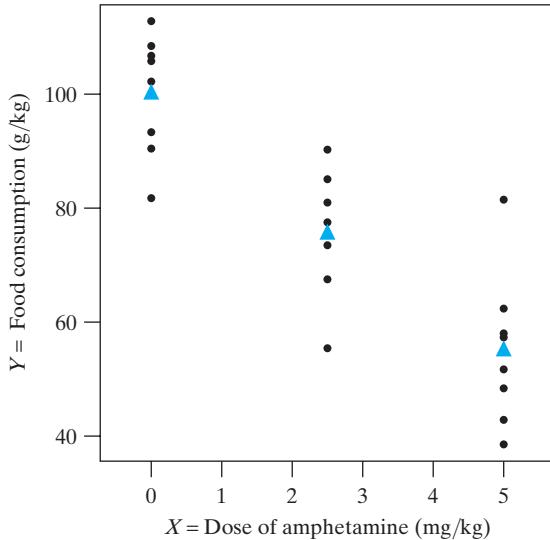
Sometimes we are able to calculate a sample average,  $\bar{y}$ , for each of several  $X$  values. A graph of  $\bar{y}|X$  is known as a **graph of averages**, since it shows the (observed) average of  $Y$  for different values of  $X$ .

---

**Example  
12.3.5**

**Amphetamine and Food Consumption** Figure 12.3.5 is a graph of averages for the food consumption data in Table 12.1.1, showing the average  $y$  value for each of the 3 levels of  $X$ . Note that the 3  $\bar{y}$ 's almost lie on a line. This supports the use of the linear model with these data.

**Figure 12.3.5** Graph of averages ( $\Delta$ ) for food consumption data from Example 12.1.1 with the original data plotted as black dots



If the  $\bar{y}$ 's in a graph of averages fall exactly on a line, then that line is the regression line and  $\mu_{Y|X}$  is estimated with  $\bar{y}|X$ . Usually, however, the  $\bar{y}$ 's are not perfectly collinear. In this case, the regression line is a *smoothed* version of the graph of averages, resulting in a fitted model in which all of the estimates of  $\mu_{Y|X}$  fall on a line. By smoothing the graph of averages into a line, we use information from *all* the observations to estimate  $\mu_{Y|X}$  at any level of  $X$ .

---

**Example  
12.3.6**

**Amphetamine and Food Consumption** If we apply the preceding regression formulas to the food consumption data in Table 12.1.1, we obtain  $b_0 = 99.3$  and  $b_1 = -9.01$ . Thus, the estimate of  $\mu_{Y|X=0}$  is 99.3 g/kg. This estimate differs slightly from  $\bar{y}|X = 0$ , which is 100.0 g/kg. The estimate 99.3 makes use of (1) the 8  $y$  values when  $X = 0$  (which averaged to 100.0) and (2) the linear trend established by the other 16 data points, which showed higher food consumption associated with lower doses. Likewise,  $\mu_{Y|X=2.5}$  is  $99.3 - 9.01 \times 2.5 = 76.78$  g/kg, which differs slightly from  $\bar{y}|X = 2.5$ , which is 75.5 g/kg, and  $\mu_{Y|X=5}$  is  $99.3 - 9.01 \times 5 = 54.25$  g/kg, which differs slightly from  $\bar{y}|X = 5$ , which is 55.0 g/kg.

The idea of smoothing the graph of averages into a straight line carries over to the setting in which we have only a single observation at each level of  $X$ , as is the case with the rice arsenic example. When we draw a line through a set of  $(X, Y)$  data, we are expressing a belief that the underlying dependence of the mean value of  $Y$  on  $X$  is smooth, even though the data may show the relationship only roughly. Linear regression is one formal way of providing a smooth description of the data.

## The Residual Sum of Squares

We now consider a statistic that describes the scatter of the points about the fitted regression line. The equation of the fitted line is  $y = b_0 + b_1x$ . Thus, for each observed  $x_i$  in our data there is a predicted  $Y$  value of

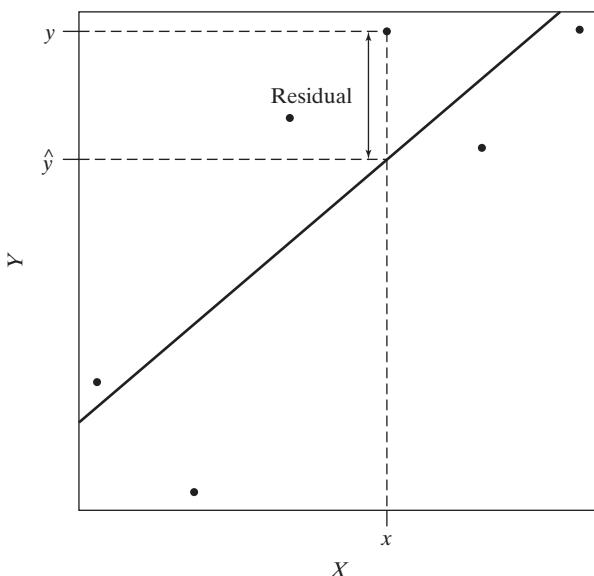
$$\hat{y}_i = b_0 + b_1x_i$$

Also associated with each observed pair  $(x_i, y_i)$  is a quantity called a **residual**, defined as

$$e_i = y_i - \hat{y}_i$$

Figure 12.3.6 shows  $\hat{y}$  and the residual for a typical data point  $(x_i, y_i)$ . It can be shown that the sum of the residuals, taking into account their signs, is always zero, because of “balancing” of data points above and below the fitted regression line. The *magnitude* (absolute value) of each residual is the vertical distance of the data point from the fitted line.

**Figure 12.3.6**  $\hat{y}$  and the residual for a typical data point  $(x, y)$



Note that a residual is calculated in terms of *vertical* distance. In using the regression model  $\hat{y} = b_0 + b_1x$  we are thinking of the variable  $X$  as a predictor and the variable  $Y$  as a response that depends on  $X$ . We care primarily about how close each observed value,  $y_i$ , is to its predicted value,  $\hat{y}_i$ . Thus, we measure vertical distance from each point to the fitted line. A summary measure of the distances of the data points from the regression line is the **residual sum of squares**, or **SS(resid)**, which is defined as follows:

### Residual Sum of Squares

$$\text{SS(resid)} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

It is clear from the definition that the residual sum of squares will be small if the data points all lie very close to the line.

The following example illustrates the computation of SS(resid).

**Example  
12.3.7**

**Arsenic in Rice** For the rice arsenic data, Table 12.3.3 indicates how  $\text{SS}(\text{resid})$  would be calculated from its definition. The values displayed are abbreviated to improve readability.

**Table 12.3.3** Calculation of  $\text{SS}(\text{resid})$  for a portion of the rice arsenic data

Obs #	$x$	$y$	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8.3	186.2	176.2 ...	10.0 ...	99.50 ...
2	11.8	115.5	167.6 ...	-52.1 ...	2716.00 ...
3	14.3	87.9	161.2 ...	-73.3 ...	5373.93 ...
4	18.7	217.2	150.2 ...	67.0 ...	4492.74 ...
5	19.7	213.8	147.8 ...	66.0 ...	4356.67 ...
6	21.2	150.0	144.0 ...	6.0 ...	35.53 ...
7	23.0	136.2	139.4 ...	-3.2 ...	10.26 ...
8	25.1	148.3	134.1 ...	14.2 ...	200.46 ...
9	26.4	153.4	130.8 ...	22.6 ...	512.49 ...
:	:	:	:	:	:
27	38.3	69.0	101.0 ...	-32.1 ...	1028.99 ...
28	41.1	132.8	94.0 ...	38.8 ...	1503.19 ...
29	45.2	96.6	83.6 ...	12.9 ...	167.11 ...
30	44.9	84.5	84.5 ...	0.0 ...	0.00 ...
31	45.7	51.7	82.5 ...	-30.8 ...	948.51 ...
32	51.8	58.6	67.1 ...	-8.5 ...	71.69 ...
Sum				0.0	41727.11 = $\text{SS}(\text{resid})$

## The Least-Squares Criterion

Many different criteria can be proposed to define the straight line that “best” fits a set of data points. The classical criterion is the least-squares criterion:

### Least-Squares Criterion

The “best” straight line is the one that minimizes the residual sum of squares.

The formulas given for  $b_0$  and  $b_1$  were derived from the least-squares criterion by applying calculus to solve the minimization problem. (The derivation is given in Appendix 12.1.) The fitted regression line is also called the “least-squares line.”

The least-squares criterion may seem arbitrary and even unnecessary. Why not fit a straight line by eye with a ruler? Actually, unless the data lie nearly on a straight line, it can be surprisingly difficult to fit a line by eye. The least-squares criterion provides an answer that does not rely on individual judgment and that (as we shall see in Sections 12.4 and 12.5) can be usefully interpreted in terms of estimating the distribution of  $Y$  values for each fixed  $X$ . Furthermore, we will see in Section 12.8 that the least-squares criterion is a versatile concept, with applications far beyond the simple fitting of straight lines.

## The Residual Standard Deviation

A summary of the results of the linear regression analysis should include a measure of the closeness of the data points to the fitted line. A measure derived from the residual sum of squares and easier to interpret is the **residual standard deviation**, denoted  $s_e$ , which is defined as follows:

### Residual Standard Deviation

$$s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n - 2}} = \sqrt{\frac{\text{SS(resid)}}{n - 2}}$$

The residual standard deviation tells how far above or below the regression line points tend to be. Thus, the residual standard deviation specifies how far off predictions made using the regression model tend to be. Notice the factor in the denominator  $n - 2$ , rather than the usual  $n - 1$ . The following example illustrates the calculation of  $s_e$ .

---

**Example**  
**12.3.8**

**Arsenic in Rice** For the rice arsenic data, we use SS(resid) from Example 12.3.7 to calculate

$$s_e = \sqrt{\frac{41727.11}{32 - 2}} = \sqrt{1390.90} = 37.30 \text{ } \mu\text{g/kg}$$

Thus, predictions for the concentrations of arsenic in rice based on the regression model tend to err by about 37.30  $\mu\text{g/kg}$  on average. ■

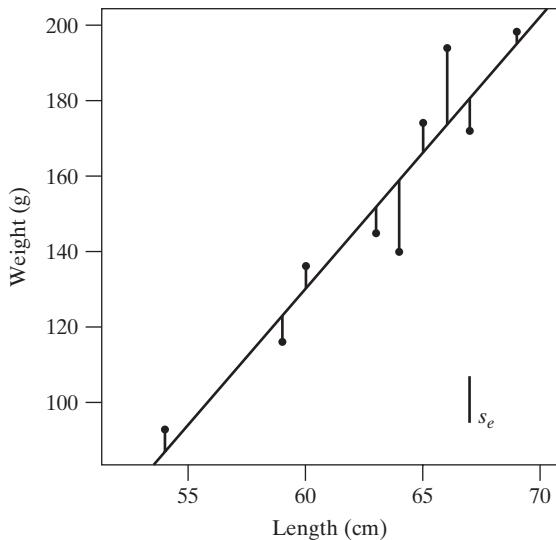
Note that the formula for  $s_e$  is closely analogous to the formula for  $s_y$ :

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

Both these SDs measure variability in  $Y$ , but the residual SD measures variability around the *regression line* and the ordinary SD measures variability around the mean,  $\bar{y}$ . Roughly speaking,  $s_e$  is a measure of the typical vertical distance of the data points from the regression line. (Notice that the unit of measurement of  $s_e$  is the same as that of  $Y$ —for instance,  $\mu\text{g/kg}$  in the case of the rice arsenic data or grams in the case of the snake data from Example 12.2.1.) Figure 12.3.7 shows the scatterplot and regression line for the snake data from Example 12.2.1 with the residuals represented as vertical lines and the residual SD indicated as a vertical ruler line. Note that the residual SD roughly indicates the magnitude of a typical residual. Finding the equation of this line and the residual standard deviation appears as an exercise at the end of this section.

In many cases,  $s_e$  can be given a more definite quantitative interpretation. Recall from Section 2.6 that for a “nice” data set, we expect roughly 68% of the observations to be within 1 SD of the mean (and similarly for 95%, 2 SDs). Recall also that these rules work best if the data follow approximately a normal distribution. Similar interpretations hold for the residual SD: For “nice” data sets that are not too small, we expect roughly 68% of the observed  $Y$ ’s to be within  $\pm s_e$  of the regression line. In other words, we expect roughly 68% of the data points to be within a vertical distance of  $s_e$  above and below the regression line (and similarly for 95%,

**Figure 12.3.7** Weight versus length of nine snakes showing the residuals and a line segment denoting the magnitude of the residual SD

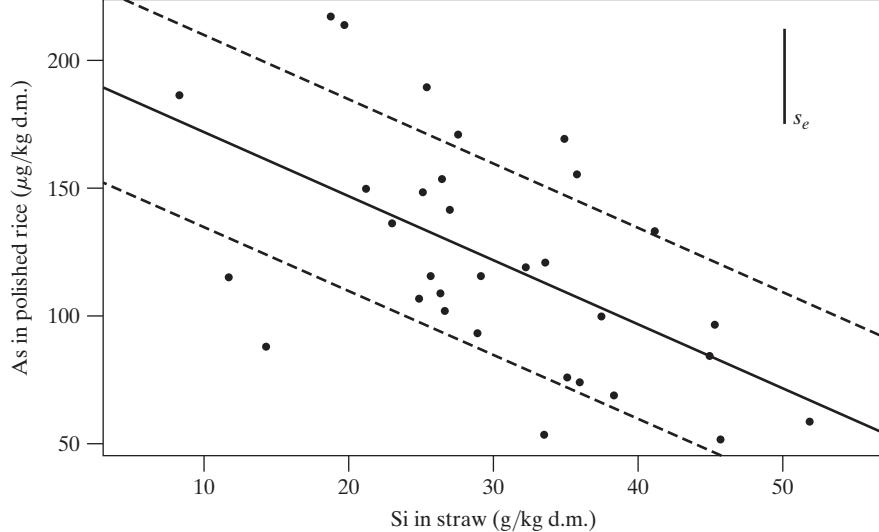


$\pm 2s_e$ ). These rules work best if the residuals follow approximately a normal distribution. The rice arsenic data we've been working with are well-suited to illustrate the 68% rule.

**Example  
12.3.9**

**Arsenic in Rice** For the rice arsenic data, the fitted regression line is  $\hat{y} = 197.17 - 2.51x$  and the residual standard deviation is  $s_e = 37.30$ . Figure 12.3.8 shows the data and the regression line. The dashed lines are a vertical distance of  $s_e$  from the regression line. Of the 32 data points, 22 are within the dashed lines; thus,  $22/32$  or  $\approx 69\%$  of the observed  $Y$ 's are within  $\pm 1s_e$  of the regression line. ■

**Figure 12.3.8** Arsenic in rice versus silicon in straw for 32 rice plants. The dashed lines are a vertical distance of  $s_e$  from the regression line



## The Coefficient of Determination

We have said that the magnitude of  $r$  describes the tightness of the linear relationship between  $X$  and  $Y$  and have seen how its value is related to the slope of the regression line. When squared, it also provides an additional and very interpretable

summary of the regression relationship. The **coefficient of determination**,  $r^2$ , describes the proportion of the variance in  $Y$  that is explained by the linear relationship between  $Y$  and  $X$ . This interpretation follows from the following fact (proved in Appendix 12.2).

### Fact 12.3.1: Approximate Relationship of $r$ to $s_e$ and $s_y$

The correlation coefficient  $r$  obeys the following approximate relationship:

$$r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

(The approximation in Fact 12.3.1 is best for large  $n$ , but it holds reasonably well even for  $n$  as small as 10.) The numerator,  $s_y^2 - s_e^2$ , can be roughly interpreted as the total variance in  $Y$  explained by the regression line: It is the difference between the variance in  $Y$  and the residual variance—the variance left over after fitting the regression line to the data. If the line fits the data very well, then  $s_e^2$  will be close to zero so this numerator will be close to  $s_y^2$ ; in this case  $r^2$  will be close to 1. At the other extreme, if the line is a very poor fit, then  $s_e^2$  will be close to  $s_y^2$  and the numerator will be close to 0; in this case  $r^2$  will be close to 0. The denominator,  $s_y^2$ , is the variance of  $Y$ ; thus the ratio,  $r^2$ , is the proportion of the variance of  $Y$  that is explained by the regression relationship between  $Y$  and  $X$ . Note that because  $-1 \leq r \leq 1$ ,  $0 \leq r^2 \leq 1$ . The following examples illustrate the interpretation and an application of  $r^2$  in context.

#### Example 12.3.10

**Arsenic in Rice** For the rice arsenic data, we found  $r = -0.566$ , so  $r^2 = 0.320$  or 32.0%. Thus, 32% of the variance in rice arsenic concentration is explained by the linear relationship between rice arsenic concentration and straw silicon concentration. ■

#### Example 12.3.11

**Amphetamine and Food Consumption** The standard deviation of food consumption for our entire sample of 24 rats (i.e., combining rats across all three doses of amphetamine) was  $s_y = 21.84$  g/kg. Further, suppose  $r^2$  was given to be 0.739. What is the estimated standard deviation of food consumption for rats given 4-mg/kg doses of amphetamine? That is, what is the value of  $s_{Y|X=4}$ ?

To answer this question we first must recognize that the value of  $X$  is irrelevant; the residual standard deviation  $s_e$  describes the standard deviation of  $Y$  values for any given  $X$  value, and therefore for  $X = 4$ . Thus, we need to find the value of  $s_e$ . From Fact 12.3.1 we have

$$r^2 \approx 1 - \frac{s_e^2}{s_y^2}$$

After a little algebra, we find that the (approximate) standard deviation of food consumption for rats given 4-mg/kg doses of amphetamine is

$$s_e \approx s_y \sqrt{1 - r^2} = 21.84 \sqrt{1 - 0.739} = 11.16 \text{ g/kg}$$

## Exercises 12.3.1–12.3.10

**12.3.1** In a study of protein synthesis in the oocyte (developing egg cell) of the frog *Xenopus laevis*, a biologist injected individual oocytes with radioactively labeled leucine. At various times after injection, he made radioactivity measurements and calculated how much of the leucine had been incorporated into protein. The results are given in the accompanying table; each leucine value is the content of labeled leucine in two oocytes. All oocytes were from the same female.<sup>13</sup>

	TIME	LEUCINE
0	0.02	
10	0.25	
20	0.54	
30	0.69	
40	1.07	
50	1.50	
60	1.74	
Mean	30.00	0.830
SD	21.60	0.637
	$r = 0.993$	
	$SS(\text{resid}) = 0.035225$	

- (a) Plot the data. Does there appear to be a relationship between  $X$  and  $Y$ ? Is it linear or nonlinear? Weak or strong?
- (b) Use linear regression to estimate the rate of incorporation of the labeled leucine.
- (c) Draw the regression line on your graph.
- (d) Calculate the residual standard deviation.

**12.3.2** In an investigation of the physiological effects of alcohol (ethanol), 15 mice were randomly allocated to three treatment groups, each to receive a different oral dose of alcohol. The dosage levels were 1.5, 3.0, and 6.0 gm alcohol per kg body weight. The body temperature of each mouse was measured immediately before the alcohol was given and again 20 minutes afterward. The accompanying table shows the drop (before minus after) in body temperature for each mouse. (The negative value  $-0.1$  refers to a mouse whose temperature rose rather than fell).<sup>14</sup>

ALCOHOL		DROP IN BODY TEMPERATURE (°C)					
		DOSE (gm/kg)	LOG(DOSE) $X$	INDIVIDUAL VALUES ( $Y$ )			
1.5	0.176	0.2	1.9	-0.1	0.5	0.8	0.66
3.0	0.477	4.0	3.2	2.3	2.9	3.8	3.24
6.0	0.778	3.3	5.1	5.3	6.7	5.9	5.26

- (a) Plot the mean drop in body temperature versus dose. Plot the mean drop in body temperature versus  $\log(\text{dose})$ . Which plot appears more nearly linear?
- (b) Plot the individual  $(x, y)$  data points [where  $X = \log(\text{dose})$ ].
- (c) For the regression of  $Y$  on  $X = \log(\text{dose})$  preliminary calculations yield the following:  $\bar{x} = 0.477$ ,  $\bar{y} = 3.05333$ ,  $s_x = 0.25439$ ,  $s_y = 2.13437$ ,  $r = 0.91074$ . Calculate the fitted regression line and the (approximate) residual standard deviation.
- (d) Draw the regression line on your graph.
- (e) Is this study an example of an observational study or an experiment? How can you tell?
- (f) Could data from this study be used to determine whether or not alcohol lowers body temperature? Briefly explain.

**12.3.3** Consider the cob weight data from Exercise 12.2.5.

- (a) Use the summaries in Exercise 12.2.5 to calculate the fitted regression line and approximate residual standard deviation.
- (b) Interpret the value of the slope of the regression line,  $b_1$ , in the context of this setting.
- (c)  $SS(\text{resid}) = 1337.3$ . Use this value to compute the residual standard deviation. How does it compare to the approximate value determined in part (a)?
- (d) Interpret the value of  $s_e$  in the context of this setting.
- (e) What proportion of the variation in cob weights is explained by the linear relationship between cob weight and density?

**12.3.4** Consider the Fungus growth data from Exercise 12.2.6.

- (a) Calculate the linear regression of  $Y$  on  $X$ .
- (b) Plot the data and add the regression line to your graph. Does the line appear to fit the data well?
- (c)  $SS(\text{resid}) = 16.7812$ . Use this to compute  $s_e$ . What are the units of  $s_e$ ?
- (d) Draw a ruler line on your graph to show the magnitude of  $s_e$ . (See Figure 12.3.8).

**12.3.5** Consider the Energy Expenditure data from Exercise 12.2.7.

- (a) Calculate the linear regression of  $Y$  on  $X$ .
- (b) Plot the data and add the regression line to your graph. Does the line appear to fit the data well?
- (c) Interpret the value of the slope of the regression line,  $b_1$ , in the context of this setting.
- (d)  $SS(\text{resid}) = 21026.1$ . Use this to compute  $s_e$ . What are the units of  $s_e$ ?

**12.3.6** The rowan (*Sorbus aucuparia*) is a tree that grows in a wide range of altitudes. To study how the tree adapts to its varying habitats, researchers collected twigs with attached buds from 12 trees growing at various altitudes in North Angus, Scotland. The buds were brought back to the laboratory and measurements were made of the dark respiration rate. The accompanying table shows the altitude of origin (in meters) of each batch of buds and the dark respiration rate (expressed as  $\mu\text{l}$  of oxygen per hour per mg dry weight of tissue).<sup>15</sup>

ALTITUDE OF ORIGIN $X$ (m)	RESPIRATION RATE $Y$ ( $\mu\text{l}/\text{hr} \times \text{mg}$ )
90	0.11
230	0.20
240	0.13
260	0.15
330	0.18
400	0.16
410	0.23
550	0.18
590	0.23
610	0.26
700	0.32
790	0.37
Mean	433.333
SD	214.617
$r = 0.88665$	
$\text{SS(resid)} = 0.013986$	

- (a) Calculate the linear regression of  $Y$  on  $X$ .
- (b) Plot the data and the regression line.
- (c) Interpret the value of the slope of the regression line,  $b_1$ , in the context of this setting.
- (d) Calculate the residual standard deviation.

**12.3.7** Scientists studied the relationship between the length of the body of a bullfrog and how far it can jump. Eleven bullfrogs were included in the study. The results are given in the table.<sup>16</sup>

- (a) Calculate the linear regression of  $Y$  on  $X$ .
- (b) Interpret the value of the slope of the regression line,  $b_1$ , in the context of this setting.
- (c) What proportion of the variation in maximum jump distances can be explained by the linear relationship between jump distance and frog length?
- (d) Calculate the residual standard deviation and specify the units.
- (e) Interpret the value of the residual standard deviation in the context of this setting.

BULLFROG	LENGTH $X$ (mm)	MAXIMUM JUMP $Y$ (cm)
1	155	71.0
2	127	70.0
3	136	100.0
4	135	120.0
5	158	103.3
6	145	116.0
7	136	109.2
8	172	105.0
9	158	112.5
10	162	114.0
11	162	122.9
Mean	149.6364	103.9909
SD	14.4725	17.9415
$r = 0.28166$		
$\text{SS(resid)} = 2,963.61$		

**12.3.8** The peak flow rate of a person is the fastest rate at which the person can expel air after taking a deep breath. Peak flow rate is measured in units of liters per minute and gives an indication of the person's respiratory health. Researchers measured peak flow rate and height for each of a sample of 17 men. The results are given in the table.<sup>17</sup>

SUBJECT	HEIGHT $X$ (cm)	PEAK FLOW RATE $Y$ (l/min)
1	174	733
2	183	572
3	176	500
4	169	738
5	183	616
6	186	787
7	178	866
8	175	670
9	172	550
10	179	660
11	171	575
12	184	577
13	200	783
14	195	625
15	176	470
16	176	642
17	190	856
Mean	180.4118	660.0000
SD	8.5591	117.9952
$r = 0.32725$		
$\text{SS(resid)} = 198,909$		

- (a) Calculate the linear regression of  $Y$  on  $X$ .
- (b) What proportion of the variation in flow rate is explained by the linear regression of flow rate on height?
- (c) For each subject, calculate the predicted peak flow rate, using the regression equation from part (a).
- (d) For each subject, calculate the residual, using the results from part (c).
- (e) Calculate  $s_e$  and specify the units.
- (f) What percentage of the data points are within  $\pm s_e$  of the regression line? That is, what percentage of the 17 residuals are in the interval  $(-s_e, s_e)$ ?

**12.3.9** For each of the following data sets, prepare a plot like Figure 12.3.8, showing the data, the fitted regression

line, and two lines whose vertical distance above and below the regression line is  $s_e$ . What percentage of the data points are within  $\pm s_e$  of the regression line? What percentage of the data points do you expect to find within  $\pm s_e$  of the regression line? How do these values compare?

- (a) The body temperature data of Exercise 12.3.2.
- (b) The corn yield data of Exercise 12.3.3.

**12.3.10** Suppose a large sample of  $(x, y)$  pairs were used to fit the regression of  $Y$  on  $X$ . Now suppose we observed 100 further  $(x, y)$  pairs. About how many of these new observations would you expect to be farther than  $2s_e$  from the regression line?

## 12.4 Parametric Interpretation of Regression: The Linear Model

One use of regression analysis is simply to provide a concise description of the data. The quantities  $b_0$  and  $b_1$  locate the regression line and  $s_e$  describes the scatter of the points about the line.

For many purposes, however, data description is not enough. In this section we consider inference from the data to a larger population. In previous chapters we have spoken of one or several populations of  $Y$  values. Now, to encompass the  $X$  variable as well, we need to expand the notion of a population.

### Conditional Populations and Conditional Distributions

A **conditional population** of  $Y$  values is a population of  $Y$  values associated with a fixed, or given, value of  $X$ . Within a conditional population we may speak of the **conditional distribution** of  $Y$ . The mean and standard deviation of a conditional population distribution are denoted as

$$\mu_{Y|X} = \text{Population mean } Y \text{ value for a given } X$$

$$\sigma_{Y|X} = \text{Population SD of } Y \text{ values for a given } X$$

(Note that the “given” symbol “|” is the same one used for conditional probability in Chapters 3 and 10.) The following example illustrates this notation.

---

**Example  
12.4.1**

**Amphetamine and Food Consumption** In the rat experiment introduced in Example 12.1.1, the response variable  $Y$  was food consumption and the three values of  $X$  (dose) were  $X = 0$ ,  $X = 2.5$ , and  $X = 5$ . In Example 12.3.5 we examined the graph of averages and considered the food consumption data as three independent samples (as for an ANOVA). In the ANOVA context we denote the three population means as  $\mu_1$ ,  $\mu_2$ , and  $\mu_3$ . In regression notation these means would be denoted as

$$\mu_{Y|X=0} \quad \mu_{Y|X=2.5} \quad \mu_{Y|X=5}$$

respectively. Similarly, the three population standard deviations, which would be denoted as  $\sigma_1$ ,  $\sigma_2$ , and  $\sigma_3$  in an ANOVA context, would be denoted as

$$\sigma_{Y|X=0} \quad \sigma_{Y|X=2.5} \quad \sigma_{Y|X=5}$$

respectively. In other words, the symbols

$$\mu_{Y|X} \text{ and } \sigma_{Y|X}$$

represent the mean and standard deviation of food consumption values for rats that are given dose  $X$  of amphetamine. ■

In observational studies, conditional distributions pertain to subpopulations rather than experimental treatment groups, as in the following example.

### Example 12.4.2

**Height and Weight of Young Men** Consider the variables

$$X = \text{Height}$$

and

$$Y = \text{Weight}$$

for a population of young men. The conditional means and standard deviations are

$$\mu_{Y|X} = \text{Mean weight of men who are } X \text{ inches tall}$$

$$\sigma_{Y|X} = \text{SD of weights of men who are } X \text{ inches tall}$$

Thus,  $\mu_{Y|X}$  and  $\sigma_{Y|X}$  are the mean and standard deviation of weight in the *subpopulation* of men whose height is  $X$ . Of course, there is a different subpopulation for each value of  $X$ . ■

## The Linear Model

When we conduct a linear regression analysis, we think of  $Y$  as having a distribution that depends on  $X$ . The analysis can be given a parametric interpretation if two conditions are met. These conditions, which constitute the **linear model**, are given in the following box.

### The Linear Model

1. *Linearity.*  $Y = \mu_{Y|X} + \varepsilon$  where  $\mu_{Y|X}$  is a linear function of  $X$ ; that is

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

Thus,  $Y = \beta_0 + \beta_1 X + \varepsilon$ .

2. *Constancy of standard deviation.*  $\sigma_{Y|X}$  does not depend on  $X$ . We denote this constant value as  $\sigma_\varepsilon$ .

In the linear model  $Y = \beta_0 + \beta_1 X + \varepsilon$ , the  $\varepsilon$  term represents **random error**. We include this term in the model to reflect the fact that  $Y$  varies, even when  $X$  is fixed. The variability of  $Y$  for a fixed value of  $X$  is measured by the conditional standard deviation of  $Y$ ,  $\sigma_{Y|X}$ . But, because the linear model stipulates that this standard deviation is the same for every value of  $X$ , we commonly use the notation  $\sigma_\varepsilon$  to represent this standard deviation and refer to it as the standard deviation of the random error.

The following two examples show the meaning of the linear model.

**Example 12.4.3**

**Amphetamine and Food Consumption** For the rat food consumption experiment, the linear model asserts that (1) the population mean food consumption is a linear function of dose, and that (2) the population standard deviation of food consumption values is the same for all doses. Notice that the second condition is closely analogous to the condition in ANOVA that the population SDs are equal:  $\sigma_1 = \sigma_2 = \sigma_3$ . The linear model also allows for the fact that there is variability in  $Y$  when  $X$  is fixed. For example, there were 8 observations for which  $X = 5$ . The 8  $y$ -values averaged 55.0, but none of the observations was equal to 55.0; there was substantial variability within the 8  $y$ -values. This variability is quantified by the SD of 13.3. ■

**Example 12.4.4**

**Height and Weight of Young Men** We consider an idealized fictitious population of young men whose joint height and weight distribution fits the linear model exactly. For our fictitious population we will assume that the conditional means and SDs of weight given height are as follows:

$$\begin{aligned}\mu_{Y|X} &= -145 + 4.25X \\ \sigma_\varepsilon &= 20\end{aligned}$$

Thus, the regression parameters of the population are  $\beta_0 = -145$  and  $\beta_1 = 4.25$ . (This fictitious population resembles that of U.S. 17-year-olds.)<sup>18</sup> Thus, the model is  $Y = -145 + 4.25X + \varepsilon$ .

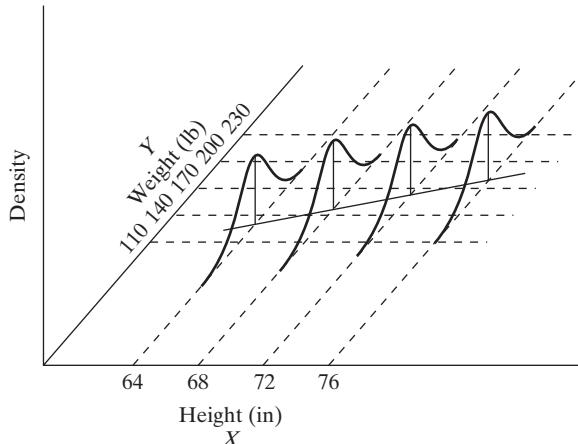
Table 12.4.1 shows the conditional means and SDs of  $Y$  = weight for a few selected values of  $X$  = height. Figure 12.4.1 shows the conditional distributions of  $Y$  given  $X$  for these selected subpopulations.

**Table 12.4.1** Conditional means and SDs of weight given height in a population of young men\*

Height (in) $X$	Mean weight (lb) $\mu_{Y X}$	Standard deviation of weights (lb) $\sigma_{Y X}$
64	127	20
68	144	20
72	161	20
76	178	20

\*Note that all values of  $\sigma_{Y|X}$  are the same; they equal  $\sigma_\varepsilon = 20$ .

**Figure 12.4.1** Conditional distributions of weight given height in a population of young men



Note, for example, that if height = 68 (in), then the mean weight is 144 (lb) and the SD of the weights is 20 (lb). For this subpopulation,  $Y = 144 + \varepsilon$ . If a particular young man who is 68 inches tall weighs 145 pounds, then  $\varepsilon = 1$  for him. If another 68-inch-tall young man weighs 140 pounds, then  $\varepsilon = -4$  in his case. Of course,  $\beta_0$ ,  $\beta_1$ , and  $\varepsilon$  are generally not observable. This example is fictitious. ■

**Remark.** Actually, the term *regression* is not confined to linear regression. In general, the relationship between  $\mu_{Y|X}$  and  $X$  is called the *regression of  $Y$  on  $X$* . The linearity assumption asserts that the regression of  $Y$  on  $X$  is linear rather than, for instance, a curvilinear function.

## Estimation in the Linear Model

Consider now the analysis of a set of  $(X, Y)$  data. Suppose we assume that the linear model is an adequate description of the true relationship of  $Y$  and  $X$ . Suppose further that we are willing to adopt the following **random subsampling model**:

### Random Subsampling Model

For each observed pair  $(x, y)$ , we regard the value  $y$  as having been sampled at random from the conditional population of  $Y$  values associated with the  $X$  value  $x$ .

Within the framework of the linear model and the random subsampling model, the quantities  $b_0$ ,  $b_1$ , and  $s_e$  calculated from a regression analysis can be interpreted as estimates of population parameters:

$b_0$  is an estimate of  $\beta_0$

$b_1$  is an estimate of  $\beta_1$

$s_e$  is an estimate of  $\sigma_\varepsilon$

### Example 12.4.5

**Length and Weight of Snakes** From the summaries of the snake data of Example 12.2.1 and 12.2.2, we can compute the following regression coefficients  $b_0 = -301$ ,  $b_1 = 7.19$ , and  $s_e = 12.5$  (computing these yourself from the provided summaries would be a good exercise). Thus,

-301 is our estimate of  $\beta_0$

7.19 is our estimate of  $\beta_1$

12.5 is our estimate of  $\sigma_\varepsilon$  ■

The application of the linear model to the snake data has yielded two benefits. First, the slope of the regression line, 7.19 gm/cm, is an estimate of a morphological parameter (“weight per unit length”), which is of potential biological interest in characterizing the population of snakes. Second, we have obtained an estimate (12.5 g) of the variability of weight among snakes of fixed length, even though no direct estimate of this variability was possible because no two of the observed snakes were the same length.

## Interpolation in the Linear Model

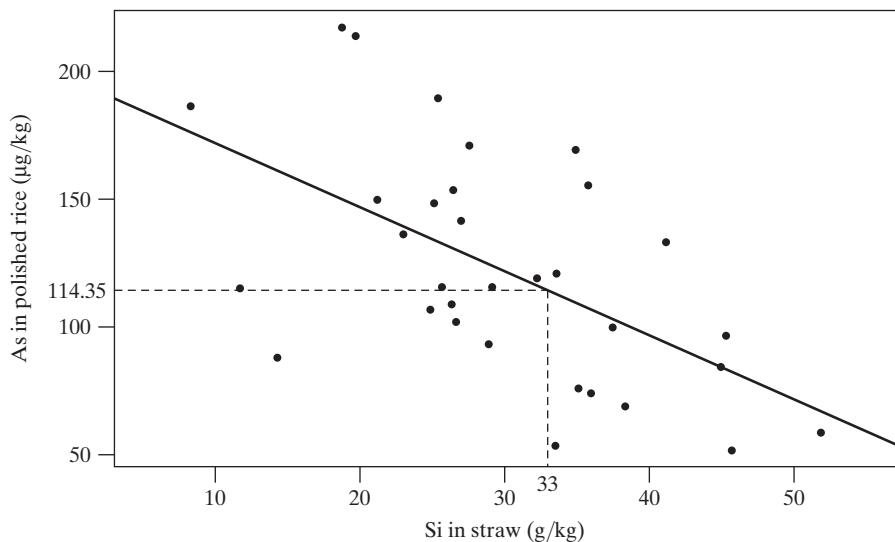
In Section 12.3 we regarded the regression line as a line of averages. The idea of smoothing the graph of averages into a straight line can be extended to the setting in which we have only a single observation at each level of  $X$ . When we draw a line through a set of  $(X, Y)$  data, we are expressing a belief that the underlying dependence of  $Y$  on  $X$  is smooth, even though the data may show the relationship only roughly. Linear regression is one formal way of providing a smooth description of the data as illustrated in the following example.

### Example 12.4.6

**Arsenic in Rice** What are the mean and standard deviation of arsenic concentrations in rice for plants with straw silicon concentrations of 33 g/kg? None of our observed plants had a straw silicon concentration of 33 g/kg. If there were some observations with this much silicon, we could average the associated arsenic concentrations to obtain one answer to our question, but because there is an apparent linear relationship between  $X$  and  $Y$ , we can use the line to obtain an even better estimate of the mean rice arsenic concentration that uses all of the data. In Example 12.3.4 we found the regression equation to be  $\hat{y} = 197.17 - 2.51x$  and  $s_e = 37.30$ . Thus the estimated mean arsenic concentration for straw with 33 g/kg silicon is  $197.17 - 2.51 \times 33 = 114.35 \mu\text{g}/\text{kg}$  with a standard deviation of  $s_e = 37.30 \mu\text{g}/\text{kg}$ . Figure 12.4.2 shows the interpolation graphically. ■

**Figure 12.4.2**

Concentrations of arsenic in rice versus silicon in straw for 32 rice plants



Note that estimation of the mean uses the linearity condition of the linear model, while estimation of the standard deviation uses the condition of constant standard deviation. In some situations only the linearity condition may be plausible, and then only the mean would be estimated.

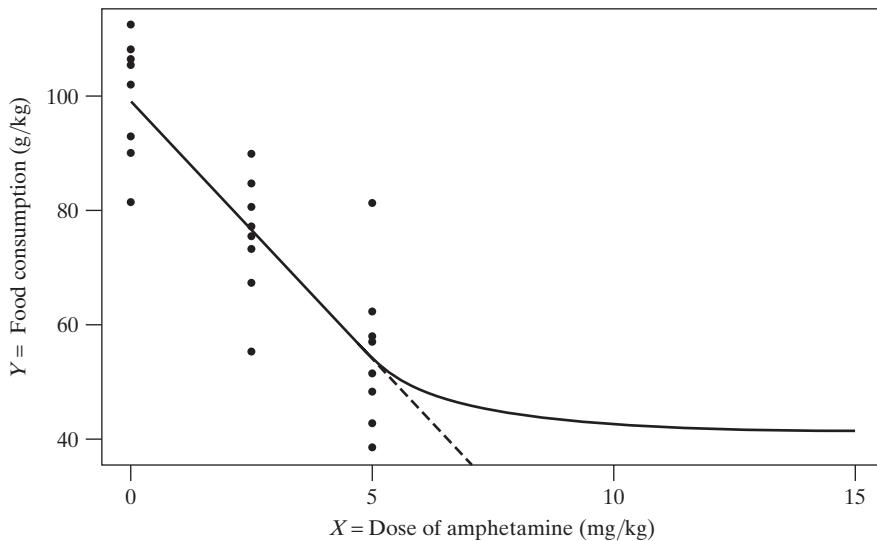
Example 12.4.6 is an example of **interpolation**, because the  $X$  values we chose ( $X = 33$  for the rice arsenic and 3.5 for the food consumption examples) were within the range of observed values of  $X$ . By contrast, **extrapolation** is the use of a regression line (or other curve) to predict  $Y$  for values of  $X$  that are outside the range of the data. Extrapolation should be avoided whenever possible, because there is usually no assurance that the relationship between  $\mu_{Y|X}$  and  $X$  remains linear for  $X$  values outside the range of those observed. Many biological relationships are linear for only part of the possible range of  $X$  values. The following is an example.

**Example  
12.4.7**

**Amphetamine and Food Consumption** The dose-response relationship for the rat food consumption experiment of Example 12.1.1 looks approximately like Figure 12.4.3.<sup>19</sup> The data cover only the linear portion of the relationship. Clearly it would be unwise to extrapolate the fitted line out to  $X = 10$  or  $X = 15$ . ■

**Figure 12.4.3**

Dose-response curve (mean response versus dose) for rat food consumption experiment



### Prediction and the Linear Model

Consider the setting of using height,  $X$ , to predict weight,  $Y$ , for a large group of young men for whom the average weight is 150 pounds. Suppose a young man is chosen at random and we must predict his weight.

1. If we don't know anything about the height of the man, then the best estimate we can give of his weight is the overall average weight,  $\bar{y} = 150$ .
2. Suppose we learn that the man's height is 76 inches. If we know that the average weight of all 76-inch-tall men in the group is 180 pounds, then we can use this conditional average,  $\bar{y}|x = 76$ , as our prediction of the man's weight. We expect this prediction, which essentially is using the graph of averages (but without smoothing), to be more accurate than the one given in part 1.
3. Suppose we learn that the man's height is 76 inches and we also know that the least-squares regression equation is  $Y = -140 + 4.3X$ . Then we can use the value  $x = 76$  to get a prediction, which would be  $-140 + 4.3 \times 76 = 186.8$ .

Is the prediction in 3 better than the prediction made in 2? Since using the regression equation amounts to smoothing the graph of averages, we expect prediction 3 to be better than prediction 2 *to the extent that we believe that there is a linear relationship between height and weight*. Prediction 3 has the advantage of using information from all the data points, not just those for which  $x = 76$ . Method 3 also has the advantage of allowing for predictions when the  $x$  value (the height) is not one that is in the original data set (as discussed in the preceding subsection “Interpolation in the Linear Model”), so that  $\bar{y}|x$  is not known. However, method 3 will give poor predictions if the linear relationship does not hold. Thus it is very important to think about such relationships, and to explore them graphically, before using a regression model.

## Exercises 12.4.1–12.4.9

**12.4.1** For the data in Exercise 12.2.6 there were two observations for which  $X = 0$ . The average response ( $Y$  value) for these points is  $\frac{33.3 + 31.0}{2} = 32.15$ . However, the intercept of the regression line,  $b_0$ , is not 32.15. Why not? Why is  $b_0$  a better estimate of the average fungus growth when laetisaric acid concentration is zero than 32.15?

**12.4.2** Refer to the body temperature data of Exercise 12.3.2. Assuming that the linear model is applicable, estimate the mean and the standard deviation of the drop in body temperature that would be observed in mice given alcohol at a dose of 2 gm/kg. [Tip: Is the  $X$  variable dose or  $\log(\text{dose})$ ?]

**12.4.3** Refer to the cob weight data of Exercises 12.2.5 and 12.3.3. Assume that the linear model holds.

- (a) Estimate the mean cob weight to be expected in a plot containing (i) 100 plants; (ii) 120 plants.
- (b) Assume that each plant produces one cob. How much grain would we expect to get from a plot containing (i) 100 plants? (ii) 120 plants?

**12.4.4** (Continuation of Exercise 12.4.3). For the cob weight data,  $\text{SS}(\text{resid}) = 1,337.3$ . Estimate the standard deviation of cob weight in plots containing (i) 100 plants; (ii) 120 plants.

**12.4.5** Refer to the fungus growth data of Exercise 12.2.6. For these data,  $\text{SS}(\text{resid}) = 16.7812$ . Assuming

that the linear model is applicable, find estimates of the mean and standard deviation of fungus growth at a laetisaric acid concentration of 15  $\mu\text{g}/\text{ml}$ .

**12.4.6** Refer to the energy expenditure data of Exercise 12.2.7. Assuming that the linear model is applicable, estimate the 24-hour energy expenditure of a man whose fat-free mass is 55 kg.

**12.4.7** Refer to the Ca pump activity of Exercise 12.2.8. For these data  $\text{SS}(\text{resid}) = 21,984,623$ .

- (a) Assuming that the linear model is applicable, estimate the mean and standard deviation basal Ca pump activity for children born to mothers with a hair Hg level of 3  $\mu\text{g}/\text{g}$ .
- (b) Using the values computed in part (a) to support your answer, would it be surprising for a mother with a hair Hg level of 3  $\mu\text{g}/\text{g}$  to give birth to a child with a basal Ca pump activity above 4000 nmol/mg/hr?

**12.4.8** Refer to the bullfrog data of Exercise 12.3.7. Assuming that the linear model is applicable, estimate the maximum jump length of a bullfrog whose body length is 150 mm.

**12.4.9** Refer to the peak flow data of Exercise 12.3.8. Assuming that the linear model is applicable, find estimates of the mean and standard deviation of peak flow for men 180 cm tall.

## 12.5 Statistical Inference Concerning $\beta_1$

The linear model provides interpretations of  $b_0$ ,  $b_1$ , and  $s_e$  that take them beyond data description into the domain of statistical inference. In this section we consider inference about the true slope  $\beta_1$  of the regression line. The methods are based on the condition that the conditional population distribution of  $Y$  for each value of  $X$  is a normal distribution. This is equivalent to stating that in the linear model of  $Y = \beta_0 + \beta_1 X + \varepsilon$ , the  $\varepsilon$  values come from a normal distribution.

### The Standard Error of $b_1$

Within the context of the linear model,  $b_1$  is an estimate of  $\beta_1$ . Like all estimates calculated from data,  $b_1$  is subject to sampling error. The standard error of  $b_1$  is calculated as follows:

#### Standard Error of $b_1$

$$\text{SE}_{b_1} = \frac{s_e}{s_x \sqrt{n - 1}}$$

The following example illustrates the calculation of  $\text{SE}_{b_1}$ .

**Example  
12.5.1**

**Length and Weight of Snakes** For the snake data, we found in Table 12.2.2 that  $n = 9$ ,  $s_x = 4.637$ , and in Example 12.4.5 that  $s_e = 12.5$ . The standard error of  $b_1$  is

$$\text{SE}_{b_1} = \frac{12.5}{4.637\sqrt{9 - 1}} = 0.9531$$

To summarize, the slope of the fitted regression line (from Example 12.4.5) is

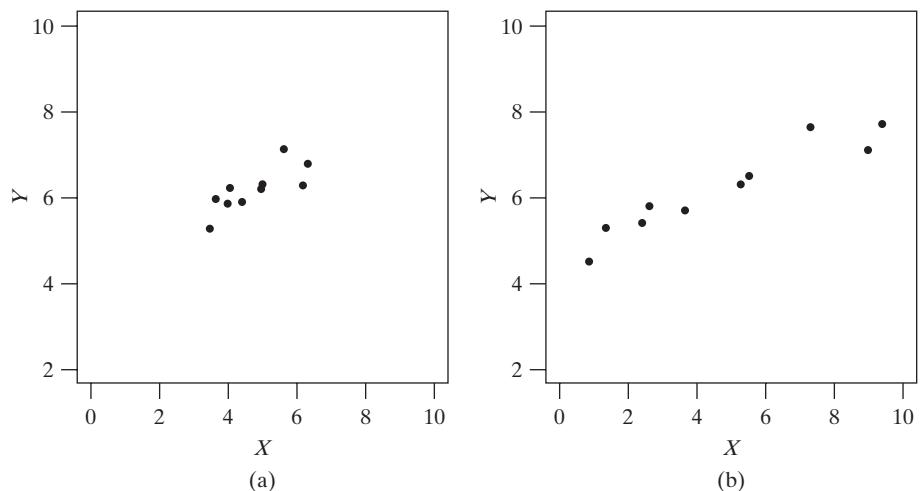
$$b_1 = 7.19 \text{ gm/cm}$$

and the standard error of this slope is

$$\text{SE}_{b_1} = 0.95 \text{ gm/cm}$$

**Structure of the SE.** Let us see how the standard error of  $b_1$  depends on various aspects of the data. In the same way that  $\text{SE}_{\bar{Y}}$  depends on the variability in the  $Y$  data ( $s_y$ ) and the sample size ( $n$ ),  $\text{SE}_{b_1}$  depends on the scatter of the data about the regression line ( $s_e$ ) and the size of the sample ( $n$ ). The formula for  $\text{SE}_{b_1}$  supports our intuition showing that data with less scatter about the regression line (smaller  $s_e$ ) and larger sample sizes (larger  $n$ ) produce more precise estimates of  $\beta_1$  (i.e., a smaller  $\text{SE}_{b_1}$ ). While variability in  $Y$  and sample size are the only two factors that affect our ability to estimate a population mean precisely ( $\text{SE}_{\bar{Y}}$ ), there is a third factor that is important for precise estimation of  $\beta_1$ : the variability of the  $X$  data. The more spread out our  $X$  values (larger  $s_x$ ), the more precise our estimate of  $\beta_1$  will be. The dependence on the spread in the  $X$  values is illustrated in Figure 12.5.1, which shows two data sets with the same value of  $s_e$  and the same value of  $n$ , but different values of  $s_x$ . Imagine using a ruler to fit a straight line by eye; it is intuitively clear that the data in case (b)—with the larger  $s_x$ —would determine the slope of the line more precisely.

**Figure 12.5.1** Two data sets with the same value of  $n$  and of  $s_e$  but different  $s_x$ : (a) smaller  $s_x$  and (b) larger  $s_x$



As another way of thinking about this, imagine holding your arms out in front of you, extending the index finger on each hand, and balancing a meter stick on your two fingers. If you move your hands far apart from each other, balancing the meter stick is easy—this is like case (b). However, if you move your hands close together, balancing the meter stick becomes more difficult—this is like case (a). Having the base of support spread out increases stability. Likewise, having the  $x$  values spread out decreases the standard error of the slope.

**Implications for Design.** The previous discussion implies that, for the purpose of gaining precise information about  $\beta_1$ , it is best to have the values of  $X$  as widely dispersed as possible. This fact can guide the experimenter when the design of the experiment includes choosing values of  $X$ . Other factors also play a role, however. For instance, if  $X$  is the dose of a drug, the criterion of widely dispersed  $X$ 's would lead to using only two dosages, one very low and one very high. But in practice an experimenter would want to have at least a few observations at intermediate doses, to verify that the relation is actually linear within the range of the data.

### Confidence Interval for $\beta_1$

In many studies the quantity  $\beta_1$  is a biologically meaningful parameter and a primary aim of the data analysis is to estimate  $\beta_1$ . A confidence interval for  $\beta_1$  can be constructed by the familiar method based on the SE and Student's  $t$  distribution. For instance, a 95% confidence interval is constructed as

$$b_1 \pm t_{0.025} \text{SE}_{b_1}$$

where the critical value  $t_{0.025}$  is determined from Student's  $t$  distribution with

$$\text{df} = n - 2$$

Intervals with other confidence coefficients are constructed analogously; for instance, for a 90% confidence interval one would use  $t_{0.05}$ .

---

**Example  
12.5.2**

**Length and Weight of Snakes** Let us use the snake data to construct a 95% confidence interval for  $\beta_1$ . We found that  $b_1 = 7.19186$  and  $\text{SE}_{b_1} = 0.9531$ . There are  $n = 9$  observations; we refer to Table 4 with  $\text{df} = 9 - 2 = 7$ , and obtain

$$t_{7,0.025} = 2.365$$

The confidence interval is

$$7.19186 \pm 2.365 \times 0.9531$$

or

$$4.94 \text{ gm/cm} < \beta_1 < 9.45 \text{ gm/cm}$$

We are 95% confident that the true slope of the regression of weight on length for this snake population is between 4.94 gm/cm and 9.45 gm/cm; this is a rather wide interval because the sample size is not very large. ■

### Testing the Hypothesis $H_0: \beta_1 = 0$

In some investigations it is not a foregone conclusion that there is any linear relationship between  $X$  and  $Y$ . It then may be relevant to consider the possibility that any apparent trend in the data is illusory and reflects only sampling variability. In this situation it is natural to formulate the null hypothesis

$$H_0: \mu_{Y|X} \text{ does not depend on } X$$

Within the linear model, this hypothesis can be translated as

$$H_0: \beta_1 = 0$$

A  $t$  test of  $H_0$  is based on the test statistic\*

$$t_s = \frac{b_1 - 0}{\text{SE}_{b_1}}$$

Critical values are obtained from Student's  $t$  distribution with

$$\text{df} = n - 2$$

The following example illustrates the application of this  $t$  test.

**Example  
12.5.3**

**Blood Pressure and Platelet Calcium** The blood pressure and platelet calcium data from Example 12.2.3 are shown in Figure 12.5.2. Calculations from the data yield  $\bar{x} = 94.50000$ ,  $\bar{y} = 107.86840$ ,  $s_x = 8.04968$ ,  $s_y = 16.07780$ , from which we can calculate†

$$b_0 = -2.2009 \text{ and } b_1 = 1.16475$$

The residual sum of squares is 6311.7618.

Thus,

$$s_e = \sqrt{\frac{6311.76}{38 - 2}} = 13.24 \text{ and } \text{SE}_{b_1} = \frac{13.24}{8.04968\sqrt{38 - 1}} = 0.2704$$

The values of  $b_0$ ,  $b_1$ , SS(resid), and  $\text{SE}_{b_1}$  are generally found using computer software. The following computer output is typical:

The regression equation is

Platelet Calcium =  $-2.2 + 1.16$  Blood Pressure

Predictor	Coef	SE Coef	T	P
Constant	-2.20	25.65	-0.09	0.932
Blood Pressure	1.1648	0.2704	4.31	0.000

S = 13.2411 R - Sq = 34.0% R - Sq(adj) = 32.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3252.6	3252.6	18.55	0.000
Residual Error	36	6311.8	175.3		
Total	37	9564.3			

We will test the null hypothesis

$$H_0: \beta_1 = 0$$

against the nondirectional alternative

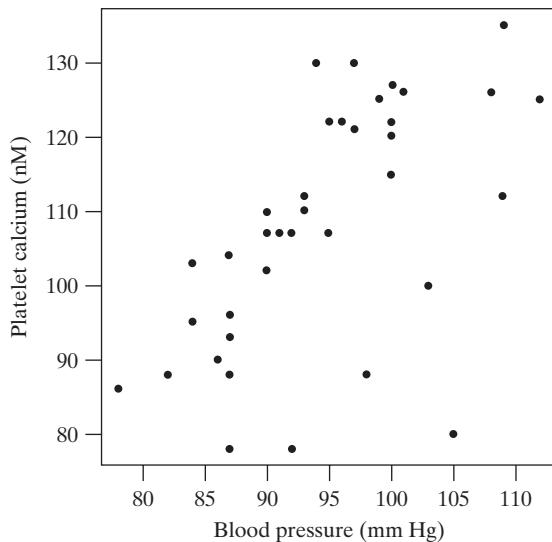
$$H_A: \beta_1 \neq 0$$

---

\*We include the “-0” in the numerator of the test statistic to remind us that we are comparing our estimated (observed) slope,  $b_1$ , to the slope we'd expect to observe if the null hypothesis were true. In the exercises we will consider a situation for which the hypothesized slope may be a value other than zero.

†As the following values are intermediate calculations used in the regression, we include more digits than one would typically display in a summary.

**Figure 12.5.2** Blood pressure and platelet calcium for 38 persons with normal blood pressure



These hypotheses are translations, within the linear model, of the verbal hypotheses

$H_0$ : Mean platelet calcium is not linearly related to blood pressure

$H_A$ : Mean platelet calcium is linearly related to blood pressure

(Note: “Linearly related” does not necessarily refer to causal dependence as we have discussed in Section 12.2.)

Let us choose  $\alpha = 0.05$ . The test statistic is

$$t_s = \frac{1.16475}{0.2704} = 4.308$$

From Table 4 with  $df = n - 2 = 36 \approx 40$ , we find  $t_{40,0.0005} = 3.551$ . Thus, we find  $P\text{-value} < 0.001$  and we reject  $H_0$ . The data provide sufficient (and very strong) evidence to conclude that the true slope of the regression of platelet calcium on blood pressure in this population is positive (that is,  $\beta_1 > 0$ ). ■

Note that the test on  $\beta_1$  does not ask *whether* the relationship between  $\mu_{Y|X}$  and  $X$  is linear. Rather, the test asks whether, *assuming* that the linear model holds, we can conclude that the slope is nonzero. It is therefore necessary to be careful in phrasing the conclusion from this test. For instance, the statement “There is a significant linear trend” could easily be misunderstood.\*

As is the case with other hypothesis tests, if we wish to use a directional alternative hypothesis we follow the two-step procedure of (1) checking that the specified direction is correct (which in a regression setting means checking that the slope of the regression line has the correct + or - sign) and (2) cutting the nondirectional  $P$ -value in half if this condition is met.

---

\*There are tests that can (in some circumstances) test whether the true relationship is linear. Furthermore, there are tests that can test for a linear component of trend without assuming that the relationship is linear. These tests are beyond the scope of this book.

## Exercises 12.5.1–12.5.9

**12.5.1** Refer to the leucine data given in Exercise 12.3.1.

- Construct a 95% confidence interval for  $\beta_1$ .
- Interpret the confidence interval from part (a) in the context of this setting.

**12.5.2** Refer to the body temperature data of Exercise 12.3.2. For these data,  $s_e = 0.91472$ . Construct a 95% confidence interval for  $\beta_1$ .

**12.5.3** Refer to the cob weight data of Exercise 12.2.5. For these data,  $SS(\text{resid}) = 1,337.3$ .

- Construct a 95% confidence interval for  $\beta_1$ .
- Interpret the confidence interval from part (a) in the context of this setting.

**12.5.4** Refer to the fungus growth data of Exercise 12.2.6. For these data,  $SS(\text{resid}) = 16.7812$ .

- Calculate the standard error of the slope,  $SE_{\beta_1}$ .
- Consider the null hypothesis that laetisaric acid has no effect on growth of the fungus. Assuming that the linear model is applicable, formulate this as a hypothesis about the true regression line, and test the hypothesis against the alternative that laetisaric acid inhibits growth of the fungus. Let  $\alpha = 0.05$ .

**12.5.5** Refer to the energy expenditure data of Exercise 12.2.7. For these data,  $SS(\text{resid}) = 21,026.1$ .

- Construct a 95% confidence interval for  $\beta_1$ .
- Construct a 90% confidence interval for  $\beta_1$ .

**12.5.6** Refer to the basal Ca pump data from Exercise 12.2.8. For these data,  $s_e = 548.78$ .

- Construct a 95% confidence interval for  $\beta_1$ .
- What do you think about a claim that that  $\beta_1$  is less than  $-800$  (nmol/mg/hr)/( $\mu$ g/g)? Use your interval from part (a) to support your answer.

- What do you think about a claim that  $\beta_1$  is less than  $800$  (nmol/mg/hr)/( $\mu$ g/g) in magnitude? Use your interval from part (a) to support your answer.

**12.5.7** Refer to the respiration data of Exercise 12.3.6. Assuming that the linear model is applicable, test the null hypothesis of no relationship against the alternative that trees from higher altitudes tend to have higher respiration rates. Let  $\alpha = 0.05$ .

**12.5.8** The following computer output is from fitting a regression model to the snake length data of Example 12.2.2. Use this output to construct a 95% confidence interval for  $\beta_1$ .

The regression equation is

$$\text{Weight} = -301 + 7.19 \text{Length}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	-301.09	60.19	-5.00	0.000
Length	7.1919	0.9531	7.55	0.000

$s = 12.50$  R-sq = 89.1% R-sq(adj) = 87.5%

### Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	8896.3	8896.3	56.94	0.000
Error	7	1093.7	156.2		
Total	8	9990.0			

**12.5.9** Refer to the peak flow data of Exercise 12.3.8. Assume that the linear model is applicable.

- Test the null hypothesis of no relationship against the alternative that peak flow is related to height. Use a nondirectional alternative with  $\alpha = 0.10$ .
- Repeat the test from part (a), but this time use the directional alternative that peak flow tends to increase with height. Again let  $\alpha = 0.10$ .

## 12.6 Guidelines for Interpreting Regression and Correlation

Any set of  $(X, Y)$  data can be submitted to a regression analysis and values of  $b_0, b_1, s_e$ , and  $r$  can be calculated. But these quantities require care in interpretation. In this section we discuss guidelines and cautions for interpretation of linear regression and correlation. We first consider the use of regression and correlation for purely descriptive purposes and then turn to inferential uses.

### When Is Linear Regression Descriptively Inadequate?

Linear regression and correlation may provide inadequate description of a data set if any of the following features is present:

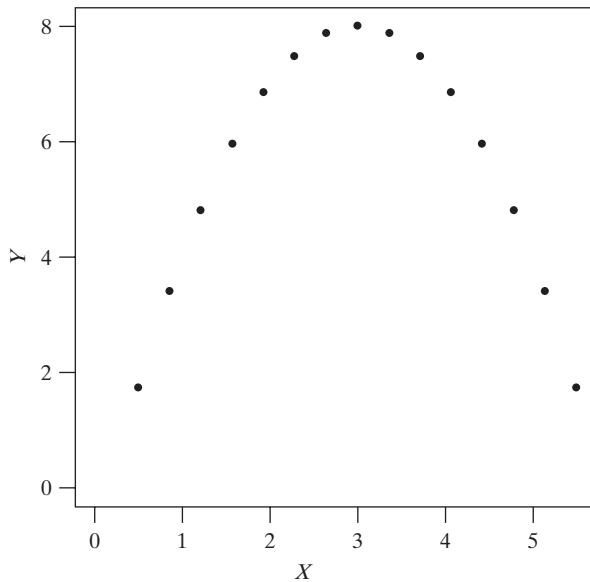
- curvilinearity
- outliers
- influential points

We briefly discuss each of these.

If the dependence of  $Y$  on  $X$  is actually curvilinear rather than linear, the application of linear regression and correlation can be very misleading. The following example shows how this can happen.

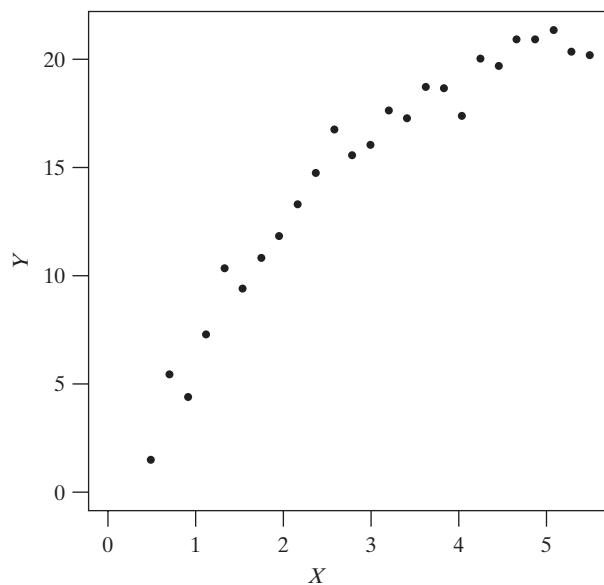
**Example  
12.6.1**

**Figure 12.6.1** Data for which  $X$  and  $Y$  are uncorrelated but have a strong curvilinear relationship



Generally, the consequences of curvilinearity are that (1) the fitted line does not adequately represent the data; (2) the correlation is misleadingly small; (3)  $s_e$  is inflated. Of course, Example 12.6.1 is an extreme case of this distortion. A data set with mild, but still noticeable, curvilinearity is shown in Figure 12.6.2.

**Figure 12.6.2** Data displaying mild curvilinearity

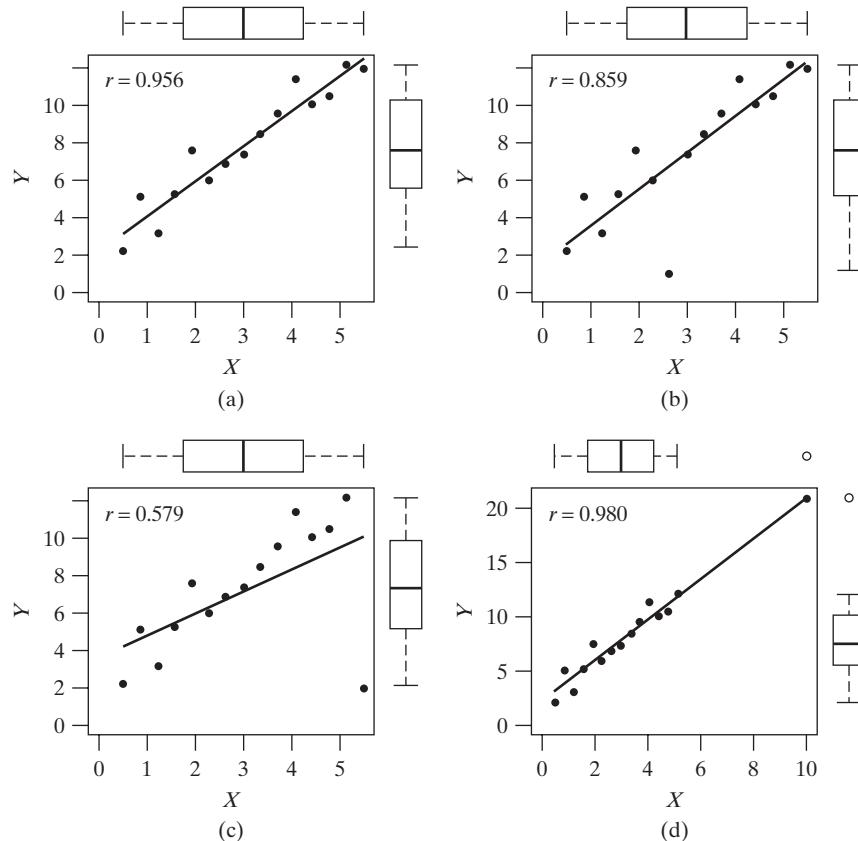


**Outliers** in a regression setting are data points that are unusually far from the linear trend formed by the data. Outliers can distort regression analysis in two ways: (1) by inflating  $s_e$  and reducing correlation; and (2) by unduly influencing the regression line. Note that a point can be an outlier in a scatterplot without being an outlier in either the distribution of  $X$  values or the distribution of  $Y$  values as we shall see in the following example.

Figure 12.6.3 displays a data set with a variety of outliers. Figure 12.6.3(a) displays a data set with no outliers, while (b) and (c) show data with regression outliers—they have points that fall far from the regression line. In plot (b) the outlying point does not appear to affect the slope of the regression line very much, but it does increase the residual standard deviation,  $s_e$ , and reduce correlation. The outlying point in plot (c) appears to greatly affect the slope of the estimated regression line; it also increases  $s_e$  and reduces the correlation. While the unusual point in plot (d) is an outlier with respect to the  $X$  (and  $Y$ ) distribution, it is not an outlier in the regression context as it does not fall far from the regression line.

**Leverage points** are points that have the potential to greatly influence the slope of the fitted regression model. The further a point's  $X$  value is from the center of the  $X$  distribution, the more leverage that point has on the overall regression model. *Having* and actually *exerting* leverage are two different things, however. Figure 12.6.3 plots (c) and (d) display examples of leverage points. In plot (c) the leverage point is shown to actually exert its leverage on the line, tipping the regression from the bulk of the data. A point that has a large effect on the regression model is called an **influential point**. Plot (d) shows a leverage point (because of the extreme  $X$  value) that is not influential because the regression line does not get pulled away from the trend in the bulk of the data. Note that the outlier in plot (b) is not consid-

**Figure 12.6.3** Different effects of outliers on the regression line. Boxplots of the  $X$  and  $Y$  data appear in the margins of each scatterplot. (a) A data set with no outliers; (b) the same data except for one outlier in the middle of the  $X$  values; (c) the same data except for one outlier at the high end of the  $X$  values (a point with leverage and influence); and (d) the same data except for one point that is an outlier with respect to the  $X$  (and  $Y$ ) distribution, but not with respect to the regression line (a point with leverage, but little influence)



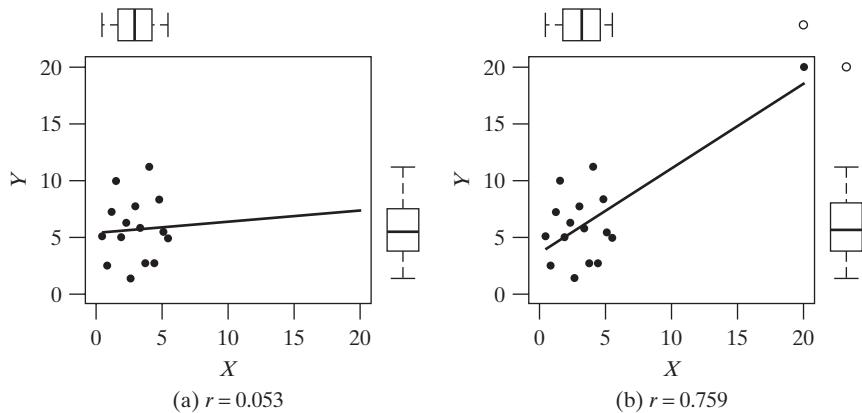
ered a leverage point—its ability to affect the slope of the line is weak as its  $X$  value is near the center of the  $X$  distribution.

Influential points can also greatly affect (increase or decrease) the size of the correlation coefficient. In Figure 12.6.3, the influential point in (c) lowered the correlation from  $r = 0.956$  in (a) to  $r = 0.579$ . Example 12.6.3 shows a situation for which the correlation is increased by the presence of an influential point.

Figure 12.6.4 (a) shows a data set and a regression line. Figure 12.6.4 (b) shows the same data set, but with an influential point added. Including the influential point in the data set changes the regression line noticeably. Although the influential point is an outlier in the  $X$  and  $Y$  distributions, it is not a regression outlier since the residual for this point is not very large.

The correlation coefficient for the data in Figure 12.6.4(a) is  $r = 0.053$ . Adding the influential point to the data set changes the correlation to  $r = 0.759$  for the data in Figure 12.6.4(b).

**Figure 12.6.4** The effect of an influential point on the regression line. (a) A data set; (b) the same data with an influential point added



## Conditions for Inference

The quantities  $b_0$ ,  $b_1$ ,  $s_e$ , and  $r$  can be used to describe a scatterplot that shows a linear trend. However, statistical inference based on these quantities depends on certain conditions concerning the design of the study, the parameters, and the conditional population distributions. We summarize these conditions and then discuss guidelines and cautions concerning them.

1. *Design conditions.* We have discussed two sampling models for regression and correlation:
  - (a) Random subsampling model: For each observed  $X$ , the corresponding observed  $Y$  is viewed as randomly chosen from the conditional population distribution of  $Y$  values for that  $X$ .\*
  - (b) Bivariate random sampling model: Each observed pair  $(X, Y)$  is viewed as randomly chosen from the joint population distribution of bivariate pairs  $(X, Y)$ .

In either sampling model, each observed pair  $(X, Y)$  must be independent of the others. This means that the experimental design must not include any pairing, blocking, or hierarchical structure.

---

\*If the  $X$  variable includes measurement error, then  $X$  in the linear model must be interpreted as the measured value of  $X$  rather than some underlying “true” value of  $X$ . A linear model involving the “true” value of  $X$  leads to a different kind of regression analysis.

2. *Conditions concerning parameters.* The linear model states that
  - (a)  $\mu_{Y|X} = \beta_0 + \beta_1 X$ .
  - (b)  $\sigma_e$  does not depend on  $X$ .
3. *Condition concerning population distributions.* The confidence interval and  $t$  test are based on the conditional population distribution of  $Y$  for each fixed  $X$  having a normal distribution.

The random subsampling model is required if  $b_0$ ,  $b_1$ , and  $s_e$  are to be viewed as estimates of the parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma_e$  mentioned in the linear model. The bivariate random sampling model is required if  $r$  is to be viewed as an estimate of a population parameter  $\rho$ . It can be shown that if the bivariate random sampling model is applicable, then the random subsampling model is also applicable. Thus, regression parameters can always be estimated if correlation can be estimated, but not vice versa.

## Guidelines Concerning the Sampling Conditions

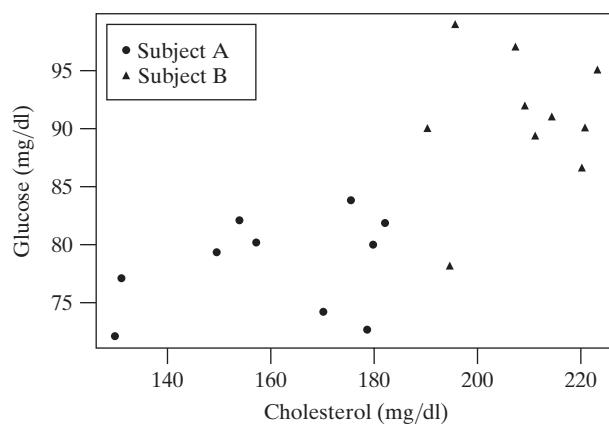
Departures from the sampling conditions not only affect the validity of formal techniques such as the confidence interval for  $\beta_1$ , but can also lead to faulty interpretation of the data even if no formal statistical analysis is performed. Two errors of interpretation that sometimes occur in practice are (1) failure to take into account dependency in the observations, and (2) insufficient caution in interpreting  $r$  when the  $X$ 's do not represent a random sample.

The following two examples illustrate studies with dependent observations.

### Example 12.6.2

**Serum Cholesterol and Serum Glucose** A data set consists of 20 pairs of measurements on serum cholesterol ( $X$ ) and serum glucose ( $Y$ ) in humans. However, the experiment included only two subjects; each subject was measured on 10 different occasions. Because of the dependency in the data, it is not correct to naively treat all 20 data points alike. Figure 12.6.5 illustrates the difficulty; the figure shows that there is no evidence of any correlation between  $X$  and  $Y$ , except for the modest fact that the subject who has larger  $X$  values happens also to have larger  $Y$  values. Clearly it would be impossible to properly interpret the scatterplot if all 20 points were plotted with the same symbol. By the same token, application of regression or correlation formulas to the 20 observations would be seriously misleading.<sup>20</sup>

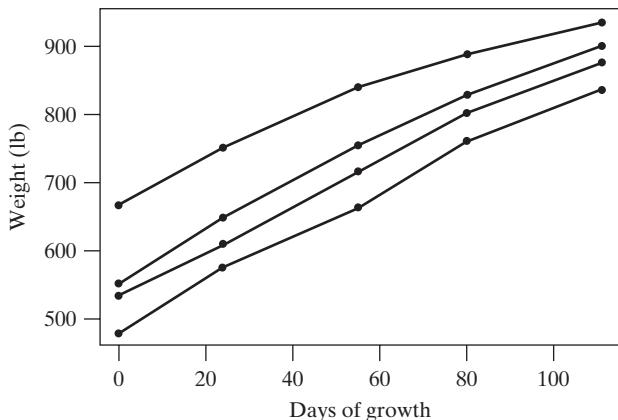
**Figure 12.6.5** Twenty observations of  $X$  = serum cholesterol and  $Y$  = serum glucose in humans



### Example 12.6.3

**Growth of Beef Steers** Figure 12.6.6 shows 20 pairs of measurements on the weight ( $Y$ ) of beef steers at various times ( $X$ ) during a feeding trial. The data represent four animals, each weighed at five different times; observations on the same animal are joined by lines in the figure. An ordinary regression analysis on the 20 data points

**Figure 12.6.6** Twenty observations of  $X = \text{days}$  and  $Y = \text{weight in steers}$ . Data for individual animals are joined by lines



would ignore the information carried in the lines and would yield inflated SEs and weak tests. Similarly, an ordinary scatterplot (without the lines) would be an inadequate representation of the data.<sup>21</sup>

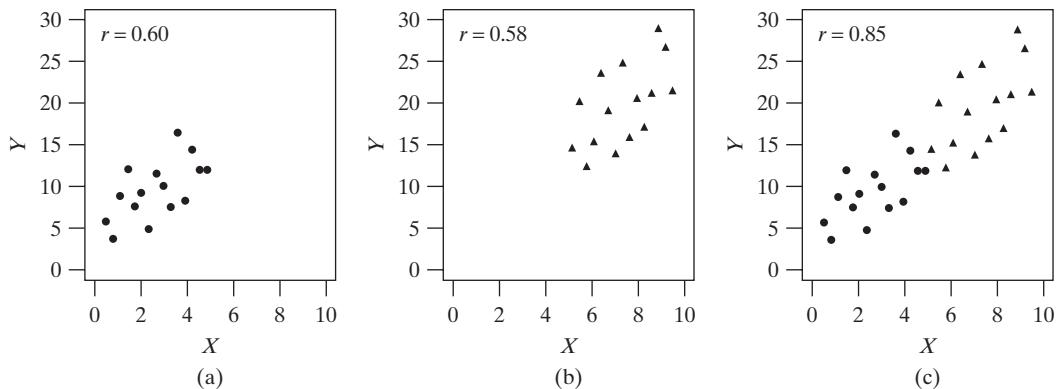
In Example 12.6.2, ignoring the dependency in the observations would lead to *overinterpretation* of the data—that is, concluding that a relationship exists when there is actually very little evidence for it. By contrast, ignoring the dependency in Example 12.6.3 would lead to *underinterpretation* of the data—that is, insufficiently extracting the “signal” from the “noise.”

In interpreting the correlation coefficient  $r$ , one should recognize that  $r$  is influenced by the degree of spread in the values of  $X$ . If the regression quantities  $b_0$ ,  $b_1$ , and  $s_e$  are unchanged, *more spread in the  $X$  values leads to a stronger correlation (larger magnitude of  $r$ )*. The following example shows how this happens.

**Example  
12.6.4**

Figure 12.6.7 shows fictitious data that illustrate how  $r$  can be affected by the distribution of  $X$ . The data points in parts (a) and (b) have been plotted together in part (c). The regression line is nearly the same in all three scatterplots, but notice that  $X$  and  $Y$  appear more highly correlated in (c) than in either (a) or (b). The contrasting appearance of the scatterplots is reflected in the correlation coefficients; in fact,  $r = 0.60$  for (a),  $r = 0.58$  for (b), but  $r = 0.85$  for (c).

The fact that  $r$  depends on the distribution of  $X$  does not mean that  $r$  is invalid as a descriptive statistic. But it does mean that, when the values of  $X$  cannot be viewed as a random sample,  $r$  must be interpreted cautiously. For instance, suppose two



**Figure 12.6.7** Dependence of  $r$  on the distribution of  $X$ . The data of (a) and (b) are plotted together in (c)

experimenters conduct separate studies of response ( $Y$ ) to various doses ( $X$ ) of a drug. Each of them could calculate  $r$  as a description of her or his own data, but they should *not* expect to obtain *similar* values of  $r$  unless they both use the same choice of doses ( $X$  values). By contrast, they might reasonably expect to obtain similar regression lines and similar residual standard deviations, regardless of their choice of  $X$  values, as long as the dose-response relationship remains the same throughout the range of doses used.

**Labeling  $X$  and  $Y$ .** If the bivariate random sampling model is applicable, then the investigator is free to decide which variable to label  $X$  and which to label  $Y$ . Of course, for calculation of  $r$  the labeling does not matter. For regression calculations, the decision depends on the purpose of the analysis. The regression of  $Y$  on  $X$  yields (within the linear model) estimates of  $\mu_{Y|X}$ —that is, the population mean  $Y$  value for fixed  $X$ . Similarly, the regression of  $X$  on  $Y$  is aimed at estimating  $\mu_{X|Y}$ —that is, the mean  $X$  value for fixed  $Y$ . These approaches do not lead to the same regression line because they are directed at answering different questions. An intuitive example follows.

---

**Example  
12.6.5**

**Height and Weight of Young Men** For the population of young men described in Example 12.4.4, the mean weight of young men 76" (6'4") tall is 178 lb. Now consider this question: What would be the mean height of young men who weigh 178 lb? There is no reason that the answer should be 76". Intuition suggests that the answer should be less than 76"—and in fact it is about 71". ■

## Guidelines Concerning the Linear Model and Normality Condition

The test and confidence interval for  $\beta_1$  are based on the linear model and the condition of normality. The interpretation of these inferences can be seriously degraded if the linearity condition is not met; after all, we have seen earlier in this section that even the descriptive usefulness of regression is reduced if curvilinearity or outliers are present.

In addition to linearity, the linear model specifies that  $\sigma_e$  is the same for all the observations. A common pattern of departure from this condition is a trend for larger means to be associated with larger SDs. Mild nonconstancy of the SDs does not seriously affect the interpretation of  $b_0$ ,  $b_1$ ,  $SE_{b_1}$ , and  $r$  (although it does invalidate the interpretation of  $s_e$  as a pooled estimate of a common SD).

## Residual Plots

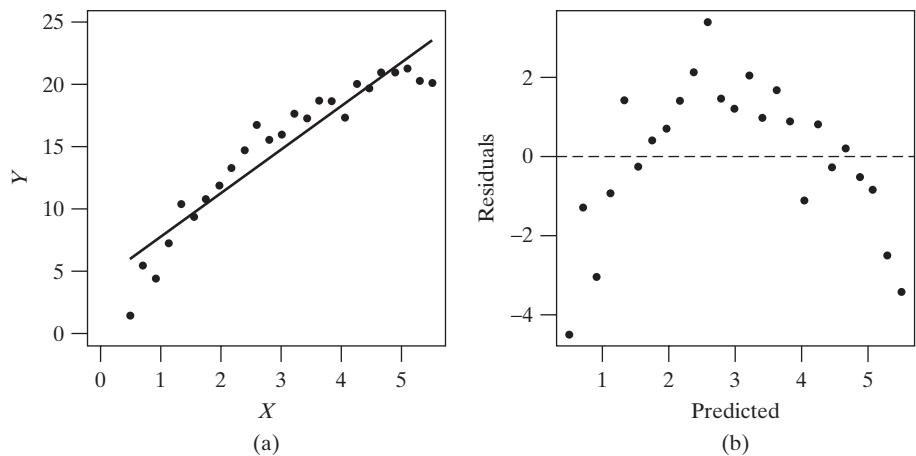
Formal statistical tests for curvilinearity, unequal standard deviations, nonnormality, and outliers are beyond the scope of this book. However, the single most useful instrument for detecting these features is the human eye, aided by scatterplots. For instance, notice how easily the eye detects the mild curvilinearity in Figure 12.6.2 and the outlier in Figure 12.6.3(b). Notice also in Figure 12.6.3(b) that examination of the marginal distributions of  $X$  and  $Y$  separately would not have revealed the outlier.

In addition to scatterplots of  $Y$  versus  $X$ , it is often useful to look at various displays of the residuals. A scatterplot of each residual ( $y_i - \hat{y}_i$ ) against  $\hat{y}_i$  is called a **residual plot**. Residual plots are very useful for detecting curvature; they can also reveal trends in the conditional standard deviation. Figure 12.6.8 shows the data from Figure 12.6.2 together with a residual plot of those data.

A residual plot shows the data after the linear trend has been removed, which makes it easier to see nonlinear patterns in the data. The curvature in Figure 12.6.8(a) is apparent, but it is much more visible in the residual plot of Figure 12.6.8(b).

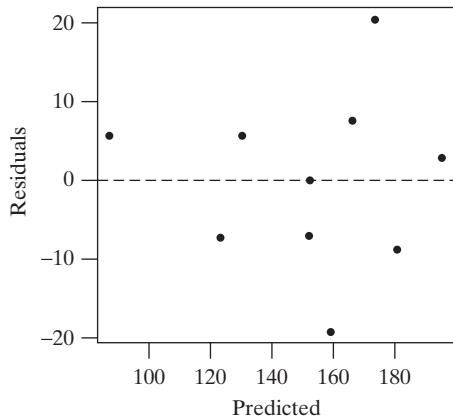
If the linear model holds, with no outliers, then the fitted regression line captures the trend in the data, leaving a random pattern in the residual plot. Thus, we

**Figure 12.6.8** (a) Data displaying mild curvilinearity with linear regression line; (b) a residual plot of the data



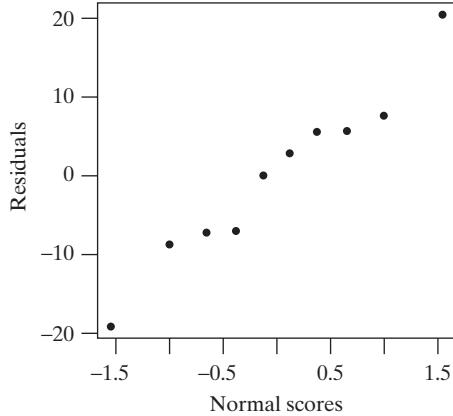
hope to see no striking pattern in a residual plot. For example, Figure 12.6.9 shows a residual plot of the snake data of Example 12.2.1. The lack of unusual features in this plot supports the use of a regression model for these data.

**Figure 12.6.9** Residual plot of the snake data



If the condition of normality is met, then the distribution of the residuals should look roughly like a normal distribution.\* A normal probability plot of the residuals provides a useful check of the normality condition. The normal probability plot of the snake data in Figure 12.6.10 is fairly linear, which supports the use of the  $t$  test and the confidence interval presented in Section 12.5.

**Figure 12.6.10** Normal probability plot of the snake data



\*This is the basis for the 68% and 95% interpretations of  $s_e$  given in Section 12.3.

## The Use of Transformations

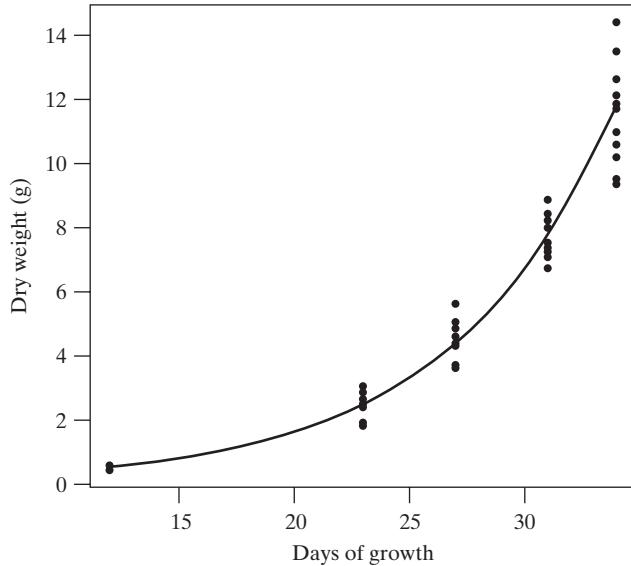
If the conditions of linearity, constancy of standard deviation, and normality are not met, a remedy that is sometimes useful is to transform the scale of measurement of either  $Y$ , or  $X$ , or both. The following example illustrates the use of a logarithmic transformation.

### Example 12.6.6

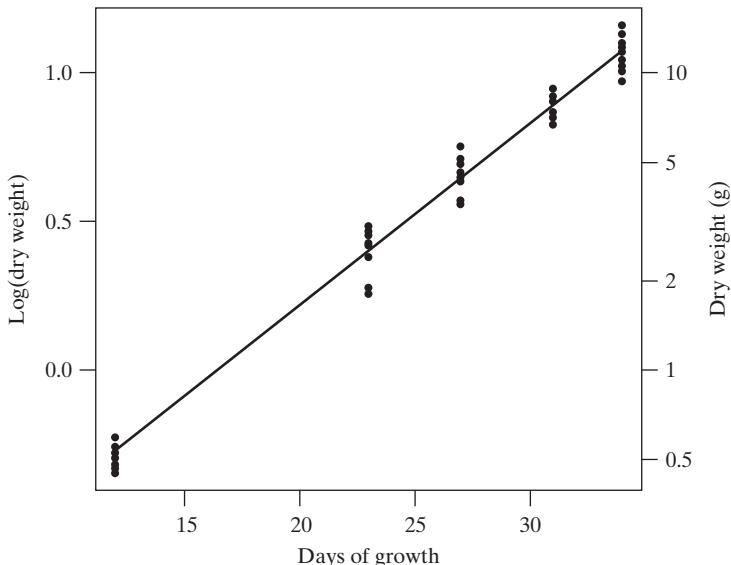
**Growth of Soybeans** A botanist placed 60 one-week-old soybean seedlings in individual pots. After 12 days of growth, she harvested, dried, and weighed 12 of the young soybean plants. She weighed another 12 plants after 23 days of growth, and groups of 12 plants each after 27 days, 31 days, and 34 days. Figure 12.6.11 shows the 60 plant weights plotted against days of growth; a smooth curve connects the group means. It is easy to see from Figure 12.6.11 that the relationship between mean plant weight and time is curvilinear rather than linear and that the conditional standard deviation is not constant but is strongly increasing.<sup>22</sup>

Figure 12.6.12 shows the logarithms (base 10) of the plant weights, plotted against days of growth together with the regression line. Notice that the logarithmic

**Figure 12.6.11** Weight of soybean plants plotted against days of growth



**Figure 12.6.12**  
Log(weight) of soybean plants plotted against days of growth



transformation has simultaneously straightened the curve and more nearly equalized the standard deviations. It would not be unreasonable to assume that the linear model is valid for the variables  $Y = \log(\text{dry weight})$  and  $X = \text{days of growth}$ . Table 12.6.1 shows the means and standard deviations before and after the logarithmic transformation. Note especially the effect of the transformation on the equality of the SDs.

**Table 12.6.1** Summary of soybean growth data in original scale and after log transformation

Days of growth	Number of plants	Dry weight (gm)		Log(dry weight)	
		Mean	SD	Mean	SD
12	12	0.50	0.06	-0.31	0.055
23	12	2.63	0.37	0.42	0.062
27	12	4.67	0.70	0.67	0.066
31	12	7.57	1.19	0.87	0.069
34	12	11.20	1.62	1.04	0.064

## Exercises 12.6.1–12.6.9

**12.6.1** In a metabolic study, four male swine were tested three times: when they weighed 30 kg, again when they weighed 60 kg, and again when they weighed 90 kg. During each test, the experimenter analyzed feed intake and fecal and urinary output for 15 days, and from these data calculated the nitrogen balance, which is defined as the amount of nitrogen incorporated into body tissue per day. The results are shown in the accompanying table.<sup>23</sup>

ANIMAL NUMBER	BODY WEIGHT	NITROGEN BALANCE (gm/day)		
		30 kg	60 kg	90 kg
1		15.8	21.3	16.5
2		16.4	20.8	18.2
3		17.3	23.8	17.8
4		16.4	22.1	17.5
Mean		16.48	22.00	17.50

Suppose these data are analyzed by linear regression. With  $X$  = body weight and  $Y$  = nitrogen balance, preliminary calculations yield  $\bar{x} = 60$  and  $\bar{y} = 18.7$ . The slope is  $b_1 = 0.017$ , with standard error  $SE_{b_1} = 0.032$ . The  $t$  statistic is  $t_s = 0.53$ , which is not significant at any reasonable significance level. According to this analysis, there is insufficient evidence to conclude that nitrogen balance depends on body weight under the conditions of this study.

The above analysis is flawed in two ways. What are they? (Hint: Look for ways in which the conditions for inference are not met. There may be several minor departures

from the conditions, but you are asked to find two major ones. No calculation is required.)

**12.6.2** For measuring the digestibility of forage plants, two methods can be used: The plant material can be fermented with digestive fluids in a glass container, or it can be fed to an animal. In either case, digestibility is expressed as the percentage of total dry matter that is digested. Two investigators conducted separate studies to compare the methods by submitting various types of forage to both methods and comparing the results. Investigator A reported a correlation of  $r = 0.8$  between the digestibility values obtained by the two methods, and investigator B reported  $r = 0.3$ . The apparent discrepancy between these results was resolved when it was noted that one of the investigators had tested only varieties of canary grass (whose digestibilities ranged from 56% to 65%), whereas the other investigator had used a much wider spectrum of plants, with digestibilities ranging from 35% for corn stalks to 72% for timothy hay.<sup>24</sup>

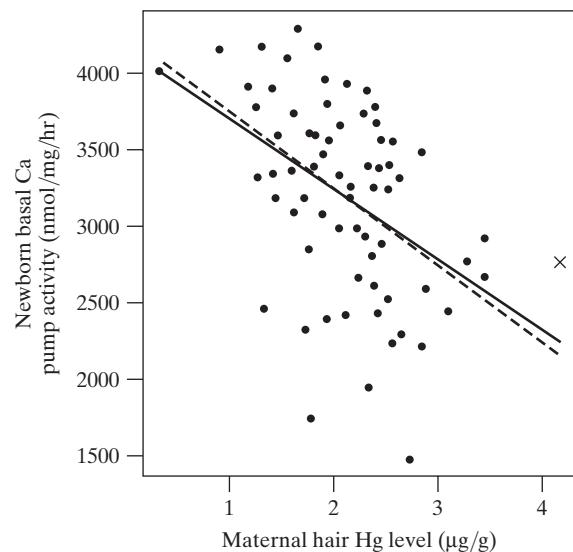
Which investigator (A or B) used only canary grass? How does the different choice of test material explain the discrepancy between the correlation coefficients?

**12.6.3** Refer to the energy expenditure data of Exercise 12.2.7. Each subject's expenditure value ( $Y$ ) is the average of two measurements made on different occasions. It might be proposed that it would be better to use the two measurements as separate data points, thus yielding 14 observations rather than 7. If this proposed approach were used, one of the conditions for inference would be highly doubtful. Which one, and why?

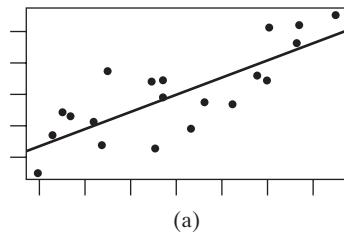
**12.6.4** Refer to the fungus growth data of Exercise 12.2.6. In that exercise the investigator found  $r = -0.98754$ . Suppose a second investigator were to replicate the experiment, using concentrations of 0, 2, 4, 6, 8, and 10 mg, with two petri dishes at each concentration. Would you predict that the value of  $r$  calculated by this second investigator would be about the same as that found in Exercise 12.2.6, smaller in magnitude, or larger in magnitude? Explain.

**12.6.5** In the following scatterplot of the Ca pump data of Exercise 12.2.8, one of the points is marked with an “ $\times$ .” In addition, there are two regression lines on the plot: The solid line includes all of the data and the dashed line omits the point marked “ $\times$ .”

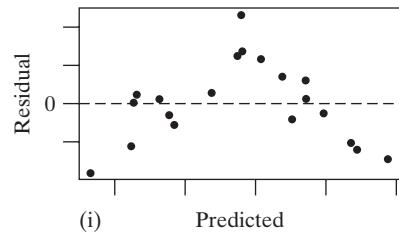
- (a) Would we consider the point marked “ $\times$ ” an outlier? Explain.
- (b) Would we consider the point marked “ $\times$ ” a leverage point? Explain.
- (c) Noting the very small change in the slopes of the dashed and solid lines, would we consider the point marked “ $\times$ ” an influential observation? Explain.



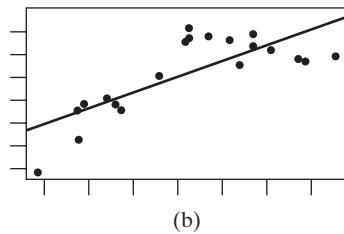
**12.6.6** The following three residual plots, (i), (ii), and (iii), were generated after fitting regression lines to the following three scatterplots, (a), (b), and (c). Which residual plot goes with which scatterplot? How do you know?



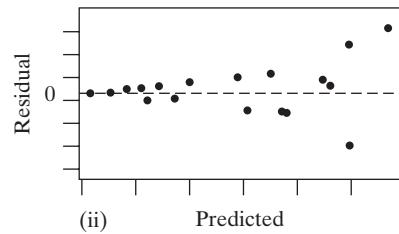
(a)



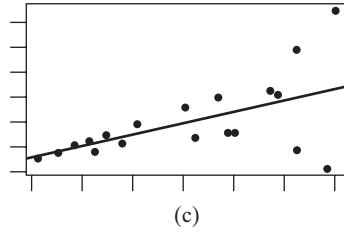
(i) Predicted



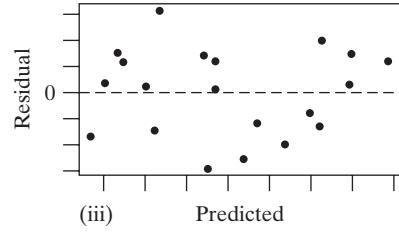
(b)



(ii) Predicted

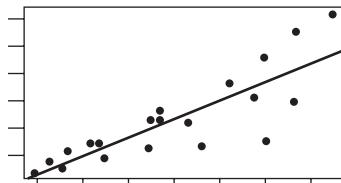


(c)

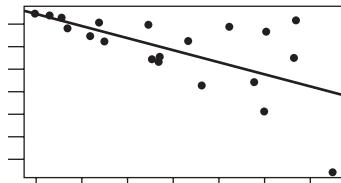


(iii) Predicted

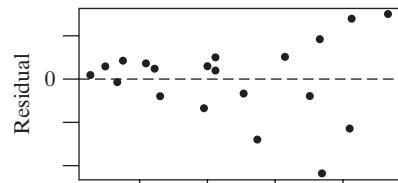
**12.6.7** The following two residual plots, (i), and (ii), were generated after fitting regression lines to the two scatterplots (a) and (b). Which residual plot goes with which scatterplot? How do you know?



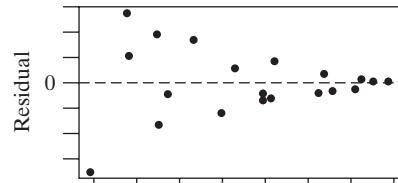
(a)



(b)

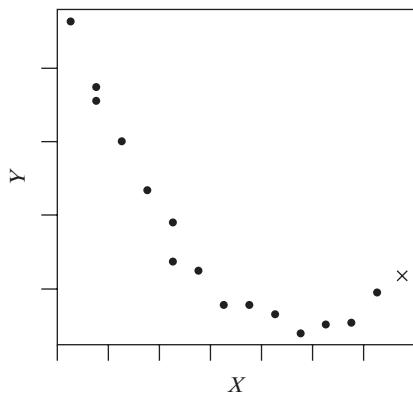


(i) Predicted



(ii) Predicted

**12.6.8** Sketch the residual plot that would be produced by fitting a regression line to the following scatterplot. One of the points is plotted with an “ $\times$ .” Indicate this point on the residual plot.



**12.6.9** (Computer exercise) Researchers measured the diameters of 20 trees in a central Amazon rain forest and used  $^{14}\text{C}$ -dating to determine the ages of these trees. The data are given in the following table.<sup>25</sup> Consider the use of diameter,  $X$ , as a predictor of age,  $Y$ .

DIAMETER (cm)	AGE (yr)	DIAMETER (cm)	AGE (yr)
180	1372	115	512
120	1167	140	512
100	895	180	455
225	842	112	352
140	722	100	352
142	657	118	249
139	582	82	249
150	562	130	227
110	562	97	227
150	552	110	172

- Make a scatterplot of  $Y = \text{age}$  versus  $X = \text{diameter}$  and fit a regression line to the data.
- Make a residual plot from the regression in part (a). Then make a normal probability plot of the residuals. How do these plots call into question the use of a linear model and regression inference procedures?
- Take the logarithm of each value of age. Make a scatterplot of  $Y = \log(\text{age})$  versus  $X = \text{diameter}$  and fit a regression line to the data.
- Make a residual plot from the regression in part (c). Next, make a normal probability plot of the residuals. Based on these plots, does a regression model using a log scale, from part (c), seem appropriate?

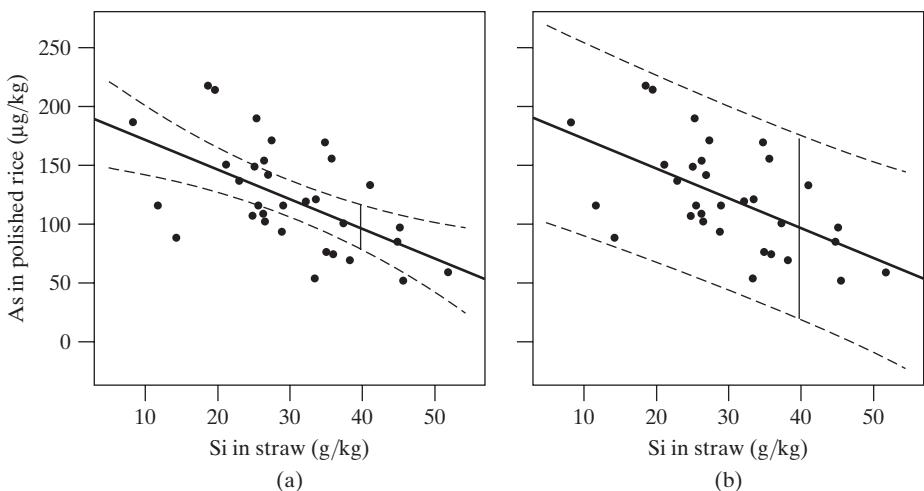
## 12.7 Precision in Prediction (Optional)

In Section 12.4 we learned that one very practical use of regression is prediction. In this section we shall distinguish between the prediction of the *mean*  $Y$  value for a particular  $X$  value and the prediction of a *single*  $Y$  value for a particular  $X$  value. In particular, we will compare the precisions of these two very different types of predictions.

## Confidence and Prediction Intervals

In Example 12.4.6 we used a regression line to make a prediction:  $\hat{y} = 197.17 - 2.51x$ . Using this line again we could predict the *mean* arsenic concentration in rice from plants with straw silicon concentrations of 40 g/kg to be  $\hat{y} = 197.17 - 2.51(40) = 96.77 \mu\text{g}/\text{kg}$ . What if instead of estimating the mean arsenic concentration of all plants with this silicon concentration, we wanted to predict *the* arsenic concentration of *a* particular plant whose straw silicon concentration was 40 g/kg? Our estimate would still be the same,  $\hat{y} = 96.77 \mu\text{g}/\text{kg}$ . That is, whether we are estimating the mean  $Y$  value or a single  $Y$  value for a particular value of  $X$ , we use the regression line in the same manner. However, the precisions of these estimates are very different.

Predicting a single  $Y$  value is much less precise than predicting the mean  $Y$  value because in addition to the uncertainty in the regression line (e.g., uncertainty in our estimates of the slope and intercept of the line), there is also uncertainty due to the inherent variability in  $Y$  values that have the same value of  $X$ . For example, there is variability among the rice arsenic concentrations for all plants with straw silicon concentrations of 40 g/kg (in fact we estimate this variability to be  $s_e$ ). The two graphs in Figure 12.7.1 illustrate the differences in our prediction precisions for the two types of estimates.



**Figure 12.7.1** 95% confidence and prediction bands for arsenic concentrations of rice. Plot (a) shows a 95% confidence band for the predicted mean arsenic concentrations and the 95% confidence interval for the predicted mean arsenic concentration when straw silicon is 40 g/kg. Plot (b) shows a 95% prediction band for predicted arsenic concentrations and the 95% prediction interval for the predicted arsenic concentration when straw silicon is 40 g/kg.

Figure 12.7.1 (a) displays a band representing all 95% confidence intervals for predicting mean arsenic levels as well as the specific interval for  $X = 40 \text{ g}/\text{kg}$  marked by the vertical line. The confidence band reflects the uncertainty associated with estimating the slope and intercept of the regression line. Notice that the intervals are narrower (more precise) for straw silicon concentrations near the center of the data set and much wider near the extreme  $X$  values. We are 95% confident that the population regression line  $\beta_0 + \beta_1x$  lies within this band. The widening of the intervals on the end is a reflection of our uncertainty in our estimate of the slope of the regression line. The width of the band in the middle expresses our uncertainty of the overall height of the regression line (vis-à-vis  $b_0$ ).

In contrast, Figure 12.7.1 (b) displays a band representing all 95% prediction intervals for predicting individual arsenic levels. The specific prediction interval for  $X = 40$  is marked by the vertical line. Note how much wider this band is in (b) than in (a). Example 12.7.1 illustrates the use of confidence and prediction intervals for prediction in regression.

**Example  
12.7.1**

**Arsenic Concentrations in Rice** Figure 12.7.1 shows that for rice with straw silicon concentrations of 40 g/kg, the 95% confidence interval for the mean arsenic concentration is about 75 to 125  $\mu\text{g}/\text{kg}$ . In other words, we are 95% confident that the mean arsenic concentration of rice from plants with straw silicon concentrations of 40 g/kg is 75 to 125  $\mu\text{g}/\text{kg}$ . On the other hand, using the prediction interval we estimate that 95% of plants with straw silicon of 40 g/kg will have rice arsenic concentrations roughly between 25 and 175  $\mu\text{g}/\text{kg}$ . ■

Recall that the regression line can be interpreted as a “line of averages,” and individuals will necessarily fall from this average. These graphs show us that we are much less certain about saying, “rice from plants with  $X$  amount of straw silicon will have  $Y$  amount of arsenic” than we are about saying “rice from plants with  $X$  amount of straw silicon will, *on average*, have  $Y$  amount of arsenic.”

## Computing the Intervals

Consider predicting  $\mu_{Y|X=x^*}$  or  $Y|X = x^*$ ; that is, predicting the mean or actual  $Y$  value when  $X = x^*$ . A 95% confidence interval for  $\mu_{Y|X=x^*}$  is given by

$$\hat{y} \pm t_{0.025} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

and a 95% prediction interval for  $Y|X = x^*$  is given by

$$\hat{y} \pm t_{0.025} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

with the critical value  $t_{0.025}$  determined from Student’s  $t$  distribution with  $\text{df} = n - 2$ .

While these two formulas are very similar, note the extra “1” under the radical sign in the prediction interval formula. This “1” factors in the added variability associated with trying to make a prediction for an individual rather than for a population mean.

As we have seen in Figure 12.7.1, both confidence and prediction intervals are wider when we are making predictions far from the center of our data. Both formulas

account for this additional uncertainty through the term  $\frac{(x^* - \bar{x})^2}{(n-1)s_x^2}$ . This term will be

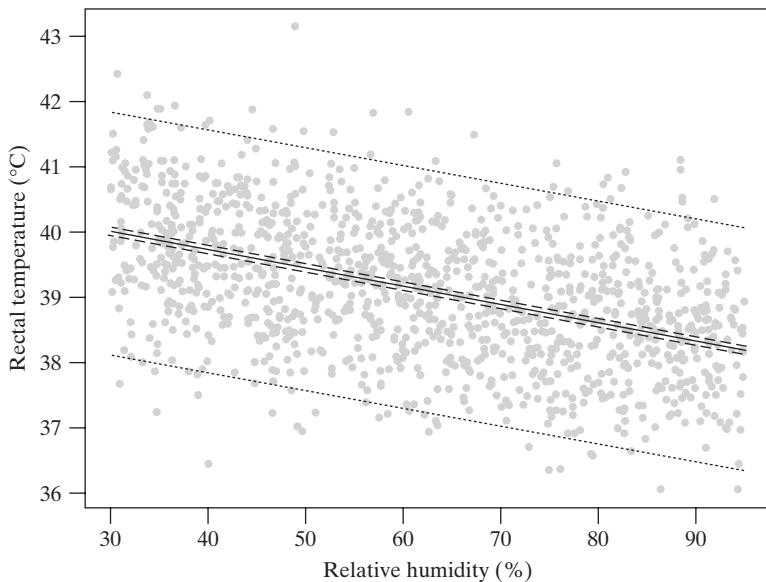
large when  $x^*$  is far from  $\bar{x}$  and thus increase the width of the interval. Note that when  $x^* = \bar{x}$  the confidence interval formula can reduce to a very familiar form:  $\hat{y} \pm t_{0.025} \left( \frac{s_e}{\sqrt{n}} \right)$ , which looks very similar to the formula for a confidence interval for a population mean from Chapter 6.

Most statistical software can compute and display confidence and prediction bands quite easily.

## Exercises 12.7.1–12.7.3

**12.7.1** In a study of heat stress on cows, researchers measured the rectal temperature ( $^{\circ}\text{C}$ ) ( $Y$ ) and relative humidity (%) ( $X$ ) for 1,280 lactating cows.<sup>26</sup> The following graph displays the data and regression line (solid line). There are two other pairs of lines on this graph: dashed and dotted. One pair of lines shows the 95% confidence band and the other shows the 95% prediction band.

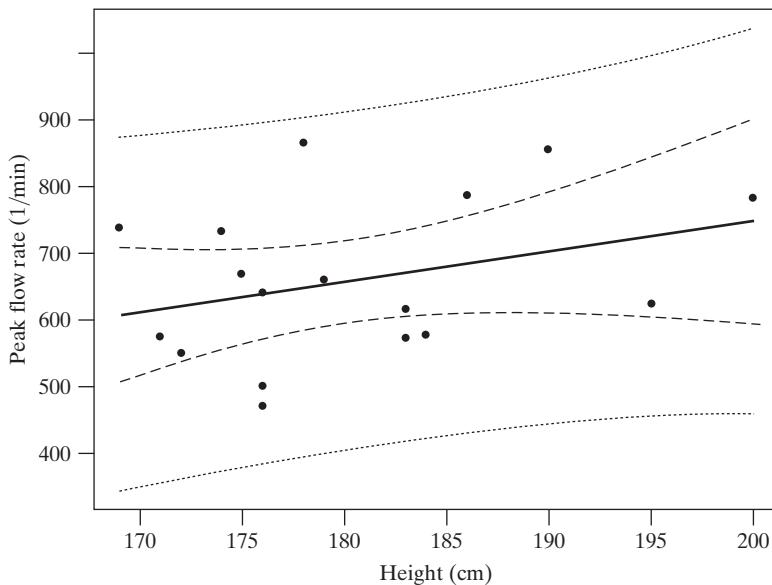
- Which pair of lines shows the confidence band? What does this band tell us?
- Which pair of lines shows us the prediction band? What does this band tell us?
- If the data set were smaller, describe what would happen to these bands. Would we have narrower or wider bands around the regression line?



**12.7.2** (Continuation of 12.7.1) Suppose 5,000 additional cows were included in the sample and a similar plot of the data, regression line, confidence and prediction bands were made of this new larger sample. Would

the prediction band get much narrower? Explain your reasoning.

**12.7.3** The following graph displays the regression line and 95% confidence and prediction bands for the peak respiration flow data from Exercise 12.3.8.



- (a) Using the graph to justify your answer, would it be very surprising to find a 195-cm-tall individual with a peak flow rate above 900 l/min?
- (b) Using the graph to justify your answer, would it be surprising to find a large group of 195-cm-tall individuals to have a mean peak flow rate above 900 l/min?

## 12.8 Perspective

To put the methods of Chapter 12 in perspective, we will discuss their relationship to methods described in earlier chapters, and to methods that might be included in a second statistics course. We begin by relating regression to the methods of Chapters 7 and 11.

### Regression and the *t* Test

When there are several  $Y$  values for each of two values of  $X$ , one could analyze the data with a two-sample  $t$  test or with a regression analysis. Each approach uses the data to estimate the conditional mean of  $Y$  for each fixed  $X$ ; these parameters are estimated by the fitted line  $b_0 + b_1x$  in the regression approach and by the individual sample means  $\bar{Y}$  in the  $t$  test approach. To test the null hypothesis of no dependence of  $Y$  on  $X$ , each approach translates the null hypothesis into its own terms. The following example illustrates the approaches.

---

**Example  
12.8.1**

**Toluene and the Brain** In Chapter 7 we analyzed data on norepinephrine (NE) concentrations in the brains of six rats exposed to toluene and of five control rats. The data are reproduced in Table 12.8.1.

**Table 12.8.1** NE concentrations (ng/gm)

	Toluene	Control
543	535	
523	385	
431	502	
635	412	
564	387	
549		
<i>n</i>	6	5
$\bar{y}$	540.83	444.20
<i>s</i>	66.12	69.64

In Chapter 7 the null hypothesis

$$H_0: \mu_1 - \mu_2 = 0$$

was tested using the (unpooled) two-sample  $t$  test. The test statistic was

$$t_s = \frac{(540.83 - 444.20) - 0}{41.195} = 2.346$$

These data could be analyzed using a pooled  $t$  test (or, equivalently, with analysis of variance). The pooled variance is

$$s_{\text{pooled}}^2 = \frac{(6 - 1)66.12^2 + (5 - 1)69.64^2}{(6 + 5 - 2)} = 4584.24 = 67.71^2$$

and the pooled SE is

$$\text{SE}_{\text{pooled}} = 67.71 \sqrt{\frac{1}{6} + \frac{1}{5}} = 41.00$$

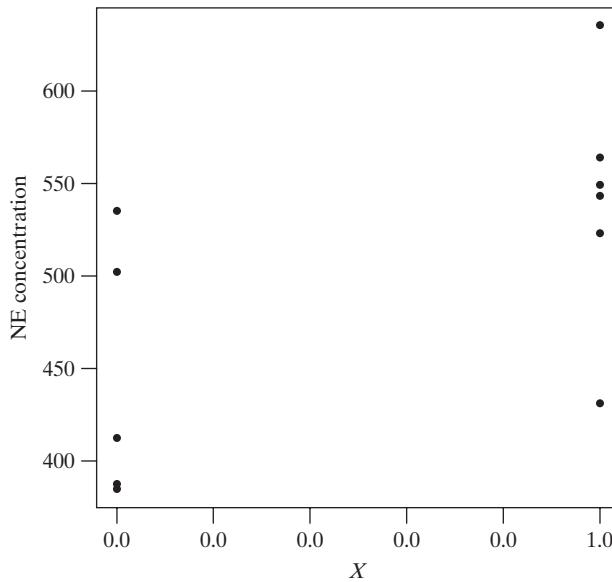
This leads to a test statistic of

$$t_s = \frac{(540.83 - 444.20) - 0}{41.00} = 2.357$$

which is not much different than the unpooled  $t$  test result.

These data can also be analyzed with a regression model. To use regression, we define an **indicator variable**—a variable that indicates group membership—as follows. Let  $X = 0$  for observations in the control group and let  $X = 1$  for observations in the toluene group. Then we can present the data graphically with a scatterplot, as in Figure 12.8.1.

**Figure 12.8.1** NE concentration data.  $X = 0$  represents the control group;  $X = 1$  represents the toluene group



We can analyze the data in the scatterplot with the linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

which states that  $\mu_{Y|X} = \beta_0 + \beta_1 X$ .

The linear model states that for rats in the control group, the (population) mean NE concentration is given by

$$\mu_{Y|X=0} = \beta_0 + \beta_1(0) = \beta_0$$

And, for rats in the toluene group, NE concentration is given by

$$\mu_{Y|X=1} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$

The difference between the two group means is  $\beta_1$ . Thus, the null hypothesis

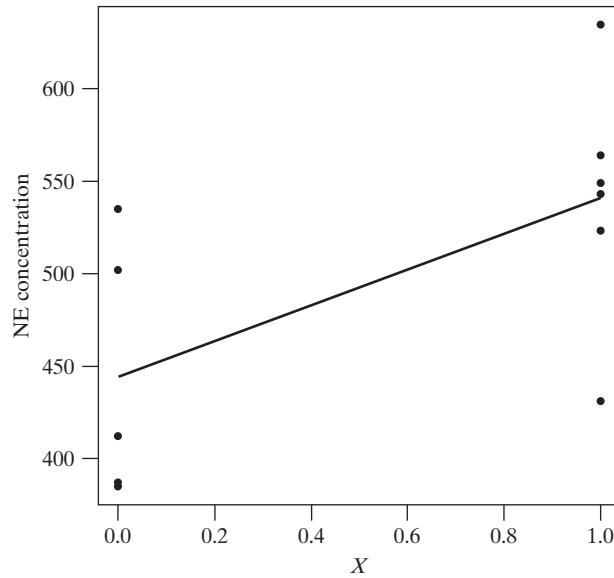
$$H_0: \mu_{Y|X=0} - \mu_{Y|X=1} = 0$$

is equivalent to the null hypothesis

$$H_0: \beta_1 = 0$$

The fitted regression line is  $\hat{y} = 444.2 + 96.63x$ . Note that when  $X = 0$ , the fitted regression line gives a value of  $\hat{y} = 444.2$ , which is the sample mean of the control group. When  $X = 1$ , the fitted regression line gives a value of  $\hat{y} = 444.2 + 96.63 = 540.83$ , which is the sample mean of the toluene group. That is, the sample value of the slope is equal to the change in the sample means when going from the control group ( $X = 0$ ) to the toluene group ( $X = 1$ ), as shown in Figure 12.8.2.

**Figure 12.8.2** NE concentration data with regression line added



The test statistic for testing the hypothesis  $H_0: \beta_1 = 0$  is

$$t_s = \frac{96.63}{41.0} = 2.36$$

This is identical to the previous pooled two-sample  $t$  test statistic. (Note that the regression analysis assumes that  $\sigma_{Y|X} = \sigma_\epsilon$  is constant. Thus, regression is similar to the pooled  $t$  test, rather than the unpooled  $t$  test.) The following computer output shows the coefficients for the fitted regression line as well as the  $t$  statistic.

The regression equation is

$$\text{NE} = 444 + 96.6X$$

Predictor	Coef	SE Coef	T	P
-----------	------	---------	---	---

Constant	444.20	30.28	14.67	0.000
----------	--------	-------	-------	-------

X	96.63	41.00	2.36	0.043
---	-------	-------	------	-------

$S = 67.7049 \quad R-\text{Sq} = 38.2\% \quad R-\text{Sq}(\text{adj}) = 31.3\%$

#### Analysis of Variance

Source	DF	SS	MS	F	P
--------	----	----	----	---	---

Regression	1	25467	25467	5.56	0.043
------------	---	-------	-------	------	-------

Residual Error	9	41256	4584		
----------------	---	-------	------	--	--

Total	10	66723			
-------	----	-------	--	--	--



The following example compares the regression approach and the two-sample approach to a data set for which (unlike Example 12.8.1)  $X$  varies within as well as between the samples.

**Example  
12.8.2**

**Blood Pressure and Platelet Calcium** In Example 12.5.3 we described blood pressure ( $X$ ) and platelet calcium ( $Y$ ) measurements on 38 subjects. Actually, the study included two groups of subjects: 38 volunteers with normal blood pressure, selected from hospital lab personnel and other nonpatients, and 45 patients with a diagnosis of high blood pressure. Table 12.8.2 summarizes the platelet calcium measurements in the two groups and Figure 12.8.3 shows the blood pressure and calcium measurements for all 83 subjects.<sup>4</sup>

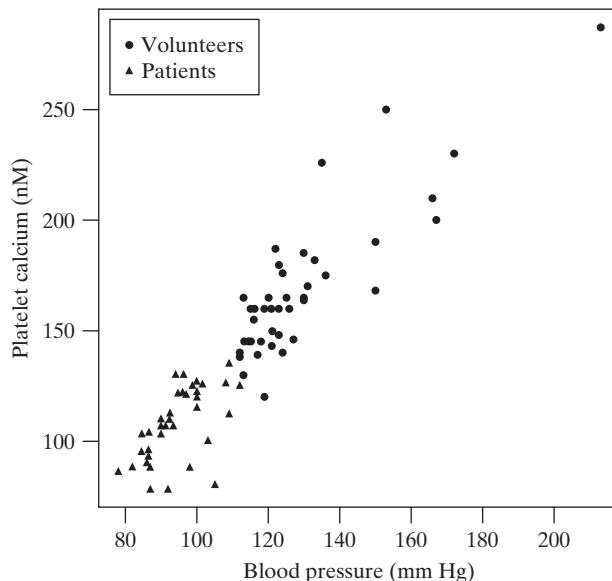
Two ways to analyze the data are (1) as two independent samples and (2) by regression analysis. To test for a relationship between blood pressure and platelet calcium (1) a two-sample  $t$  test of  $H_0: \mu_1 = \mu_2$  can be applied to Table 12.8.2; (2) a regression  $t$  test of  $H_0: \beta_1 = 0$  can be applied to the data in Figure 12.8.3. The two-sample  $t$  statistic (unpooled) is  $t_s = 11.2$  and the regression  $t$  statistic is  $t_s = 20.8$ . Both of these are highly significant, but the latter is more so because the regression analysis extracts more information from the data.

For these data, the regression approach is more enlightening and convincing than the two-sample approach. Figure 12.8.3 suggests that platelet calcium is correlated with blood pressure, not only between, but also within the two groups. Relevant regression analyses would include (1) testing for a correlation within each group separately (as in Examples 12.2.3 and 12.5.3); (2) testing for an overall correlation (as in the previous paragraph); (3) testing whether the regression lines in the two groups are identical (using methods not described in this book).

**Table 12.8.2** Platelet calcium (nM) in two groups of subjects

	Normal blood pressure	High blood pressure
$\bar{y}$	107.9	168.2
$s$	16.1	31.7
$n$	38	45

**Figure 12.8.3** Blood pressure and platelet calcium for 83 subjects



Formal testing aside, notice the advantage of the scatterplot as a tool for understanding the data and for communicating the results. Figure 12.8.3 provides eloquent testimony to the reality of the relationship between blood pressure and platelet calcium. (We emphasize once again, however, that a “real” relationship is not necessarily a causal relationship. Further, even if the relationship is causal, the data do not indicate the direction of causality—that is, whether high calcium causes high blood pressure or vice versa.\*)

Example 12.8.2 illustrates a general principle: If quantitative information on a variable  $X$  is available, it is usually better to use that information than to ignore it.

### Extensions of Least Squares

We have seen that the classical method of fitting a straight line to data is based on the least-squares criterion. This versatile criterion can be applied to many other statistical problems. For instance, in **curvilinear regression**, the least-squares criterion is used to fit curvilinear relationships such as

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Another application is **multiple regression and correlation**, in which the least-squares criterion is used to fit an equation relating  $Y$  to several  $X$  variables— $X_1, X_2$ , and so on; for instance,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

The following example illustrates both curvilinear and multiple regression.

\*In fact, the authors of the study remark that “It remains possible . . . that an increased intracellular calcium concentration is a consequence rather than a cause of elevated blood pressure.”

**Example  
12.8.3**

**Serum Cholesterol and Blood Pressure** As part of a large health study, various measurements of blood pressure, blood chemistry, and physique were made on 2,599 men.<sup>27</sup> The researchers found a positive correlation between blood pressure and serum cholesterol ( $r = 0.23$  for systolic blood pressure). But blood pressure and serum cholesterol are also related to age and physique. To untangle the relationships, the researchers used the method of least squares to fit the following equation:

$$Y = b_0 + b_1X_1 + c_1X_1^2 + b_2X_2 + b_3X_3 + b_4X_4$$

where

$Y$  = Systolic blood pressure

$X_1$  = Age

$X_2$  = Serum cholesterol

$X_3$  = Blood glucose

$X_4$  = Ponderal index (height divided by the cube root of weight)

Note that the regression is curvilinear with respect to age ( $X_1$ ) and linear in the other  $X$  variables.

By applying multiple regression and correlation analysis, the investigators determined that there is little or no correlation between blood pressure and serum cholesterol, after accounting for any relationship between blood pressure and age and ponderal index. They concluded that the observed correlation between serum cholesterol and blood pressure was an indirect consequence of the correlation of each of these with age and physique. ■

## Nonparametric and Robust Regression and Correlation

We have discussed the classical least-squares methods for regression and correlation analysis. There are also many excellent modern methods that are not based on the least-squares criterion. Some of these methods are *robust*—that is, they work well even if the conditional distributions of  $Y$  given  $X$  have long straggly tails or outliers. The nonparametric methods assume little or nothing about the form of dependence—linear or curvilinear—of  $Y$  on  $X$ , or about the form of the conditional distributions.

## Analysis of Covariance

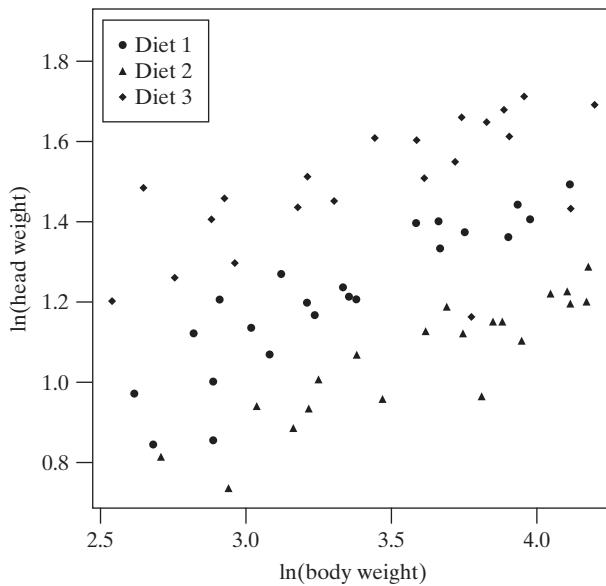
Sometimes regression ideas can add greatly to the power of a data analysis, even if the relationship between  $X$  and  $Y$  is not of primary interest. The following is an example.

**Example  
12.8.4**

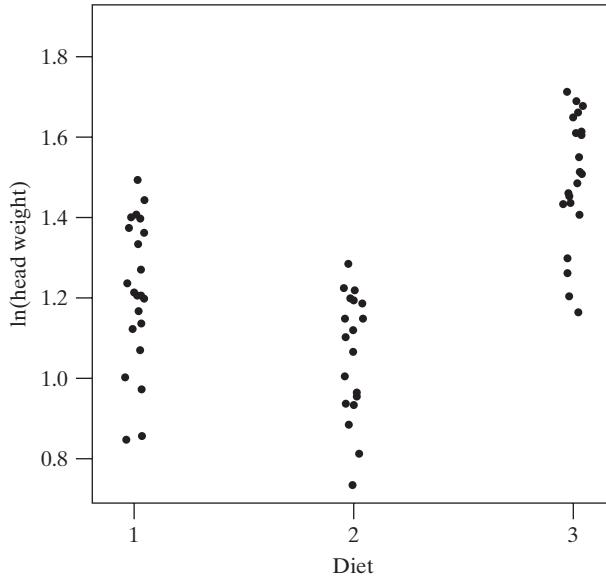
**Caterpillar Head Size** Can diet affect the size of a caterpillar's head? Such an effect is plausible, because a caterpillar's chewing muscles occupy a large part of the head. To study the effect of diet, a biologist raised caterpillars (*Pseudaletia unipuncta*) on three different diets: diet 1, an artificial soft diet; diet 2, soft grasses; and diet 3, hard grasses. He measured the weight of the head and of the entire body in the final stage of larval development. The results are shown in Figure 12.8.4, where  $Y = \ln(\text{head weight})$  is plotted against  $X = \ln(\text{body weight})$ , with different sym-

bols for the three diets.<sup>28</sup> Note that the effect of diet is striking; there is virtually no overlap between the three groups of points. But if we were to ignore  $X$  and consider  $Y$  only, as displayed in Figure 12.8.5, the effect of diet would be much less pronounced.

**Figure 12.8.4** Head weight versus body weight (on logarithmic scales) for caterpillars on three different diets



**Figure 12.8.5** Head weight (on a logarithmic scale) for caterpillars on three different diets



Example 12.8.4 shows how comparison of several groups with respect to a variable  $Y$  can be strengthened by using information on an auxiliary variable  $X$  that is correlated with  $Y$ . A classical method of statistical analysis for such data is **analysis of covariance**, which proceeds by fitting regression lines to the  $(X, Y)$  data. But even without this formal technique, an investigator can often clarify the interpretation of data simply by constructing a scatterplot like Figure 12.8.4. Plotting the data against

$X$  has the visual effect of removing that part of the variability in  $Y$  which is accounted for by  $X$ , causing the treatment effect to stand out more clearly against the residual background variation.

## Logistic Regression

Regression and correlation are used to analyze the relationship between two quantitative variables,  $X$  and  $Y$ . Sometimes data arise in which a quantitative variable  $X$  is used to predict the response of a categorical variable  $Y$ . For example, we might wish to use  $X$  = cholesterol level as a predictor of whether or not a person has heart disease. Here we could define a variable  $Y$  as 1 if a person has heart disease and 0 otherwise. We could then study how  $Y$  depends on  $X$ . When the response variable is dichotomous, as in this case, a technique known as **logistic regression** can be used to model the relationship. For example, logistic regression could be used to model how the probability of heart disease depends on blood pressure.

Example 12.8.5 provides a more detailed look at the use of logistic regression.

---

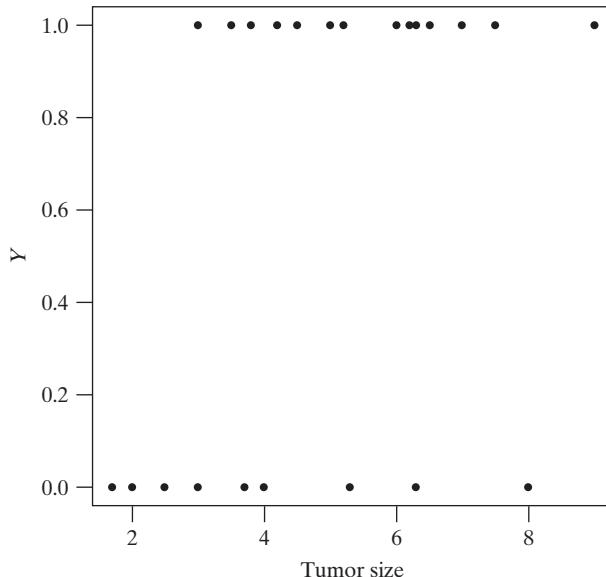
### Example 12.8.5

**Esophageal Cancer** Esophageal cancer is a serious and very aggressive disease. Scientists conducted a study of 31 patients with esophageal cancer in which they studied the relationship between the size of the tumor that a patient had and whether or not the cancer had spread (metastasized) to the lymph nodes of the patient. In this study the response variable is dichotomous:  $Y = 1$  if the cancer had spread to the lymph nodes and  $Y = 0$  if not. The predictor variable is the size (recorded as the maximum dimension, in cm) of the tumor found in the esophagus. The data are given in Table 12.8.3 and plotted in Figure 12.8.6.<sup>29</sup>

**Table 12.8.3** Esophageal cancer data

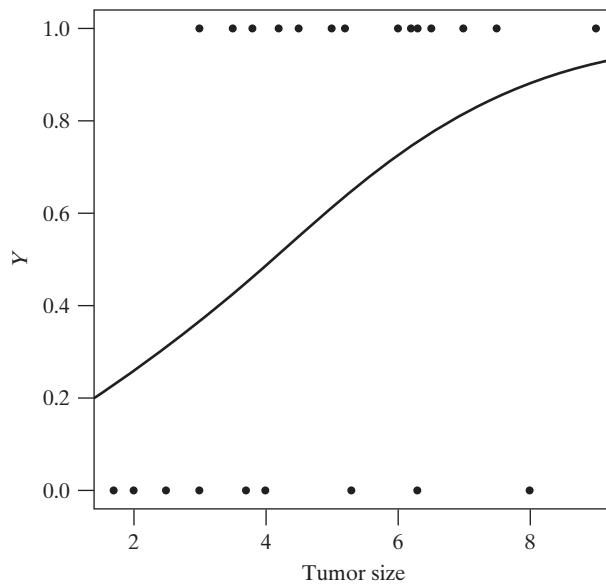
Patient number	Tumor size (cm), $X$	Lymph node metastasis, $Y$	Patient number	Tumor size (cm), $X$	Lymph node metastasis, $Y$
1	6.5	1	17	6.2	1
2	6.3	0	18	2.0	0
3	3.8	1	19	9.0	1
4	7.5	1	20	4.0	0
5	4.5	1	21	3.0	1
6	3.5	1	22	6.0	1
7	4.0	0	23	4.0	0
8	3.7	0	24	4.0	0
9	6.3	1	25	4.0	0
10	4.2	1	26	5.0	1
11	8.0	0	27	9.0	1
12	5.2	1	28	4.5	1
13	5.0	1	29	3.0	0
14	2.5	0	30	3.0	1
15	7.0	1	31	1.7	0
16	5.3	0			

**Figure 12.8.6** Lymph node metastasis,  $Y$ , as a function of tumor size,  $X$



The idea of logistic regression is to model the relationship between  $X$  and  $Y$  by fitting a response curve that is always between 0 and 1. With values bound between 0 and 1, the logistic regression model can be used to estimate the probability  $Y = 1$  (e.g., metastasis) for a given value of  $X$  (e.g., tumor size). Thus, unlike linear regression, in which we model  $Y$  as a linear function of  $X$  (which does not remain between 0 and 1), with logistic regression we model the relationship between  $X$  and  $Y$  as having an “S” shape, as shown in Figure 12.8.7.

**Figure 12.8.7** Lymph node metastasis,  $Y$ , as a function of tumor size,  $X$ , with smooth curve added

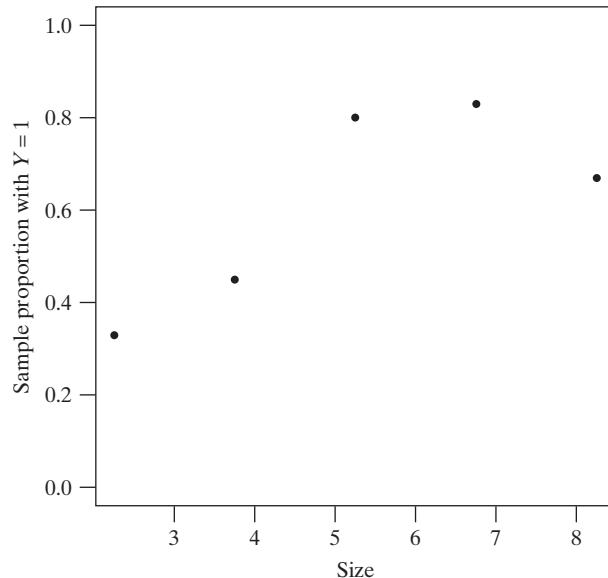


One way to begin understanding the data is to form groups on the basis of size,  $X$ , and calculate for each group the proportion of the  $Y$  values that are 1’s. (This is somewhat analogous to finding the graph of averages described in Section 12.3, except that here we group together data points with differing  $X$  values.) Table 12.8.4 provides such a summary, which is shown graphically in Figure 12.8.8. Note that the

**Table 12.8.4** Esophageal cancer data in groups

Size range	Points with $Y = 1$	Points with $Y = 0$	Fraction $Y = 1$	Proportion $Y = 1$
(1.5, 3.0]	2	4	2/6	0.33
(3.0, 4.5]	5	6	5/11	0.45
(4.5, 6.0]	4	1	4/5	0.80
(6.0, 7.5]	5	1	5/6	0.83
(7.5, 9.0]	2	1	2/3	0.67

**Figure 12.8.8** Sample proportion of patients with lymph node metastasis ( $Y = 1$ ) for patients grouped by tumor size,  $X$



proportion of 1's (that is, the proportion of patients for whom the cancer has metastasized) increases as tumor size increases (except for the last category of (7.5, 9], which has only three cases).

We can fit a smooth, continuous function to the data, to smooth out the percentages in the last column of Table 12.8.4. We can also impose the condition that the function be monotonically increasing, meaning that the probability of metastasis ( $Y = 1$ ) strictly increases as tumor size increases. To do this, we use a computer to fit a **logistic response function**.\* The fitted logistic response function for the esophageal cancer data is

$$\Pr \{Y = 1\} = \frac{e^{-2.086 + 0.5117 \times \text{size}}}{1 + e^{-2.086 + 0.5117 \times \text{size}}}$$

For example, suppose the size of a tumor is 4.0 cm. Then the predicted probability that the cancer has metastasized is

$$\frac{e^{-2.086 + 0.5117(4)}}{1 + e^{-2.086 + 0.5117(4)}} = \frac{e^{-0.0392}}{1 + e^{-0.0392}} = \frac{0.96156}{1 + 0.96156} = 0.49$$

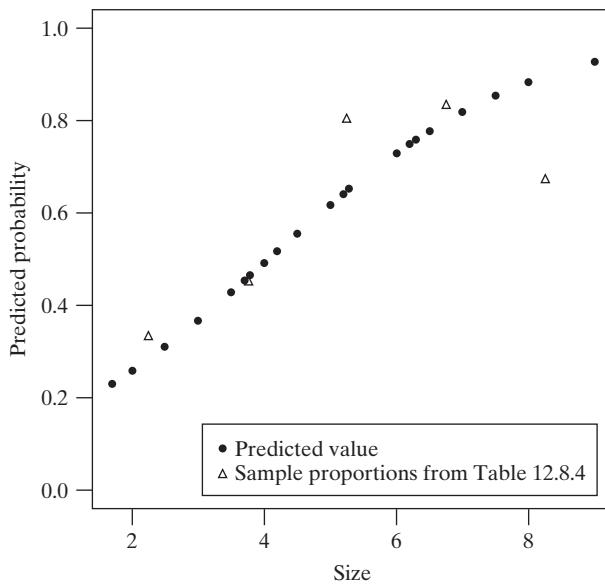
\*Fitting a logistic model is quite a bit more complicated than is fitting a linear regression model. A technique known as maximum likelihood estimation is commonly used, with the help of a computer.

On the other hand, suppose the size of a tumor is 8.0 cm. Then the predicted probability that the cancer has metastasized is

$$\frac{e^{-2.086 + 0.5117 \times 8}}{1 + e^{-2.086 + 0.5117 \times 8}} = \frac{e^{2.0076}}{1 + e^{2.0076}} = \frac{7.4454}{1 + 7.4454} = 0.88$$

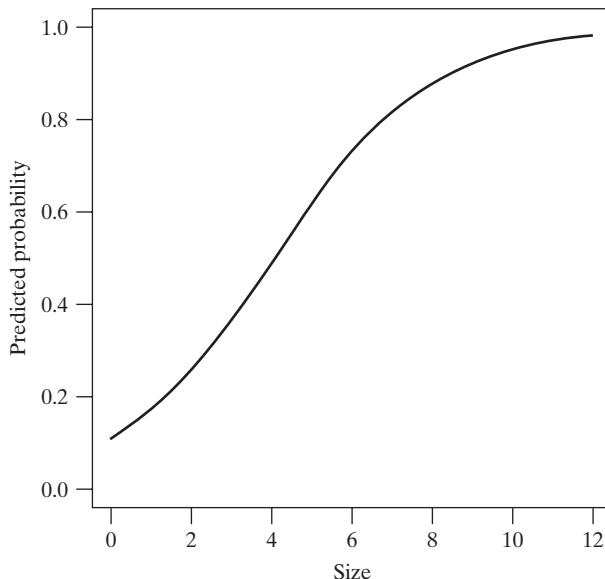
We can calculate a predicted probability that  $Y = 1$  for each value of  $X$ . Figure 12.8.9 shows a graph of such predictions, which have, generally speaking, an S shape. ■

**Figure 12.8.9** Predicted probability that  $Y = 1$  as a function of tumor size,  $X$  with sample proportions from Table 12.8.4



The S shape of the logistic curve is easier to see if we extend the range of  $X$ , as shown in Figure 12.8.10. As  $X$  grows, the logistic curve approaches, but never exceeds, 1. Likewise, if we were to extend the curve into the region where  $X$  is less than zero we would see that as  $X$  gets smaller and smaller, the logistic curve approaches,

**Figure 12.8.10** Logistic response function for the cancer data, shown over a larger range



but never drops below, 0. (Of course, in the setting of Example 12.8.5 it does not make sense to talk about tumor sizes that are negative. Thus, we only show the logistic curve for positive values of  $X$ .)

In general, if we have a logistic response function

$$\Pr \{Y = 1\} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

with  $b_1$  positive, then as  $X$  grows,  $\Pr \{Y = 1\}$  approaches one and as  $X$  gets smaller,  $\Pr \{Y = 1\}$  approaches zero. Thus, unlike a linear regression model, a logistic curve stays between zero and one, which makes it appropriate for modeling a response probability.

## 12.9 Summary of Formulas

For convenient reference, we summarize the formulas presented in Chapter 12.

### Correlation Coefficient

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

Fact 12.3.1:  $r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$

### Fitted Regression Line

$$\hat{y} = b_0 + b_1 x$$

where

$$b_1 = r \times \left( \frac{s_y}{s_x} \right)$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Residuals:

$$y_i - \hat{y}_i \quad \text{where} \quad \hat{y}_i = b_0 + b_1 x_i$$

Residual Sum of Squares:

$$\text{SS(resid)} = \sum (y_i - \hat{y}_i)^2$$

Residual Standard Deviation:

$$s_e = \sqrt{\frac{\text{SS(resid)}}{n - 2}}$$

**Inference**

Standard Error of  $b_1$ :

$$\text{SE}_{b_1} = \frac{s_e}{s_x \sqrt{n - 1}}$$

95% confidence interval for  $\beta_1$ :

$$b_1 \pm t_{0.025} \text{SE}_{b_1}$$

Test of  $H_0: \beta_1 = 0$  or  $H_0: \rho = 0$ :

$$t_s = \frac{b_1}{\text{SE}_{b_1}} = r \sqrt{\frac{n - 1}{1 - r^2}}$$

Critical values for the test and confidence interval are determined from Student's  $t$  distribution with  $\text{df} = n - 2$ .

**Prediction**

A 95% confidence interval for  $\mu_{Y|X=x^*}$  is given by

$$\hat{y} \pm t_{0.025} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n - 1)s_x^2}}$$

A 95% prediction interval for  $Y|X = x^*$  is given by

$$\hat{y} \pm t_{0.025} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n - 1)s_x^2}}$$

Critical values for intervals are determined from Student's  $t$  distribution with  $\text{df} = n - 2$ .

**Exercises 12.S.1–12.S.22**

**12.S.1** In a study of the Mormon cricket (*Anabrus simplex*), the correlation between female body weight and ovary weight was found to be  $r = 0.836$ . The standard deviation of the ovary weights of the crickets was 0.429 g. Assuming that the linear model is applicable, estimate the standard deviation of ovary weights of crickets whose body weight is 4 g.<sup>30</sup>

**12.S.2** In a study of crop losses due to air pollution, plots of Blue Lake snap beans were grown in open-top field chambers, which were fumigated with various concentrations of sulfur dioxide. After a month of fumigation, the plants were harvested and the total yield of bean pods was recorded for each chamber. The results are shown in the table.<sup>31</sup>

$X = \text{SULFUR DIOXIDE CONCENTRATION (ppm)}$				
	0	0.06	0.12	0.30
$Y = \text{yield (kg)}$	1.15	1.19	1.21	0.65
	1.30	1.64	1.00	0.76
	1.57	1.13	1.11	0.69
Mean	1.34	1.32	1.11	0.70

Preliminary calculations yield the following results.

$$\bar{x} = 0.12 \quad \bar{y} = 1.117$$

$$s_X = 0.11724 \quad s_Y = 0.31175$$

$$r = -0.8506 \quad \text{SS(resid)} = 0.2955$$

- Calculate the linear regression of  $Y$  on  $X$ .
- Plot the data and draw the regression line on your graph.
- Calculate  $s_e$ . What are the units of  $s_e$ ?

**12.S.3** Refer to Exercise 12.S.2.

- Assuming that the linear model is applicable, find estimates of the mean and the standard deviation of yields of beans exposed to 0.24 ppm of sulfur dioxide.
- Which condition of the linear model appears doubtful for the snap bean data?

**12.S.4** Refer to Exercise 12.S.2. Consider the null hypothesis that sulfur dioxide concentration has no effect on yield. Assuming that the linear model holds, formulate this as a hypothesis about the true regression line. Use the data to test the hypothesis against a directional alternative. Let  $\alpha = 0.05$ .

**12.S.5** Another way to analyze the data of Exercise 12.S.2 is to take each treatment mean as the observation  $Y$ ; then the data would be summarized as in the accompanying table.

SULFUR DIOXIDE $X$ (ppm)		MEAN YIELD $Y$ (kg)
0.00		1.34
0.06		1.32
0.12		1.11
0.30		0.70
Mean	0.1200	1.1175
SD	0.12961	0.29714
$r = -0.98666$		
SS(resid) = 0.007018		

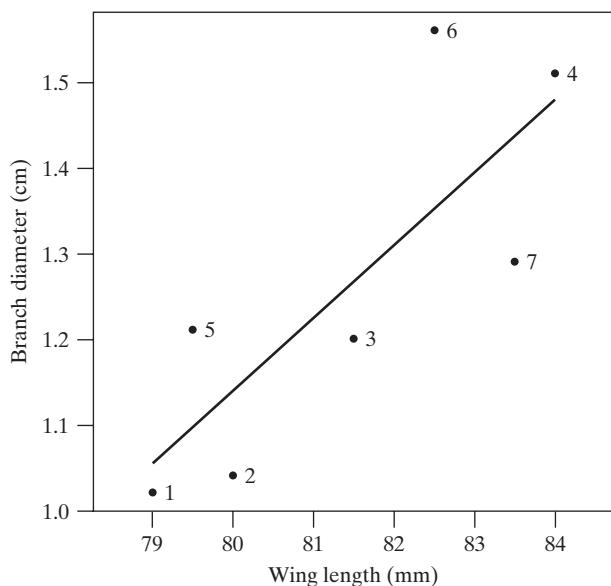
- (a) For the regression of mean yield on  $X$ , calculate the regression line and the residual standard deviation, and compare with the results of Exercise 12.S.2. Explain why the discrepancy is not surprising.
- (b) What proportion of the variability in mean yield is explained by the linear relationship between mean yield and sulfur dioxide? Using the data in Exercise 12.S.5, what proportion of the variability in individual chamber yield is explained by the linear relationship between individual chamber yield and sulfur dioxide? Explain why the discrepancy is not surprising.

**12.S.6** In a study of the tufted titmouse (*Parus bicolor*), an ecologist captured seven male birds, measured their wing lengths and other characteristics, and then marked and released them. During the ensuing winter, he repeatedly observed the marked birds as they foraged for insects and seeds on tree branches. He noted the branch diameter on each occasion, and calculated (from 50 observations) the average branch diameter for each bird. The results are shown in the table.<sup>32</sup>

BIRD	WING LENGTH $X$ (mm)	BRANCH DIAMETER $Y$ (cm)
1	79.0	1.02
2	80.0	1.04
3	81.5	1.20
4	84.0	1.51
5	79.5	1.21
6	82.5	1.56
7	83.5	1.29
Mean	81.429	1.2614
SD	1.98806	0.21035
$r = 0.80335$		
SS(resid) = 0.09415		

- (a) Calculate  $s_e$  and specify the units. Verify the approximate relationship between  $s_Y$  and  $s_e$ , and  $r$ .
- (b) Do the data provide sufficient evidence to conclude that the diameter of the forage branches chosen by male titmice is correlated with their wing length? Test an appropriate hypothesis against a nondirectional alternative. Let  $\alpha = 0.05$ .
- (c) The test in part (a) was based on 7 observations, but each branch diameter value was the mean of 50 observations. If we were to test the hypothesis of part (a) using the raw numbers, we would have 350 observations rather than only 7. Why would this approach not be valid?

**12.S.7** (*Continuation of 12.S.6*) A scatterplot and fitted regression line of the data from Exercise 12.S.6 follow. The individual birds are labeled in the plot.



- (a) Which bird/point has the largest regression residual?
- (b) Which bird(s)/points(s) have the most leverage?
- (c) Are there any birds/points that are influential?
- (d) Invent your own bird observation of  $x = \text{wing length}$  and  $y = \text{branch diameter}$  that would be an example of a regression outlier.
- (e) Invent your own bird observation of  $x = \text{wing length}$  and  $y = \text{branch diameter}$  that would be an example of a leverage point.

**12.S.8** Exericise 12.3.7 deals with data on the relationship between body length and jumping distance of bullfrogs. A third variable that was measured in that study was the mass of each bullfrog. The following table shows these data.<sup>16</sup>

BULLFROG	LENGTH X (mm)	MASS Y (g)
1	155	404
2	127	240
3	136	296
4	135	303
5	158	422
6	145	308
7	136	252
8	172	533.8
9	158	470
10	162	522.9
11	162	356
Mean	149.636	373.427
SD	14.4725	104.2922

Preliminary calculations yield the following results:

$$r = 0.90521 \quad \text{SS(resid)} = 19642$$

- Calculate the linear regression of  $Y$  on  $X$ .
- Interpret the value of the slope of the regression line,  $b_1$ , in the context of this setting.
- Calculate and interpret the value of  $s_e$  in the context of this setting.
- Calculate and interpret the value of  $r^2$  in the context of this problem.

**12.S.9** (*Continuation of 12.S.8*). A residual plot and normal probability plot from the linear regression of  $Y$  on  $X$  based on the bullfrog mass data in Exercise 12.S.8 follow.

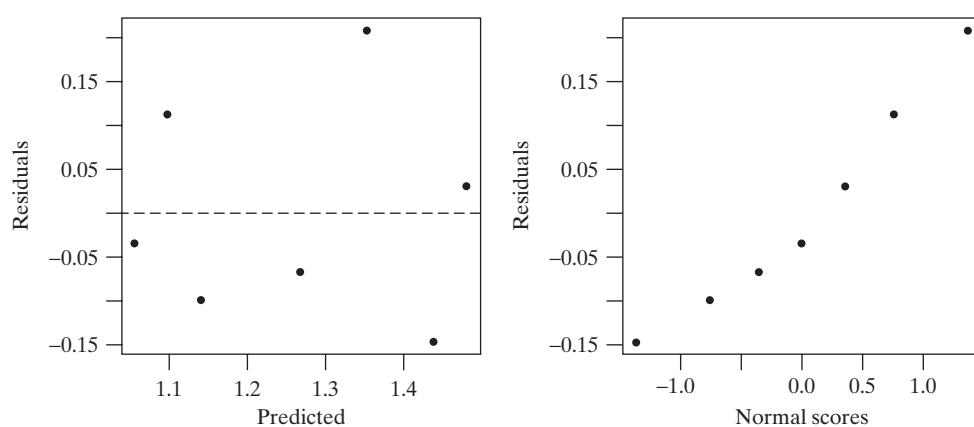
Use these plots to comment on the required conditions for inference in regression. Is there any reason to substantially doubt that these conditions are met?

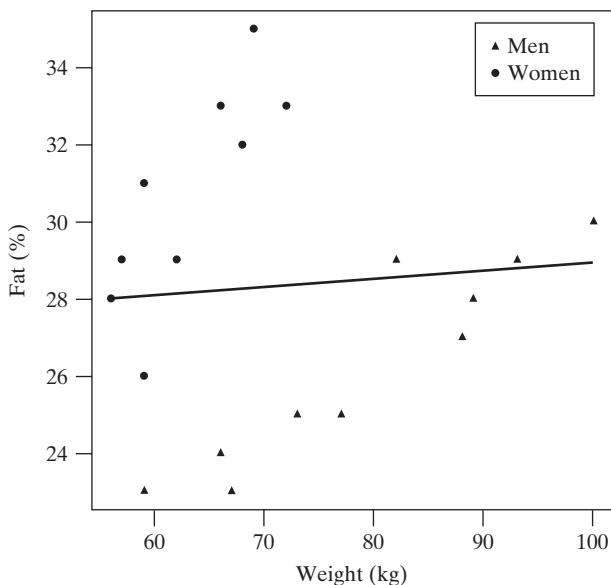
**12.S.10** An exercise physiologist used skinfold measurements to estimate the total body fat, expressed as a percentage of body weight, for 19 participants in a physical fitness program. The body fat percentages and the body weights are shown in the table.<sup>33</sup>

PARTICIPANT	WEIGHT		FAT		
	X (kg)	Y (%)	PARTICIPANT	X (kg)	Y (%)
1	89	28	11	57	29
2	88	27	12	68	32
3	66	24	13	69	35
4	59	23	14	59	31
5	93	29	15	62	29
6	73	25	16	59	26
7	82	29	17	56	28
8	77	25	18	66	33
9	100	30	19	72	33
10	67	23			

Actually, participants 1 to 10 are men, and participants 11 to 19 are women. A summary and graph of the data for men, women, and both sexes combined into a single sample follow.

MEN ( $n = 10$ )	WOMEN ( $n = 9$ )	BOTH SEXES ( $n = 19$ )
$\bar{x} = 79.40$	$\bar{x} = 63.1$	$\bar{x} = 71.68$
$\bar{y} = 26.30$	$\bar{y} = 30.67$	$\bar{y} = 28.37$
$s_X = 13.2430$	$s_X = 5.7975$	$s_X = 13.1320$
$s_Y = 2.6269$	$s_Y = 2.8723$	$s_Y = 3.4835$
$r = 0.9352$	$r = 0.8132$	$r = 0.0780$

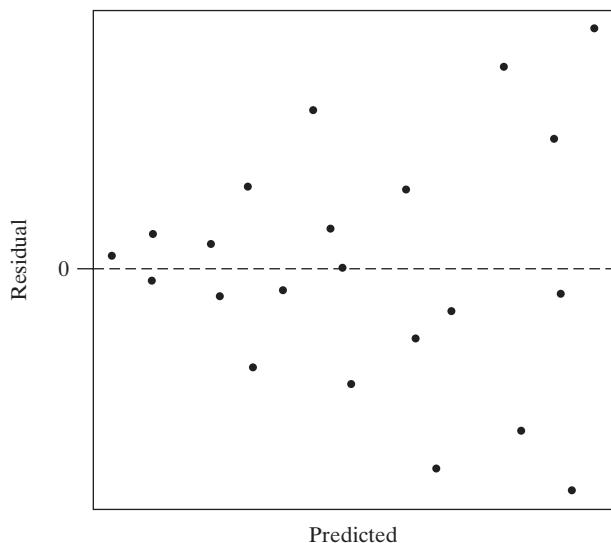




- (a) Compute the regression equations for the males and females separately.
- (b) The equation to the fitted regression line for both sexes combined, which is shown on the plot, is  $\hat{y} = 26.88 + 0.021x$ . How does the slope of this line compare to the slopes you computed in part (a)? Can you explain the discrepancy?
- (c) Examine the correlation coefficients for (i) the males, (ii) the females, and (iii) both sexes combined. Do these values agree with your reasoning provided in part (b)?

**12.S.11** Refer to the respiration rate data of Exercise 12.3.6. Construct a 95% confidence interval for  $\beta_1$ .

**12.S.12** The following plot is a residual plot from fitting a regression model to some data. Make a sketch of the scatterplot of the data that led to this residual plot. (Note: There are two possible scatterplots—one in which  $b_1$  is positive and one in which  $b_1$  is negative.)



**12.S.13** Biologists studied the relationship between embryonic heart rate and egg mass for 20 species of birds. They found that heart rate,  $Y$ , has a linear relationship with the logarithm of egg mass,  $X$ . The data are given in the following table.<sup>34</sup>

SPECIES	EGG MASS (g)	LOG-(EGG MASS) X	HEART RATE Y (beats/min)
Zebra finch	0.96	-0.018	335
Bengalese finch	1.10	0.041	404
Marsh tit	1.39	0.143	363
Bank swallow	1.42	0.152	298
Great tit	1.59	0.201	348
Varied tit	1.69	0.228	356
Tree sparrow	2.09	0.320	335
Budgerigar	2.19	0.340	314
House martin	2.25	0.352	357
Japenese bunting	2.56	0.408	370
Red-cheeked starling	4.14	0.617	358
Cockatiel	5.08	0.706	300
Brown-eared bulbul	6.40	0.806	333
Domestic pigeon	17.10	1.233	247
Fantail pigeon	19.70	1.294	267
Homing pigeon	19.80	1.297	230
Barn owl	20.10	1.303	219
Crow	20.50	1.312	297
Cattle egret	27.50	1.439	251
Lanner falcon	41.20	1.615	242
Mean	9.94	0.690	311

For these data the fitted regression equation is

$$\hat{y} = 368.06 - 82.452x$$

and

$$SS(\text{resid}) = 15748.6$$

- (a) Interpret the value of the intercept of the regression line,  $b_0$ , in the context of this setting.
- (b) Interpret the value of the slope of the regression line,  $b_1$ , in the context of this setting.
- (c) Calculate  $s_e$  and specify the units.
- (d) Interpret the value of  $s_e$  in the context of this setting.

**12.S.14** (Computer exercise) The accompanying table gives two data sets: (A) and (B). The values of  $X$  are the same for both data sets and are given only once.

(A)			(B)		
X	Y	Y	X	Y	Y
0.61	0.88	0.96	2.56	1.97	1.20
0.93	1.02	0.97	2.74	2.02	3.59
1.02	1.12	0.07	3.04	2.26	3.09
1.27	1.10	2.54	3.13	2.27	1.55
1.47	1.44	1.41	3.45	2.43	0.71
1.71	1.45	0.84	3.48	2.57	3.05
1.91	1.41	0.32	3.79	2.53	2.54
2.00	1.59	1.46	3.96	2.73	3.33
2.27	1.58	2.29	4.12	2.92	2.38
2.33	1.66	2.51	4.21	2.96	3.08

- (a) Generate scatterplots of the two data sets.
- (b) For each data set (i) estimate  $r$  visually and (ii) calculate  $r$ .
- (c) For data set (a), multiply the values of  $X$  by 10, and multiply the values of  $Y$  by 3 and add 5. Recalculate  $r$  and compare with the value before the transformation. How is  $r$  affected by the linear transformation?
- (d) Find the equations of the regression lines and verify that the regression lines for the two data sets are virtually identical (even though the correlation coefficients are very different).
- (e) Draw the regression line on each scatterplot.
- (f) Construct a scatterplot in which the two data sets are superimposed, using different plotting symbols for each data set.

**12.S.15** (*Computer exercise*) This exercise shows the power of scatterplots to reveal features of the data that may not be apparent from the ordinary linear regression calculations. The accompanying table gives three fictitious data sets, A, B, and C. The values of  $X$  are the same for each data set, but the values of  $Y$  are different.<sup>35</sup>

DATA SET:	A	B	C
X	Y	Y	Y
10	8.04	9.14	7.46
8	6.95	8.14	6.77
13	7.58	8.74	12.74
9	8.81	8.77	7.11
11	8.33	9.26	7.81
14	9.96	8.10	8.84
6	7.24	6.13	6.08
4	4.26	3.10	5.39
12	10.84	9.13	8.15
7	4.82	7.26	6.42
5	5.68	4.74	5.73

- (a) Verify that the fitted regression line is almost exactly the same for all three data sets. Are the residual standard deviations the same? Are the values of  $r$  the same?
- (b) Construct a scatterplot for each of the data sets. What does each plot tell you about the appropriateness of linear regression for the data set?
- (c) Plot the fitted regression line on each of the scatterplots.

**12.S.16** (*Computer exercise*) In a pharmacological study, 12 rats were randomly allocated to receive an injection of amphetamine at one of two dosage levels or an injection of saline. Shown in the table is the water consumption of each animal (ml water per kg body weight) during the 24 hours following injection.<sup>36</sup>

DOSE OF AMPHETAMINE (ml/kg)		
0	1.25	2.5
122.9	118.4	134.5
162.1	124.4	65.1
184.1	169.4	99.6
154.9	105.3	89.0

- (a) Calculate the regression line of water consumption on dose of amphetamine, and calculate the residual standard deviation.
- (b) Construct a scatterplot of water consumption against dose.
- (c) Draw the regression line on the scatterplot.
- (d) Use linear regression to test the hypothesis that amphetamine has no effect on water consumption against the alternative that amphetamine tends to reduce water consumption. (Use  $\alpha = 0.05$ .)
- (e) Use analysis of variance to test the hypothesis that amphetamine has no effect on water consumption. (Use  $\alpha = 0.05$ .) Compare with the result of part (d).
- (f) What conditions are necessary for the validity of the test in part (d) but not for the test in part (e)?
- (g) Calculate the pooled standard deviation from the ANOVA, and compare it with the residual standard deviation calculated in part (a).

**12.S.17** (*Computer exercise*) Consider the Amazon tree data from Exercise 12.6.9. The researchers in this study were interested in how age,  $Y$ , is related to  $X$  = “growth rate,” where growth rate is defined as diameter/age (i.e., cm of growth per year).

- (a) Create the variable “growth rate” by dividing each diameter by the corresponding tree age.
- (b) Make a scatterplot of  $Y$  = age versus  $X$  = growth rate and fit a regression line to the data.

- (c) Make a residual plot from the regression in part (b). Then make a normal probability plot of the residuals. How do these plots call into question the use of a linear model and regression inference procedures?
- (d) Take the logarithm of each value of age and of each value of growth rate. Make a scatterplot of  $Y = \log(\text{age})$  versus  $X = \log(\text{growth rate})$  and fit a regression line to the data.
- (e) Make a residual plot from the regression in part (d). Then make a normal probability plot of the residuals. Based on these plots, does a regression model in log scale, from part (d), seem appropriate?

**12.S.18** (Computer exercise) Researchers measured the blood pressures of 22 students in two situations: when the students were relaxed and when the students were taking an important examination. The table lists the systolic and diastolic pressures for each student in each situation.<sup>37</sup>

DURING EXAM		RELAXED	
SYSTOLIC PRESSURE (mm Hg)	DIASTOLIC PRESSURE (mm Hg)	SYSTOLIC PRESSURE (mm Hg)	DIASTOLIC PRESSURE (mm Hg)
132	75	110	70
124	170	90	75
110	65	90	65
110	65	110	80
125	65	100	55
105	70	90	60
120	70	120	80
125	80	110	60
135	80	110	70
105	80	110	70
110	70	85	65
110	70	100	60
110	70	120	80
130	75	105	75
130	70	110	70
130	70	120	80
120	75	95	60
130	70	110	65
120	70	100	65
120	80	95	65
120	70	90	60
130	80	120	70

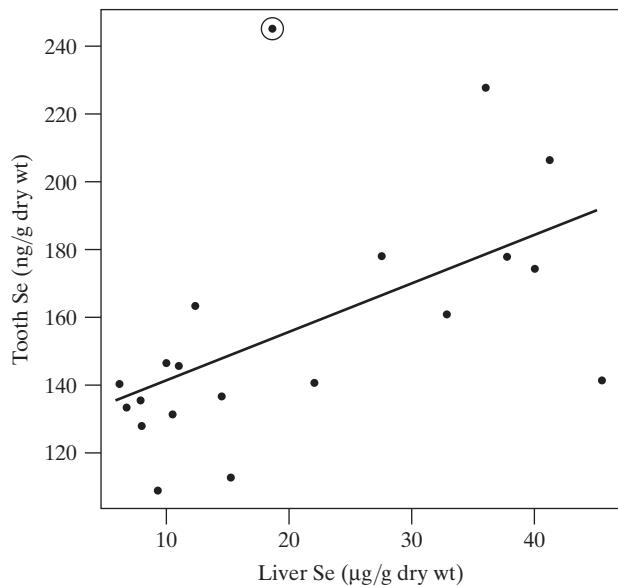
- (a) Compute the change in systolic pressure by subtracting systolic pressure when relaxed from systolic pressure during the exam; call this variable  $X$ .
- (b) Repeat part (a) for diastolic pressure. Call the resulting variable  $Y$ .

- (c) Make a scatterplot of  $Y$  versus  $X$  and fit a regression line to the data.
- (d) Make a residual plot from the regression in part (c).
- (e) Note the outlier in the residual plot [and on the scatterplot from part (c)]. Delete the outlier from the data set. Then repeat parts (c) and (d).
- (f) What is the fitted regression model (after the outlier has been removed)?

**12.S.19** (Continuation of 12.S.18) Consider the data from Exercise 12.S.18, part (f).

- (a) Construct a 95% confidence interval for  $\beta_1$ .
- (b) Interpret the confidence interval from part (a) in the context of this setting.

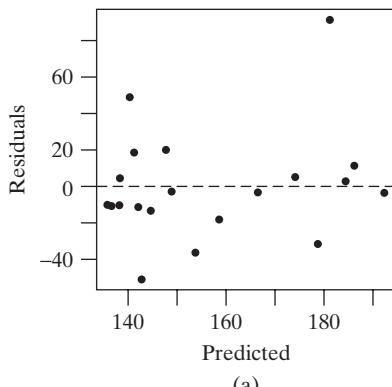
**12.S.20** Selenium (Se) is an essential element which has been shown to play an important role in protecting marine mammals against the toxic effects of mercury (Hg) and other metals. It has been suggested that metal concentrations in marine mammal teeth can potentially be used as bioindicators for body burden. Twenty Belugas (*Delphinapterus leucas*) were harvested from the Mackenzie Delta, Northwest Territories, in 1996 and 2002, as part of an annual traditional Inuit hunt. Tooth and liver Se concentrations are reported in the table, summarized, and graphed.<sup>38</sup>



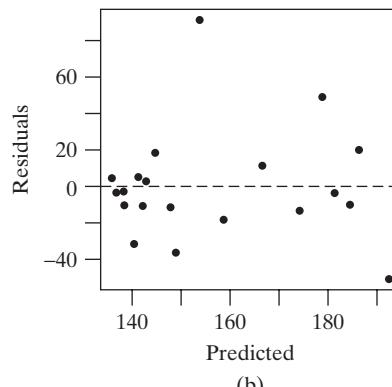
- (a) Can we regard the sample correlation between Tooth ( $Y$ ) and Liver ( $X$ ) selenium,  $r = 0.53726$ , as an estimate of the population correlation coefficient? Briefly explain.
- (b) If the circled point were removed from the data set, would the sample correlation listed in part (a) increase, decrease, or stay about the same?

WHALE	LIVER SE ( $\mu\text{g/g}$ )	TOOTH SE (ng/g)	WHALE	LIVER SE ( $\mu\text{g/g}$ )	TOOTH SE (ng/g)
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

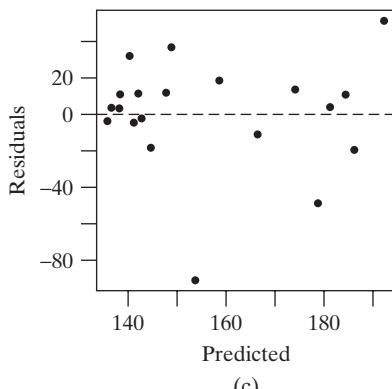
- (c) If the roles of  $X$  and  $Y$  were reversed (i.e.,  $Y = \text{Liver}$  and  $X = \text{Tooth Selenium}$ ), would the sample correlation listed in part (a) increase, decrease, or stay about the same?  
 (d) Is the circled point on the plot a leverage and/or influential point? Explain briefly.  
 (e) Is the circled point on the plot an outlier?



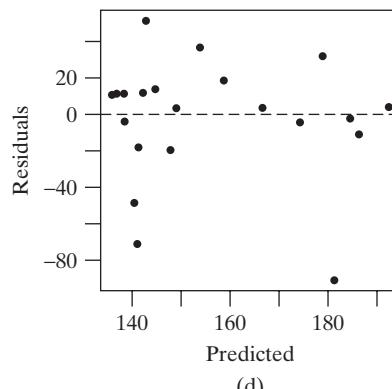
(a)



(b)



(c)



(d)

- 12.S.21** (*Continuation of 12.S.20*) The following are summary statistics for the Selenium data in Exercise 12.S.20.
- $$\bar{x} = 20.684 \quad \bar{y} = 156.599$$
- $$s_x = 13.4489 \quad s_y = 36.0586$$
- $$r = 0.53726 \quad \text{SS(resid)} = 17,573.3$$

- (a) Calculate the regression line of Tooth Selenium on Liver Selenium.  
 (b) Compute a 95% confidence interval for the slope of the regression line.  
 (c) Interpret the interval computed in part (b) in the context of the problem.  
 (d) Using the interval computed in part (b), is it reasonable to believe that the slope is as small as 0.25 (ng/g)/( $\mu\text{g/g}$ )?

- 12.S.22** (*Continuation of 12.S.20 and 12.S.21*) Referring to the data plotted in Exercise 12.S.20, which of the following is a residual plot resulting from fitting the regression line in Exercise 12.S.21, part (a)? Justify your choice.

- 12.S.23** (*Continuation of 12.S.20*) The whales observed in this study were harvested during a traditional Inuit hunt in two particular years. What are we assuming about the captured whales to justify our analyses of these data in the preceding problems?

# A SUMMARY OF INFERENCE METHODS

## Objectives

In this chapter we summarize inference methods presented throughout the text. We will

- show the process of choosing an appropriate inference technique from among those presented in earlier chapters.
- consider several examples of choosing an inference method.

### 13.1 Introduction

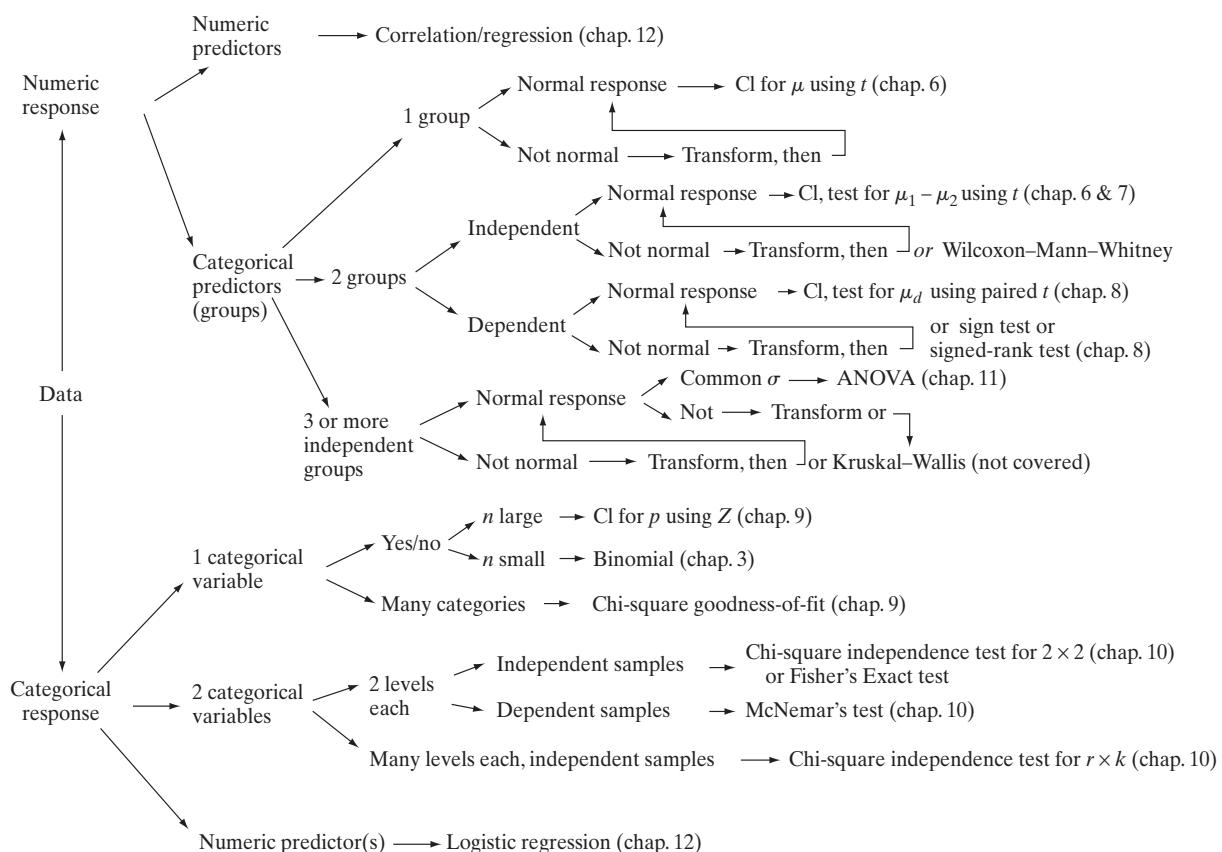
In Chapters 2 and 6 through 12 we introduced many statistical methods for visually and numerically summarizing data and for making inferences. Statistics students are often overwhelmed by the number and variety of procedures that have been presented. What a statistician sees as a clearly arranged set of tools for analyzing data can appear as a blur to the novice. In this chapter we present a variety of examples that demonstrate the analysis process from exploration and summary to inference using some of the methods presented in earlier chapters. With the examples, we also provide some guidelines that are useful in deciding how to make an inference from a given set of data.

When presented with a set of data, it is useful to ask a series of questions:

1. *What question were the researchers attempting to answer when they collected these data?* Data analysis is done for a purpose: to extract information and to aid decision making. When looking at data, it helps to bear in mind the purpose for which the data were collected. For example, were the researchers trying to compare groups, perhaps patients given a new drug and patients given a placebo? Were they trying to see how two quantitative variables are related, so that they can use one variable to make predictions of the other? Were they checking whether a hypothesized model gives accurate predictions of the probabilities associated with a categorical variable? A good understanding of why the data were collected often clarifies the next question:
2. *What is the response variable in the study?* For example, if the researchers were concerned with the effect of a medication on blood pressure, then the likely response variable is  $Y = \text{change in blood pressure of an individual}$  (a continuous numeric variable). If they were concerned with whether or not a medication cures an illness, then the response variable is categorical with two levels: yes if a person is cured, no if a person is not cured, or maybe even categorical with three or more ordered levels: fully cured, improved, no improvement.

3. *What predictor variables, if any, were involved?* For example, if a new drug is being compared to a placebo, then the predictor variable is group membership: A patient is either in the group that gets the new drug or else the patient is in the placebo group. If height is used to predict weight, then height is the predictor (and weight is the response variable). Sometimes there is no predictor variable. For example, a researcher might be interested in the distribution of cholesterol levels in adults. In this case, the response variable is cholesterol level, but there is no predictor variable. (One might argue that there is a predictor: whether or not someone is an adult. If we wished to compare cholesterol levels of adults to those of children, then whether or not someone is an adult would be a predictor. But if there is no comparison to be made, so that everyone in the study is part of the same group (adults), then it is not accurate to speak of a predictor *variable*, since group membership does not vary from person to person.)

The answers to these questions help frame the analysis to be conducted. Sometimes the analysis will be entirely descriptive and will not include any statistical inference, such as when the data are not collected by way of a random sample. Even when a statistical inference is called for, there is generally more than one way to proceed. Two statisticians analyzing the same set of data may use somewhat different methods and draw different conclusions. However, there are commonly used statistical procedures in various situations. The flowchart given in Figure 13.1.1 helps to organize the inference methods that have been presented in this book.



**Figure 13.1.1** A flowchart of inference methods

To use this flowchart, we start by asking whether the response variable is quantitative or categorical. We then consider the type of predictor variables in the study and whether the samples collected are independent of one another or are dependent (e.g., matched pairs). Many of the methods, such as the confidence interval for a population mean presented in Chapter 6, depend on the data being from a population that has a normal distribution. (This condition is less important for large samples than it is for small samples, due to the Central Limit Theorem.) Nonnormal data can often be transformed to approximate normality and normal-based methods then applied. If such transformation fails to achieve approximate normality, then nonparametric methods, such as the Wilcoxon-Mann-Whitney test or the Wilcoxon Signed-Rank test, can be used.

Note that the flowchart only directs attention to the collection of inference methods presented in the previous chapters; this is not an exhaustive list. Beware of the Mark Twain fallacy: “When your only tool is a hammer, every problem looks like a nail.” Not every statistical inference problem can be addressed with the methods presented here. In particular, these methods center on consideration of parameters, such as a population mean,  $\mu$ , or proportion,  $p$ . Sometimes researchers are interested in other aspects of distributions, such as the 75th percentile. When in doubt about how to proceed in an analysis, consult a statistician.

## Exploratory Data Analysis

No matter what type of analysis is being considered, it is always a good idea to start by making one or more graphs of the data. The choice of graphics depends on the type of data being analyzed. For example, when comparing two samples of quantitative data, side-by-side dotplots or boxplots are informative—both as a visual comparison of the two samples and for assessing whether or not the data satisfy the normality condition. When analyzing categorical data, bar charts are useful. When dealing with two quantitative variables, scatterplots are helpful.

Bear in mind that a statistical analysis is intended to help us understand the scientific problem at hand. Thus, conclusions should be stated in the context of the scientific study. In Section 13.2 we present some examples of data sets and the kinds of analyses that might be performed on them.

## 13.2 Data Analysis Examples

In this section we consider several data sets and the kinds of analyses that are appropriate for each. The three questions stated in Section 13.1 and the flowchart given in Figure 13.1.1 provide a framework for the discussion of the following examples.

---

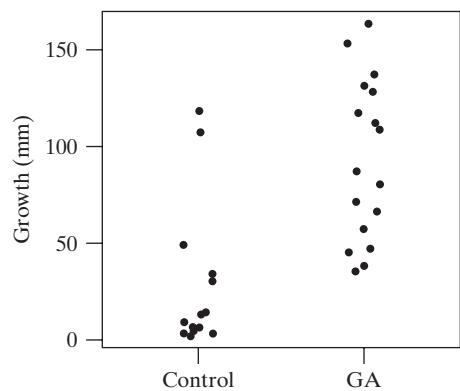
### Example 13.2.1

**Gibberellic Acid** Gibberellic acid (GA) is thought to elongate the stems of plants. Researchers conducted an experiment to investigate the effect of GA on a mutant strain of the genus *Brassica* called *ros*. They applied GA to 17 plants and applied water to 15 control plants. After 14 days they measured the growth of each of the 32 plants. For the 15 control plants the average growth was 26.7 mm, with an SD of 37.5 mm. For the 17 plants treated with GA the average growth was 92.6 mm, with an SD of 41.7 mm. The data are given in Table 13.2.1 and are graphed in Figure 13.2.1.<sup>1</sup>

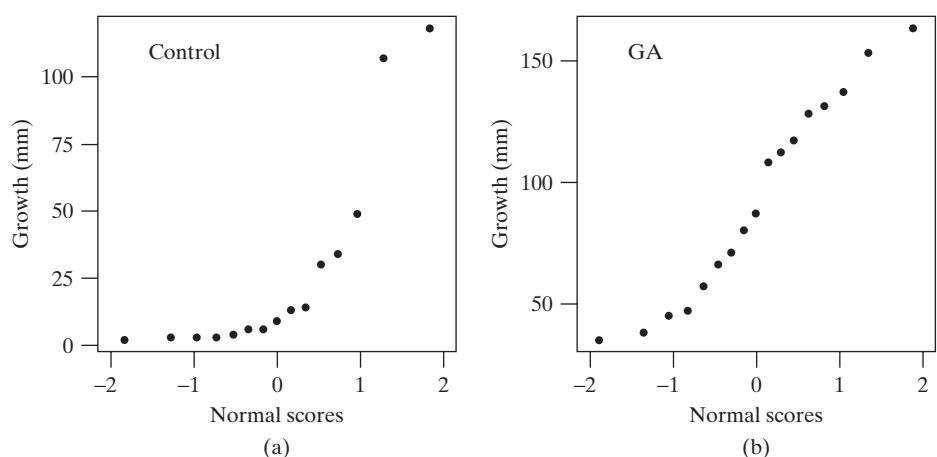
Let us turn to the three questions stated in Section 13.1. (1) In this experiment, the researchers were trying to establish whether GA affects the growth rate of *ros*; (2) the response variable is 14-day growth of *ros*, which is numeric; (3) the predictor variable is group membership (GA group or control group) and is categorical; the two groups are independent of one another.

**Table 13.2.1** Growth of *ros* plants (mm) after 14 days

Control	GA
3	71
2	87
34	117
13	80
6	112
118	66
14	128
107	153
30	131
9	45
3	38
3	137
49	57
4	163
6	47
	108
	35
Mean	26.7
SD	37.5
Mean	92.6
SD	41.7

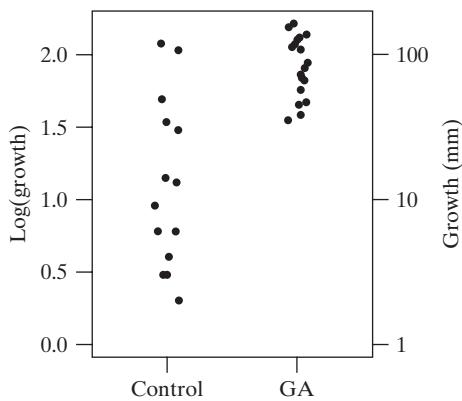
**Figure 13.2.1** Dotplots of growth of *ros* plants (mm) after 14 days

The flowchart in Figure 13.1.1 directs us to consider a two-sample *t* test, if the data are normal or can be transformed to normality, or a Wilcoxon-Mann-Whitney test. Figure 13.2.2 shows that the distribution of the control sample of data is markedly nonnormal; thus, a transformation is called for.

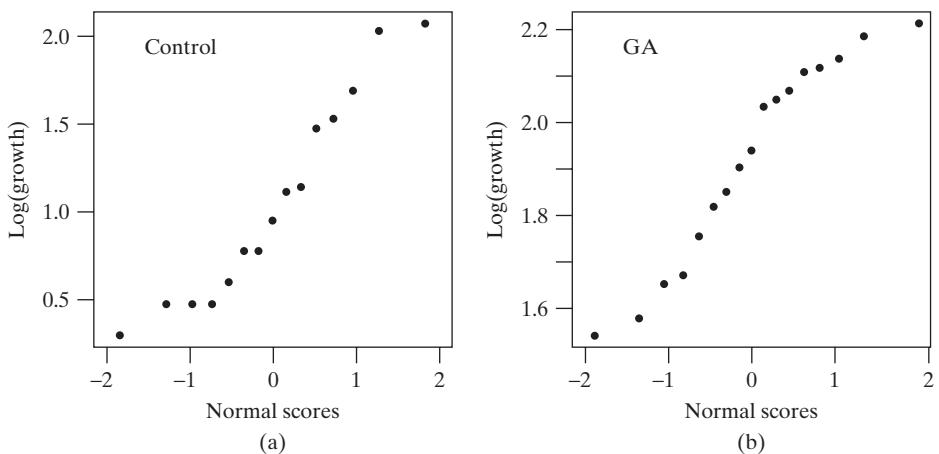
**Figure 13.2.2** Normal probability plots of (a) control data and (b) GA data

Taking logarithms of each of the observations produces the dotplots and normal probability plots in Figures 13.2.3 and 13.2.4.

**Figure 13.2.3** Dotplots of  $\log(\text{growth})$  of *ros* plants (mm) after 14 days



**Figure 13.2.4** Normal probability plots of (a) control data and (b) GA data in log scale



In log scale the data do not show marked evidence of abnormality (Shapiro-Wilk  $P$ -values for Control and GA are 0.2083 and 0.2296, respectively), so we can proceed with a two-sample  $t$  test. The standard deviations of the two samples are clearly quite different, as can be seen from Figure 13.2.3. However, an unpooled  $t$  test is still appropriate. The following computer output shows that  $t_s = -5.392$  and the  $P$ -value is very small. Thus, we have strong evidence that GA increases growth of *ros*. ■

#### Two Sample t-test

```
data: log10(Growth)
t = -5.3917, df = 17.445, p-value < 0.0001
alt. hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.1943596, -0.5234687
```

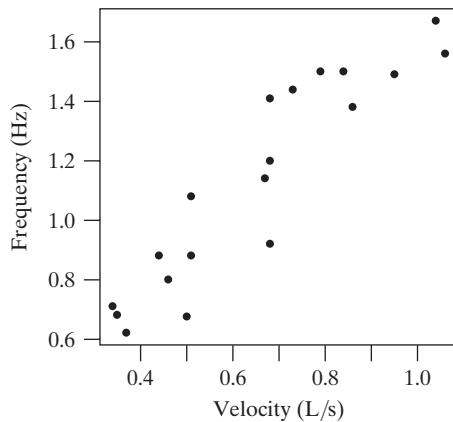
**Example  
13.2.2**

**Whale Swimming Speed** A biologist was interested in the relationship between the velocity at which a beluga whale swims and the tail-beat frequency of the whale. A sample of 19 whales was studied and measurements were made on swimming velocity, measured in units of body lengths of the whale per second (so that a value of 1.0 means that the whale is moving forward by one body length, L, per second) and tail-beat frequency, measured in units of hertz (so that a value of 1.0 means one tail-beat cycle per second).<sup>2</sup> Here are the data:

WHALE	VELOCITY (L/sec)	FREQUENCY (Hz)	WHALE	VELOCITY (L/sec)	FREQUENCY (Hz)
1	0.37	0.62	11	0.68	1.20
2	0.50	0.675	12	0.86	1.38
3	0.35	0.68	13	0.68	1.41
4	0.34	0.71	14	0.73	1.44
5	0.46	0.80	15	0.95	1.49
6	0.44	0.88	16	0.79	1.50
7	0.51	0.88	17	0.84	1.50
8	0.68	0.92	18	1.06	1.56
9	0.51	1.08	19	1.04	1.67
10	0.67	1.14			

It would be natural to ask, “When tails beats faster, do whales travel faster?” but the biologist conducting the study focused on the related question, “Does tail-beat frequency depend on velocity?” For the biologist’s question, the response variable, frequency, is numeric, and the predictor is velocity, which is also numeric. Thus, we can consider using regression analysis to study the relationship between velocity and frequency. Figure 13.2.5 is a scatterplot of the data, which shows an increasing trend in frequency as velocity increases.

**Figure 13.2.5** Scatterplot of frequency versus velocity

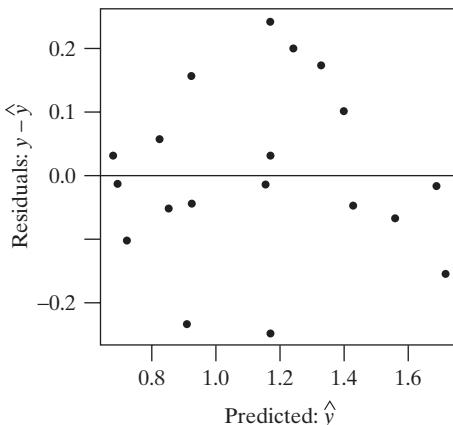


A regression model for these data is  $Y = \beta_0 + \beta_1 X + \varepsilon$ . Fitting the model to the data gives the equation  $\hat{y} = 0.19 + 1.439x$ , or Frequency = 0.19 + 1.439 × Velocity, as shown in the following computer output. Figure 13.2.6 shows the residual plot for this fit. The fact that this plot does not have any clear patterns in it supports the use of the regression model.

## Coefficients:

	Estimate	Std. Error	t value	Pr (>  t )
(Intercept)	0.1895	0.1004	1.887	0.0763
Velocity	1.4393	0.1451	9.917	1.75e-08
Residual standard error: 0.1396 on 17 degrees of freedom				
R-squared: 0.8526				

**Figure 13.2.6** Residual plot for frequency regression fit



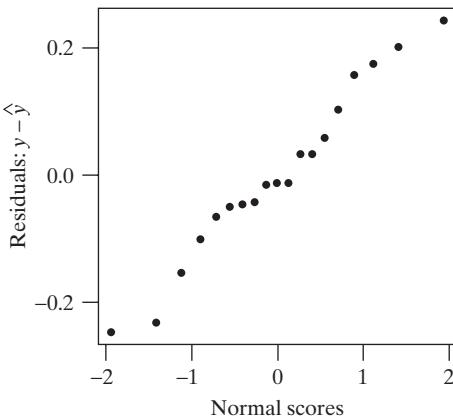
The null hypothesis

$$H_0: \beta_1 = 0$$

is tested with a  $t$  test, as shown in the regression output. A normal probability plot of the residuals, given in Figure 13.2.7, supports the use of the  $t$  test here, since it indicates that the distribution of the 19 residuals is consistent with what we would expect to see if the random errors came from a normal distribution. The  $t$  statistic has 17 degrees of freedom and a  $P$ -value of less than 0.0001. Thus, the evidence that frequency is related to velocity is quite strong; we reject the claim that the linear trend in the data arose by chance.

Continuing the analysis, the computer output shows that  $r^2$  is 85.3%. Thus, in the sample 85.3% of the variability in frequency is accounted for by variability in velocity. (This is significantly different from zero, as indicated with the  $t$  test for  $H_0: \beta_1 = 0$ .) ■

**Figure 13.2.7** Normal probability plot of residuals for frequency regression fit

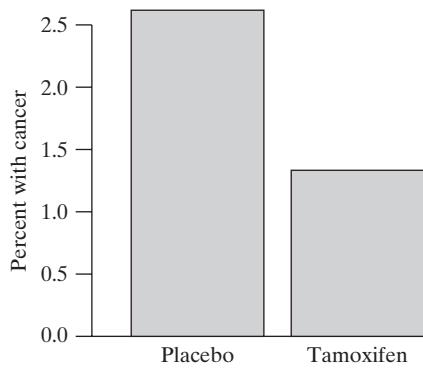


**Example  
13.2.3**

**Tamoxifen** In a randomized, double-blind experiment the drug tamoxifen was given to 6,681 women and a placebo was given to 6,707 other women. After four years there were 89 cases of breast cancer in the tamoxifen group, compared with 175 in the placebo group.<sup>3</sup>

The purpose of this experiment was to determine whether tamoxifen is effective in preventing cancer. Note that because this was an experiment, and not an observational study, we can talk in terms of a cause–effect relationship. The response variable is whether or not a woman developed cancer. The predictor variable is group membership (i.e., whether or not a woman was given tamoxifen). Figure 13.2.8 is a bar chart of the data, showing that cancer was much more common in the placebo group.

These data can be organized into a  $2 \times 2$  contingency table, such as Table 13.2.2. A chi-square test of independence yields  $\chi^2_s = 28.2$ . With 1 degree of freedom, the  $P$ -value for this test is nearly zero. There is very strong evidence that tamoxifen reduces the probability of breast cancer.



**Figure 13.2.8** Bar chart of the tamoxifen data

**Table 13.2.2** Tamoxifen data

	Treatment		Total
	Placebo	Tamoxifen	
Cancer	175	89	264
No cancer	6,532	6,592	13,124
Total	6,707	6,681	13,388

We can also construct a confidence interval with these data. Of placebo patients,

$\frac{175}{6707}$  or 2.61% developed cancer so that  $\tilde{p}_1 = \frac{175 + 1}{6707 + 2} = 0.0262$ . Of tamoxifen

patients,  $\frac{89}{6681}$  or 1.33% developed cancer so that  $\tilde{p}_2 = \frac{89 + 1}{6681 + 2} = 0.0135$ . The standard error of the difference is

$$\text{SE}_{(\tilde{p}_1 - \tilde{p}_2)} = \sqrt{\frac{(0.0262)(1 - 0.0262)}{6707 + 2} + \frac{0.0135(1 - 0.0135)}{6681 + 2}} \\ = 0.0024$$

A 95% confidence interval for  $p_1 - p_2$  is  $(0.0262 - 0.0135) \pm 1.96(0.0024)$  or  $(0.0080, 0.0174)$ . Thus, we are 95% confident that tamoxifen reduces the probability of breast cancer by between 0.80 and 1.74 percentage points.

We can also calculate the relative risk of cancer. The estimated relative risk is

$$\frac{\Pr[\text{Cancer} | \text{Tamoxifen}]}{\Pr[\text{Cancer} | \text{Placebo}]} = \frac{0.0261}{0.0133} = 1.96$$

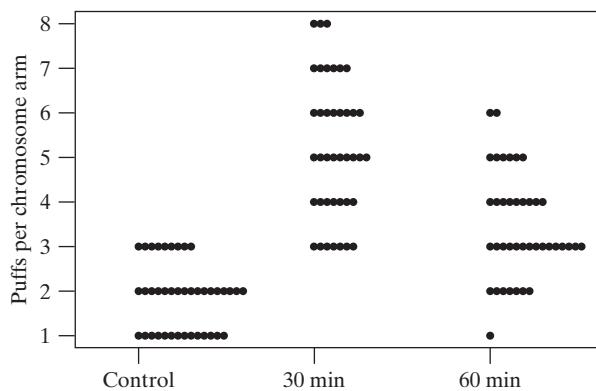
Thus, we estimate that breast cancer is 1.96 times as likely when taking placebo as when taking tamoxifen. ■

**Example  
13.2.4**

**Chromosome Puffs** Heat shock proteins (HSPs) are a type of protein produced by some organisms as protection against damage from exposure to high temperature. In the fruit fly *Drosophila melanogaster* the genes that encode HSPs are found on

chromosomes that uncoil and appear to puff out. This chromosome puffing can be seen under a microscope. A biologist counted the number of puffs per chromosomal arm from the salivary glands of 40 *Drosophila* larvae that had been heat shocked at 37 °C for 30 minutes, 40 larvae that had been heat shocked for 60 minutes, and 40 control larvae.

The purpose of this experiment was to determine the effect, if any, of heat shock on the HSPs. The response variable is the number of puffs on a chromosome arm, which is numeric. The predictor variable, group membership (control, 30 minutes, or 60 minutes), is categorical. Dotplots of the data are given in Figure 13.2.9; the data are summarized in Table 13.2.3.<sup>4</sup>



**Figure 13.2.9** Dotplots of puffs per chromosome arm for *Drosophila* heat shock experiment

Table 13.2.3 Puffs per chromosome arm for <i>Drosophila</i> heat shock experiment			
Group	n	Mean	SD
Control	40	1.88	0.76
30 min.	40	5.20	1.54
60 min.	40	3.45	1.18

The dotplots suggest an effect due to heat shock (and we can speak of an effect, not just an association, because this was an experiment). This visual impression can be confirmed with an analysis of variance. The plots also show that the distributions take on only a few values each, so that the normality condition for ANOVA is not met. Since, however, the distributions appear to be reasonably symmetric, the sample sizes moderately large and equal, and the SDs are similar among the groups, we can have confidence in the ANOVA *P*-value. The following ANOVA computer output confirms that there is strong evidence against  $H_0: \mu_1 = \mu_2 = \mu_3$ . We conclude that heat shock does, indeed, increase the number of puffs per chromosome arm.

	Df	Sum Sq	Mean Sq	F value	Pr (> F)
Group	2	221.32	110.658	76.757	< 0.0001
Residuals	117	168.68	1.442		
Total	110	390.00			

As an extension of the ANOVA, we could consider a contrast that compares the control mean to the average of the two heat shock means. ■

### Example 13.2.5

**Therapeutic Touch** Therapeutic touch (TT) is a form of alternative medicine in which a practitioner manipulates the human energy field of the patient. However, many persons have questioned the ability of TT practitioners to detect the human energy field—and whether the human energy field even exists. An experimenter tested

the abilities of 28 TT practitioners as follows. A screen was set up between the experimenter and the practitioner, who sat on opposite sides of a table. The practitioner extended his or her hands under the screen and rested them, palms up, on the table. The researcher tossed a coin to choose one of the practitioner's hands. The experimenter then held her right hand, palm down, above the chosen hand of the practitioner. The practitioner was then asked to identify which hand had been chosen, as a test of whether the practitioner could detect a human energy field extending from the hand of the experimenter.

Each of the 28 TT practitioners was tested 10 times. The number of correct detections, "hits," in 10 trials varied from 1 to 8, with an average of 4.4. There were 123 hits in the 280 total trials.<sup>5</sup> Table 13.2.4 shows the distribution of hits among the 28 tested practitioners.

**Table 13.2.4** Distribution of hits per ten trials in therapeutic touch experiment

Number of hits	Number of practitioners
0	0
1	1
2	1
3	8
4	5
5	7
6	2
7	3
8	1
9	0
10	0
Total	28

The goal of this experiment was to determine the ability of TT practitioners to detect the human energy field. The response variable is a yes/no (categorical) variable: yes for a hit and no for a miss. There is no predictor variable here, there is just a single group of 28 TT practitioners who were tested.

Let  $p$  denote the probability of a hit in one of the trials of the experiment. The natural null hypothesis is  $H_0: p = 0.5$ . One way to analyze the data would be to conduct a chi-square goodness-of-fit test of  $H_0$  using the 280 total trials, with a directional alternative  $H_A: p > 0.5$ . The  $P$ -value for this test is greater than 0.50, since the data do not deviate from  $H_0$  in the direction specified by  $H_A$ .

One might argue that  $p$  might be greater than 0.5 for some TT practitioners, but perhaps not for all of them. If  $p$  is not the same for each TT practitioner (whether or not  $p$  is 0.5 for anyone), then the chi-square goodness-of-fit test using all 280 trials is not appropriate, since the 280 trials are not independent of one another. However, the data for each of the 28 practitioners could be analyzed separately. A binomial model could be used in these analyses, since the sample size of  $n = 10$  is rather small. The binomial probabilities are given in Table 13.2.5. The probability of 8 or more hits in 10 trials, for a binomial with  $p = 0.5$ , is  $0.04395 + 0.00977 + 0.00098 = 0.0547$ . Thus, if the data from each of the 28 practitioners were analyzed separately, testing  $H_0: p = 0.5$  versus  $H_A: p > 0.5$ , the

**Table 13.2.5** Observed and expected numbers (if  $p = 0.5$ ) of hits per ten trials in the therapeutic touch experiment

Number of hits	Binomial probability	Observed number, O	Expected number, E
0	0.00098	0	0.027
1	0.00977	1	0.273
2	0.04395	1	1.231
3	0.11719	8	3.281
4	0.20508	5	5.742
5	0.24609	7	6.891
6	0.20508	2	5.742
7	0.11710	3	3.281
8	0.04395	1	1.231
9	0.00977	0	0.273
10	0.00098	0	0.027
Total	1.00000	28	27.999

smallest of the 28  $P$ -values would be 0.0547, and again provides no significant evidence in support of  $H_A$ .\*

A different way to conduct the analysis is to investigate whether the 280 observations presented in Table 13.2.4 are consistent with a binomial model. In particular, we can check the model that states that  $Y$  has a binomial distribution, with  $n = 10$  and  $p = 0.5$ , where  $Y$  is the number of hits in 10 trials. (This is similar to the analysis presented in Section 3.9.) A goodness-of-fit test can be used here. Table 13.2.5 shows the observed numbers (from Table 13.2.4) and expected numbers for each of the 11 possible outcomes. (The expected numbers don't sum to 28 due to round-off error.)

The chi-square statistic is  $\chi_s^2 = \sum \frac{(o_i - e_i)^2}{e_i} = 11.7$ . The test statistic has 10 degrees of freedom, since there are 11 categories in the model. The  $P$ -value for this test is 0.306, which is quite large. Thus, the data are consistent with a binomial distribution for which  $p = 0.5$  (i.e., the TT practitioners might as well have tossed coins to choose a hand, rather than trying to detect the human energy field of the experimenter). (Note: These data do not disprove the existence of the human energy field; they only fail to provide evidence for its existence). ■

## Brief Examples

We will now consider some examples for which we will identify the type of analysis that is appropriate, but we won't conduct the analysis.

### Example 13.2.6

**Seastars** Researchers measured the length of the longest ray on each of over 200 members of the species *Phataria unifascialis* (a seastar found in the waters of the Gulf of California, Mexico). For a sample of 184 individuals found near Loreto, the

\*Considering the material in optional Section 11.9 on multiple comparisons, note that if we were to consider all 28 of these tests, we ought to require a great deal of evidence before rejecting  $H_0$ . Using the Bonferroni correction, we would require that an individual  $P$ -value be less than  $\alpha_{cw} = 0.05/28 = 0.0018$  before rejecting  $H_0$ .

average length was 6.78 cm, with an SD of 1.21 cm. For a sample of 77 individuals found near Bahia de Los Angeles, the average length was 8.13 cm, with an SD of 1.33 cm.<sup>6</sup>

The response variable is numeric and there are two independent groups. Thus, a two-sample *t* test is appropriate, along with a confidence interval for the difference in population means. (*Note:* The normality condition is not essential, since the sample sizes are quite large.) ■

---

**Example  
13.2.7**

**Twins** Researchers in Finland studied the physical activity levels of hundreds of sets of same-sex twins. In 1975 they classified subjects into the physical activity categories “exerciser” and “sedentary.” They kept track of the health of the subjects through 1994, by which time there were several pairs of twins for whom one twin was alive, but the other had died. In this group there were 49 “sedentary” twins who were living, but whose “exerciser” twin pair was dead. There were 76 “exerciser” twins who were living, but whose “sedentary” twin pair was dead.<sup>7</sup>

The response variable in this observational study is whether or not a subject is alive, a categorical variable. The predictor is also categorical: whether the person is “sedentary” or is an “exerciser.” Since the data are paired, McNemar’s test is appropriate. ■

---

**Example  
13.2.8**

**Soil Samples** Researchers took eight soil samples at each of six locations in Mediterranean pastures. They divided the samples into four pairs and put the soil in pots. One pot from each pair was watered continuously, while the other pot was watered for 13 days, then not watered for 18 days, and then watered again for 30 days. The researchers recorded the number of germinations in each pot during the experiment.<sup>8</sup>

This example is similar to Example 13.2.6, in that there are two samples to be compared and the response variable is numeric. However, the samples here are paired, so a paired analysis (Chapter 8) is called for. If the 24 sample differences show a normal distribution, then a paired *t* test or confidence interval could be used; if not, a transformation could be tried, or a Wilcoxon signed-rank or sign test could be used. ■

---

**Example  
13.2.9**

**Vaccinations** In 1996 there was an outbreak of the disease varicella in a child care center in Georgia. Some of the children had been vaccinated against varicella, but others had not. Varicella occurred in 9 out of 66 vaccinated children and in 72 out of 82 unvaccinated children.<sup>9</sup>

The response and predictor variables in this experiment are both categorical. The data could be arranged into a  $2 \times 2$  contingency table and analyzed with a chi-square test of independence. The difference in sample proportions is obviously quite large. However, this is an observational study and not an experiment. Thus, we cannot conclude that the difference in proportions is entirely due to the effect of the vaccine, since the effects of other variables, such as economic status, are confounded with the effect of the vaccine. ■

---

**Example  
13.2.10**

**Estrogen and Steroids** Plasma estrone plus estradiol (Plasma E<sub>1+2</sub>) steroid levels were measured in women given estrogen (Premarin) and in a control group of women. The women given estrogen were divided into three treatment groups. One group was given a daily dose of 0.625 mg, one group was given 1.25 mg, and the third group

was given 2.5 mg. The researchers noted that the plasma steroid levels were not normally distributed, but became so after a logarithm transformation was applied. In log scale, the data are given in Table 13.2.6.<sup>10</sup>

**Table 13.2.6** Log ng/100 ml plasma  $E_{1+2}$  concentration for estrogen study

Group	n	Mean	SD
Control	30	2.01	0.27
0.625	16	2.10	0.31
1.25	24	2.34	0.39
2.5	21	2.20	0.24

The response variable in this experiment, log(plasma  $E_{1+2}$  concentration), is numeric. It has already been transformed to normality. There are four independent groups to be compared, so an analysis of variance is appropriate. A contrast that compares the control to the average of the three treatment groups would also be useful. ■

**Example  
13.2.11**

**Damselflies** A researcher captured male damselflies and randomly assigned them to one of three groups. For those in the first group the sizes of red spots on the wing were artificially enlarged with red ink. For those in the second group the wing spots were enlarged with clear ink. The third group served as a control. The damselflies were then released into a contained area. The numbers surviving in each of the three groups 22 days later were determined. There were 312 damselflies in each of the three groups. After 22 days there were 41 survivors in the “artificially enlarged with red ink” group, 49 survivors in the “enlarged with clear ink” group, and 57 survivors in the control group.<sup>11</sup>

The response variable in this experiment, survival, is categorical, as is the predictor variable, ink status/type. These data could be arranged into a  $2 \times 3$  contingency table and analyzed with a chi-square test of independence. ■

**Example  
13.2.12**

**Tobacco Use Prevention** In the Hutchinson Smoking Prevention Project 40 school districts in the state of Washington were formed into 20 pairs on the basis of size, location, and prevalence of high school tobacco use as of the beginning of the study. In each pair, one district was randomly assigned to be in an intervention group and the other was assigned to the control group. If a school district was in the intervention group, then the third-grade students in the district were given a curriculum on preventing tobacco use and the teachers in the district were given special training to help students refrain from smoking. This was repeated one year later with the next new cohort of third-grade students. All the students were then followed for several years. A primary outcome measurement of the study was whether or not students were smoking two years after graduating from high school.

The experimental unit here is an entire school district, so it is natural to use as the response variable the percentage of students from a district who smoke, a numeric variable. The predictor is categorical: intervention group or control group. There are two groups, which are paired together by the design of the experiment. Out of the 20 pairs, there were 13 pairs in which the smoking rate was higher in the control district and 7 pairs in which the smoking rate was higher in the intervention district.<sup>12</sup> A sign test could be used to analyze these data. ■

## Exercises 13.2.1–13.2.22

**13.2.1** Researchers conducted a randomized, double-blind, clinical trial in which some patients with schizophrenia were given the drug clozapine and others were given haloperidol. After one year 61 of 163 patients in the clozapine group showed clinically important improvement in symptoms, compared with 51 out of 159 in the haloperidol group.<sup>13</sup> Identify the type of statistical method that is appropriate for these data, but do not actually conduct the analysis.

**13.2.2** Consider the data of Exercise 13.2.1. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

**13.2.3** A biologist collected data on the height (in inches) and peak expiratory flow (PEF—a measure of how much air a person can expire, measured in l/min) for 10 women.<sup>14</sup> Here are the data:

SUBJECT	HEIGHT	PEF	SUBJECT	HEIGHT	PEF
1	63	410	6	62	360
2	63	440	7	67	380
3	66	450	8	64	380
4	65	510	9	65	360
5	64	340	10	67	570

Is PEF related to height? Identify the type of statistical method that is appropriate for these data and this question, but do not actually conduct the analysis.

**13.2.4** Consider the data of Exercise 13.2.3. Maria is 1 inch taller than Anika. Using the information from Exercise 13.2.3, how much greater would you predict Maria's PEF to be than Anika's?

**13.2.5** A geneticist self-pollinated pink-flowered snapdragon plants and produced 97 progeny with the following colors: 22 red plants, 52 pink plants, and 23 white plants.<sup>15</sup> The purpose of this experiment was to investigate a genetic model that states that the probabilities of red, pink, and white are 0.25, 0.50, and 0.25. Identify the type of statistical method that is appropriate for these data, but do not actually conduct the analysis.

**13.2.6** Consider the data of Exercise 13.2.5. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

**13.2.7** The effect of diet on heart disease has been widely studied. As part of this general area of investigation, researchers were interested in the short-term effect of diet on endothelial function, such as the effect on triglyceride

level. To study this, they designed an experiment in which 20 healthy subjects were given, in random order, a high-fat breakfast and a low-fat breakfast at 8 A.M., following a 12-hour fast, on days one week apart from each other. Serum triglyceride levels were measured on each subject before each breakfast and again four hours after each breakfast.<sup>16</sup> If you had access to all of the measurements collected in this experiment, how would you analyze the data?

**13.2.8** Biologists were interested in the distribution of trees in a wooded area. They intended to use the number of trees per 100-square meter plot as their unit of measurement. However, they were concerned that the shapes of the plots might affect the data collection. To investigate the possibility, they counted the numbers of trees in square plots, round plots, and rectangular plots. The data are shown in the following table.<sup>17</sup> What type of analysis is appropriate for these data?

PLOT SHAPE		
SQUARE	ROUND	RECTANGULAR
5	5	10
5	7	2
5	5	3
8	2	12
8	4	9
7	4	5
4	4	3
9	7	6
9	7	5
7	10	3
5	9	8
2	2	9
8	7	3
Mean	6.3	5.6
SD	2.14	2.47
		3.27

**13.2.9** Consider the data of Exercise 13.2.8. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

**13.2.10** A sample of 15 patients was randomly split into two groups as part of a double-blind experiment to compare two pain relievers.<sup>18</sup> The 7 patients in the first group were given Demerol and reported the following numbers of hours of pain relief:

2, 6, 4, 13, 5, 8, 4

The 8 patients in the second group were given an experimental drug and reported the following numbers of hours of pain relief.

0, 8, 1, 4, 2, 2, 1, 3

How might these data be analyzed?

**13.2.11** Consider the data of Exercise 13.2.10. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

**13.2.12** A researcher was interested in the relationship between forearm length and height. He measured the forearm lengths and heights of a sample of 16 women and obtained the following data.<sup>19</sup> How might these data be (i) visualized and (ii) analyzed?

FOREARM		FOREARM	
HEIGHT (CM)	LENGTH (CM)	HEIGHT (CM)	LENGTH (CM)
163	25.5	157	26
161	26	178	27
151	25	163	24.5
163	25	161	26
166	27.2	173	28
168	26	160	24.5
170	26	158	25
163	26	170	26

**13.2.13** A randomized, double-blind, clinical trial was conducted on patients who had coronary angioplasty to compare the drug lovastatin to a placebo. The percentage of stenosis (narrowing of the blood vessels) following angioplasty was measured on 160 patients given lovastatin and on 161 patients given the placebo. For the lovastatin group the average was 46%, with an SD of 20%. For the placebo group the average was 44%, with an SD of 21%.<sup>20</sup> What type of analysis is appropriate for these data?

**13.2.14** Consider the data of Exercise 13.2.13.

- (a) Conduct an appropriate analysis of the data.
- (b) Describe a graphical procedure to visualize these data.
- (c) Discuss of how the data likely meet the necessary conditions for validity even though you do not have access to the raw data.

**13.2.15** Researchers studied persons who had received intravenous immune globulin (IGIV) to see if they had developed infections of hepatitis C virus (HCV). In part of their analysis, they considered doses of Gammagard (an IGIV product) received by 210 patients. They divided the patients into 4 groups according to the number of

doses of “Gammagard made from unscreened or first-generation anti-HCV-screened plasma.” Among 48 persons who received 0 to 3 doses, there were 4 cases of HCV infection. There were 2 cases of HCV infection among 45 persons who received 4 to 20 doses, there were 7 cases of HCV infection in the 57 persons who received between 21 and 65 doses, and there were 10 cases of HCV infection among the 51 persons who received more than 65 doses.<sup>21</sup> What type of analysis is appropriate for these data?

**13.2.16** Consider the data of Exercise 13.2.15. Conduct an appropriate analysis of the data.

**13.2.17** An experiment was conducted to study the effect of tamoxifen on patients with cervical cancer. One of the measurements made, both before and again after tamoxifen was given, was microvessel density (MVD). MVD, which is measured as number of vessels per mm<sup>2</sup>, is a measurement that relates to the formation of blood vessels that feed a tumor and allow it to grow and spread. Thus, small values of MVD are better than are large values. Data for 18 patients are shown.<sup>22</sup> How might these data be analyzed?

PATIENT	MVD	MVD	PATIENT	MVD	MVD
	BEFORE	AFTER		BEFORE	AFTER
1	98	75	10	70	60
2	100	60	11	60	65
3	82	25	12	88	45
4	100	55	13	45	36
5	93	78	14	159	144
6	119	102	15	65	27
7	70	58	16	98	90
8	78	70	17	66	16
9	104	90	18	67	53

**13.2.18** Consider the data of Exercise 13.2.17. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

**13.2.19** As part of a large experiment, researchers planted 2,400 sweetgum, 2,400 sycamore, and 1,200 green ash seedlings. After 18 years, the survival rates were 93% for the sweetgum trees, 88% for the sycamore trees, and 95% for the green ash trees.<sup>23</sup> What type of analysis is appropriate for these data?

**13.2.20** Consider the data of Exercise 13.2.19. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

**13.2.21** A group of female college students were divided into three groups according to upper body strength. Their leg strength was tested by measuring how many consecutive times they could leg press 246 pounds before exhaustion. (The subjects were allowed only one second of rest between consecutive lifts.) The data are shown in the following table.<sup>24</sup> What type of analysis is appropriate for these data?

UPPER BODY STRENGTH GROUP		
LOW	MIDDLE	HIGH
55	40	181
70	200	85
45	250	416
246	192	228
240	117	257
96	215	316
225		134
Mean	140	169
SD	93	77
		112

**13.2.22** Consider the data of Exercise 13.2.21. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

# CHAPTER APPENDICES

---

## Appendix

- |                                                                                                             |                                                              |
|-------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| <b>3.1</b> More on the Binomial Distribution<br>Formula 566                                                 | <b>6.1</b> Significant Digits 573                            |
| <b>3.2</b> Mean and Standard Deviation of the Binomial Distribution 569                                     | <b>7.1</b> How Power Is Calculated 574                       |
| <b>4.1</b> Areas of Indefinitely Extended Regions 570                                                       | <b>7.2</b> More on the Wilcoxon-Mann-Whitney Test 576        |
| <b>5.1</b> Relationship Between Central Limit Theorem and Normal Approximation to Binomial Distribution 572 | <b>9.1</b> More on Confidence Intervals for a Proportion 578 |
|                                                                                                             | <b>12.1</b> Least-Squares Formulas 580                       |
|                                                                                                             | <b>12.2</b> Derivation of Fact 12.3.1 582                    |

## Appendix 3.1

### More on the Binomial Distribution Formula

In this appendix we explain more about the reasoning behind the binomial distribution formula.

#### The Binomial Distribution Formula

We begin by deriving the binomial distribution formula for  $n = 3$ . Suppose that we conduct three independent trials and that each trial results in success (S) or failure (F). Further, suppose that on each trial the probabilities of success and failure are

$$\Pr\{S\} = p$$

$$\Pr\{F\} = 1 - p$$

There are eight possible outcomes of the three trials. Reasoning as in Example 3.6.3 shows that the probabilities of these outcomes are as follows:

OUTCOME	NUMBER OF SUCCESSES	NUMBER OF FAILURES	PROBABILITY
FFF	0	3	$(1 - p)^3$
FFS	1	2	$p(1 - p)^2$
FSF	1	2	$p(1 - p)^2$
SFF	1	2	$p(1 - p)^2$
FSS	2	1	$p^2(1 - p)$
SFS	2	1	$p^2(1 - p)$
SSF	2	1	$p^2(1 - p)$
SSS	3	0	$p^3$

Again by reasoning parallel to Example 3.6.3, these probabilities can be combined to obtain the binomial distribution formula for  $n = 3$  as shown in the table:

NUMBER OF		
SUCCESES, $j$	FAILURES, $n - j$	PROBABILITY
0	3	$1p^0(1 - p)^3$
1	2	$3p^1(1 - p)^2$
2	1	$3p^2(1 - p)^1$
3	0	$1p^3(1 - p)^0$

This distribution illustrates the origin of the binomial coefficients. The coefficient  ${}_3C_1 (= 3)$  is the number of ways in which 2 S's and 1 F can be arranged; the coefficient  ${}_3C_2 (= 3)$  is the number of ways in which 1 S and 2 F's can be arranged.

An argument similar to this shows that the general formula (for any  $n$ ) is

$$\Pr\{j \text{ successes and } n - j \text{ failures}\} = {}_nC_j p^j (1 - p)^{n-j}$$

where

${}_nC_j$  = the number of ways in which  $j$  S's and  $(n - j)$  F's can be arranged.

**Combinations** The binomial coefficient  ${}_nC_j$  is also known as the number of combinations of  $n$  items taken  $j$  at a time; it is equal to the number of different subsets of size  $j$  that can be formed from a set of  $n$  items.

## The Binomial Coefficients: A Formula

Binomial coefficients can be calculated from the formula

$${}_nC_j = \frac{n!}{j!(n - j)!}$$

where  $x!$  (“ $x$ -factorial”) is defined for any positive integer  $x$  by

$$x! = x(x - 1)(x - 2) \cdots (2)(1)$$

and  $0! = 1$ .

For example, for  $n = 7$  and  $j = 4$  the formula gives

$$\begin{aligned} {}_7C_4 &= \frac{7!}{4!3!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(3 \times 2 \times 1)} \\ &= 35 \end{aligned}$$

To see why this is correct, let us consider in detail why the number of ways of rearranging 4 S's and 3 F's should be equal to

$$\frac{7!}{4!3!}$$

Suppose 4 S's and 3 F's were written on cards, like this:

$$\boxed{S_1} \quad \boxed{S_2} \quad \boxed{S_3} \quad \boxed{S_4} \quad \boxed{F_1} \quad \boxed{F_2} \quad \boxed{F_3}$$

Temporarily we put subscripts on the S's and F's to distinguish them. First let us see how many ways there are to arrange the 7 cards in a row:

There are 7 choices for which card goes first;  
for each of these, there are 6 choices for which card goes second;  
for each of these, there are 5 choices for which card goes third;  
for each of these, there are 4 choices for which card goes fourth;  
for each of these, there are 3 choices for which card goes fifth;  
for each of these, there are 2 choices for which card goes sixth;  
for each of these, there is 1 choice for which card goes last.

It follows that there are  $7!$  ways of arranging the 7 cards. Now consider the locations of the 4 S's. There are  $4!$  ways in which the S's can be rearranged among themselves. Likewise, there are  $3!$  ways in which the F's can be rearranged among themselves. If we were to ignore the subscripts on the S's and F's, then some of the  $7!$  ways of arranging the 7 cards would be indistinguishable. Indeed, any rearrangement of the S's *among themselves* leaves the 7 card arrangement looking the same. Similarly, any rearrangement of the F's *among themselves* leaves the 7 card arrangement looking the same. Thus, the number of *distinguishable* arrangements is

$$\frac{7!}{4!3!}$$

## Appendix 3.2

### Mean and Standard Deviation of the Binomial Distribution

Suppose that  $Y$  is a binomial random variable with  $n$  trials and  $p$  as the probability of success on each trial. Then we can think of  $Y$  as the sum of  $n$  variables  $X_1, X_2, \dots, X_n$ , where each  $X_i$  is equal to either 0 or 1 (0 for a failure or 1 for a success). That is,  $Y = \sum X_i$ , with  $\Pr\{X_i = 0\} = 1 - p$  and  $\Pr\{X_i = 1\} = p$ . The  $n X_i$ 's are a random sample from a hypothetical population of  $X$ 's that has average  $\mu_X = p$  (since  $0 \times (1 - p) + 1 \times p = p$ ).

Now consider the population standard deviation,  $\sigma_X$ , for the population of  $X$ 's. Recall, from Section 2.8 that for a variable  $X$  the definition of  $\sigma$  is

$$\sigma = \sqrt{\text{Population average value of } (X - \mu)^2}$$

For the population of  $X$ 's, the mean is  $\mu_X = p$ . Thus, for this population,

$$\sigma_X = \sqrt{\text{Population average value of } (X - p)^2}.$$

In the population of  $X$ 's, the quantity  $(X - p)^2$  takes on only two possible values:

$$(X - p)^2 = \begin{cases} (0 - p)^2 & \text{if } X = 0 \\ (1 - p)^2 & \text{if } X = 1 \end{cases}$$

Furthermore, these values occur in the proportions  $(1 - p)$  and  $p$ , respectively, so that the population average value of  $(X - p)^2$  is equal to

$$(0 - p)^2 \times (1 - p) + (1 - p)^2 \times p$$

This can be simplified to

$$\begin{aligned} p^2 \times (1 - p) + (1 - p)^2 \times p &= p\{p(1 - p) + (1 - p)^2\} \\ &= p\{p - p^2 + 1 - 2p + p^2\} \\ &= p(1 - p) \end{aligned}$$

Hence, the population average value of  $(X - p)^2$  is  $p(1 - p)$ , so  $\sigma_X = \sqrt{p(1 - p)}$ .

The binomial random variable  $Y$  is  $\sum X_i$ . To find the mean and standard deviation of  $Y$ , we need two facts:

*Fact 1:* For any collection of random variables  $X_1, X_2, \dots, X_n$  the mean of  $\sum X_i = \sum(\text{mean of } X_i)$ .

*Fact 2:* For a collection of independent random variables  $X_1, X_2, \dots, X_n$  the variance of  $\sum X_i = \sum(\text{variance of } X_i)$ .

(Recall that the variance,  $\sigma^2$ , is the square of the standard deviation,  $\sigma$ .)

Using Fact 1, we see that the mean of  $Y$  is the mean of  $\sum X_i$ , which is  $\sum p$ . Thus, the mean of  $Y$  is  $\mu_Y = np$ .

Using Fact 2, the variance of  $Y$  is the variance of  $\sum X_i$ , which equals  $\sum(\text{Variance of } X_i)$  or  $np(1 - p)$ . Thus, the standard deviation of  $Y$  is  $\sigma_Y = \sqrt{np(1 - p)}$ .

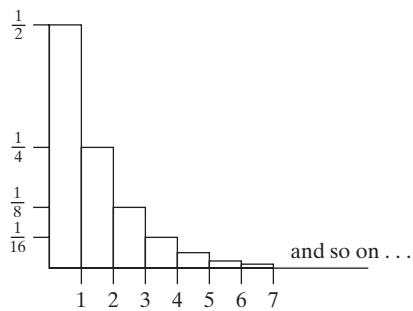
## Appendix 4.I

### Areas of Indefinitely Extended Regions

Consider the region bounded between a normal curve and the horizontal axis. Because the curve never touches the axis, the region extends indefinitely far to the left and to the right. Yet the area of the region is exactly equal to 1.0. How is it possible for an indefinitely extended region to have a finite area?

To gain insight into this paradoxical situation, consider Figure A.4.1, which shows a region that is simpler than that bounded by a normal curve. In this region, the width of each bar is 1.0; the height of the first bar is  $\frac{1}{2}$ , the second bar is half as high as the first, the third is half as high as the second, and so on. The bars form a region that is indefinitely extended. Nevertheless, we shall see that it makes sense to say that the area of the region is equal to 1.0.

**Figure A.4.1**



Let us first consider the areas of the individual bars. The area of the first bar is  $\frac{1}{2}$ , the area of the second bar is  $\frac{1}{4}$ , the third  $\frac{1}{8}$ , and so on. Now suppose that we choose a number, say  $k$ , and add up the areas of the first  $k$  bars, as follows:

BAR	HEIGHT OF BAR	CUMULATIVE TOTAL AREA
1	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{3}{4}$
3	$\frac{1}{8}$	$\frac{7}{8}$
4	$\frac{1}{16}$	$\frac{15}{16}$
⋮	⋮	⋮
$k$	$\frac{1}{2^k}$	$\frac{2^k - 1}{2^k}$

The total area of the first two bars is  $\frac{3}{4}$ , the total area of the first three bars is  $\frac{7}{8}$ , and so on. In fact, the total area of the first  $k$  bars is equal to

$$\frac{2^k - 1}{2^k} = 1 - \frac{1}{2^k}$$

If  $k$  is very large, this area is very close to 1.0. In fact, we can make the area as close to 1.0 as we wish, simply by choosing  $k$  large enough. In these circumstances it is reasonable to say that the total area of the entire, indefinitely extended region is equal to exactly 1.0.

The preceding example shows that an indefinitely extended region can have a finite area. Likewise, the total area under the normal curve is 1.0 (but the proof of this fact requires fairly advanced calculus).

## Appendix 5.1

### Relationship between Central Limit Theorem and Normal Approximation to Binomial Distribution

Consider sampling from a dichotomous population. Theorem 5.4.1 states that the sampling distribution of  $\hat{P}$ , and the equivalent binomial distribution, can be approximated by normal distributions. In this appendix we show how these approximations are related to Theorem 5.2.1 and the Central Limit Theorem.

As shown in Appendix 3.2, if  $Y$  is a binomial random variable with  $n$  trials and  $p$  as the probability of success on each trial, then we can think of  $Y$  as the sum of  $n$  variables  $X_1, X_2, \dots, X_n$ , where each  $X_i$  is equal to either 0 or 1 – 0 for a failure or 1 for a success. For a population of 0's and 1's, where the proportion of 1's is given by  $p$ , the mean is  $p$  and the standard deviation is  $\sigma = \sqrt{p(1 - p)}$ . The sample mean of  $X_1, X_2, \dots, X_n$  is  $\bar{X}$ , which is the same as the proportion of 1's in the sample (that is,  $\hat{P}$ ). Thus, the sample proportion  $\hat{P}$  can be regarded as a sample mean, and so its sampling distribution is described by Theorem 5.2.1.

From part 3 of Theorem 5.2.1 (the Central Limit Theorem), the sampling distribution of  $\hat{P}$  is approximately normal if  $n$  is large. From part 1 of Theorem 5.2.1, the mean of the sampling distribution of  $\hat{P}$  is equal to the population mean—that is,  $p$ ; this value is given in Theorem 5.4.1(b). From part 2 of Theorem 5.2.1, the standard deviation of the sampling distribution of  $\hat{P}$  is equal to

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

where  $\sigma$  represents the standard deviation of the population of 0's and 1's, which is  $\sqrt{p(1 - p)}$ . Thus, the standard deviation of the sampling distribution of  $\hat{P}$  is equal to

$$\frac{\sqrt{p(1 - p)}}{\sqrt{n}} = \sqrt{\frac{p(1 - p)}{n}}$$

which is the value given in Theorem 5.4.1(b).

Note that the binomial distribution is just a rescaled version of the sampling distribution of  $\hat{P}$ :  $\hat{P} = \frac{Y}{n}$ , so  $Y = n\hat{P}$ . It follows that the binomial distribution also can be approximated by a normal curve with suitably rescaled mean and standard deviation. The mean of  $\hat{P}$  is  $p$  and the SD of  $\hat{P}$  is

$$\sqrt{\frac{p(1 - p)}{n}}$$

The rescaled mean is  $np$  and the rescaled standard deviation is

$$n\sqrt{\frac{p(1 - p)}{n}} = \sqrt{np(1 - p)}$$

which are as given in Theorem 5.4.1(a).

## Appendix 6.1 Significant Digits

In this appendix we review the concept of significant digits. Let us begin with an example.

Suppose a university president reports that there are 23,000 students at the university. How many significant digits are in the number

23,000?

When the number is expressed this way—in ordinary rather than scientific notation—it is not really possible to tell for sure how many significant digits it has. Does the president *really* mean

23,000 rather than 23,001 or 22,999?

If she does, then all five of the digits are significant. If (as is probable) she really means

23,000 rather than 22,000 or 24,000

then only the 2 and the 3 are significant digits, since only those digits are known with certainty; the three 0's in 23,000 are placeholders. Scientific notation removes the ambiguity:

$2.3 \times 10^4$  has 2 significant digits

$2.3000 \times 10^4$  has 5 significant digits

As the preceding example illustrates, you can clarify how many significant digits are in a number by expressing the number in scientific notation. Here are some examples:

ORDINARY NOTATION	SCIENTIFIC NOTATION	NUMBER OF SIGNIFICANT DIGITS
60,700	$6.07 \times 10^4$	3
60,700	$6.0700 \times 10^4$	5
60.7	$6.07 \times 10^1$	3
60.70	$6.070 \times 10^1$	4
0.0607	$6.07 \times 10^{-2}$	3
0.06070	$6.070 \times 10^{-2}$	4

In the preceding numbers, note that the interior zero (between 6 and 7) is always a significant digit; the leading zeros (before the 6) are not significant; the terminal zeros (after the 7) are significant in scientific notation and ambiguous in ordinary notation. Digits other than zero are always significant.

Here are some examples of rounding a number to two significant digits:

NUMBER	ROUNDED TO TWO SIGNIFICANT DIGITS
60,700	61,000 (that is, $6.1 \times 10^4$ )
60.7	61
0.0607	0.061
0.0592	0.059
0.0596	0.060

## Appendix 7.1

### How Power Is Calculated

The required sample sizes given in Table 5 were determined by calculating the power of the  $t$  test. For large samples, an appropriate power calculation can be based on the normal curve (Table 3). In this appendix we indicate how such an approximate calculation is done.

Recall that the power is the probability of rejecting  $H_0$  when  $H_A$  is true. In order to calculate power, therefore, we need to know the sampling distribution of  $t_s$  when  $H_A$  is true. For large samples, the sampling distribution can be approximated by a normal curve, as shown by the following theorem.

**Theorem A.1** Suppose we choose independent random samples, each of size  $n$ , from normal populations with means  $\mu_1$  and  $\mu_2$  and a common standard deviation  $\sigma$ . If  $n$  is large, the sampling distribution of  $t_s$  can be approximated by a normal distribution with

$$\text{Mean} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}} = \sqrt{\frac{n}{2}} \left( \frac{\mu_1 - \mu_2}{\sigma} \right)$$

and

$$\text{Standard deviation} = 1$$

To illustrate the use of Theorem A.1 for power calculations, suppose we are considering a one-tailed  $t$  test with  $\alpha = 0.025$ . The hypotheses are

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 > \mu_2$$

If we want a power of 0.80 for an effect size of 0.4, then Table 5 recommends samples of size of  $n = 100$ . Let us confirm this recommendation using Theorem A.1.

If  $H_0$  is true, so that  $\mu_1 = \mu_2$ , then the sampling distribution of  $t_s$  is approximately a normal distribution with mean equal to 0 and SD equal to 1. This is the null distribution of  $t_s$ ; it is shown as the dashed curve in Figure A.7.1.

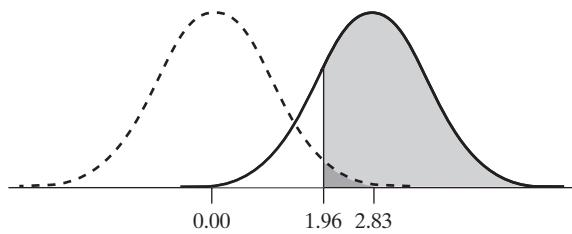


Figure A.7.1

Suppose that in fact  $H_A$  is true, that the effect size is

$$\frac{\mu_1 - \mu_2}{\sigma} = 0.4$$

and that we are using samples of size  $n = 100$ . Then, according to Theorem A.1, the sampling distribution of  $t_s$  will be approximately a normal distribution with SD equal to 1 and mean equal to

$$\sqrt{\frac{n}{2}} \left( \frac{\mu_1 - \mu_2}{\sigma} \right) = \sqrt{\frac{100}{2}}(0.4) = 2.83$$

This distribution is the solid curve in the figure.

For  $n_1 = n_2 = 100$ , we have  $df \approx \infty$ , so from Table 4 the critical value is equal to 1.96. Thus, the  $P$ -value would be less than 0.025 and we would reject  $H_0$  if

$$t_s > 1.96$$

Using the dotted curve, the probability of this event is equal to 0.025; this is shown in the figure as the dark gray area. Using the solid curve, the probability that  $t_s > 1.96$  includes all the shaded area in the figure. The shaded area can be determined from Table 3 using

$$Z = \frac{1.96 - 2.38}{1} = -0.87$$

From Table 3, the area is  $0.8078 \approx 0.81$ . Thus, we have shown that, for  $n_1 = n_2 = 100$ ,

$$\text{if } \frac{\mu_1 - \mu_2}{\sigma} = 0.4, \text{then } \Pr\{\text{reject } H_0\} \approx 0.81$$

We have found that the power against the specified alternative is approximately equal to 0.81; this agrees well with Table 5, which claims that the power is equal to 0.80.

If we were concerned with a two-tailed test at  $\alpha = 0.05$ , the critical value would again be 1.96, and so the power would again be approximately equal to 0.81, because the area under the solid curve corresponding to the left-hand tail of the dashed curve is negligible.

Of course, in constructing Table 5, one begins with the specified power (0.80) and determines  $n$ , rather than the other way around. This “inverse” problem can be solved using an approach similar to the foregoing. In the figure, the shaded area (0.80) would be given; this would determine the  $Z$  value and in turn determine  $n$ , once the effect size is specified.

## Appendix 7.2

### More on the Wilcoxon-Mann-Whitney Test

In Section 7.10 we saw how critical values for the Wilcoxon-Mann-Whitney test are related to the null distribution of  $K_1$ ,  $K_2$ , and  $U_s$ . In this appendix we indicate how these null distributions can be determined by simple counting methods.

Let us consider the sample sizes of  $n = 5$  and  $n' = 4$ . In Figure 7.10.3 (page 286), the  $Y_1$ 's and  $Y_2$ 's are plotted as dots. To save space, let us now represent the data in a more compact way: We will represent each  $Y_1$  by a “1” and each  $Y_2$  by a “2.” Thus, the arrangement in Figure 7.10.3(a) (where the  $Y_1$ 's are entirely to the left of the  $Y_2$ 's) would be represented as

1 1 1 1 1 2 2 2 2

For sample sizes  $n = 5$  and  $n' = 4$ , there are 126 possible arrangements of the  $Y_1$ 's and  $Y_2$ 's. Here is a partial list of those arrangements and the associated values of  $K_1$  and  $K_2$ :

NUMBER	ARRANGEMENT	$K_1$	$K_2$
1	1 1 1 1 1 2 2 2 2	0	20
2	1 1 1 1 2 1 2 2 2	1	19
3	1 1 1 1 2 2 1 2 2	2	18
4	1 1 1 2 1 1 2 2 2	2	18
5	1 1 2 1 1 1 2 2 2	3	17
6	1 1 1 2 1 2 1 2 2	3	17
7	1 1 1 1 2 2 2 1 2	3	17
8	1 2 1 1 1 1 2 2 2	4	16
9	1 1 2 1 1 2 1 2 2	4	16
10	1 1 1 2 1 2 2 1 2	4	16
11	1 1 1 2 2 1 1 2 2	4	16
12	1 1 1 1 2 2 2 2 1	4	16
. . . and so on . . .			
126	2 2 2 2 1 1 1 1 1	20	0

To determine the null distributions from this list, we need to know the likelihood of the various arrangements, assuming that  $H_0$  is true. According to  $H_0$ , all 9 observations ( $Y$ 's) were drawn at random from the same population. Under this assumption, it can be shown that the 126 arrangements are all *equally likely*. Because of this simple and elegant fact, the null distribution of  $K_1$  and  $K_2$  (and therefore  $U_s$ ) can be determined by straightforward counting. Working from the above list, we find the following probabilities:

$K_1$	$K_2$	PROBABILITY
0	20	$\frac{1}{126}$
1	19	$\frac{1}{126}$
2	18	$\frac{2}{126}$
3	17	$\frac{3}{126}$
4	16	$\frac{5}{126}$
. . . and so on . . .		
20	0	$\frac{1}{126}$
Total 1		

These probabilities constitute the null distribution of  $K_1$  and  $K_2$ —plotted in Figure 7.10.4(a). For instance, the first probability in the null distribution is

$$\Pr\{K_1 = 0, K_2 = 20\} = \frac{1}{126} \approx 0.008$$

as stated in Section 7.10.

Why is the Wilcoxon-Mann-Whitney test distribution free? The reason should be clear from the preceding discussion. If the two population distributions are the same, then all possible arrangements of the  $Y$ 's are equally likely and the specific shape of the population distributions does not matter (except, of course, that we have assumed that there would be no ties; the null distributions are altered if ties are possible).

The Wilcoxon-Mann-Whitney null distribution can always be determined by straightforward counting such as illustrated above (although for larger sample sizes the counting is very tedious and approximate methods are used instead). The number of possible arrangements for samples of size  $n$  and  $n'$  is equal to

$$\frac{(n + n')!}{n!n'!}$$

For example, for sample sizes 5 and 4 (as before), we find

$$\frac{9!}{5!4!}$$

(To see why this formula works, refer to the discussion of the formula for binomial coefficients at the end of Appendix 3.1; the reasoning is exactly parallel.)

## Appendix 9.1

### More on Confidence Intervals for a Proportion

In this appendix we present some of the technical details behind the confidence interval for a proportion introduced in Section 9.2. For a more complete discussion of these ideas, see the paper by Agresti and Coull that is given as Note 3 for Chapter 9.

Suppose we want to develop a  $100(1 - \alpha)\%$  confidence interval for a proportion  $p$ ; for example, a 95% confidence level corresponds to  $\alpha = 0.05$ . A common procedure is to use

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

This is called the *Wald confidence interval*.

Another way to construct a confidence interval is to find all values of  $p$  such that

$$-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1 - p)/n}} \leq z_{\alpha/2} \quad (1)$$

(This is known as “inverting a hypothesis test”; hypothesis testing for proportions is discussed in Section 9.4.) The basic idea here is that the sampling distribution of  $\hat{p}$  can be approximated by a normal distribution and that we should take values that correspond to the middle  $100(1 - \alpha)\%$  (e.g., the middle 95%) of the normal distribution.

In the following presentation we will let  $z$  denote  $z_{\alpha/2}$ . Solving inequality (1) for  $p$  gives an interval of the form

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p}) + \frac{z^2}{4n}}{n}}}{1 + \frac{z^2}{n}}$$

This is called the *score confidence interval*.

Most books present the Wald confidence interval (without giving it that name), since it is much more simple in form than the score confidence interval. However, the Wald confidence interval has poor coverage properties: A nominal 95% Wald confidence interval might actually cover  $p$  only 80% of the time, rather than 95% of the time. The score confidence interval has excellent coverage properties but is quite complex.

The formulation of the *Wilson confidence interval* presented in Section 9.2 is based on approximating the score interval. Using algebra, it can be shown that the midpoint of the score interval is

$$\hat{p} \left( \frac{n}{n + z^2} \right) + \frac{1}{2} \left( \frac{z^2}{n + z^2} \right) \quad (2)$$

which is a weighted average of  $\hat{p}$  and  $\frac{1}{2}$ , with weights  $\frac{n}{n+z^2}$  and  $\frac{z^2}{n+z^2}$ . Note that as  $n$  increases, more weight is given to  $\hat{p}$ ; for small  $n$ , more weight is given to  $\frac{1}{2}$ . Because  $\hat{p} = \frac{y}{n}$ , the midpoint given by formula (2) is

$$\frac{\frac{y}{n}}{\frac{y}{n} + \frac{z^2}{n}} + \frac{\frac{z^2}{n}}{\frac{y}{n} + \frac{z^2}} = \frac{\frac{1}{2}z^2}{\frac{y}{n} + \frac{z^2}}$$

For a 95% confidence interval,  $z = 1.96 \approx 2$ , so that the midpoint is approximated by

$$\frac{y}{n+4} + \frac{2}{n+4} = \frac{y+2}{n+4}$$

which we called  $\tilde{p}$  in Section 9.1.

The standard error that is used in the Wald confidence interval is  $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ .

This is based on the fact that the variance of  $Y$  is equal to  $np(1-p)$ , as derived in Appendix 3.2, so that the variance of  $\hat{P}$  is equal to  $\frac{np(1-p)}{n^2}$  or  $\frac{p(1-p)}{n}$ . Hence

the standard deviation of  $\hat{P}$  is equal to  $\sqrt{\frac{p(1-p)}{n}}$ , as discussed in Appendix 5.1; the standard error uses the sample proportion  $\hat{p}$  in place of the unknown value of  $p$ . Likewise, the variance of  $\tilde{P}$  is equal to  $\frac{np(1-p)}{(n+z^2)^2}$ , which is approximately  $\frac{p(1-p)}{n+z^2}$ .

Hence the standard error for the Wilson confidence interval is given as  $\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+z^2}}$ , for a 95% confidence interval,  $z \approx 2$ , giving

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

As a closing note, we mention that some authors advocate the use of an “exact” confidence interval, based on the binomial distribution. However, as Agresti and Coull show, the exact confidence interval is quite conservative, which is one reason that it is not widely used. (It is also quite complex.)

## Appendix 12.1

### Least-Squares Formulas

In this appendix we show that the least-squares criterion leads to the formulas

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

We will make use of the fact that the minimum of a quadratic function  $Q(x) = Ax^2 + Bx + C$  occurs at  $x = -\frac{B}{2A}$ .

### Preliminary Result

Given a set of data  $y_1, y_2, \dots, y_n$ , the number  $c$  that minimizes the quantity  $\sum_{i=1}^n (y_i - c)^2$  is the mean  $\bar{y}$ .

To see this, first expand  $(y_i - c)^2$  to get  $y_i^2 - 2y_i c + c^2$  and then distribute the summation to get

$$\sum_{i=1}^n (y_i - c)^2 = \sum_{i=1}^n y_i^2 - 2c \sum_{i=1}^n y_i + nc^2$$

The last expression is a quadratic function  $Q(c) = Ac^2 + Bc + C$  where  $A = n$  and  $B = -2 \sum_{i=1}^n y_i$ . Therefore the minimum occurs at  $c = -\frac{B}{2A} = -\frac{-2 \sum_{i=1}^n y_i}{2n} = \bar{y}$ .

Recall that if  $\hat{y}_i = b_0 + b_1 x_i$  is the least-squares regression line then  $b_0$  and  $b_1$  are the values that minimize the residual sum of squares, given by

$$\text{SS(resid)} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$$

We can write  $y_i - (b_0 + b_1 x_i)$  as  $(y_i - b_1 x_i) - b_0$ . Next, we apply the preliminary result with  $(y_i - b_1 x_i)$  in the place of  $y_i$  and  $b_0$  in the place of  $c$ . Thus, the minimum of the residual sum of squares occurs when

$$b_0 = \text{the mean of } (y_i - b_1 x_i) = \bar{y} - b_1 \bar{x} \quad (1)$$

Thus, the least-squares line goes through the point of averages,  $(\bar{x}, \bar{y})$ .

Substituting this value of  $b_0$  into the residual sum of squares gives

$$\begin{aligned} \text{SS(resid)} &= \sum_{i=1}^n (y_i - (\bar{y} - b_1 \bar{x}) - b_1 x_i)^2 = \sum_{i=1}^n (y_i - \bar{y} + b_1 \bar{x} - b_1 x_i)^2 \\ &= \sum_{i=1}^n [(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \end{aligned}$$

This can be expanded as

$$\begin{aligned} \text{SS(resid)} &= \sum_{i=1}^n [(y_i - \bar{y})^2 - 2b_1(x_i - \bar{x})(y_i - \bar{y}) + b_1^2(x_i - \bar{x})^2] \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

Thus, the residual sum of squares is a quadratic function in  $b_1$  with  $A = \sum_{i=1}^n (x_i - \bar{x})^2$  and  $B = -2\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ . Hence, the minimum of  $\text{SS}(\text{resid})$  occurs at  $-\frac{B}{2A}$  or

$$\begin{aligned} b_1 &= -\frac{-2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{2 \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)}{(n - 1)s_x^2} s_x s_y \\ &= r \left( \frac{s_y}{s_x} \right) \end{aligned} \tag{2}$$

Together equations (1) and (2) give the formulas for the coefficients of the least-squares regression line.

## Appendix 12.2

### Derivation of Fact 12.3.1

By definition

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

Define  $\text{SPXY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ ,  $\text{SSX} = \sum_{i=1}^n (x_i - \bar{x})^2$ , and  $\text{SSY} = \sum_{i=1}^n (y_i - \bar{y})^2$ . Thus

$$s_x^2 = \frac{\text{SSX}}{n-1} \quad \text{and} \quad s_y^2 = \frac{\text{SSY}}{n-1}$$

$$\text{Then } r = \frac{1}{n-1} \frac{\text{SPXY}}{s_x s_y} \text{ so } r^2 = \frac{\text{SPXY}^2}{(\text{SSX})(\text{SSY})}.$$

Recall that  $b_1 = r \left( \frac{s_y}{s_x} \right)$  and  $b_0 = \bar{y} - b_1 \bar{x}$ . Thus

$$b_1 = \frac{1}{n-1} \frac{\text{SPXY}}{s_x s_y} \left( \frac{s_y}{s_x} \right) = \frac{1}{n-1} \frac{\text{SPXY}}{s_x^2} = \frac{\text{SPXY}}{\text{SSX}}$$

The definition of a residual is  $y - \hat{y}$ , but this is the same as  $y - (b_0 + b_1 x) = y - ([\bar{y} - b_1 \bar{x}] + b_1 x) = (y - \bar{y}) - b_1(x - \bar{x})$ .

Thus,

$$\begin{aligned} \text{SS(resid)} &= \sum_{i=1}^n ((y_i - \bar{y}) - b_1(x_i - \bar{x}))^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \text{SSY} - 2b_1 \text{SPXY} + b_1^2 \text{SSX} \\ &= \text{SSY} - 2 \frac{\text{SPXY}}{\text{SSX}} \text{SPXY} + \frac{\text{SPXY}^2}{\text{SSX}} \\ &= \text{SSY} - \frac{\text{SPXY}^2}{\text{SSX}} \end{aligned}$$

Thus, we have established that  $\frac{\text{SPXY}^2}{\text{SSX}} = \text{SSY} - \text{SS(resid)}$ .

Thus,

$$r^2 = \frac{\text{SPXY}^2}{(\text{SSX})(\text{SSY})} = \frac{\text{SSY} - \text{SS(resid)}}{\text{SSY}}$$

So

$$r^2 = \frac{\frac{\text{SSY}}{n-1} - \frac{\text{SS(resid)}}{n-1}}{\frac{\text{SSY}}{n-1}} = \frac{s_y^2 - s_e^2 \left( \frac{n-2}{n-1} \right)}{s_y^2} = \frac{s_y^2 - s_e^2 f}{s_y^2} \approx \frac{s_y^2 - s_e^2}{s_y^2}$$

$$\text{where } f = \frac{n-2}{n-1}$$

The factor  $f$  is close to 1 unless  $n$  is quite small. Here are some values of  $f$ :

$n$	$f$
10	0.89
15	0.93
20	0.95

The approximation is reasonably good if  $n \geq 10$ .

# CHAPTER NOTES

---

## Chapter 1

1. Nicolle, J. (1961). *Louis Pasteur: The Story of His Major Discoveries*. New York: Basic Books. p. 170. © 1961 by Jacques Nicolle. © 1961 English translation Hutchinson & Co. (Publishers) Ltd. Reprinted by permission of Perseus Books Group.
2. Mizutani, T., and Mitsuoka, T. (1979). Effect of intestinal bacteria on incidence of liver tumors in gnotobiotic C3H/He male mice. *Journal of the National Cancer Institute* **63**, 1365–1370.
3. Tripepi, R. R., and Mitchell, C. A. (1984). Metabolic response of river birch and European Birch roots to hypoxia. *Plant Physiology* **76**, 31–35. Raw data courtesy of the authors.
4. Adapted from Potkin, S. G., Cannon, H. E., Murphy, D. L., and Wyatt, R. J. (1978). Are paranoid schizophrenics biologically different from other schizophrenics? *New England Journal of Medicine* **298**, 61–66. The data are approximate, having been reconstructed from the histograms and summary information given by Potkin et al. Reprinted by permission of the *New England Journal of Medicine*.
5. Wolfson, J. L. (1987). Impact of *Rhizobium* nodules on *Sitona hispidulus*, the clover root curculio. *Entomologia Experimentalis et Applicata* **43**, 237–243. Data courtesy of the author. The experiment actually included 11 dishes.
6. Webb, P. (1981). Energy expenditure and fat-free mass in men and women. *American Journal of Clinical Nutrition* **34**, 1816–1826.
7. The headline appeared on page 2 of the Sunday edition of *The New York Times*, 16 July 1911.
8. Allen, L. S., and Gorski, R. A. (1992). Sexual orientation and the size of the anterior commissure in the human brain. *Proceedings of the National Academy of Science* **89**, 7199–7202. The data are approximate, having been reconstructed from the dotplots and summary information given by Allen and Gorski. Regarding the first concern mentioned in Example 1.2.2, the authors were mindful of the effect that the two largest observations could have on their conclusions and calculated the average for the homosexual men a second time, after deleting these two values. As for the second concern, the authors calculated the averages for those who had AIDS and those who did not in each group of men. They found that AIDS is associated with smaller, not larger, AC areas, so that when only persons without AIDS are compared, the difference between homosexual and heterosexual men is even larger than the difference found in the full data set.
9. Hakansson, K., Rovio, S., Helkala, E. L., Vilska, A. R., Winblad, B., Soininen, H., Nissinen, A., Mohammed, A. H., and Kivipelto, M. (2009). Association between mid-life marital status and cognitive function in later life: Population based cohort study. *BMJ* **339**, b2462.
10. Bradstreet, T. E. (1992). Favorite data sets from early phases of drug research—part 2. *Proceedings of the Section on Statistical Education of the American Statistical Association*, 219–223.
11. Benson, H., and Friedman, R. (1996). Harnessing the power of the placebo effect and renaming it “remembered wellness.” In *Annual Review of Medicine* **47**, 193–199. Annual Reviews, Inc., Palo Alto, Calif.
12. Sandler, A. D., Sutton, K. A., DeWeese, J., Girardi, M. A., Sheppard, V., and Bodfish, J. W. (1999). Lack of benefit of a single dose of synthetic human secretin in the treatment of autism and pervasive developmental disorder. *New England Journal of Medicine* **341**, 1801–1806. The improvement in the placebo group was somewhat better than the improvement in the secretin group for the response variable of change in total Autism Behavior Checklist score, but the *P*-value for the difference was 0.11.
13. Butler, C., and Steptoe, A. (1986). Placebo response: An experimental study of asthmatic volunteers. *British Journal of Clinical Psychology* **25**, 173–183.
14. Barsamian, E. M. (1977). The rise and fall on internal mammary artery ligation in the treatment of angina pectoris and the lessons learned: In Bunker, J. P., Barnes, B. A., and Mosteller, F. (eds.), *Costs, Risks, and Benefits of Surgery*. New York: Oxford University Press, pp. 212–220.
15. Chalmers, T. C., Celano, P., Sacks, H. S., and Smith, H. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine* **309**, 1358–1361.
16. The Coronary Drug Project Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New England Journal of Medicine* **303**, 1038–1041. Several variables were measured on each subject at the start of the experiment. Adjusting for the effects of these covariates within the placebo group only slightly reduces the difference in mortality rates between adherers and non-adherers. Thus, differences in overall health explain only a small part of the “adherent versus nonadherent” mortality rate difference.

17. Diehl, H. S., Baker, A. B., and Cowan, D. W. (1938). Cold vaccines: An evaluation based on a controlled study. *Journal of the American Medical Association* **111**, 1168–1173.
18. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* **34**, 585–612.
19. Sacks, H., Chalmers, T. C., and Smith, H. (1982). Randomized versus historical controls for clinical trials. *American Journal of Medicine* **72**, 233–240.
20. Dublin, L. I. (1957). *Water fluoridation: Facts, not myths*. New York: Public Affairs Committee, Inc.
21. Sandler, R. S., Zorich, N. L., Filloon, T. G., Wiseman, H. B., Lietz, D. J., Brock, M. H., Royer, M. G., and Miday, R. K. (1999). Gastrointestinal symptoms in 3181 volunteers ingesting snack foods containing olestra or triglycerides. A 6-week randomized, placebo-controlled trial. *Annals of Internal Medicine* **130**, 253–261.
22. Moertel, C. G., Fleming, T. R., Creagan, E. T., Rubin, J., O'Connell, M. J., and Ames, M. M. (1985). High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy. *New England Journal of Medicine* **312**, 137–141.
23. Pauling, L., and Cameron, E. (1976). Supplemental ascorbate in the supportive treatment of cancer: Prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Sciences* **73**, 3685–3789.
24. *Cleveland Plain Dealer*, 23 October 1997, page 15-A.
25. Huber, Ann M. (2005). Moisture requirements for the germination of early seedling survival of *Cirsium loncholepis*. Master's thesis in Biological Sciences, California Polytechnic State University.
26. Personal communication from L. Vredevoe regarding an ongoing research project (2009).
27. Parks, N. J., Krohn, K. A., Mathis, C. A., Chasko, J. H., Geiger, K. R., Gregor, M. E., and Peek, N. F. (1981). Nitrogen-13-labelled nitrite and nitrate: Distribution and metabolism after intratracheal administration. *Science* **212**, 58–61.
28. *Cleveland Plain Dealer*, 25 June 1991, page 3-A.
29. Hull, H. F., Bettinger, C. J., Gallaher, M. M., Keller, N. M., Wilson, J., and Mertz, G. J. (1988). Comparison of HIV-antibody prevalence in patients consenting to and declining HIV-antibody testing in an STD clinic. *Journal of the American Medical Association*, **260**, 935–938.
30. Hoover, D. W., and Milich, R. (1994). Effects of sugar ingestion expectancies on mother-child interactions. *Journal of Abnormal Psychology* **22**, 501–515.
31. Krummel, D. A., Seligson, F. H., and Guthrie, H. A. (1996). Hyperactivity: Is candy causal? *Critical Reviews in Food Science and Nutrition* **36**, 31–47.

## Chapter 2

1. Stewart, R. N., and Arisumi, T. (1966). Genetic and histogenic determination of pink bract color in poinsettia. *Journal of Heredity* **57**, 217–220.
2. Haselgrove, C., Straker, L., Smith, A., O'Sullivan, P., Perry, M., and Sloan, N. (2008). Perceived school bag load, duration of carriage, and method of transport to school are associated with spinal pain in adolescents: An observational study. *Australian Journal of Physiotherapy* **54**, 193–200.
3. Data obtained online from *The World Factbook* produced by the CIA at [www.cia.gov/library/publications/the-world-factbook/](http://www.cia.gov/library/publications/the-world-factbook/)
4. Unpublished data courtesy of C. M. Cox and K. J. Drewry.
5. Unpublished data courtesy of W. F. Jacobson.
6. Unpublished data collected at Oberlin College by J. Witmer.
7. Knoll, A. E., and Barghoorn, E. S. (1977). Archean microfossils showing cell division from the Swaziland system of South Africa. *Science* **198**, 396–398.
8. Nurse, C. A. (1981). Interactions between dissociated rat sympathetic neurons and skeletal muscle cells developing in cell culture. II. Synaptic mechanisms. *Developmental Biology* **88**, 71–79.
9. Topinard, P. (1888). Le poids de l'encephale d'après les registres de Paul Broca. *Memoires Societe d'Anthropologie Paris*, 2nd series, **3**, 1–41. The data shown are a subset of the data published by Topinard.
10. Johannsen, W. (1903). *Ueber Erblichkeit in Populationen und in reinen Linien*. Jena: G. Fischer. Data reproduced in Strickberger, M. W. (1976). *Genetics*, New York: Macmillan, p. 277; and Peters, J. A. (ed.) (1959). *Classic Papers in Genetics*, Englewood Cliffs, NJ: Prentice Hall, p. 23.
11. Unpublished data courtesy of W. F. Jacobson.

12. Simpson, G. G., Roe, A., and Lewontin, R. C. (1960). *Quantitative Zoology*. New York: Harcourt, Brace. p. 51.
13. Adapted from Potkin, S. G., Cannon, H. F., Murphy, D. L., and Wyatt, R. J. (1978). Are paranoid schizophrenics biologically different from other schizophrenics? *New England Journal of Medicine* **298**, 61–66. The data given are approximate, having been reconstructed from the histogram and summary information given by Potkin et al. Reprinted by permission of the *New England Journal of Medicine*.
14. Peters, H. G., and Bademan, H. (1963). The form and growth of stellate cells in the cortex of the guinea-pig. *Journal of Anatomy (London)* **97**, 111–117.
15. Data courtesy of R. F. Jones, Indiana State Dairy Association, Inc.
16. Unpublished data courtesy of D. J. Honor and W. A. Vestre.
17. Hepp, J., Buck, C., and Catalano, J. (2007). Analysis of three corn hybrids to determine if more expensive varieties are worth their cost. Unpublished manuscript, Oberlin College.
18. Bruce, D., Harvey, D., Hamerton, A. E., and Bruce, L. (1913). Morphology of various strains of the trypanosome causing disease in man in Nyasaland. I. The human strain. *Proceedings of the Royal Society of London, Series B* **86**, 285–302. See also Pearson, K. (1914). On the probability that two independent distributions of frequency are really samples of the same population, with reference to recent work on the identity of trypanosome strains. *Biometrika* **10**, 85–143.
19. Shields, D. R. (1981). The influence of niacin supplementation on growing ruminants and *in vivo* and *in vitro* rumen parameters. Ph.D. thesis, Purdue University. Raw data courtesy of the author and D. K. Colby.
20. Gwynne, D. T. (1981). Sexual difference theory: Mormon crickets show role reversal in mate choice. *Science* **213**, 779–780. Copyright 1981 by the AAAS. Raw data courtesy of the author.
21. Unpublished data courtesy of M. A. Morse and G. P. Carlson.
22. Adapted from Anderson, J. W., Story, L., Sieling, B., Chen, W. L., Petro, M. S., and Story, J. (1984). Hypocholesterolemic effects of oat-bran or bean intake for hypercholesterolemic men. *American Journal of Clinical Nutrition* **40**, 1146–1155. There were actually 20 men in the study.
23. Unpublished data courtesy of C. H. Noller.
24. Luria, S. F., and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511.
25. Fictitious but realistic data. See Roberts, J. (1975). Blood pressure of persons 18–74 years, United States, 1971–72. U.S. National Center for Health Statistics, *Vital and Health Statistics*, Series 11, No. 150. Washington, D.C.: U.S. Department of Health, Education and Welfare.
26. Unpublished data collected from a sample of Oberlin College students.
27. Unpublished data courtesy of M. Kimmel.
28. Unpublished data courtesy of F. Delgado.
29. Kitts, C., Moline, M., Schaffner, A., Samadpour, M., MacNiel, K., and Duffield, S. (2002). *Identifying the Sources of Escherichia coli Contamination in Crassostrea gigas from the Morro Bay Estuary*. Technical Report for the National Estuary Program and the California Central Coast Regional Water Quality Control Board.
30. Kinghorn, A., Humphries, M., Outridge, P., and Chan H. M. (2008). Teeth as biomonitor of selenium concentrations in tissues of beluga whales (*Delphinapterus leucas*). *Science of the Total Environment* **402**, 43–50.
31. Govind, C. K., and Pearce, J. (1986). Differential reflex activity determines claw and closer muscle asymmetry in developing lobsters. *Science* **233**, 354–356. Copyright 1986 by the AAAS.
32. Adapted from Gerdes, N. (2001). Morphological and life history variation in three populations of golden-mantled ground squirrels along a Pacific coast transect. Master's thesis California Polytechnic State University, San Luis Obispo.
33. Adapted from Barclay, A. M., and Crawford, R. M. M. (1984). Seedling emergence in the rowan (*Sorbus aucuparia*) from an altitudinal gradient. *Journal of Ecology* **72**, 627–636. Reprinted with permission of Blackwell Scientific Publications Limited.
34. Fictitious but realistic data. Based on Beyl, C. A., and Mitchell, C. A. (1977). Characterization of mechanical stress dwarfing in chrysanthemum. *Journal of the American Society for Horticultural Science* **102**, 591–594.
35. Based on a subset of the data in Tuddenham, R. D., and Snyder, M. M. (1954). Physical growth of California boys and girls from birth to age 18. *Calif. Publ. Child Develop.* **1**, 183–364. Data as reported in Weisberg, S. (1985). *Applied Linear Regression*, 2nd ed. New York: Wiley.

- 36.** Nelson, L. A. (1980). Report of the Indiana Beef Evaluation Program, Inc. Purdue University, West Lafayette, Indiana.
- 37.** Data collected by J. Witmer at a statistics workshop at Johns Hopkins University, July 1995.
- 38.** Day, K. M., Patterson, F. L., Luetkemeier, O. W., Ohm, H. W., Polizotto, K., Roberts, J. J., Shaner, G. E., Huber, D. M., Finney, R. F., Foster, J. F., and Gallun, R. L. (1980). Performance and adaptation of small grains in Indiana. Station Bulletin No. 290. West Lafayette, Ind., Agricultural Experiment Station of Purdue University. Raw data provided courtesy of W. F. Nyquist.
- 39.** Tripepi, R. R., and Mitchell, C. A. (1984). Metabolic response of river birch and European birch roots to hypoxia. *Plant Physiology* **76**, 31–35. Raw data courtesy of the authors.
- 40.** Ogilvie, R. I., Macleod, S., Fernandez, P., and McCullough, W. (1974). Timolol in essential hypertension. In B. Magnani (ed.). *Beta-Adrenergic Blocking Agents in the Management of Hypertension and Angina Pectoris*. New York: Raven Press. pp. 31–43.
- 41.** Unpublished data courtesy of J. F. Nash and J. E. Zabik.
- 42.** Schall, J. J., Bennett, A. F., and Putnam, R. W. (1982). Lizards infected with malaria: Physiological and behavioral consequences. *Science* **217**, 1057–1059. Copyright 1982 by the AAAS. Raw data courtesy of J. J. Schall.
- 43.** Fictitious but realistic data. Each observation is the average of several measurements made on the same woman at different times. See Royston, J. P., and Abrams, R. M. (1980). An objective method for detecting the shift in basal body temperature in women. *Biometrics* **36**, 217–224.
- 44.** Adapted from data in Cicirelli, M. F., Robinson, K. R., and Smith, L. D. (1983). Internal pH of *Xenopus* oocytes: A study of the mechanism and role of pH changes during meiotic maturation. *Developmental Biology* **100**, 133–146.
- 45.** Adapted from Royston, J. P., and Abrams, R. M. (1980). An objective method for detecting the shift in basal body temperature in women. *Biometrics* **36**, 217–224.
- 46.** Adapted from data provided courtesy of L. A. Nelson.
- 47.** Ikin, E. W., Prior, A. M., Race, R. R., and Taylor, G. L. (1939). The distribution of the A<sub>1</sub>A<sub>2</sub>BO blood groups in England. *Annals of Eugenics (London)* **9**, 409–411. Reprinted with permission of Cambridge University Press.
- 48.** Borg, S., Kvande, H., and Sedvall, G. (1981). Central norepinephrine metabolism during alcohol intoxication in addicts and healthy volunteers. *Science* **213**, 1135–1137. Copyright 1981 by the AAAS. Raw data courtesy of S. Borg.
- 49.** Fictitious but realistic population. Adapted from LeClerc, E. L., Leonard, W. H., and Clark, A. G. (1962). *Field Plot Technique*. Minneapolis: Burgess.
- 50.** Selawry, O. S. (1974). The role of chemotherapy in the treatment of lung cancer. *Seminars in Oncology* **1**, No. 3, 259–272.
- 51.** Hayes, H. K., East, E. M., and Bernhart, E. G. (1913). *Connecticut Agricultural Experiment Station Bulletin* **176**. Data reproduced in Strickberger, M. W. (1976). *Genetics*. New York: Macmillan, p. 288.
- 52.** Unpublished data courtesy of J. Y. Latimer and C. A. Mitchell.
- 53.** Connolly, K. (1968). The social facilitation of preening behaviour in *Drosophila melanogaster*. *Animal Behaviour* **16**, 385–391.
- 54.** The results of similar assays are reported in Pascholati, S. F., and Nicholson, R. L. (1983). *Helminthosporum maydis* suppresses expression of resistance to *Helminthosporum carbonum* in corn. *Phytopathologische Zeitschrift* **107**, 97–105. Unpublished data courtesy of the investigators.
- 55.** Richens, A., and Ahmad, S. (1975). Controlled trial of valproate in severe epilepsy. *British Medical Journal* **4**, 255–256.
- 56.** Fleming, W. E., and Baker, F. E. (1936). A method for estimating populations of larvae of the Japanese beetle in the field. *Journal of Agricultural Research* **53**, 319–331. Data reproduced in *Statistical Ecology*, Vol. 1 (1971). University Park: Pennsylvania State University Press, p. 327.
- 57.** Chiarotti, R. M. (1972). An investigation of the energy expenditure of women squash players. Master's thesis, Pennsylvania State University. Raw data courtesy of R. M. Lyle (nee Chiarotti).
- 58.** Mastey, J. (1983). Innervation of the equine small intestine. Master's thesis, Purdue University. Raw data courtesy of the author.
- 59.** Fictitious but realistic data. Adapted from data presented in Falconer, D. S. (1981). *Introduction to Quantitative Genetics*, 2nd ed. New York: Longman, Inc., p. 97.
- 60.** Dow, T. G. B., Rooney, P. J., and Spence, M. (1975). Does anaemia increase the risks to the fetus caused by smoking in pregnancy? *British Medical Journal* **4**, 253–254.

61. Christophers, S. R. (1924). The mechanism of immunity against malaria in communities living under hyper-endemic conditions. *Indian Journal of Medical Research* **12**, 273–294. Data reproduced in Williams, C. B. (1964). *Patterns in the Balance of Nature*. London: Academic Press. p. 243.
62. Data taken from *Climatological Data, Ohio*, and *Local Climatological Data, Cleveland, Ohio*; National Oceanic and Atmospheric Administration, U.S. Dept. of Commerce.
63. These data were published on page 8-A of the *Cleveland Plain Dealer*, 6 February 1997, from information compiled by the United Network for Organ Sharing. The mortality rate and volume variables are averages over a four-year period beginning in October 1987. There are 31 hospitals in the low-volume group and 76 in the high-volume group.
64. Erne, P., Bolli, P., Buergisser, E., and Buehler, F. R. (1984). Correlation of platelet calcium with blood pressure. *New England Journal of Medicine* **310**, 1084–1088. Reprinted by permission. Raw data courtesy of F. R. Buehler. The original data set had 47 subjects; we have omitted 9 patients with “borderline” high blood pressure.

## Chapter 3

1. Based on an article by the Neonatal Inhaled Nitric Oxide Study Group (1997). See Inhaled nitric oxide in full-term and nearly full-term infants with hypoxic respiratory failure. *New England Journal of Medicine* **336**, 597–604.
2. Fictitious but realistic population. Adapted from Hubbs, C. L., and Schultz, L. P. (1932). *Cottus tubulatus*, a new sculpin from Idaho. *Occasional Papers of the Museum of Zoology, University of Michigan* **242**, 1–9. Data reproduced in Simpson, G. G., Roe, A., and Lewontin, R. C. (1960). *Quantitative Zoology*. New York: Harcourt, Brace. p. 81.
3. [www.bloodbook.com/world-abo.html](http://www.bloodbook.com/world-abo.html)
4. This table is a modified version of data adapted from Ammon, O. (1899). *Zur Anthropologie der Badener*. Jena: G. Fischer. Ammon’s data appear in Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* **49**, 732–764. The numbers in the table have been rounded to aid the exposition.
5. Unpublished data courtesy of Diana Zumas and Lisa Yasuhara, Oberlin College.
6. Adapted from Taira, D. A., Safran, D. G., Seto, T. B., Rogers, W. H., and Tarlov, A. R. (1997).

The relationship between patient income and physician discussion of health risk behaviors. *Journal of the American Medical Association* **278**, 1412–1417.

7. The population is fictitious but resembles the population of American women aged 18–24, excluding known or suspected diabetics, as reported in Gordon, T. (1964). Glucose tolerance of adults, United States 1960–62. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 2. Washington, D.C.: U.S. Department of Health, Education and Welfare.
8. Meyer, W. H. (1930). Diameter distribution series in even-aged forest stands. *Yale University School of Forestry Bulletin* **28**. The curve is fitted in Bliss, C. I., and Reinker, K. A. (1964). A lognormal approach to diameter distributions in even-aged stands. *Forest Science* **10**, 350–360.
9. Pearson, K. (1914). On the probability that two independent distributions of frequency are really samples of the same population, with reference to recent work on the identity of trypanosome strains. *Biometrika* **10**, 85–143. Reprinted by permission of the Biometrika Trustees.
10. Adapted from unpublished data courtesy of Gloria Zender, Oberlin College.
11. Fictitious but realistic situation. Based on data given by Lack, D. (1948). Natural selection and family size in the starling. *Evolution* **2**, 95–110. Data reproduced by Riclefs, R. E. (1973). *Ecology*. Newton, Mass.: Chiron Press. p. 37.
12. Adapted from unpublished data courtesy of Marni Hansill, Oberlin College.
13. Halpine, T., and Kerr, S. J. (1986). Mutant allele frequencies in the cat population of Omaha, Nebraska. *The Journal of Heredity* **77**, 460–462.
14. This is one of the crosses performed by Gregor Mendel in his classic studies of heredity; heterozygous plants (which are yellow seeded because yellow is dominant) are crossed with each other.
15. Fictitious but realistic value. See Hutchison, J. G. P., Johnston, N. M., Plevey, M. V. P., Thangkhiew, I., and Aidney, C. (1975). Clinical trial of Mebendazole, a broad-spectrum anthelmintic. *British Medical Journal* **2**, 309–310.
16. Fictitious but realistic population. Adapted from Owen, D. F. (1963). Polymorphism and population density in the African land snail, *Limicolaria martensiana*. *Science* **140**, 666–667.

17. Mathews, T. J., and Hamilton, B. E. (2005). Trend analysis of the sex ratio at birth in the United States. *National Vital Statistics Reports* **53**, No. 20. Hyattsville Md.; National Center for Health Statistics. The sex ratio varies slightly over time and by race.
18. Adapted from discussion in Galen, R. S., and Gambino, S. R. (1980). *Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses*. New York: Wiley. pp. 71–74.
19. This would be true for some central-city populations. See Annest, J. L., Mahaffey, K. R., Cox, D. H., and Roberts, J. (1982). Blood lead levels for persons 6 months–74 years of age: United States, 1976–80. *U.S. National Center for Health Statistics, Advance Data from Vital and Health Statistics*, No. 79. Hyattsville, Md.; U.S. Department of Health and Human Services.
20. Geissler, A. (1889). Beitrage zur Frage des Geschlechtsverhaltnisses der Geborenen. *Zeitschrift des K. Sachsischen Statistischen Bureaus* **35**, 1–24. Data reproduced by Edwards, A. W. F. (1958). An analysis of Geissler's data on the human sex ratio. *Annals of Human Genetics* **23**, 6–15. The data are also discussed by Stern, C. (1960). *Human Genetics*. San Francisco: Freeman.
21. Haseman, J. K., and Soares, E. R. (1976). The distribution of fetal death in control mice and its implications on statistical tests for dominant lethal effects. *Mutation Research* **41**, 277–288.
22. Data courtesy of S. N. Postlethwaite.
23. Adapted from Looker, A., et al. (1997). Prevalence of iron deficiency in the United States. *Journal of the American Medical Association* **277**, 973–976.
24. Fictitious but realistic situation. See Krebs, C. J. (1972). *Ecology: The Experimental Analysis of Distribution and Abundance*. New York: Harper & Row. p. 142.
25. See Mather, K. (1943). *Statistical Analysis in Biology*. London: Methuen. p. 38.
26. The technique is described in Waid, W. M., Orne, E. C., Cook, M. R., and Orne, M. T. (1981). Meprobamate reduces accuracy of physiological detection of deception. *Science* **212**, 71–73.
27. Fictitious but realistic population, closely resembling the population of males aged 45–59 years as described in Roberts, J. (1975). Blood pressure of persons 18–74 years, United States, 1971–72. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 150. Washington, D.C.: U.S. Department of Health, Education and Welfare.

## Chapter 4

1. Data from the 2003–2004 National Health and Nutrition Examination Survey, which can be found at [www.donofinquiry.com/nhanes/source/choose.php](http://www.donofinquiry.com/nhanes/source/choose.php)
2. Ikeme, A. I., Roberts, C., Adams, R. L., Hester, P. Y., and Stadelman, W. J. (1983). Effects of supplementary water-administered vitamin D<sub>3</sub> on egg shell thickness. *Poultry Science* **62**, 1120–1122. The normal curve was fitted to raw data provided courtesy of W. J. Stadelman and A. I. Ikeme.
3. Hengstenberg, R. (1971). Das Augenmuskelsystem der Stubenfliege *Musca domestica*. 1. Analyse der "clock-spikes" und ihrer Quellen. *Kybernetik* **2**, 56–57.
4. Adapted from Magath, T. B., and Betkson, J. (1960). Electronic blood-cell counting. *American Journal of Clinical Pathology* **34**, 203–213. Actually, the percentage error is somewhat less for high counts and somewhat more for low counts. Described in Coulter Electronics (1982). *Performance Characteristics and Specifications for Coulter Counter Model S-560*. Hialeah, Fl: Coulter Electronics.
5. Fictitious but realistic population. Adapted from data given by Hildebrand, S. F., and Schroeder, W. C. (1927). Fishes of Chesapeake Bay. *Bulletin of the United States Bureau of Fisheries* **43**, Part 1, p. 88. The fish are young of the year, observed in October; they are quite small. (The distribution of lengths in older populations is not approximately normal.)
6. Adapted from Pearl, R. (1905). Biometrical studies on man. I. Variation and correlation in brain weight. *Biometrika* **4**, 13–104.
7. Adapted from Swearingen, M. L., and Halt, D. A. (1976). Using a "blank" trial as a teaching tool. *Journal of Agronomic Education* **5**, 3–8. The standard deviation given in this problem is realistic for an idealized "uniform" field, in which yield differences between plots are due to local random variation rather than large-scale and perhaps systematic variation.
8. Adapted from Coulter Electronics (1982). *Performance Characteristics and Specifications for the Coulter Counter Model S-560*. Hialeah, Fl: Coulter Electronics.
9. Unpublished data courtesy of Susan Whitehead, Oberlin College.
10. Data taken from [www.athlinks.com/results/50228/97027/u1/2008-Rome-Marathon.aspx](http://www.athlinks.com/results/50228/97027/u1/2008-Rome-Marathon.aspx)

11. Unpublished data courtesy of Kaelyn Stiles, Oberlin College.
12. Unpublished data courtesy of Paul Harnik and Lydia Ries, Oberlin College.
13. Summary weather information derived from [www.centralcoastweather.net](http://www.centralcoastweather.net)
14. Summary weather information derived from [www.wrcc.dri.edu/cgi-bin/cliMAIN.pl?akjune](http://www.wrcc.dri.edu/cgi-bin/cliMAIN.pl?akjune)
15. Long, E. C. (1976). *Liquid Scintillation Counting Theory and Techniques*. Irvine, Calif.: Beckman Instruments. The distribution is actually a discrete distribution called a Poisson distribution; however, a Poisson distribution with large mean is approximately normal.
16. Fictitious but realistic population, based on data of Emerson, R. A., and East, E. M. (1913). Inheritance of quantitative characters in maize. *Nebraska Experimental Station Research Bulletin* **2**. Data reproduced by Mather, K. (1943). *Statistical Analysis in Biology*. London: Methuen. pp. 29, 34. Modern hybrid corn is taller and less variable than this population.
17. These percentiles are based on data in the National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics Centers for Disease Control and Prevention. The following URL provides a link to the data table: [www.cdc.gov/nchs/about/major/nhanes/hgtfem.pdf](http://www.cdc.gov/nchs/about/major/nhanes/hgtfem.pdf)
18. This is the standard reference distribution for Stanford-Binet scores. See Sattler, J. M. (1982). *Assessment of Children's Intelligence and Special Abilities*, 2nd ed. Boston: Allyn and Bacon. p. 19 and back cover.
19. Unpublished data courtesy of Forrest Crawford and Yvonne Piper, Oberlin College.
20. D.C.: U.S. Department of Health and Human Services. The normality assumption may or may not be realistic.
21. Adapted from data given in Sebens, K. P. (1981). Recruitment in a sea anemone population; juvenile substrate becomes adult prey. *Science* **213**, 785–787.
22. Fictitious but realistic data. Adapted from distribution given for men aged 45–59 in Roberts, J. (1975). Blood pressure of persons 18–74 years, United States, 1971–72. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 150. Washington, D.C.: U.S. Department of Health, Education and Welfare.
23. Based on data in Roberts, J. D., et al. (1997). Inhaled nitric oxide and persistent pulmonary hypertension of the newborn. *New England Journal of Medicine* **336**, 605–610.
24. The distribution in Figure 5.3.1 is based on data given in Zeleny, C. (1922). The effect of selection for eye facet number in the white bar-eye race of *Drosophila melanogaster*. *Genetics* **7**, 1–115. The data are displayed in Falconer, D. S. (1981). *Introduction to Quantitative Genetics*, 2nd ed. New York: Longman. p. 97.
25. The distribution in Figure 5.3.3 is adapted from data described by Bradley, J. V. (1980). Nonrobustness in one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society* **15** (1), 29–32, used by permission of the Psychonomic Society, Inc.; and Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *American Statistician* **31**, 147–150. Bradley's distribution included additional peaks, because sometimes the subject fumbled the button more than once on a single trial.
26. Fictitious but realistic situation, adapted from data given in Bradley, D. D., Krauss, R. M., Petitte, D. B., Ramcharan, S., and Wingard, I. (1978). Serum high-density lipoprotein cholesterol in women using oral contraceptives, estrogens, and progestins. *New England Journal of Medicine* **299**, 17–20.
27. Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology* **3**, 430–454.
28. Strickberger, M. W. (1976). *Genetics*, 2nd ed. New York: Macmillan. p. 206.
29. [www.cureresearch.com/artic/other\\_important\\_stds\\_niaid\\_fact\\_sheet\\_niaid.htm](http://www.cureresearch.com/artic/other_important_stds_niaid_fact_sheet_niaid.htm)

## Chapter 5

1. Data from the 2003–2004 National Health and Nutrition Examination Survey, which can be found at [www.dnafinquiry.com/nhanes/source/choose.php](http://www.dnafinquiry.com/nhanes/source/choose.php)
2. Fictitious but realistic population. See Example 2.2.11.
3. The mean and standard deviation are realistic for American women aged 25–34. See O'Brien, R. J., and Drizd, T. A. (1981). Basic data on spirometry in adults 25–74 years of age: United States, 1971–75. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 222. Washington,

13. Fictitious but realistic situation. See Waugh, G. D. (1954). The occurrence of *Mytilicola intestinalis* (Steuer) on the east coast of England. *Journal of Animal Ecology* **23**, 364–367.
14. Mosteller, F., and Tukey, J. W. (1977). *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley. p. 25.
15. Fictitious but realistic population, resembling the population of young American men aged 18–24, as described in Abraham, S., Johnson, C. L., and Najjar, M. F. (1979). Weight and height of adults 18–74 years of age: United States 1971–1974. *U.S. National Center for Health Statistics, Series 11*, No. 211. Washington, D.C.: U.S. Department of Health, Education and Welfare.
16. The mean and standard deviation are realistic, based on unpublished data provided courtesy of J. Y. Ustimer and C. A. Mitchell. The normality assumption may or may not be realistic.
17. The mean and standard deviation are realistic, based on unpublished data provided courtesy of S. Newman and D. L. Harris. The normality assumption may or may not be realistic.
18. Adapted from Cherney, J. H., Volenec, J. J., and Nyquist, W. E. (1985). Sequential fiber analysis of forage as influenced by sample weight. *Crop Science* **25**, No. 6 (Nov./Dec. 1985), 1113–1115 (Table 1). By permission of the Crop Science Society of America, Inc. Raw data courtesy of W. E. Nyquist.
19. Dice, L. R. (1932). Variation in the geographic race of the deer mouse, *Peromyscus maniculatus bairdii*. *Occasional Papers of the Museum of Zoology, University of Michigan*, No. 239. Data reproduced in Simpson, G. G., Roe, A., and Lewontin, R. C. (1960). *Quantitative Zoology*. New York: Harcourt, Brace, p. 79.
20. Bodor, N., and Simpkins, J. W. (1983). Redox delivery system for brain-specific, sustained release of dopamine. *Science* **221**, 65–67.
21. Student (W. S. Gosset) (1908). The probable error of a mean. *Biometrika* **6**, 1–25.
22. The Writing Group for the PEPI Trial (1996). Effects of hormone therapy on bone mineral density. *Journal of the American Medical Association* **276**, 1389–1396. This study compared change in bone mineral density over 36 months for four medications and a placebo. (Hip bone mineral density was measured at the beginning of the experiment and again 36 months later.) Only the data for those women who adhered to the experimental protocol are used in the example. Standard deviations are calculated based on the standard errors reported in the article.
23. Data collected by Denise D'Abundo, Oberlin College, April 1991.
24. Bockman, D. E., and Kirby, M. L. (1984). Dependence of thymus development on derivatives of the neural crest. *Science* **233**, 498–500. Copyright 1984 by the AAAs.
25. Brown, S. A., Riviere, J. E., Coppoc, G. L., Hinsman, E. J., Carlton, W. W., and Steckel, R. R. (1985). Single intravenous and multiple intramuscular dose pharmacokinetics and tissue residue profile of gentamicin in sheep. *American Journal of Veterinary Research* **46**, 69–74. Raw data courtesy of S. A. Brown and G. L. Coppoc.
26. Lobstein, D. D. (1983). A multivariate study of exercise training effects on beta-endorphin and emotionality in psychologically normal, medically healthy men. Ph.D. thesis, Purdue University. Raw data courtesy of the author.
27. Nicholson, R. L., and Moraes, W. B. C. (1980). Survival of *Colletotrichum graminicola*: Importance of the spore matrix. *Phytopathology* **70**, 255–261.

## Chapter 6

1. Data provided by Dennis Frey, California Polytechnic State University, San Luis Obispo.
2. Newman, S., Everson, D. O., Gunsett, F. C., and Christian, R. E. (1984). Analysis of two- and three-way crosses among Ramhouillet, Targhee, Columbia, and Suffolk sheep for three preweaning traits. Unpublished manuscript. Raw data courtesy of S. Newman.
3. Adapted from the following two papers. Potkin, S. G., Cannon, H. E., Murphy, D. L., and Wyatt, R. J. (1978). Are paranoid schizophrenics biologically different from other schizophrenics? *New England Journal of Medicine* **298**, 61–66. Murphy, D. L., Wright, C., Buchsbaum, M., Nichols, A., Costa, J. L., and Wyatt, R. J. (1976). Platelet and plasma amine oxidase activity in 680 normals: Sex and age differences and stability over time. *Biochemical Medicine* **16**, 254–265. The data displayed are fictitious but realistic, having been reconstructed from the histograms and summary information given by Potkin et al. and Murphy et al.
4. Based on data reported in Rea, T. M., Nash, J. F., Zabik, J. E., Born, G. S., and Kessler, W. V. (1984). Effects of toluene inhalation on brain biogenic amines in the rat. *Toxicology* **31**, 143–150.
5. Based on an experiment by M. Morales.

16. Adapted from Morris, J. G., Gripe, W. S., Chapman, H. L., Jr., Walker, D. F., Armstrong, J. B., Alexander, J. D., Jr., Miranda, R., Sanchez, A., Jr., Sanchez, B., Blair-West, J. R., and Denton, D. A. (1984). Selenium deficiency in cattle associated with Heinz bodies and anemia. *Science* **223**, 491–492. Copyright 1984 by the AAAS.
17. Shaffer, P. L., and Rock, G. C. (1983). Tufted apple budmoth (*Lepidoptera: Tortricidae*): Effects of constant daylengths and temperatures on larval growth rate and determination of larval-pupal ecdysis. *Environmental Entomology* **12**, 76–80.
18. Bishop, N. J., Morley, R., Day, J. P., and Lucas, A. L. (1997). Aluminum neurotoxicity in preterm infants receiving intravenous-feeding solutions. *New England Journal of Medicine* **336**, 1557–1561.
19. Kaufman, J. S., Reda, D. J., Fye, C. L., Goldfarb, D. S., Henderson, W. G., Kleinman, J. G., and Vaamonde, C. A. (1998). Subcutaneous compared with intravenous epoetin in patients receiving hemodialysis. *New England Journal of Medicine* **339**, 578–583.
20. Based on data provided by C. H. Noller.
21. This is roughly the SD for the U.S. population of middle-aged men. See Moore, F. E., and Gordon, T. (1973). Serum cholesterol levels in adults, United States 1960–62. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 22. Washington, D.C.: U.S. Department of Health, Education and Welfare.
22. Pappas, T., and Mitchell, C. A. (1984). Effects of seismic stress on the vegetative growth of *Glycine max* (L.) Merr. cv. Wells II. *Plant, Cell and Environment* **8**, 143–148.
23. Noll, S. L., Waibel, P. E., Cook, R. D., and Witmer, J. A. (1984). Biopotency of methionine sources for young turkeys. *Poultry Science* **63**, 2458–2470.
24. Schaeffer, J., Andrysiak, T., and Ungerleider, J. T. (1981). Cognition and long-term use of ganja (cannabis). *Science* **213**, 465–466.
25. Desai, R. (1982). An anatomical study of the canine male and female pelvic diaphragm and the effect of testosterone on the status of the levator ani of male dogs. *Journal of the American Animal Hospital Association* **18**, 195–202.
26. Nicholson, R. L., and Moraes, W. B. C. (1980). Survival of *Colletotrichum graminicola*: Importance of the spore matrix. *Phytopathology* **70**, 255–261. Raw data courtesy of R. L. Nicholson.
27. The probabilities in Table 6.5.2 were estimated by computer simulation carried out by M. Samuels and R. P. Becker. The standard error of each probability estimate is less than 0.0015. The sources of the parent distributions are given in Notes 7 and 8 to Chapter 5.
28. Burnett, A., and Haywood, A. (1997). A statistical analysis of differences in sediment yield over time on the West Branch of the Black River. Unpublished manuscript, Oberlin College.
29. Hessell, E. A., Johnson, D. D., Ivey, T. D., and Miller, D. W. (1980). Membrane vs bubble oxygenator for cardiac operations. *Journal of Thoracic and Cardiovascular Surgery* **80**, 111–122.
30. Peters, H. G. and Bademan, H. (1963). The form and growth of stellate cells in the cortex of the guinea-pig. *Journal of Anatomy (London)* **97**, 111–117.
31. Kaneto, A., Kosaka, K., and Nakao, K. (1967). Effects of stimulation of the vagus nerve on insulin secretion. *Endocrinology* **80**, 530–536. Copyright © 1967 by the Endocrine Society.
32. Simmons, F. J. (1943). Occurrence of superparasitism in *Nemeritis canescens*. *Revue Canadienne de Biologie* **2**, 15–40. Data reproduced in Williams, C. B. (1964). *Patterns in the Balance of Nature*. London: Academic Press. p. 223.
33. These data are diameters at breast height of American Sycamore trees in the floodplain of the Vermilion River. Data collected Emily Norland, Oberlin College, March 1995.
34. Adapted from Sanders, K. (2004). A quantitative, vegetative, and reproductive comparison of *Centromadia parryi* ssp. *congdonii* in two locations. Master's Thesis, California Polytechnic State University.
35. Hunter, A., and Terasaki, T. (1993). Statistical analysis comparing vital capacities of brass majors in the Conservatory and a normal population. Unpublished manuscript, Oberlin College. All subjects were men, age 18–21, with heights between 175 and 183 cm. Because vital capacity is related to height, the raw data were adjusted slightly, using linear regression, to control for the effect of height.
36. Chang, K. (2005). Randomized controlled trial of Coblation versus electrocautery tonsillectomy. *American Academy of Otolaryngology—Head and Neck Surgery* **132**, 273–280.
37. Knight, S. L., and Mitchell, C. A. (1983). Enhancement of lettuce yield by manipulation of light and nitrogen nutrition. *Journal of the American Society for Horticultural Science* **108**, 750–754. Raw data courtesy of the authors. (The actual sample sizes were equal; some observations have been omitted from the exercise.)

38. O'Marra, S. (1996). Antibacterial soaps: Myth or reality. Unpublished manuscript, Oberlin College. The primary purpose of this study was to assess the effectiveness of antibacterial soaps. A solution made from antibacterial soap killed all *E. coli*, in contrast to the non-antibacterial soap and the control. The soap solution was a 1:4 solution of soap and water.
39. Ahern, T. (1998). Statistical analysis of EIN plants treated with ancyminidol and H<sub>2</sub>O. Unpublished manuscript, Oberlin College. The mutant strain EIN (e-longated in-ternode) of *Brassica* was used in this experiment. The data presented here are a randomly selected subset of the full data set.
40. Hagerman, A. E., and Nicholson, R. L. (1982). High-performance liquid chromatographic determination of hydroxycinnamic acids in the maize mesocotyl. *Journal of Agricultural and Food Chemistry* **30**, 1098–1102. Reprinted with permission. Copyright 1982 American Chemical Society.
41. Patel, C., Marmot, M. M., and Terry, D. J. (1981). Controlled trial of biofeedback-aided behavioral methods in reducing mild hypertension. *British Medical Journal* **282**, 2005–2008.
42. Lipsky, J. J., Lewis, J. C., and Novick, W. J., Jr. (1984). Production of hypoprothrombinemia by Moxalactam and 1-methyl-5-thiotetrazole in rats. *Antimicrobial Agents and Chemotherapy* **25**, 380–381.
43. Long, T. F., and Murdock, L. L. (1983). Stimulation of blowfly feeding behavior by octopaminergic drugs. *Proceedings of the National Academy of Sciences* **80**, 4159–4163. Raw data courtesy of the authors and L. C. Sudlow.
44. Gwynne, D. T. (1981). Sexual difference theory: Mormon crickets show role reversal in mate choice. *Science* **213**, 779–780. Copyright 1981 by the AAAS. Data provided courtesy of the author.
45. Appel, L. J., et al. (1997). A clinical trial of the effects of dietary patterns on blood pressure. *New England Journal of Medicine* **336**, 1117–1124.
46. Crawford, F., and Piper, Y. (1999). How does caffeine influence heart rate? Unpublished manuscript, Oberlin College. There were 10 subjects in the caffeine group, but an outlier was deleted from the data.
47. Gent, A. (1999). Unpublished data collected at Oberlin College. The colors of light were created using gels: thin pieces of colored plastic used in theater lighting.
48. Parks, N. J., Krohn, K. A., Mathis, C. A., Chasko, J. H., Geiger, K. R., Gregor, M. E., and Peek, N. F. (1981). Nitrogen-13-labelled nitrite and nitrate: Distribution and metabolism after intratracheal administration. *Science* **212**, 58–61. Copyright 1981 by the AAAS. Raw data courtesy of N. J. Parks.
49. Krick, J. A. (1982). Effects of seeding rate on culm diameter and the inheritance of culm diameter in soft red winter wheat (*Triticum aestivum* L. em Thell). Master's thesis, Department of Agronomy, Purdue University. Raw data courtesy of J. A. Krick and H. W. Ohm. Each diameter is the mean of measurements taken at six prescribed locations on the stem.
50. Data collected by Deborah Ignatoff, Oberlin College, spring 1997.
51. Bailey, J., and Marshall, J. (1970). The relationship of the post-ovulatory phase of the menstrual cycle to total cycle length. *Journal of Biosocial Science* **2**, 123–132.
52. Nansen, C., Tchabi, A., and Meikle, W. G. (2001). Successional sequence of forest types in a disturbed dry forest reserve in southern Benin, West Africa. *Journal of Tropical Ecology* **17**, 525–539.
53. Unpublished data courtesy of W. F. Jacobson.
54. Dale, E. M., and Housley, T. L. (1986). Sucrose synthase activity in developing wheat endosperms differing in maximum weight. *Plant Physiology* **82**, 7–10. Raw data courtesy of the authors.
55. See Note 23 of Chapter 3.
56. Adapted from data courtesy of the Morro Bay National Estuary Foundation, 2009.
57. Graph created from data included in Erne, P., Bolli, P., Buergisser, E., and Buehler, F. R. (1984). Correlation of platelet calcium with blood pressure. *New England Journal of Medicine* **310**, 1084–1088. Reprinted by permission. Raw data courtesy of F. R. Buehler.
58. Urban L. E., et al. (2010). The accuracy of stated energy contents of reduced-energy, commercially prepared foods. *J. Am. Diet Assoc.* **110**, 116–123.

## Chapter 7

1. Kotler, D. (2000). A comparison of aerobics and modern dance training on health-related fitness in college women. Unpublished manuscript, Oberlin College.
2. Pappas, T., and Mitchell, C. A. (1985). Effects of seismic stress on the vegetative growth of *Glycine max* (L.) Merr. cv. Wells II. *Plant, Cell and Environment* **8**, 143–148. Reprinted with permission of Blackwell Scientific Publications Limited. Raw data courtesy of the authors. The original experiment included many treatments and more than nine observations per group; only a subset of the data is presented here, for simplicity.

3. Unpublished data courtesy of J. A. Henricks and V. J. K. Liu.
4. Rea, T. M., Nash, J. F., Zabik, J. E., Born, G. S., and Kessler, W. V. (1984). Effects of toluene inhalation on brain biogenic amines in the rat. *Toxicology* **31**, 143–150. Raw data courtesy of J. F. Nash and J. E. Zabik.
5. Sagan, C. (1977). *The Dragons of Eden*. New York: Ballantine. p. 7.
6. Lemenager, R. P., Nelson, L. A., and Hendrix, K. S. (1980). Influence of cow size and breed type on energy requirements. *Journal of Animal Science* **51**, 566–576. Some of the animals *lost* weight during the 78 days, so that the mean weight gains are based on both positive and negative values.
7. Adapted from Miyada, V. S. (1978). Uso da levedura seca de distilarias de alcool de cana de acucar na alimentacao de suinos em crescimento e acabamento. Master's thesis, University of Sao Paulo, Brazil.
8. Kalsner, S., and Richards, R. (1984). Coronary arteries of cardiac patients are hyperreactive and contain stores of amines: A mechanism for coronary spasm. *Science* **223**, 1435–1437. Copyright 1984 by the American Association for the Advancement of Science (AAAS).
9. Adapted from Dybas, H. S., and Lloyd, M. (1962). Isolation by habitat in two synchronized species of periodical cicadas (Homoptera, Cicadidae, *Magicicada*). *Ecology* **43**, 444–459.
10. Namdar, M., Koepfli, P., Grathwohl, R., Siegrist, P. T., Klainguti, M., Schepis, T., Deleloye, R., Wyss, C. A., Gaemperli, O., and Kaufmann, P. A. (2006). Caffeine decreases exercise-induced myocardial flow reserve. *Journal of the American College of Cardiology* **47**, 405–410. Raw data read from Figure 1.
11. Bockman, D. E., and Kirby, M. L. (1984). Dependence of thymus development on derivatives of the neural crest. *Science* **223**, 498–500. Copyright 1984 by the AAAS.
12. Triepel, R. R., and Mitchell, C. A. (1984). Metabolic response of river birch and European birch roots to hypoxia. *Plant Physiology* **76**, 31–35. Raw data courtesy of the authors.
13. Lamke, L. O., and Liljedahl, S. O. (1976). Plasma volume changes after infusion of various plasma expanders. *Resuscitation* **5**, 93–102.
14. Anderson, J. W., Story, L., Sieling, B., Chen, W. J. L., Petro, M. S., and Story, J. (1984). Hypocholesterolemic effects of oat-bran or bean intake for hypercholesterolemic men. *The American Journal of Clinical Nutrition* **40**, 1146–1155.
15. Ahne, A., and Myers, S. (1999). The effect of Miracle Grow on radish growth. Unpublished manuscript, Oberlin College. The data presented here are a subset of the full data set. (The means and standard deviations for the full data set are similar to those for the subset presented here. In particular, the sample mean for the control group is greater than for the fertilizer group.)
16. Borg, E. (2008). A comparison of *Orconectes rusticus* and *O. sanbornii* weight. Unpublished manuscript, Oberlin College.
17. Heald, F. (1974). Hematocrit values of youths 12–17 years, United States. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 146. Washington, D.C.: U.S. Department of Health, Education and Welfare. Actually, the data were obtained by a sampling scheme more complicated than simple random sampling.
18. Long, T. F., and Murdock, L. L. (1983). Stimulation of blowfly feeding behavior by octopaminergic drugs. *Proceedings of the National Academy of Sciences* **80**, 4159–4163. Raw data courtesy of the authors and L. C. Sudlow.
19. Yerushalmy, J. (1971). The relationship of parents' cigarette smoking to outcome of pregnancy—implications as to the problem of inferring causation from observed associations. *American Journal of Epidemiology* **93**, 443–456.
20. Gould, S. J. (1981). *The Mismeasure of Man*. New York: Norton. pp. 50ff. The SDs were estimated from the ranges reported by Gould.
21. Yerushalmy, J. (1972). Infants with low birth weight born before their mothers started to smoke cigarettes. *American Journal of Obstetrics and Gynecology* **112**, 277–284.
22. Anderson, G. D., Blidner, I. N., McClemont, S., and Sinclair, J. C. (1984). Determinants of size at birth in a Canadian population. *American Journal of Obstetrics and Gynecology* **150**, 236–244.
23. Mochizuki, M., Marno, T., Masuko, K., and Ohtsu, T. (1984). Effects of smoking on fetoplacental-maternal system during pregnancy. *American Journal of Obstetrics and Gynecology* **149**, 413–420.
24. Wainright, R. L. (1983). Change in observed birth weight associated with a change in maternal cigarette smoking. *American Journal of Epidemiology* **117**, 668–675.
25. Moore, R. M., Diamond, E. L., and Cavalieri, R. L. (1988). The relationship of birth weight and intrauterine diagnostic ultrasound exposure. *Obstetrics and Gynecology* **71**, 513–517.

26. Waldenstrom, U., Nilsson, S., Fall, O., Axelsson, O., Eklund, G., Lindeberg, S., and Sjodin, Y. (1988). Effects of routine one-stage ultrasound screening in pregnancy: A randomized clinical trial. *Lancet* (10 Sept.), 585–588.
27. National Center for Health Statistics. Data are taken from Table LCWK9 ([www.cdc.gov/nchs/datawh/statab/unpubd/mortabs/lcwk9\\_10.htm](http://www.cdc.gov/nchs/datawh/statab/unpubd/mortabs/lcwk9_10.htm)).
28. Cook, L. S., Daling, J. R., Voigt, L. F., deHart, M. P., Malone, K. E., Stanford, J. L., Weiss, N. S., Brinton, L. A., Gammon, M. D., and Brogan, D. (1997). Characteristics of women with and without breast augmentation. *Journal of the American Medical Association* **277**, 1612–1617.
29. LaCroix, A. Z., Mead, L. A., Liang, K., Thomas, C. B., and Pearson, T. A. (1986). Coffee consumption and the incidence of coronary heart disease. *New England Journal of Medicine* **315**, 977–982.
30. Yerushalmy, J., and Hilleboe, H. E. (1957). Fat in the diet and mortality from heart disease. *New York State Journal of Medicine* **57**, 2343–2354. Reprinted by permission. Copyright by the Medical Society of the State of New York.
31. *Cleveland Plain Dealer*, 10 February 1999, page 17-A.
32. David, R. J., and Collins, J. W. (1997). Differing birth weight among infants of U.S.-born blacks, African-born blacks, and U.S.-born whites. *New England Journal of Medicine* **337**, 1209–1214. Low birth weight means a weight of less than 1500 g, which the authors referred to as “very low birth weight” in the article.
33. Gwilym, S., Howard, D. P. J., Davies, N., and Willett, K. (2005). Harry Potter casts a spell on accident prone children. *British Medical Journal* **331**, 1505–1506. The authors note that the weather on “Harry Potter” weekends was good and was not appreciably different from weather on other weekends.
34. Adapted from data provided courtesy of D. R. Shields and D. K. Colby. See Shields, D. R. (1981). The influence of niacin supplementation on growing ruminants and *in vivo* and *in vitro* rumen parameters. Ph.D. thesis, Purdue University.
35. Schall, J. J., Bennett, A. F., and Putnam, R. W. (1982). Lizards infected with malaria: Physiological and behavioral consequences. *Science* **217**, 1057–1059. Copyright 1982 by the AAAS.
36. Agosti, E., and Camerota, G. (1965). Some effects of hypnotic suggestion on respiratory function. *International Journal of Clinical and Experimental Hypnosis* **13**, 149–156.
37. Adapted from Knight, S. L., and Mitchell, C. A. (1983). Enhancement of lettuce yield by manipulation of light and nitrogen nutrition. *Journal of the American Society for Horticultural Science* **108**, 750–754.
38. Rickard, I. J. (2008). Offspring are lighter at birth and smaller in adulthood when born after a brother versus a sister in humans. *Evolution and Human Behavior* **29**, 196–200.
39. Unpublished data courtesy of J. L. Wolfson.
40. Fictitious but realistic data.
41. Massey, R. L. (2010). A randomized trial of rocking-chair motion on the effect of postoperative ileus duration in patients with cancer recovering from abdominal surgery. *Applied Nursing Research* **23**, 59–64.
42. Shima, J. S. (2001). Recruitment of a coral reef fish: Roles of settlement, habitat, and postsettlement losses. *Ecology* **82**, 2190–2199. Raw data courtesy of the author.
43. Adapted from Williams, G. Z., Widdowson, G. M., and Penton, J. (1978). Individual character of variation in time-series studies of healthy people. II. Differences in values for clinical chemical analytes in serum among demographic groups, by age and sex. *Clinical Chemistry* **24**, 313–320.
44. Fictitious but realistic data. See Abraham, S., Johnson, C. L., and Najjar, M. F. (1979). Weight and height of adults 18–74 years of age, United States 1971–74. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 211. Washington, D.C.: U.S. Department of Health, Education and Welfare.
45. Example communicated by D. A. Holt.
46. Petrie, B., and Segalowitz, S. J. (1980). Use of fetal heart rate, other perinatal and maternal factors as predictors of sex. *Perceptual and Motor Skills* **50**, 871–874. Copyright 1980 by Ammons Scientific, Ltd. Reproduced with permission of Ammons Scientific, Ltd. via Copyright Clearance Center.
47. Hagerman, A. E., and Nicholson, R. L. (1982). High-performance liquid chromatographic determination of hydroxycinnamic acids in the maize mesocotyl. *Journal of Agricultural and Food Chemistry* **30**, 1098–1102. Copyright 1982 American Chemical Society. Reprinted with permission.
48. Ressler, S. (1977) AnthroKids—Anthropometric data of children. Data are taken from the file individuals.csv at [ovrt.nist.gov/projects/anthrokids/](http://ovrt.nist.gov/projects/anthrokids/)

49. Adapted from Williams, G. Z., Widdowson, G. M., and Penton, J. (1978). Individual character of variation in time-series studies of healthy people. II. Difference in values for clinical chemical analytes in serum among demographic groups, by age and sex. *Clinical Chemistry* **24**, 313–320. Reprinted by permission.
50. Hamill, P. V. V., Johnston, F. E., and Lemeshow, S. (1973). Height and weight of youths 12–17 years, United States. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 124. Washington, D.C.: U.S. Department of Health, Education and Welfare.
51. Phelan, S., and Schaffner, A. (2009). NIH grant proposal: Prevention of postpartum weight retention in low-income WIC women.
52. Roberts, J. (1975). Blood pressure of persons 18–74 years, United States, 1971–72. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 150. Washington, D.C.: U.S. Department of Health, Education and Welfare. However, the distribution of systolic blood pressure is more skewed (see Exercise 5.2.18).
53. Pearson, E. S., and Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika* **62**, 223–241.
54. Mena, E. A., Kossovsky, N., Chu, C., and Hu, C. (1995). Inflammatory intermediates produced by tissues encasing silicone breast implants. *Journal of Investigative Surgery* **8**, 31–42. [Note: There were two control groups in this study. The control group included in this analysis is “patients undergoing reverse augmentation mammoplasty” (the “scar” group discussed in the article). Also, the authors neglected to transform the data before conducting a *t* test. Thus, they got a large *P*-value, although they noted that the two groups looked quite different.]
55. Fictitious but realistic data. Based on unpublished data provided by Bill Plummer.
56. Fierer, N. (1994). Statistical analysis of soil respiration rates in a light gap and surrounding old-growth forest. Unpublished manuscript, Oberlin College.
57. Noether, G. E. (1967). *Elements of Nonparametric Statistics*. New York: Wiley.
58. It is sometimes stated that the validity of the Mann–Whitney test requires that the two population distributions have the same shape and differ only by a shift. This is not correct. The computations underlying Table 6 require only that the common population distribution (under the null hypothesis) be continuous. A further property, technically called *consistency* of the test, requires that the two distributions be *stochastically ordered*, which is the technical way of saying that one of the variables has a consistent tendency to be larger than the other. In fact, the title of Mann and Whitney’s original paper is “On a test of whether one of two random variables is stochastically larger than the other” (*Annals of Mathematical Statistics* **18**, 1947). In Section 7.12 we discuss the requirement of stochastic ordering, calling it an “implicit assumption.” (The confidence interval procedure mentioned at the end of Section 7.10 does require the stronger assumption that the distributions have the same shape.)
59. Zimmerman, D. W., and Zumbo, B. D. (1993). The relative power of parametric and nonparametric statistical methods in G. Keren and C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. pp. 481–517. Hillsdale, N.J.: Lawrence Erlbaum Associates. The authors use simulations to show that the Wilcoxon–Mann–Whitney test is more powerful than the *t* test in the presence of outliers, but that in the absence of outliers, the *t* test is slightly preferable for a variety of population distributions.
60. Agosti, E., and Camerota, G. (1965). Some effects of hypnotic suggestion on respiratory function. *International Journal of Clinical and Experimental Hypnosis* **13**, 149–156.
61. Connolly, K. (1968). The social facilitation of preening behaviour in *Drosophila melanogaster*. *Animal Behaviour* **16**, 385–391.
62. Unpublished data courtesy of G. P. Carlson and M. A. Morse.
63. Lobstein, D. D. (1983). A multivariate study of exercise training effects on beta-endorphin and emotionality in psychologically normal, medically healthy men. Ph.D. thesis, Purdue University. Raw data courtesy of the author.
64. Erne, P., Bolli, P., Buergisser, E., and Buehler, F. R. (1984). Correlation of platelet calcium with blood pressure. *New England Journal of Medicine* **310**, 1084–1088. Reprinted by permission of the *New England Journal of Medicine*. Summary statistics calculated from raw data provided courtesy of F. R. Buehler.
65. Adapted from unpublished data provided by F. Delgado. The extremely high somatic cell counts probably represent cases of mastitis.
66. Pappas, T., and Mitchell, C. A. (1985). Effects of seismic stress on the vegetative growth of *Glycine max* (L.) Merr. cv. Wells II. *Plant, Cell and Environment* **8**, 143–148. Reprinted with permission of Blackwell Scientific Publications Limited. Raw data courtesy

- of the authors. The original experiment included more than two treatment groups.
67. Wee, K. (1995). Species diversity in floodplain forests. Unpublished manuscript, Oberlin College.
68. Cicirelli, M. F., Robinson, K. R., and Smith, L. D. (1983). Internal pH of *Xenopus* oocytes: A study of the mechanism and role of pH changes during meiotic maturation. *Developmental Biology* **100**, 133–146. Raw data courtesy of M. F. Cicirelli.
69. Manski, T. J., et al. (1997). Endolymphatic sac tumors: A source of morbid hearing loss in von Hippel-Lindau disease. *Journal of the American Medical Association* **277**, 1461–1466.
70. Unpublished data courtesy of J. A. Henricks and V. J. K. Liu.
71. Schall, J. J., Bennett, A. F., and Putnam, R. W. (1982). Lizards infected with malaria: Physiological and behavioral consequences. *Science* **217**, 1057–1059. Copyright 1982 by the AAAS. Raw data courtesy of J. J. Schall.
72. Unpublished data courtesy of M. B. Nichols and R. P. Maickel.
73. The neonatal inhaled nitric oxide study group (1997). Inhaled nitric oxide in full-term and nearly full-term infants with hypoxic respiratory failure. *New England Journal of Medicine* **336**, 597–604.
74. Gleason, P. P., et al. (1997). Medical outcomes and antimicrobial costs with the use of American Thoracic Society guidelines for outpatients with community-acquired pneumonia. *Journal of the American Medical Association* **278**, 32–39.
75. Hodapp, M. (1998). A Study of CDS Nutrition. Unpublished manuscript, Oberlin College.
76. Laurance, W. F., Perez-Salicrup, D., Delamonica, P., Fearside, P. M., D'Angelo, S., Jerozolinski, A., Pohl, L., and Lovejoy, T. E. (2001). Rain forest fragmentation and the structure of Amazonian liana communities. *Ecology* **82**, 105–116. The data presented were read by J. Witmer from Figure 2 in the paper and may not be completely accurate.
77. King, D. S., Sharp, R. L., Vukovich, M. D., Brown, G. A., Reifenrath, T. A., Uhl, N. L., and Parsons, K. A. Effect of oral androstenedione on serum testosterone and adaptations to resistance training in young men. *Journal of the American Medical Association* **281**, 2020–2028. Raw data courtesy of the authors. The response variable shown here is change in “maximum muscle strength,” which is the greatest weight the subject could lift. There were several other measurements taken in the experiment; generally they showed the same results seen in the lat pulldown data. A primary purpose of the experiment was to study the effect of andro on testosterone level. The researchers found that andro had no effect on serum testosterone level.
78. Fleming, M. F., Barry, K. L., Manwell, L. B., Johnson, K., and London, R. (1997). Brief physician advice for problem alcohol drinkers. *Journal of the American Medical Association* **277**, 1039–1045.
79. Conner, E. M., Sperling, R. S., Gerber, R., Kisalev, P., Scott, G., O'Sullivan, M. J., Van Dyke, R., Bey, M., Shearer, W., Jacobsen, R. L., Jimenez, E., O'Neill, E., Bazin, B., Delfraissy, J.-F., Culname, M., Coombs, R., Elkins, M., More, J., Stratton, P., and Balsley, J. (1994). Reduction of maternal-infant transmission of Human Immunodeficiency Virus Type I with zidovudine treatment. *New England Journal of Medicine* **331**, 1173–1180. Some people feel that this study should not have been conducted as a randomized experiment, since there was reason to believe that AZT would be helpful in preventing the transfer of HIV to the babies and since HIV is such a serious disease.
80. Gattinoni, L., Tognoni, G., Pesenti, A., Taccone, P., Mascheroni, D., Labarta, V., Malacrida, R., Di Giulio, P., Fumagalli, R., Pelosi, P., Brazzi, L., and Latini, R. (2001). Effect of prone positioning on the survival of patients with acute respiratory failure. *New England Journal of Medicine* **345**, 568–573.
81. Petitti, D. B., Perlman, J. A., and Sidney, S. (1987). Noncontraceptive estrogens and mortality: Long-term follow-up of women in the Walnut Creek Study. *Obstetrics & Gynecology* **70**, 289–293.

## Chapter 8

- Namdar, M., Koepfli, P., Grathwohl, R., Siegrist, P. T., Klainguti, M., Schepis, T., Delaloye, R., Wyss, C. A., Fleischmann, S. P., Gaemperli, O., and Kaufmann, P. A. (2006). Caffeine decreases exercise-induced myocardial flow reserve. *Journal of the American College of Cardiology* **47**, 405–410.
- Sargent, P. A., Sharpley, A. L., Williams, C., Goodall, E. M., and Cowen, P. J. (1997). 5-HT<sub>2C</sub> receptor activation decreases appetite and body weight in obese subjects. *Psychopharmacology* **133**, 309–312. Hunger ratings were recorded “on 10 cm visual analogue scales.”
- Unpublished data courtesy of R. Buchman. The data were collected in Oberlin, Ohio, during the spring of 2001.
- Day, K. M., Patterson, F. L., Luetkemeier, O. W., Ohm, H. W., Polizotto, K., Roberts, J. J., Shaner, G. E.,

- Huber, D. M., Finney, R. E., Foster, J. E., and Gallun, R. L. (1980). Performance and adaptation of small grains in Indiana. *Station Bulletin*, No. 290. West Lafayette, Ind.: Agricultural Experiment Station of Purdue University. Raw data provided courtesy of W. E. Nyquist. The actual trial included more than two varieties.
5. Unpublished data courtesy of C. H. Noller.
  6. Cicirelli, M. F., and Smith, L. D. (1985). Cyclic AMP levels during the maturation of *Xeno pus* oocytes. *Developmental Biology* **108**, 254–258. Raw data courtesy of M. F. Cicirelli.
  7. Judge, M. D., Aberle, E. D., Cross, H. R., and Schanbacher, B. D. (1984). Thermal shrinkage temperature of intramuscular collagen of bulls and steers. *Journal of Animal Science* **59**, 706–709. Raw data courtesy of the authors and E. W. Mills.
  8. Swedo, S. E., Leonard, H. L., Rapoport, J. L., Lenane, M. C., Goldberger, E. L., and Cheslow, B. S. (1989). A double-blind comparison of clomipramine and desipramine in the treatment of trichotillomania (hair pulling). *New England Journal of Medicine* **321**, 497–501.
  9. Unpublished data courtesy of A. Ladavac. The data were collected in Oberlin, Ohio, in November 1996.
  10. In a study in which there is no natural pairing (for example, if identical twins are not available), one may wish to take two equal size groups and create pairs by using covariates such as age and weight. If an experiment is conducted in which members of a pair are randomly assigned to opposite treatment groups, then a paired data analysis has good properties. However, if an observational study is conducted (so that there is no random assignment within pairs), then a paired analysis, such as a paired *t* test, will tend to underestimate the true variability of the difference being studied and the true Type I error rate of a *t* test will be greater than the nominal level of the test. For discussion, see David, H. A., and Gunnink, J. L. (1997). The paired *t* test under artificial pairing. *The American Statistician* **51**, 9–12.
  11. Schriewer, H., Guennewig, V., and Assmann, G. (1983). Effect of 10 weeks endurance training on the concentration of lipids and lipoproteins as well as on the composition of high-density lipoproteins in blood serum. *International Journal of Sports Medicine* **4**, 109–115. Reprinted with permission of Georg Thieme Verlag KG.
  12. Data from experiments reported in several papers, for example, Fout, G. S., and Simon, E. H. (1983). Antiviral activities directed against wild-type and interferon-sensitive mengovirus. *Journal of General Virology* **64**, 1543–1555. Raw data courtesy of E. H. Simon. The unit of measurement is proportional to the number of plaques formed by the virus on a monolayer of mouse cells. Because they are obtained by a serial dilution technique, the measurements have varying numbers of significant digits; the final zeroes of the three-digit numbers are not significant digits.
  13. Adapted from Batchelor, J. R., and Hackett, M. (1970). HL-A matching in treatment of burned patients with skin allografts. *Lancet* **2**, 581–583.
  14. Sallan, S. E., Cronin, C., Zelen, M., and Zinberg, N. E. (1980). Antiemetics in patients receiving chemotherapy for cancer. *New England Journal of Medicine* **302**, 135–138. Reprinted by permission.
  15. Koh, K. K., Mincemoyer, R., Bui, M. N., Csako, G., Pucino, F., Guetta, V., Waclawiw, M., and Cannon, R. O. (1997). Effects of hormone replacement therapy on fibrinolysis in postmenopausal women. *New England Journal of Medicine* **336**, 683–690. Raw data courtesy of K. K. Koh.
  16. Rosenzweig, M. R., Bennett, E. L., and Diamond, M. C. (1972). Brain changes in response to experience. *Scientific American* **226**, No. 2, 22–29. Also Bennett, E. L., Diamond, M. C., Krech, D., and Rosenzweig, M. R. (1964). Chemical and anatomical plasticity of brain. *Science* **146**, 610–619. Copyright 1964 by the American Association for the Advancement of Science.
  17. Richens, A., and Ahmad, S. (1975). Controlled trial of valproate in severe epilepsy. *British Medical Journal* **4**, 255–256.
  18. Wiedenmann, R. N., and Rabenold, K. N. (1987). The effects of social dominance between two subspecies of dark-eyed juncos, *Junco hyemalis*. *Animal Behavior* **35**, 856–864. Raw data courtesy of the authors.
  19. Masty, J. (1983). Innervation of the equine small intestine. Master's thesis, Purdue University. Raw data courtesy of the author.
  20. Golden, C. J., Gruber, B., Blose, I., Berg, R., Coffman, J., and Block, S. (1981). Difference in brain densities between chronic alcoholic and normal control patients. *Science* **211**, 508–510. Raw data courtesy of C. J. Golden. Copyright 1981 by the AAAS.
  21. Data from Namdar, M., et al. The experiment described in Example 8.1.1 was conducted under simulated high altitude, whereas the experiment described in Exercise 8.5.7 was conducted under conditions that mimic being at sea level.
  22. Patel, C., Marmot, M. G., and Terry, D. J. (1981). Controlled trial of biofeedback-aided behavioural

- methods in reducing mild hypertension. *British Medical Journal* **282**, 2005–2008.
23. Forde, O. H., Knutsen, S. F., Arnesen, E., and Thelle, D. S. (1985). The Tromso heart study: Coffee consumption and serum lipid concentrations in men with hypercholesterolaemia: A randomised intervention study. *British Medical Journal* **290**, 893–895. (The sample sizes are unequal because the 25 no-coffee men actually represent three different treatment groups, which followed the same regimen for the first five weeks of the study and different regimens thereafter.)
  24. Dalvit, S. P. (1981). The effect of the menstrual cycle on patterns of food intake. *American Journal of Clinical Nutrition* **34**, 1811–1815.
  25. Unpublished data courtesy of D. J. Honor and W. A. Vestre.
  26. Sesin, G. P. (1984). Pharmacokinetic dosing of Tobramycin sulfate. *American Pharmacy* **NS24**, 778. Vakoutis, J., Stein, G. E., Miller, P. B., and Clayman, A. E. (1981). Aminoglycoside monitoring program. *American Journal of Hospital Pharmacy* **38**, 1477–1480. Copyright 1981, American Society of Hospital Pharmacists, Inc. All rights reserved. Reprinted with permission.
  27. Jovan, S. (2000). Catnip bonanza. *Stats*, No. 27, 25–27.
  28. Dale, E. M., and Housley, T. L. (1986). Sucrose synthase activity in developing wheat endosperms differing in maximum weight. *Plant Physiology* **82**, 7–10. Raw data courtesy of the authors.
  29. Unpublished data courtesy of M. Heithaus and D. Rogers. The samples were taken from the Vermilion River in northern Ohio during the spring of 1995.
  30. Salib, N. M. (1985). The effect of caffeine on the respiratory exchange ratio of separate submaximal arms and legs exercise of middle distance runners. Master's thesis, Purdue University.
  31. Adapted from Bodian, D. (1947). Nucleic acid in nerve-cell regeneration. *Symposia of the Society for Experimental Biology*, No. 1, *Nucleic Acid*, 163–178. Used with permission from The Society for Experimental Biology.
  32. Knowlen, G. G., Kittleson, M. D., Nachreiner, R. F., and Eyster, G. E. (1983). Comparison of plasma aldosterone concentration among clinical status groups of dogs with chronic heart failure. *Journal of the American Veterinary Medical Association* **183**, 991–996.
  33. Robinson, L. R. (1985). The effects of electrical fields on wound healing in *Notophthalmus viridescens*. Master's thesis, Purdue University. Raw data courtesy of the author and J. W. Vanable, Jr.
  34. Agosti, E., and Camerota, G. (1965). Some effects of hypnotic suggestion on respiratory function. *International Journal of Clinical and Experimental Hypnosis* **13**, 149–156. The experiment actually included a third phase.
  35. Koh, K. K., Mincemoyer, R., Bui, M. N., Csako, G., Pucino, F., Guetta, V., Waclawiw, M., and Cannon, R. O. (1997). Effects of hormone replacement therapy on fibrinolysis in postmenopausal women. *New England Journal of Medicine* **336**, 683–690. Raw data courtesy of K. K. Koh.
  36. Savin, V. J., Sharma, R., Sharma, M., McCarthy, E. T., Swan, S. K., Ellis, E., Lovell, H., Warady, B., Gunwar, S., Chonko, A. M., Artero, M., and Vincenti, F. (1996). Circulating factor associated with increased glomerular permeability to albumin in recurrent focal segmental glomerulosclerosis. *New England Journal of Medicine* **334**, 878–883. Raw data courtesy of V. J. Savin.

## Chapter 9

1. White, A. S., Godard, R. D., Belling, C., Kasza, V., and Beach, R. L. (2010). Beverages obtained from soda fountain machines in the U.S. contain microorganisms, including coliform bacteria. *International Journal of Food Microbiology* **137**, 61–66.
2. From the National Survey of Family Growth (2002). U.S. Dept. of Health and Human Services, *Vital and Health Statistics*, Series 23, No. 25. Data are taken from Table 75.
3. Agresti, A., and Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126. The authors show that 95% confidence intervals based on  $\tilde{p}$  are superior to other commonly used confidence intervals. They also note that if one uses  $\tilde{p}$ , then it is not necessary to construct tables or rules for how large the sample size needs to be in order for the confidence interval to have good coverage properties.
4. Couch, F. J., et al. (1997). *BRCA1* mutations in women attending clinics that evaluate the risk of breast cancer. *New England Journal of Medicine* **336**, 1409–1415.
5. Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Statistical Science* **4**, 298–306. The ECMO data are discussed in greater detail in Section 10.4.
6. Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9**, 97–113.
7. Adapted from McCloskey, R. V., Goren, R., Bissett, D., Bentley, J., and Tutlane, V. (1982). Cefotaxime in

- the treatment of infections of the skin and skin structure. *Reviews of Infectious Diseases* **4**, Supp., S444–S447.
8. Adapted from Petras, M. L. (1967). Studies of natural populations of *Mus*. III. Coat color polymorphisms. *Canadian Journal of Genetic Cytology* **9**, 287–296.
  9. Miller, C. L., Pollock, T. M., and Clewer, A. D. F. (1974). Whooping-cough vaccination: An assessment. *The Lancet* **ii**, 510–513.
  10. Erskine, A. G., and Socha, W. W. (1978). *The Principles and Practices of Blood Grouping*. St. Louis: Mosby, p. 209.
  11. Curtis, H. (1983). *Biology*, 4th ed. New York: Worth, p. 908.
  12. Mourant, A. E., Kopec, A. C., and Domaniewska-Sobczak, K. (1976). *The Distribution of Human Blood Groups and Other Polymorphisms*, 2nd ed. London: Oxford University Press, p. 44.
  13. Based on an experiment described in Oellerman, C. M., Patterson, F. L., and Gallun, R. L. (1983). Inheritance of resistance in “Luso” wheat to Hessian fly. *Crop Science* **23**, 221–224.
  14. Cogswell, M. E., Looker, A. C., Pfeiffer, C. M., Cook, J. D., Lacher, D. A., Beard, J. L., Lynch, S. R., and Grummer-Strawn, L. M. (2009). Assessment of iron deficiency in US preschool children and nonpregnant females of childbearing age: National Health and Nutrition Examination Survey 2003–2006. *American Journal of Clinical Nutrition* **89**, 1334–1342.
  15. Hayes, D. L., et al. (1997). Interference with cardiac pacemakers by cellular telephones. *New England Journal of Medicine* **336**, 1473–1479. The data cited are for CDMA telephones. Although interference was recorded in 15.7% of the tests, a much smaller percentage of the tests caused symptoms that were clinically significant.
  16. Duggan, D. J., Gorospe, J. R., Fanin, M., Hoffman, E. P., and Angelini, C. (1997). Mutations in the sarcoglycan genes in patients with myopathy. *New England Journal of Medicine* **336**, 618–624.
  17. Rabenold, K. R., and Rabenold, P. P. (1985). Variation in altitudinal migration, winter segregation, and site tenacity in two subspecies of dark-eyed juncos in the Southern Appalachians. *The Auk* **102**, 805–819.
  18. Fictitious but realistic data based on a personal communication with F. Villablanca.
  19. Saeidi, G., and Rowland, G. G. (1997) The inheritance of variegated seed color and palmitic acid in flax. *Journal of Heredity* **88**, 466–468.
  20. Phillips, D. P., and Smith, D. G. (1990). Postponement of death until symbolically meaningful occasions. *Journal of the American Medical Association* **263**, 1947–1951. For comparison purposes, the authors also examined deaths among elderly Jewish women during the same time period; they did not find any excess of deaths after the festival for this comparison group.
  21. Sinnott, E. W., and Durham, G. B. (1922). Inheritance in the summer squash. *Journal of Heredity* **13**, 177–186.
  22. Adapted from Gould, J. L. (1985). How bees remember flower shapes. *Science* **227**, 1492–1494. Figure copyright 1985 by the American Association for the Advancement of Science; used by permission.
  23. Adapted from 1983 birth data for West Lafayette, Indiana.
  24. Bateson, W., and Saunders, E. R. (1902). *Reports to the Evolution Committee of the Royal Society* **1**, 1–160. Feather color and comb shape are controlled independently; white feather is dominant and small comb is dominant. The parents in the experiment were first-generation hybrids ( $F_1$ ) and thus were necessarily double heterozygotes.
  25. This is a realistic value. See Exercise 3.6.6.
  26. Jakkula, L. R., Knault, D. A., and Gorbet, D. W. (1997). Inheritance of a shriveled seed trait in peanut. *Journal of Heredity* **88**, 47–51. The data are taken from Table 5 of the paper.
  27. Adapted from Mantel, N., Bohidar, N. R., and Ciminera, J. L. (1977). Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of inter-litter information. *Cancer Research* **37**, 3863–3868. (A more powerful analysis, which uses the partially informative triplets, is described in the paper.)
  28. Adapted from Jacobs, G. H. (1978). Spectral sensitivity and colour vision in the ground-dwelling sciurids: Results from golden mantled ground squirrels and comparisons for five species. *Animal Behaviour* **26**, 409–421. See also Jacobs, G. H. (1981). *Comparative Color Vision*, Academic Press.
  29. Petrij, F., van Veen, K., Mettler, M., and Bruckmann, V. (2001). A second acromelanistic allelomorph at the albino locus of the Mongolian gerbil (*Meriones unguiculatus*). *Journal of Heredity* **92**, 74–78. The gerbils we call “brown” are referred to as “Siamese” by the authors.

30. Kaitz, M. (1992). Recognition of familiar individuals by touch. *Physiology and Behavior* **52**, 565–567.
31. Fawole, I. (2001). Genetic analysis of mutations at loci controlling leaf form in cowpea (*Vigna unguiculata* [L.] Walp.). *Journal of Heredity* **92**, 43–50. These data come from the 1993a generation listed in Table 8 of the article. The types we call I, II, and III are identified by the authors as trifoliolate, trifoliolate orbicular, and unifoliolate orbicular.
32. Baur, E., Fischer, E., and Lenz, F. (1931). *Human Heredity*, 3rd ed. New York: Macmillan, p. 52.
33. This is typical for U.S. populations. See, for example, MacCready, R. A., and Mannin, M. C. (1951). A typing study of one hundred and fifty thousand bloods. *Journal of Laboratory and Clinical Medicine* **37**, 634–636.
34. Fictitious but realistic situation. See Krebs, C. J. (1972). *Ecology: The Experimental Analysis of Distribution and Abundance*. New York: Harper and Row.
35. Cheatum, F. L., and Severinghaus, C. W. (1950). Variations in fertility of white-tailed deer related to range conditions. *Transactions of the North American Wildlife Conference* **15**, 170–189.
36. Ziegler, E. E., Nelson, S. E., and Jeter, J. M. (2009). Iron supplementation of breastfed infants from an early age. *American Journal of Clinical Nutrition* **89**, 525–532.
37. Fischer, C., and Fischer, U. (1997). Analysis of cork taint in wine and cork material at olfactory sub-threshold levels by solid phase microextraction. *Journal of Agricultural and Food Chemistry* **45**, 1995–1997.
38. Freeland, W. J. (1981). Parasitism and behavioral dominance among male mice. *Science* **213**, 461–462. Copyright 1981 by the AAAS.
39. Collins, R. L. (1970). The sound of one paw clapping: An inquiry into the origin of left-handedness. In Lindzey, G., and Thiessen, D. D. (eds.). *Contributions to Behavior-Genetic Analysis: The Mouse as Prototype*. Appleton-Century-Crofts.
40. Fawole, I. op cit. These data are from Table 3 of the paper.
41. Floersheim, G. L., Weber, O., Tschumi, P., and Ulbrich, M. (1983). Research cited in *Scientific American* **248** (April 1983), No. 4, p. 75.
42. Fuchs, J. A., Smith, J. D., and Bird, L. S. (1972). Genetic basis for an 11:5 dihybrid ratio observed in *Gossypium hirsutum*. *Journal of Heredity* **63**, 300–303. The genetic basis for the 13:3 and 11:5 ratios is explained in Strickberger, M. W. (1976). *Genetics*, 2nd ed. New York: Macmillan, pp. 206–208.
43. Adapted from Goodyear, C. P. (1970). Terrestrial and aquatic orientation in the starhead top-minnow, *Fundulus noti*. *Science* **168**, 603–605. Copyright 1970 by the AAAS.
44. See Batschelet, E. (1981). *Circular Statistics in Biology*. Academic Press.
45. Unpublished data courtesy J. L. Wolfson, collected at Bard College in 1997.
46. Paris, H. S. (1997). Genes for developmental fruit coloration of acorn squash. *Journal of Heredity* **88**, 52–56. The experiment included crossing Table Queen squash (TQE) with Vegetable Spaghetti (VSP). The data presented in the exercise are from a back-cross of VSP with a TQE × VSP cross.
47. Chen, P., Ma, G., Buss, G. R., Gunduz, I., Roane, C. W., and Tolin, S. A. (2001). Inheritance and allelism tests of Raiden soybean for resistance to soybean mosaic virus. *Journal of Heredity* **92**, 51–55. The “resistant” classification includes both resistant and systemic necrotic plants.
48. Lamb, M. L., Fishbein, M., Douglas, J. M., Rhodes, F., Rogers, J., Bolan, G., Zenilman, J., Hoxworth, T., Malotte, K., Iatesta, M., Kent, C., Lentz, A., Graziano, S., Byers, R. H., and Peterman, T. A. (1998). Efficacy of risk-reduction counseling to prevent human immunodeficiency virus and sexually transmitted diseases. *Journal of the American Medical Association* **280**, 1161–1167.

## Chapter 10

1. Guyuron, B., Reed, D., Kriegler, J., Davis, J., Pashmini, N., and Amini, S. (2009). A placebo-controlled surgical trial of the treatment of migraine headaches. *Plastic and Reconstructive Surgery* **124**, 461–468.
2. Unpublished data courtesy of D. Wallace, collected at Oberlin College in the fall of 1995.
3. Brodie, E. D., Jr., and Brodie, E. D. III. (1980). Differential avoidance of mimetic salamanders by free-ranging birds. *Science* **208**, 181–182. Copyright 1980 by the AAAS.
4. Karban, R., Adamchak, R., and Schnathorst, W. C. (1987). Induced resistance and interspecific competition between spider mites and a vascular wilt fungus. *Science* **235**, 678–680. Copyright 1987 by the AAAS.
5. Inskip, P. D., Targone, R. E., Hatch, E. E., Wilcosky, T. C., Shapiro, W. R., Selker, R. G., Fine, H. A., Black, P. M., Loeffler, J. S., and Linet, M. S. (2001). Cellular-telephone use and brain tumors. *New England Journal of Medicine* **344**, 79–86. The data are taken from Table 4 of the paper.

6. Turnbull, D. M., Rawlins, M. D., Weightman, D., and Chadwick, D. W. (1982). A comparison of phenytoin and valproate in previously untreated adult epileptic patients. *Journal of Neurology, Neurosurgery, and Psychiatry* **45**, 55–59.
7. Unpublished data courtesy of W. Singleton and K. Hendrix.
8. Mizutani, T., and Mitsuoka, T. (1979). Effect of intestinal bacteria on incidence of liver tumors in gnotobiotic C3H/He male mice. *Journal of the National Cancer Institute* **63**, 1365–1370.
9. Selawry, O. S. (1974). The role of chemotherapy in the treatment of lung cancer. *Seminars in Oncology* **1**, 259–272.
10. Kannus, P., Parkkari, J., Niemi, S., Pasanen, M., Palvanen, M., Jarvinen, M., and Vuori, I. (2000). Prevention of hip fracture in elderly people with use of a hip protector. *New England Journal of Medicine* **343**, 1506–1513.
11. Cohen, S., Doyle, W. J., Skoner, D. P., Rabin, B. S., and Gwaltney, J. M. (1997). Social ties and susceptibility to the common cold. *Journal of the American Medical Association* **277**, 1940–1944.
12. Sherman, D. G., Atkinson, R. P., Chippendale, T., Levin, K. A., Ng, K., Futrell, N., Hsu, C. Y., and Levy, D. E. (2000). Intravenous ancrod for treatment of acute ischemic stroke. *Journal of the American Statistical Association* **283**, 2395–2403.
13. Gueguen, N. (2009). The receptivity of women to courtship solicitation across the menstrual cycle: A field experiment. *Biological Psychology* **80**, 321–324.
14. Adapted from Ammon, O. (1899). *Zur Anthropologie der Badener*. Jena: G. Fischer. Ammon's data appear in Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* **49**, 732–764. Light hair was blonde or red; dark hair was brown or black. Light eyes were blue, grey, or green; dark eyes were brown.
15. Cruz-Coke, R. (1970). *Color Blindness; An Evolutionary Approach*. Springfield, Ill.: Thomas.
16. Bill-Alexson, A., et al. for the Scandinavian Prostate Cancer Study Group No. 4 (2005). Radical prostatectomy versus watchful waiting in early prostate cancer. *New England Journal of Medicine* **352**, 1977–1984.
17. Adapted from Porac, C., and Coren, S. (1981). *Lateral Preferences and Human Behavior*. New York: Springer-Verlag. The frequencies given are approximate, having been deduced from percent-ages on pages 36 and 45. People with neutral preference were counted as Left.
18. Upton, G., and Fingleton, B. (1985). *Spatial Data Analysis by Example: Point Pattern and Quantitative Data*, Vol. 1. New York: Wiley, p. 230. Adapted from Diggle, P. J. (1979). Statistical methods for spatial point patterns in ecology, pp. 95–150 in *Spatial and Temporal Analysis in Ecology*, R. M. Cormack and J. K. Ord (eds.). Fairland, Md.: International Cooperative Publishing House.
19. Based on an article by the Writing Group for Bypass Angioplasty Revascularization Investigation (BARI) Investigators (1997). See five-year clinical and functional outcome comparing bypass surgery and angioplasty in patients with multivessel coronary disease. *Journal of the American Medical Association* **277**, 715–721.
20. These data are fictitious, but the proportions of left-handed males and females are realistic and the independence of the twins is in agreement with published data. See Porac, C., and Coren, S. (1981). *Lateral Preferences and Human Behavior*. New York: Springer-Verlag, p. 36; and Morgan, M. C., and Corballis, M. J. (1978). On the biological basis of human laterality: I. Evidence for a maturational left-right gradient. *The Behavioral and Brain Sciences* **2**, p. 274.
21. Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Statistical Science* **4**, 298–306. There is controversy surrounding this experiment. An earlier experiment using a nonstandard randomization scheme had shown ECMO to be highly effective. Thus, some statisticians question whether this second experiment was necessary. For a discussion of these issues see the articles on pages 306–340 that follow the Ware article in *Statistical Science* **4**.
22. Remus, J. K., and Zahren, L. (1995). An investigation of the influenza virus at Oberlin College. Unpublished manuscript, Oberlin College. This study actually involved more students than are reported here. For simplicity, we restrict attention to the 41 students who had at least two colds during the 1994–1995 school year.
23. Hurt, R. D., Sachs, D. P. L., Glover, E. D., Offord, K. P., Johnston, J. A., Dale, L. C., Khayrallah, M. A., Schroeder, D. R., Glover, P. N., Sullivan, C. R., Croghan, I. T., and Sullivan, P. M. (1997). A comparison of sustained-release bupropion and placebo for smoking cessation. *New England Journal of Medicine* **337**, 1195–1202.
24. Unpublished data courtesy of B. Rogers, collected at the Oberlin College Conservatory of Music in the spring of 1991.

25. Soutou, B., Juhl, H., Hackenbruck, J., Rockseisen, M., Klomp, H.-J., Raulais, D., Vigny, M., and Wellstein, A. (1998). Relationship between serum concentrations of the growth factor pleiotrophin and pleiotrophin-positive tumors. *Journal of the National Cancer Institute* **90**, 1468–1473.
26. Bent, S., Kane, C., Katsuto, S., Neuhaus, J., Hudes, E. S., Goldberg, H., and Avins, A. L. (2006). Saw palmetto for benign prostatic hyperplasia. *New England Journal of Medicine* **354**, 557–566.
27. Hogarty, G. E., Kornblith, S. J., Greenwald, D., DiBarry, A. L., Cooley, S., Ulrich, R. F., Carter, M., and Flesher, S. (1997). Three-year trials of personal therapy among schizophrenic patients with or independent of family, I: Description of study and effects on relapse rates. *American Journal of Psychiatry* **154**, 1504–1513.
28. Dreitz, V. J. (2009). Parental behavior of a precocial species: Implications for juvenile survival. *Journal of Applied Ecology* **46**, 870–878.
29. Adapted from Ammon, O. (1899). *Zur Anthropologie der Badener*. Jena: G. Fischer. Ammon's data appear in Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* **49**, 732–764.
30. Clegg, D. O., et al. (2006). Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis. *New England Journal of Medicine* **354**, 795–808.
31. Inglesfield, C., and Begon, M. (1981). Open-ground individuals and population structure in *Drosophila subobscura* Collin. *Biological Journal of the Linnean Society* **15**, 259–278.
32. Aird, I., Bentall, H. H., Mehigan, J. A., and Roberts, J. A. F. (1954). The blood groups in relation to peptic ulceration and carcinoma of colon, rectum, breast, and bronchus: An association between the ABO blood groups and peptic ulceration. *British Medical Journal* **ii**, 315–321.
33. Govind, C. K., and Pearce, J. (1986). Differential reflex activity determines claw and closer muscle asymmetry in developing lobsters. *Science* **233**, 354–356. Copyright 1986 by the AAAS.
34. LeBars, P. L., Katz, M. M., Berman, N., Itil, T. M., Freedman, A. M., and Schatzberg, A. F. (1997). A placebo-controlled, double-blind, randomized trial of an extract of Gingko biloba for dementia. *Journal of the American Medical Association* **278**, 1327–1332.
35. Unpublished data courtesy of L. Solimine.
36. Hudson, J. I., McElroy, S. L., Raymond, N. C., Crow, S., Keck, P. E., Carter, W. P., Mitchell, J. E., Strakowski, S. M., Pope, H. G., Coleman, B. S., and Jonas, J. M. (1998). Fluvoxamine in the treatment of binge-eating disorder: A multicenter placebo-controlled, double-blind trial. *American Journal of Psychiatry* **155**, 1756–1762. The response variable has ordered categories, so there are more powerful methods, beyond the scope of this text, that can be used to analyze the data.
37. Boden, W. E., O'Rourke, R. A., Teo, K. K., Hartigan, P. M., Maron, D. J., Kostuk, W. J., et al. COURAGE Trial Research Group. (2007). Optimal medical therapy with or without PCI for stable coronary disease. *New England Journal of Medicine* **356**, 1503–1516.
38. Wolfson, J. L. (1987). Impact of *Rhizobium* nodules on *Sitona hispidulus*, the clover root curculio. *Entomologia Experimentalis et Applicata* **43**, 237–243. Data courtesy of the author. The experiment actually included 11 dishes.
39. Adapted from Paige, K. N., and Whitham, T. G. (1985). Individual and population shifts in flower color by scarlet gilia: A mechanism for pollinator tracking. *Science* **227**, 315–317. The raw data given are fictitious but have been constructed to agree with the summary statistics given by Paige and Whitham.
40. Brusco, N. K., Shields, N., Taylor, N. F., and Paratz, J. (2007). A Saturday physiotherapy service may decrease length of stay in patients undergoing rehabilitation in hospital: A randomised controlled trial. *Australian Journal of Physiotherapy* **53**, 75–81.
41. Beck, S. L., and Gavin, D. L. (1976). Susceptibility of mice to audiogenic seizures is increased by handling their dams during gestation. *Science* **193**, 427–428. Copyright 1976 by the AAAS.
42. Pittet, P. G., Acheson, K. J., Wuersch, P., Maeder, E., and Jequier, E. (1981). Effects of an oral load of partially hydrolyzed wheatflour on blood parameters and substrate utilization in man. *The American Journal of Clinical Nutrition* **34**, 2438–2445.
43. Agresti, A., and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician* **54**, 280–288. Agresti and Caffo conduct a series of simulations which show that adding 1 to each cell results in good coverage properties when the sample sizes,  $n_1$  and  $n_2$ , are as small as 10. Unpublished calculations done by J. Witmer show that these good properties are also obtained when  $n_1$  and  $n_2$  are as small as 5, provided  $p_1$  and  $p_2$  are not both close to 0 or both close to 1, in which case the interval becomes quite conservative (i.e., the coverage rate approaches 100% for a nominal 95% confidence interval).

44. Agresti, A. Personal communication.
45. Lyles, K. W., et al. for the HORIZON Recurrent Fracture Trial (2007). Zolendronic acid and clinical hip fractures and mortality after hip fracture. *New England Journal of Medicine* **357**, 1799–1809.
46. Saunders, M. C., Dick, J. S., Brown, I. M., McPherson, K., and Chalmers, I. (1985). The effects of hospital admission for bed rest on the duration of twin pregnancy: A randomised trial. *The Lancet* **ii**, 793–795.
47. Lader, M., and Scotto, J.-C. (1998). A multicentre double-blind comparison of hydroxyzine, buspirone and placebo in patients with generalized anxiety disorder. *Psychopharmacology* **139**, 402–406. Improvement is taken to be a 50% or greater reduction in Hamilton Anxiety (HAM-A) score. There was a third treatment group in this study, which we are ignoring here.
48. Nesheim, S. R., Shaffer, N., Vink, P., Thea, D. M., Palumbo, P., Greenberg, B., Weedon, J., and Simmons, R. J. (1996). Lack of increased risk for perinatal human immunodeficiency virus transmission to subsequent children born to infected women. *Pediatric Infectious Disease Journal* **15**, 886–890.
49. Collaborative Group for the Study of Stroke in Young Women (1973). Oral contraception and increased risk of cerebral ischemia or thrombosis. *New England Journal of Medicine* **288**, 871–878. Reprinted by permission.
50. Johnson, S. K., and Johnson, R. E. (1972). Tonsillectomy history in Hodgkin's disease. *New England Journal of Medicine* **287**, 1122–1125.
51. Rillich, J., Buhl, E., Schildberger, K., and Stevenson, P. A. (2009). Female crickets are driven to fight by the male courting and calling songs. *Animal Behavior* **77**, 737–742.
52. Sidney, S., Tekawa, I. S., and Friedman, G. D. (1993). A prospective study of cigarette tar yield and lung cancer. *Cancer Causes and Control* **4**, 3–10.
53. The steering committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine* **318**, 262–264.
54. Witsberger, T. H., Villamil, J. A., Schultz, L. G., Hahn, A. W., and Cook, J. L. (2008). Prevalence of and risk factors for hip dysplasia and cranial cruciate ligament deficiency in dogs. *Journal of the American Veterinary Medical Association* **232**, 1818–1824.
55. Zwerling, C., Whitten, P. S., Davis, C. S., and Sprince, N. L. (1997). Occupational injuries among workers with disabilities. *Journal of the American Medical Association* **278**, 2163–2166. In this study an injury means an occupational injury in the year preceding when the person was interviewed.
56. Kernan, W. N., Viscoli, C. M., Brass, L. M., Broderick, J. P., Brott, T., Feldmann, E., Morgenstern, L. B., Wilterdink, J. L., and Horwitz, R. I. (2000). Phenylpropanolamine and the risk of hemorrhagic stroke. *New England Journal of Medicine* **343**, 1826–1832.
57. Cohen, M., Demers, C., Gurkinkel, E. P., Turpie, A. G. G., Fromell, G. J., Goodman, S., Langer, A., Calif, R. M., Fox, K. A. A., Premmereur, J., and Bigonzi, F. (1997). A comparison of low-molecular-weight heparin with unfractionated heparin for unstable coronary artery disease. *New England Journal of Medicine* **337**, 447–452. A negative outcome here is taken to be death, myocardial infarction, or recurrent angina during the first 14 days after treatment.
58. Howard, B. V., et al. (2006). Low-fat dietary pattern and risk of cardiovascular disease. *Journal of the American Medical Association* **295**, 655–666.
59. Conover, D. O., and Kynard, B. E. (1981). Environmental sex determination: Interaction of temperature and genotype in a fish. *Science* **213**, 577–579. Copyright 1981 by the AAAS.
60. Carson, J. L., Collier, A. M., and Hu, S. S. (1985). Acquired ciliary defects in nasal epithelium of children with acute viral upper respiratory infections. *New England Journal of Medicine* **312**, 463–468. Reprinted by permission.
61. Larson, E. B., Roach, R. C., Schoene, R. B., and Hombein, T. F. (1982). Acute mountain sickness and acetazolamide. *Journal of the American Medical Association* **248**, 328–332. Copyright 1982 American Medical Association.
62. Kluger, M. J., Ringler, D. H., and Anver, M. R. (1975). Fever and survival. *Science* **188**, 166–168. Copyright 1975 by the AAAS. The original article contains a misprint, but Dr. Kluger has kindly provided the correct mortality at 40 °C.
63. Ragaz, J., Jackson, S. M., Le, N., Plenderleith, I. H., Spinelli, J. J., Basco, V. E., Wilson, K. S., Knowling, M. A., Coppin, C. M. L., Paradis, M., Coldman, A. J., and Olivotto, I. A. (1997). Adjuvant radiotherapy and chemotherapy in node-positive premenopausal women with breast cancer. *New England Journal of Medicine* **337**, 956–962.
64. Englund, J. A., Baker, C. J., Raskino, C., McKinney, R. E., Petrie, B., Fowler, M. G., Pearson, D., Gershon, A., McSherry, G. D., Abrams, E. J., Schliozberg, J.,

- and Sullivan, J. L. (1997). Zidovudine, didanosine, or both as the initial treatment for symptomatic HIV-infected children. *New England Journal of Medicine* **336**, 1704–1712. The data presented here are for an interim analysis that was conducted approximately two years into the study. As a result of the interim analysis of death rates and of rates of disease progression, the use of zidovudine alone was stopped before the end of the trial.
65. Gupta, M., and Chordhuri, A. N. R. (1980). Relationship between ABO blood groups and malaria. *Bulletin of the World Health Organization* **58**, 913–915.
  66. Shorrocks, B., and Nigro, L. (1981). Microdistribution and habitat selection in *Drosophila subobscura* collin. *Biological Journal of the Linnean Society* **16**, 293–301.
  67. Malacrida, R., Genoni, M., Maggioni, A. P., Spatato, V., Parish, S., Palmer, A., Collins, R., and Moccetti, T. (1998). A comparison of the early outcome of acute myocardial infarction in women and men. *New England Journal of Medicine* **338**, 8–14. Although the odds ratio for these data shows that men are more likely to survive than are women, the authors discuss the effect that age has on this finding. They calculate a new odds ratio after adjusting for age and other covariates (using methods that are beyond the scope of this text) and conclude that much of the difference in survival probability is due to these covariates.
  68. Mochizuki, M., Marno, T., Masuko, K., and Ohtsu, T. (1984). Effects of smoking on fetoplacental-maternal system during pregnancy. *American Journal of Obstetrics and Gynecology* **149**, 413–420.
  69. Redelmeier, D. A., and Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. *New England Journal of Medicine* **336**, 453–458. Also see Redelmeier, D. A., and Tibshirani, R. J. (1997). Is using a car phone like driving drunk? *Chance* **10**, No. 2, 5–9.
  70. Monto, A. S., Ohmit, S. E., Petrie, J. G., Johnson, E., Truscon, R., Teich, E., Rotthoff, J., Boulton, M., and Victor, J. (2009). Comparative efficacy of inactivated and live attenuated influenza vaccines. *New England Journal of Medicine* **361**, 1260–1267.
  2. Shields, D. R. (1981). The influence of niacin supplementation on growing ruminants and *in vivo* and *in vitro* rumen parameters. Ph.D. thesis, Purdue University. Adapted from raw data provided courtesy of the author and D. K. Colby.
  3. Adapted from Potkin, S. G., Cannon, H. E., Murphy, D. L., and Wyatt, R. J. (1978). Are paranoid schizophrenics biologically different from other schizophrenics? *New England Journal of Medicine*, **298**, 61–66. Reprinted by permission. The calculations are based on the data in Example 1.1.4 in this book, which are an approximate reconstruction from the histograms and summary information given by Potkin et al.
  4. Adapted from Keller, S. E., Weiss, J. M., Schleifer, S. J., Miller, N. E., and Stein, M. (1981). Suppression of immunity by stress: Effect of a graded series of stressors on lymphocyte stimulation in the rat. *Science* **213**, 1397–1400. Copyright 1981 by the AAAS. The SDs and SSs were estimated from the SEs given by Keller et al.
  5. Lobstein, D. D. (1983). A multivariate study of exercise training effects on beta-endorphin and emotionality in psychologically normal, medically healthy men. Ph.D. thesis, Purdue University. Raw data courtesy of the author.
  6. Hayden, F. G., Osterhaus, A. D., Treanor, J. J., Fleming, D. M., Aoki, F. Y., Nicholson, K. G., Bohnen, A. M., Hirst, H. M., Keene, O., and Wightman, K. (1997). Efficacy and safety of the neuraminidase inhibitor zanamivir in the treatment of influenza virus infections. *New England Journal of Medicine* **337**, 874–880. The sums of squares have been calculated from the means and SDs given in the paper.
  7. Person, A. (1999). Daffodil stem lengths. Unpublished manuscript, Oberlin College. The full data set is somewhat larger than that presented here.
  8. Kotler, D. (2000). A comparison of aerobics and modern dance training on health-related fitness in college women. Unpublished manuscript, Oberlin College.
  9. Unpublished data courtesy of H. W. Ohm.
  10. Cameron, E., and Pauling, L. (1978). Supplemental ascorbate in the supportive treatment of cancer: Re-evaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Science USA* **75**, 4538–4542.
  11. Neumann, A., Richards, A. -L., and Randa, J. (2001). Effects of acid rain on alfalfa plants. Unpublished manuscript, Oberlin College. The low acid group

## Chapter 11

1. Martinez, J. (1998). Organic practices for the cultivation of sweet corn. Unpublished manuscript, Oberlin College. For pedagogical purposes, the data presented here are a random sample from a larger study. The nematode used was *Steinernema carpocapsae*, the bacterium was *Bacillus thuringiensis*, and the wasp was *Trichogramma pretiosum*.

- was given three drops of 1.5 M HCL as well as two droppers full of water each day. For the high acid group 3.0 M HCL was used. The control group was only given water. The original data have been modified slightly for pedagogical purposes.
12. This is the design described in the following papers. Rosenzweig, M. R., Bennett, E. L., and Diamond, M. C. (1972). Brain changes in response to experience. *Scientific American* **226**, No. 2, 22–29. Bennett, E. L., Diamond, M. C., Krech, D., and Rosenzweig, M. R. (1964). Chemical and anatomical plasticity of brain. *Science* **146**, 610–619.
  13. Based on an experiment by Resh, W., and Stoughton, R. B. (1976). Topically applied antibiotics in acne vulgaris. *Archives of Dermatology* **112**, 182–184.
  14. Swearingen, M. L., and Holt, D. A. (1976). Using a “blank” trial as a teaching tool. *Journal of Agronomic Education* **5**, 3–8. Reprinted by permission of the American Society of Agronomy, Inc via Copyright Clearance Center. In fact, in order to demonstrate the variability of plot yields, the experimenters planted the same variety of barley in all 16 plots.
  15. Data adapted from McIntosh, A. R., and Townsend, C. R. (1996). Interactions between fish, grazing invertebrates and algae in a New Zealand stream: A trophic cascade mediated by fish-induced changes to grazer behavior. *Oecologia* **108**, 174–181.
  16. Pappas, T., and Mitchell, C. A. (1985). Effects of seismic stress on the vegetative growth of *Glycine max* (L.) Merr. cv. Wells II. *Plant, Cell and Environment* **8**, 143–148. Raw data courtesy of the authors. The original experiment included more than four treatments. Reprinted with permission of Plant, Cell, and Environment.
  17. Garcia-Nebot, M., Alegria, A., Barbera, R., Clemente, G., and Romero, F. (2010). Addition of milk or caseinophosphopeptides to fruit beverages to improve iron bioavailability. *Food Chemistry* **119**, 141–148.
  18. Kiesecker, J. M., Blaustein, A. R., and Belden, L. K. (2001). Complex causes of amphibian population declines. *Nature* **410**, 681–684. Sample means and standard deviation were read from Figure 2a in the article.
  19. Tripepi, R. R., and Mitchell, C. A. (1984). Metabolic response of river birch and European birch roots to hypoxia. *Plant Physiology* **76**, 31–35. Raw data courtesy of the authors.
  20. Adapted from Veterans Administration Cooperative Study Group on Antihypertensive Agents (1979). Comparative effects of ticrynafen and hydrochlorothiazide in the treatment of hypertension. *New England Journal of Medicine* **301**, 293–297. Reprinted by permission. The value of  $s_{\text{pooled}}$  was calculated from the SEs given in the paper. Copyright © 1979 Massachusetts Medical Society. All rights reserved.
  21. Knight, S. L., and Mitchell, C. A. (1983). Enhancement of lettuce yield by manipulation of light and nitrogen nutrition. *Journal of the American Society for Horticultural Science* **108**, 750–754. Calculations based on raw data provided by the authors.
  22. Fictitious but realistic data, adapted from O'Brien, R. J., and Drizd, T. A. (1981). Basic data on spirometry in adults 25–74 years of age: United States, 1971–75. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 222. Washington, D.C.: U.S. Department of Health and Human Services.
  23. U.S. Bureau of the Census. The 2008 age distribution is taken from [www.census.gov/compendia/statab/cats/population/estimates\\_and\\_projections\\_by\\_age\\_sex\\_raceethnicity.html](http://www.census.gov/compendia/statab/cats/population/estimates_and_projections_by_age_sex_raceethnicity.html). The percentages have been rounded to sum to 1.
  24. Chrisman, C. L., and Baumgartner, A. P. (1980). Micronuclei in bone-marrow cells of mice subjected to hyperthermia. *Mutation Research* **77**, 95–97. The original experiment included six treatments.
  25. Baird, J. T., and Quinlivan, L. G. (1973). Parity and hypertension. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 38. Washington, D.C.: U.S. Department of Health, Education and Welfare.
  26. U.S. Bureau of the Census (1997). *Statistical Abstract of the United States*, 1997 117th ed. Washington, D.C.: U.S. Government Printing Office.
  27. Adapted from Witelson, S. F. (1985). The brain connection: The corpus callosum is larger in left-handers. *Science* **239**, 665–668. Copyright 1985 by the AAAS. The SDs and MS (within) have been calculated from the standard errors given by Witelson. Reprinted with permission from AAAS.
  28. Adapted from Booth, D. M., and Heck, K. L., (2009). Effects of the American oyster *Crassostrea virginica* on growth rates of the seagrass *Halodule wrightii*. *Marine Ecology Progress Series* **389**, 117–126. The article has the full data set; we present only part of the data for pedagogical purposes.
  29. Latimer, J. (1985). Adapted from unpublished data provided by the investigator.

30. Adapted from Morris, J. G., Cripe, W. S., Chapman, H. L., Jr., Walker, D. F., Armstrong, J. B., Alexander, J. D., Jr., Miranda, R., Sanchez, A., Jr., Sanchez, B., Blair-West, J. R., and Denton, D. A. (1984). Selenium deficiency in cattle associated with Heinz bodies and anemia. *Science* **223**, 491–492. Copyright 1984 by the AAAS. The MS(within) is fictitious but agrees with the standard errors given by Morris et al.
31. Fictitious but realistic data. Adapted from Mizutani, T., and Mitsuoka, T. (1979). Effect of intestinal bacteria on incidence of liver tumors in gnotobiotic C3H/He male mice. *Journal of the National Cancer Institute* **63**, 1365–1370.
32. Becker, W. A. (1961). Comparing entries in random sample tests. *Poultry Science* **40**, 1507–1514.
33. Adapted from Rosner, B. (1982). Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics* **38**, 105–114. Reprinted with permission from The International Biometric Society. The medical study is reported in Berson, E. L., Rosner, B., and Simonoff, E. (1980). An outpatient population of retinitis pigmentosa and their normal relatives: Risk factors for genetic typing and detection derived from their ocular examinations. *American Journal of Ophthalmology* **89**, 763–775. The means and sums of squares have been estimated from data given by Rosner, after estimating missing values for two patients for whom only one eye was measured.
34. Heggestad, H. E., and Bennett, J. H. (1981). Photochemical oxidants potentiate yield losses in snap beans attributable to sulfur dioxide. *Science* **213**, 1008–1010. Copyright 1981 by the AAAS. Raw data courtesy of H. E. Heggestad.
35. Tardif, J., Cote, G., Lesperance, J., Bourassa, M., Lambert, J., Doucet, S., Bilodeau, L., Nattel, S., and DeGuise, P. (1997). Producol and multivitamins in the prevention of restenosis after coronary angioplasty. *New England Journal of Medicine* **337**, 365–372.
36. Walker, P., Osredkar, M., and Bilancini, S. (1999). The effect of stimuli on pillbug movement. Unpublished manuscript, Oberlin College.
37. Hoppeler, H., and Vogt, M. (2001). Muscle tissue adaptations to hypoxia. *The Journal of Experimental Biology* **204**, 3133–3139.
38. Adapted from Tajkarimi, M., Riemann, H., Hajmeer, M., Gomez, E., Razavilar, V., and Cliver, D. (2008). Ammonia disinfection of animal feeds—laboratory study. *International Journal of Food Microbiology* **122**, 23–28.

## Chapter 12

1. Unpublished data courtesy of M. B. Nichols and R. P. Maickel. The original experiment contained more than three treatment groups.
2. Bodgan, K., and Schenk, M. (2009). Evaluation of soil characteristics potentially affecting arsenic concentration in paddy rice (*Oryza sativa* L.). *Environmental Pollution* **157**, 2617–2621. 2006 data digitized from Figure 3.
3. Adapted from Andren, C., and Nilson, G. (1981). Reproductive success and risk of predation in normal and melanistic colour morphs of the adder, *Vipera berus*. *Biological Journal of the Linnean Society* **15**, 235–246. (The data are for the melanistic females; the values have been manipulated slightly to simplify the exposition.)
4. Erne, P., Bolli, P., Buergisser, E., and Buehler, F. R. (1984). Correlation of platelet calcium with blood pressure. *New England Journal of Medicine* **310**, 1084–1088. Reprinted by permission. Raw data courtesy of F. R. Buehler. To simplify the discussion, we have omitted nine patients with “borderline” high blood pressure.
5. Adapted from Spencer, D. F., Volpp, T. R., and Lembi, C. A. (1980). Environmental control of *Pithophora oedogonia* (Chlorophyceae) akinete germination. *Journal of Phycology* **16**, 424–427. The value  $r = -0.72$  was calculated from data displayed graphically by Spencer et al.
6. Albert, A. (1981). Atypicality indices as reference values for laboratory data. *American Journal of Clinical Pathology* **76**, 421–425.
7. Harding, A. J., Wong, A., Svoboda, M., Kril, J. J., and Halliday, G. M. (1997). Chronic alcohol consumption does not cause hippocampal neuron loss in humans. *Hippocampus* **7**, 78–87. The value  $r = -0.63$  was calculated from data displayed graphically by Harding et al.
8. Smith, R. D. (1978–1979). Institute of Agricultural Engineering Annual Report. Salisbury, Zimbabwe: Department of Research and Specialist Services, Ministry of Agriculture. Raw data courtesy of R. D. Smith.
9. Bowers, W. S., Hoch, H. C., Evans, P. H., and Katayama, M. (1986). Thallophytic allelopathy: Isolation and identification of laeticaric acid. *Science* **232**, 105–106. Copyright 1986 by the AAAS. Raw data courtesy of the authors.
10. Webb, P. (1981). Energy expenditure and fat-free mass in men and women. *American Journal of Clinical Nutrition* **34**, 1816–1826.

11. Huel, G., et al. (2008). Hair mercury negatively correlates with calcium pump activity in human term newborns and their mothers at delivery. *Environmental Health Perspectives* **116**, 263–267.
12. Simulated data based on typical annual ocean temperatures near Morro Bay, California.
13. Cicirelli, M. F., Robinson, K. R., and Smith, L. D. (1983). Internal pH of *Xenopus* oocytes: A study of the mechanism and role of pH changes during meiotic maturation. *Developmental Biology* **100**, 133–146. Raw data courtesy of M. F. Cicirelli.
14. Maickel, R. P., and Nash, J. F., Jr. (1985). Differing effects of short-chain alcohols on body temperature and coordinated muscular activity in mice. *Neuropharmacology* **24**, 83–89. Reprinted with permission. Copyright 1985, Pergamon Journals, Ltd. Raw data courtesy of J. F. Nash, Jr.
15. Adapted from Barclay, A. M., and Crawford, R. M. M. (1984). Seedling emergence in the rowan (*Sorbus aucuparia*) from an altitudinal gradient. *Journal of Ecology* **72**, 627–636. Reprinted with permission from John Wiley.
16. Olson, J. M., and Mardh, R. L. (1998). Activation patterns and length changes in hindlimb muscles of the bullfrog *Rana catesbeiana* during jumping. *The Journal of Experimental Biology* **201**, 2763–2777.
17. Sulcove, J. A., and Lacuesta, N. N. (1998). The effect of gender and height on peak flow rate. Unpublished manuscript, Oberlin College.
18. Hamill, P. V. V., Johnston, F. E., and Lemeshow, S. (1973). Height and weight of youths 12–17 years, United States. *U.S. National Center for Health Statistics, Vital and Health Statistics*, Series 11, No. 124. Washington, D.C.: U.S. Department of Health, Education and Welfare. The conditional distributions of weight given height are plotted in Figure 12.4.1 as normal distributions. The fictitious population agrees well with the real population (as described by Hamill et al.) in the central portion of each conditional distribution, but the real conditional distributions have shorter left tails and longer right tails than the fictitious normal conditional distributions.
19. Maickel, R. P. Personal communication.
20. Fictitious but realistic data, based on inter- and intra-individual variation as described in Williams, G. Z., Widdowson, G. M., and Penton, J. (1978). Individual character of variation in time-series studies of healthy people. II. Differences in values for clinical chemical analytes in serum among demographic groups, by age and sex. *Clinical Chemistry* **24**, 313–320.
21. Stewart, T. S., Nelson, L. A., Perry, T. W., and Martin, T. G. (1985). Unpublished data provided courtesy of T. S. Stewart.
22. Pappas, T., and Mitchell, C. A. (1985). Effects of seismic stress on the vegetative growth of *Glycine max* (L.) Merr. cv. Wells II. *Plant, Cell and Environment* **8**, 143–148. Reprinted with permission from John Wiley. Raw data courtesy of the authors.
23. Fialho, E. T., Ferreira, A. S., Freitas, A. R., and Albino, L. F. T. (1982). Energy and nitrogen balance of ration (corn-soybean meal) for male castrated and non-castrated swine of different weights and breeds (in Portuguese). *Revista Sociedade Brasileira de Zootecnia* **11**, 405–419. Raw data courtesy of E. T. Fialho.
24. Example communicated by D. A. Holt.
25. Chambers, J. Q., Higuhi, N., and Schimel, J. (1998). Ancient trees in Amazonia. *Nature* **391**, 135–136. Raw data courtesy of J. Chambers.
26. Dikmen, S., and Hansen, P. (2009). Is the temperature-humidity index the best indicator of heat stress in lactating dairy cows in a subtropical environment? *Journal of Dairy Science* **92**, 109–116. Data were simulated to produce similar results as those presented in the article.
27. Florey, C. du V., and Acheson, R. M. (1969). Blood pressure as it relates to physique, blood glucose, and serum cholesterol. *U.S. National Center for Health Statistics, Series 11*, No. 34. Washington, D.C.: U.S. Department of Health, Education and Welfare.
28. Bernays, E. A. (1986). Diet-induced head allometry among foliage-chewing insects and its importance for graminovores. *Science* **231**, 495–497. Copyright 1986 by the AAAS. Raw data courtesy of the author.
29. Hibi, K., Taguchi, M., Nakayama, H., Takase, T., Kasai, Y., Ito, K., Akiyama, S., and Nakao, A. (2001). Molecular detection of p16 promoter methylation in the serum of patients with esophageal squamous cell carcinoma. *Clinical Cancer Research* **7**, 3135–3138. There were 38 patients in the study; only the 31 patients for whom “tumor DNA was methylated” are included in this analysis.
30. Gwynne, D. T. (1981). Sexual difference theory: Mormon crickets show role reversal in mate choice. *Science* **213**, 779–780. Copyright 1981 by the AAAS. Calculations based on raw data provided courtesy of the author.
31. Heggestad, H. E., and Bennett, J. H. (1981). Photochemical oxidants potentiate yield losses in

- snap beans attributable to sulfur dioxide. *Science* **213**, 1008–1010. Copyright 1981 by the AAAS. Raw data courtesy of H. E. Heggestad.
32. Thirakhupt, K. (1985). Foraging ecology of sympatric parids: Individual and population responses to winter food scarcity. Ph.D. thesis, Purdue University. Raw data courtesy of the author and K. N. Rabenold.
  33. Unpublished data courtesy of A. H. Ismail and L. S. Verity.
  34. Tazawa, H., Pearson, J. T., Komoro, T., and Ar, A. (2001). Allometric relationships between embryonic heart rate and fresh egg mass in birds. *The Journal of Experimental Biology* **204**, 165–174.
  35. These data sets were invented by F. J. Anscombe. See Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician* **27**, 17–21.
  36. Unpublished data courtesy of M. B. Nichols and R. P Maickel. The experiment actually contained more than three groups. The data in Example 12.1.1 are from another part of the same study, using a different chemical form of amphetamine.
  37. Marazziti, D., DiMuro, A., and Castrogiovanni, P. (1992). Psychological stress and body temperature changes in humans. *Psychology & Behavior* **52**, 393–395.
  38. Kinghorn, A., Humphries, M., Outridge, P., and Chan, H. M. (2008). Teeth as biomonitor of selenium concentrations in tissues of beluga whales (*Delphinapterus leucas*). *Science of the Total Environment* **402**, 43–50. Data digitized from Figure 3.

## Chapter 13

1. Unpublished data courtesy of M. A. Johnson and F. Bretos. Data collected at Oberlin College in the fall of 1997.
2. Fish, F. E. (1998). Comparative kinematics and hydrodynamics of odontocete cetaceans: morphological and ecological correlates with swimming performance. *Journal of Experimental Biology* **201**, 2867–2877. The data presented here were read from Figure 1 in the article.
3. Fisher, B., Costantino, J. P., Wickerman, L., Redmond, C. K., Kavanah, M., Cronin, W. M., Vogel, V., Robidoux, A., Dimitrov, N., Atkins, J., Daly, M., Wieand, S., Tan-Chiu, E., Ford, L., Wolmark, N., et al. (1998). Tamoxifen for prevention of breast cancer: Report of the national surgical adjuvant breast and bowel project P-1 study. *Journal of the National Cancer Institute* **30**, 1371–1388.
4. Unpublished data courtesy of K. Pretl. Data collected at Oberlin College in the spring of 1997.
5. Rosa, L., Rosa, E., Sarner, L., and Barrett, S. (1998). A close look at therapeutic touch. *Journal of the American Medical Association* **279**, 1005–1010.
6. Morgan, M. B., and Cowles, D. L. (1996). The effects of temperature on the behaviour and physiology of *Phataria unifascialis* (Gray) (Echinodermata, Asteroidea); implications for the species' distribution in the Gulf of California, Mexico. *Journal of Experimental Marine Biology and Ecology* **208**, 13–27.
7. Kujala, U. M., Kaprio, J., Sarna, S., and Koskenvuo, M. (1998). Relationship of leisure-time physical activity and mortality: The Finnish cohort study. *Journal of the American Medical Association* **279**, 440–444. The “exerciser” category actually had two sub-categories of “occasional exerciser” and “conditioning exerciser,” which we have combined here.
8. Espigares, T., and Peco, B. (1995). Mediterranean annual pasture dynamics: Impact of autumn drought. *Journal of Ecology* **83**, 135–142.
9. Izurieta, H. S., Strebel, P. M., and Blake, P. A. (1997). Postlicensure effectiveness of varicella vaccine during an outbreak in a child care center. *Journal of the American Medical Association* **278**, 1495–1499.
10. Rose, D. P., Fern, M., Liskowski, L., and Milbrath, J. R. (1977). Effect of treatment with estrogen conjugates on endogenous plasma steroids. *Obstetrics and Gynecology* **49**, 80–82.
11. Grether, G. (1997). Survival cost of an intrasexually selected ornament in a damselfly. *Proceedings of the Royal Society of London B* **264**, 207–210.
12. Peterson, A. V., Kealey, K. A., Mann, S. L., Marek, P. M., and Sarason, I. G. (2000). Hutchinson smoking prevention project: Long-term randomized trial in school-based tobacco use prevention—results on smoking. *Journal of the National Cancer Institute* **92**, 1979–1991. The authors analyzed these data not with a sign test, but with a randomization test, similar to that presented in Section 7.1.
13. Rosenheck, R., Cramer, J., Xu, W., Thomas, J., Henderson, W., Frisman, L., Fye, C., and Charney, D., for the Department of Veterans Affairs Cooperative Study Group on Clozapine in Refractory Schizophrenia (1997). A comparison of clozapine and haloperidol in hospitalized patients with refractory schizophrenia. *New England Journal of Medicine* **337**, 809–815.

14. Unpublished data courtesy of K. Roberts. Data collected at Oberlin College in the spring of 1997.
15. Baur, E., Fischer, E., and Lenz, F. (1931). *Human Heredity*, 3rd ed. New York: Macmillan, p. 45.
16. Plotnick, G. D., Corretti, M. C., and Vogel, R. A. (1997). Effect of antioxidant vitamins on the transient impairment of endothelium-dependent brachial artery vasoactivity following a single high-fat meal. *Journal of the American Medical Association* **278**, 1682–1686.
17. Unpublished data collected by E. Lohan and M. Josephy, Oberlin College, in the fall of 1997.
18. Meier, P., Free, S. M., and Jackson, G. L. (1958). Reconsideration of methodology in studies of pain relief. *Biometrics* **14**, 330–342.
19. Unpublished data collected by J. Amundson, Oberlin College, in the fall of 1997.
20. Weintraub, W. S., Boccuzzi, S. J., Klein, J. L., Kosinski, A. S., King, S. B., Ivanhoe, R., Cedarholm, J. C., Stillabower, M. E., Talley, J. D., DeMaio, S. J., O'Neill, W. W., Frazier, J. E., Cohen-Bernstein, C. L., Robbins, D. C., Brown, C. L., Alexander, R. W., and the Lovastatin Restenosis Trial Study Group. (1994). Lack of effect of lovastatin on restenosis after coronary angioplasty. *New England Journal of Medicine* **331**, 1331–1337.
21. Bresee, J., Mast, E. E., Coleman, P. J., Baron, M. J., Schonberger, L. B., Alter, M. J., Jonas, M. M., Yu, M. W., Renzi, P. M., and Schneider, L. C. (1996). Hepatitis C virus infection associated with administration of intravenous immune globulin: A cohort study. *Journal of the American Medical Association* **276**, 1563–1567.
22. Ferrandina, G., Ranelletti, F. O., Larocca, L. M., Maggiano, N., Fruscella, E., Legge, F., Santeusanio, G., Bombonati, A., Mancuso, S., and Scambia, G. (2001). Tamoxifen modulates the expression of Ki67, apoptosis, and microvessel density in cervical cancer. *Cancer Clinical Research* **7**, 2656–2661.
23. Devine, W. D., Houston, A. E., and Tyler, D. D. (2000). Growth of three hardwood species through 18 years on a former agricultural bottomland. *Southern Journal of Applied Forestry* **24**, 159–165.
24. Unpublished data collected by S. Haaz, Oberlin College, in the spring of 1999. The division into groups is based on bench press strength, adjusted for height, weight, and whether or not the student was on an athletic team.

# STATISTICAL TABLES

---

<b>Table 1</b>	Random Digits	611
<b>Table 2</b>	Binomial Coefficients $nC_j$	615
<b>Table 3</b>	Areas Under the Normal Curve	616
<b>Table 4</b>	Critical Values of Student's $t$ Distribution	618
<b>Table 5</b>	Sample Sizes Needed for Selected Power Levels for Independent-Samples $t$ Test	619
<b>Table 6</b>	Critical Values and $P$ -Values of $U_s$ for the Wilcoxon-Mann-Whitney Test	621

<b>Table 7</b>	Critical Values and $P$ -Values of $B_s$ for the Sign Test	625
<b>Table 8</b>	Critical Values and $P$ -Values of $W_s$ for the Wilcoxon Signed-Rank Test	626
<b>Table 9</b>	Critical Values of the Chi-Square Distribution	627
<b>Table 10</b>	Critical Values of the $F$ Distribution	628
<b>Table 11</b>	Bonferroni Multipliers for 95% Confidence Intervals	638

**TABLE 1 Random Digits**

	01	06	11	16	21	26	31	36	41	46
01	06048	96063	22049	86532	75170	65711	29969	06826	39208	80631
02	25636	73908	85512	78073	19089	66458	06597	93985	14193	69366
03	61378	45410	43511	54364	97334	01267	28304	35047	38789	84896
04	15919	71559	12310	00727	54473	51547	09816	83641	72973	75367
05	47328	20405	88019	82276	33679	10328	25116	59176	64675	95141
06	72548	80667	53893	64400	81955	15163	06146	58549	75530	19582
07	87154	04130	55985	44508	37515	71689	80765	46598	45539	12792
08	68379	96636	32154	94718	22845	80265	92747	66238	58474	23783
09	89391	54041	70806	36012	30833	83132	39338	54753	00722	44568
10	15816	60231	28365	61924	66934	21243	09896	92428	51611	46756
11	29618	55219	18394	11625	27673	08117	89314	42581	36897	03738
12	30723	42988	30002	95364	45473	46107	34222	00739	84847	49096
13	54028	04975	92323	53836	76128	84762	32050	59516	40831	59687
14	40376	02036	48087	05216	26684	97959	85601	86622	70750	15603
15	64439	37357	90935	57330	79738	65361	85944	23619	30504	61564
16	83037	30144	29166	20915	53462	42573	75204	50064	08847	07082
17	71071	01636	31085	71638	77357	14256	89174	15184	81701	21592
18	67891	43187	58159	24144	29683	04276	02987	04571	18334	04291
19	52487	39499	97330	40045	47304	98528	00422	82693	87547	73525
20	67550	82107	27302	79145	73213	27217	19211	59784	63929	04609
21	86472	80165	70773	90519	49710	31921	36102	45042	04203	01439
22	08699	38051	60404	06609	98435	91560	22634	98014	43316	61099
23	59596	13000	07655	74837	81211	71530	28341	83110	72289	25180
24	31810	54868	92799	09893	97499	96509	71548	06462	40498	22628
25	71753	90756	21382	84209	95900	11119	34507	61241	17641	83147

*Continued*

**TABLE 1 Random Digits (continued)**

	51	56	61	66	71	76	81	86	91	96
01	64825	74126	86159	26710	49256	04655	06001	73192	67463	16746
02	46184	63916	89160	87844	53352	43318	70766	23625	09906	65847
03	79976	48891	69431	86571	25979	58755	08884	36704	01107	12308
04	10656	47210	48512	06805	42114	98741	51440	06070	49071	02700
05	18058	84528	56753	02623	81077	60045	06678	53748	10386	37895
06	58979	98046	88467	27762	24781	12559	98384	40926	79570	34746
07	12705	41974	14473	49872	29368	80556	95833	20766	76643	35656
08	39660	83664	18592	82388	27899	24223	36462	61582	95173	36155
09	00360	42077	84161	04464	45042	29560	37916	29889	00342	82533
10	09873	64084	34685	53542	09254	23257	14713	44295	94139	00403
11	12957	84063	79808	23633	77133	41422	26559	29131	74402	82213
12	06090	71584	48965	60201	02786	88929	19861	99361	27535	38297
13	66812	57167	28185	19708	74672	25615	61640	18955	40854	50749
14	91701	36216	66249	04256	31694	33127	67529	73254	72065	74294
15	02775	78899	36471	37098	50270	58933	91765	95157	01384	75388
16	75892	53340	92363	58300	77300	08059	63743	12159	05640	87014
17	18581	70057	82031	68349	55759	46851	33632	28855	74633	08598
18	69698	18177	52824	61742	58119	04168	57843	37870	50988	80316
19	30023	30731	00803	09336	87709	39307	09732	66031	04904	91929
20	94334	05698	97910	37850	77074	56152	67521	48973	29448	84115
21	64133	14640	28418	45405	86974	06666	07879	54026	92264	23418
22	93895	83557	17326	28030	09113	56793	79703	18804	75807	20144
23	54438	83097	52533	86245	02182	11746	58164	90520	99255	44830
24	90565	76710	42456	22612	00232	18919	24019	32254	30703	00678
25	90848	81871	24382	16218	98216	42323	75061	68261	09071	68776

**TABLE 1 Random Digits (continued)**

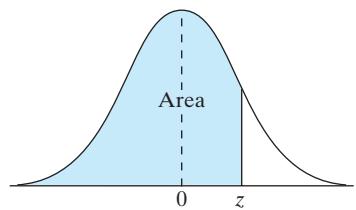
	01	06	11	16	21	26	31	36	41	46
26	17155	07370	65655	04824	53417	20737	70510	92615	89967	50216
27	36211	24724	94769	16940	43138	25260	75318	69037	95982	28631
28	94777	66946	16120	56382	58416	92391	81457	28101	69766	32436
29	52994	58881	81841	51844	75566	48567	18552	66829	91230	39141
30	84643	32635	51440	96854	35739	66440	82806	82841	56302	31640
31	95690	34873	11297	60518	72717	47616	55751	37187	31413	31132
32	64093	92948	21565	51686	40368	66151	82877	99951	85069	54503
33	89484	50055	67586	16439	96385	67868	66597	51433	44764	66573
34	70184	38164	74646	90244	83169	85276	07598	69242	90088	32308
35	75601	91867	80848	94484	98532	36183	28549	17704	28653	80027
36	99044	78699	34681	31049	40790	50445	79897	68203	11486	93676
37	10272	18347	89369	02355	76671	34097	03791	93817	43142	24974
38	69738	85488	34453	80876	43018	59967	84458	71906	54019	70023
39	93441	58902	17871	45425	29066	04553	42644	54624	34498	27319
40	25814	74497	75642	58350	64118	87400	82870	26143	46624	21404
41	29757	84506	48617	48844	35139	97855	43435	74581	35678	69793
42	56666	86113	06805	09470	07992	54079	00517	19313	53741	25306
43	26401	71007	12500	27815	86490	01370	47826	36009	10447	25953
44	40747	59584	83453	30875	39509	82829	42878	13844	84131	48524
45	99434	51563	73915	03867	24785	19324	21254	11641	25940	92026
46	50734	88330	39128	14261	00584	94266	99677	19852	49673	18680
47	89728	32743	19102	83279	68308	41160	32365	25774	39699	50743
48	71395	61945	41082	93648	99874	82577	26507	07054	29381	16995
49	50945	68182	23108	95765	81136	06792	13322	41631	37118	35881
50	36525	26551	28457	75699	74537	68623	50099	91909	23508	35751

**TABLE 1 Random Digits (continued)**

	51	56	61	66	71	76	81	86	91	96
26	41169	08175	69938	61958	72578	31791	74952	71055	40369	00429
27	84627	70347	41566	00019	24481	15677	54506	54545	89563	50049
28	67460	49111	54004	61428	61034	47197	90084	88113	39145	94757
29	99231	60774	52238	05102	71690	72215	61323	13326	01674	81510
30	95775	73679	04900	27666	18424	59793	14965	22220	30682	35488
31	42179	98675	69593	17901	48741	59902	98034	12976	60921	73047
32	91196	05878	92346	45886	31080	21714	19168	94070	77375	10444
33	18794	03741	17612	65467	27698	20456	91737	36008	88225	58013
34	88311	93622	34501	70402	12272	65995	66086	04938	52966	71909
35	17904	33710	42812	72105	91848	39724	26361	09634	50552	98769
36	05905	28509	69631	69177	39081	58818	01998	53949	47884	91326
37	23432	22211	65648	71866	49532	45529	00189	80025	68956	26445
38	29684	43229	54771	90604	48938	13663	24736	83199	41512	43364
39	26506	65067	64252	49765	87650	72082	48997	04845	00136	98941
40	08807	43756	01579	34508	94082	68736	67149	00209	76138	95467
41	50636	70304	73556	32872	07809	20787	85921	41748	10553	97988
42	32437	41588	46991	36667	98127	05072	63700	51803	77262	31970
43	32571	97567	78420	04633	96574	88830	01314	04811	10904	85923
44	28773	22496	11743	23294	78070	20910	86722	50551	37356	92698
45	65768	76188	07781	05314	26017	07741	22268	31374	53559	46971
46	68601	06488	73776	45361	89059	59775	59149	64095	10352	11107
47	98364	17663	85972	72263	93178	04284	79236	04567	31813	82283
48	95308	70577	96712	85697	55685	19023	98112	96915	50791	31107
49	68681	24419	15362	60771	09962	45891	03130	09937	15775	51935
50	30721	22371	65174	57363	37851	71554	19708	23880	86638	05880

**TABLE 2 Binomial Coefficients  $nC_j$** 

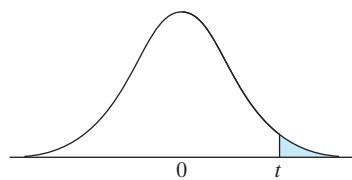
$n$	0	1	2	3	4	5	$j$	6	7	8	9	10
1	1	1										
2	1	2	1									
3	1	3	3	1								
4	1	4	6	4	1							
5	1	5	10	10	5	1						
6	1	6	15	20	15	6	1					
7	1	7	21	35	35	21	7	1				
8	1	8	28	56	70	56	28	8	1			
9	1	9	36	84	126	126	84	36	9	1		
10	1	10	45	120	210	252	210	120	45	10	1	
11	1	11	55	165	330	462	462	330	165	55	11	
12	1	12	66	220	495	792	924	792	495	220	66	
13	1	13	78	286	715	1,287	1,716	1,716	1,287	715	286	
14	1	14	91	364	1,001	2,002	3,003	3,432	3,003	2,002	1,001	
15	1	15	105	455	1,365	3,003	5,005	6,435	6,435	5,005	3,003	
16	1	16	120	560	1,820	4,368	8,008	11,440	12,870	11,440	8,008	
17	1	17	136	680	2,380	6,188	12,376	19,448	24,310	24,310	19,448	
18	1	18	153	816	3,060	8,568	18,564	31,824	43,758	48,620	43,758	
19	1	19	171	969	3,876	11,628	27,132	50,388	75,582	92,378	92,378	
20	1	20	190	1,140	4,845	15,504	38,760	77,520	125,970	167,960	184,756	

**TABLE 3 Areas Under the Normal Curve**

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

**TABLE 3 Areas Under the Normal Curve (continued)**

**TABLE 4 Critical Values of Student's t Distribution**



**TABLE 5 Sample Sizes Needed for Selected Power Levels for Independent-Samples t Test**

POWER→	SIGNIFICANCE LEVEL (TWO-TAILED TEST)																				
	$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.10$					
	0.99	0.95	0.90	0.80	0.50	0.99	0.95	0.90	0.80	0.50	0.99	0.95	0.90	0.80	0.50	0.99	0.95	0.90	0.80	0.50	
0.20																	137	0.20			
0.25																	88	0.25			
0.30																	61	0.30			
0.35																	102	0.35			
0.40																	78	0.40			
0.45																	62	0.45			
0.50																	28	0.50			
$\frac{ \mu_1 - \mu_2 }{\sigma}$	0.55																19	0.55			
	0.60	101	85	67	39												36	0.60			
	0.65	87	73	57	34												30	0.65			
	0.70	100	75	63	50	29											12	0.70			
	0.75	88	66	55	44	26											11	0.75			
	0.80	77	58	49	39	23											10	0.80			
	0.85	69	51	43	35	21											9	0.85			
	0.90	62	46	39	31	19											8	0.90			
	0.95	55	42	35	28	17											7	0.95			
	1.00	50	38	32	26	15											7	1.00			
	$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$					
	SIGNIFICANCE LEVEL (ONE-TAILED TEST)																				

*Continued*

**TABLE 5 Sample Sizes Needed for Selected Power Levels for Independent-Samples t Test (continued)**

POWER→	SIGNIFICANCE LEVEL (TWO-TAILED TEST)																					
	$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.10$						
	0.99	0.95	0.90	0.80	0.50	0.99	0.95	0.90	0.80	0.50	0.99	0.95	0.90	0.80	0.50	0.99	0.95	0.90	0.80	0.50		
1.1	42	32	27	22	13	38	28	23	19	11	32	23	19	14	8	27	19	15	12	6	1.1	
1.2	36	27	23	18	11	32	24	20	16	9	27	20	16	12	7	23	16	13	10	5	1.2	
1.3	31	23	20	16	10	28	21	17	14	8	23	17	14	11	6	20	14	11	9	5	1.3	
1.4	27	20	17	14	9	24	18	15	12	8	20	15	12	10	6	17	12	10	8	4	1.4	
1.5	24	18	15	13	8	21	16	14	11	7	18	13	11	9	5	15	11	9	7	4	1.5	
1.6	21	16	14	11	7	19	14	12	10	6	16	12	10	8	5	14	10	8	6	4	1.6	
1.7	19	15	13	10	7	17	13	11	9	6	14	11	9	7	4	12	9	7	6	3	1.7	
1.8	17	13	11	10	6	15	12	10	8	5	13	10	8	6	4	11	8	7	5		1.8	
1.9	16	12	11	9	6	14	11	9	8	5	12	9	7	6	4	10	7	6	5		1.9	
$ \mu_1 - \mu_2 $	2.0	14	11	10	8	6	13	10	9	7	5	11	8	7	6	4	9	7	6	4		2.0
$\sigma$	2.1	13	10	9	8	5	12	9	8	7	5	10	8	6	5	3	8	6	5	4		2.1
	2.2	12	10	8	7	5	11	9	7	6	4	9	7	6	5		8	6	5	4		2.2
	2.3	11	9	8	7	5	10	8	7	6	4	9	7	6	5		7	5	5	4		2.3
	2.4	11	9	8	6	5	10	8	7	6	4	8	6	5	4		7	5	4	4		2.4
	2.5	10	8	7	6	4	9	7	6	5	4	8	6	5	4		6	5	4	3		2.5
	3.0	8	6	6	5	4	7	6	5	4	3	6	5	4	4		5	4	3			3.0
3.5	6	5	5	4	3	6	5	4	4		5	4	4	3		4	3				3.5	
4.0	6	5	4	4		5	4	4	3		4	4	3			4					4.0	
	$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$						
	SIGNIFICANCE LEVEL (ONE-TAILED TEST)																					

Source: "Number of observations for  $t$ -test of difference between two means." *Research*, Volume 1 (1948), pp. 520–525. Used with permission of the Longman Group UK Ltd. and Butterworth Scientific Publications.

**TABLE 6 Critical values and P-values of  $U_s$  for the Wilcoxon-Mann-Whitney test**

Note: Because the Wilcoxon-Mann-Whitney test null distribution is discrete, this table provides selected values of the test statistic  $U_s$  in **bold type** and corresponding  $P$ -values for a non-directional alternative in *italics*. Directional  $P$ -values are found by dividing the numbers in italics in half.

<i>n</i>	<i>n'</i>	0.20	0.10	0.05	0.025	0.01	0.005
3	2	<b>6</b> 0.200					
	3	<b>8</b> 0.200	<i>9</i> 0.100				
4	2	<b>8</b> 0.133					
	3	<b>11</b> 0.114	<b>12</b> 0.057				
5	2	<b>13</b> 0.200	<b>15</b> 0.057	<b>16</b> 0.029			
	3	<b>9</b> 0.191	<b>10</b> 0.095				
6	2	<b>13</b> 0.143	<b>14</b> 0.071	<b>15</b> 0.036			
	3	<b>16</b> 0.191	<b>18</b> 0.064	<b>19</b> 0.032	<b>20</b> 0.016		
7	2	<b>20</b> 0.151	<b>21</b> 0.095	<b>23</b> 0.032	<b>24</b> 0.016	<b>25</b> 0.0079	
	3	<b>11</b> 0.143	<b>12</b> 0.071				
8	2	<b>15</b> 0.167	<b>16</b> 0.095	<b>17</b> 0.048	<b>18</b> 0.024		
	3	<b>19</b> 0.171	<b>21</b> 0.067	<b>22</b> 0.038	<b>23</b> 0.019	<b>24</b> 0.0095	
9	2	<b>23</b> 0.178	<b>25</b> 0.082	<b>27</b> 0.030	<b>28</b> 0.017	<b>29</b> 0.0087	<b>30</b> 0.0043
	3	<b>27</b> 0.180	<b>29</b> 0.093	<b>31</b> 0.041	<b>33</b> 0.015	<b>34</b> 0.0087	<b>35</b> 0.0043
10	2	<b>13</b> 0.111	<b>14</b> 0.056				
	3	<b>17</b> 0.183	<b>19</b> 0.067	<b>20</b> 0.033	<b>21</b> 0.017		
11	2	<b>22</b> 0.164	<b>24</b> 0.072	<b>25</b> 0.042	<b>26</b> 0.024	<b>28</b> 0.0061	
	3	<b>27</b> 0.149	<b>29</b> 0.073	<b>30</b> 0.048	<b>32</b> 0.018	<b>34</b> 0.0051	<b>35</b> 0.0025
12	2	<b>31</b> 0.181	<b>34</b> 0.073	<b>36</b> 0.035	<b>37</b> 0.022	<b>39</b> 0.0082	<b>40</b> 0.0047
	3	<b>36</b> 0.165	<b>38</b> 0.097	<b>41</b> 0.038	<b>43</b> 0.018	<b>45</b> 0.0070	<b>46</b> 0.0041
13	2	<b>14</b> 0.178	<b>15</b> 0.089	<b>16</b> 0.044			
	3	<b>19</b> 0.194	<b>21</b> 0.085	<b>22</b> 0.049	<b>23</b> 0.024		
14	2	<b>25</b> 0.154	<b>27</b> 0.073	<b>28</b> 0.049	<b>30</b> 0.016	<b>31</b> 0.0081	<b>32</b> 0.0040
	3	<b>30</b> 0.171	<b>32</b> 0.093	<b>34</b> 0.045	<b>36</b> 0.019	<b>38</b> 0.0062	<b>39</b> 0.0031
15	2	<b>35</b> 0.181	<b>38</b> 0.081	<b>40</b> 0.043	<b>42</b> 0.020	<b>44</b> 0.0080	<b>45</b> 0.0047
	3	<b>40</b> 0.189	<b>43</b> 0.094	<b>46</b> 0.041	<b>48</b> 0.021	<b>50</b> 0.0093	<b>52</b> 0.0037
16	2	<b>45</b> 0.195	<b>49</b> 0.083	<b>51</b> 0.050	<b>54</b> 0.021	<b>57</b> 0.0070	<b>58</b> 0.0047
	3	<b>16</b> 0.146	<b>17</b> 0.073	<b>18</b> 0.036			
17	2	<b>22</b> 0.146	<b>23</b> 0.100	<b>25</b> 0.036	<b>26</b> 0.018	<b>27</b> 0.0091	
	3	<b>27</b> 0.199	<b>30</b> 0.076	<b>32</b> 0.034	<b>33</b> 0.020	<b>35</b> 0.0056	<b>36</b> 0.0028
18	2	<b>33</b> 0.190	<b>36</b> 0.083	<b>38</b> 0.042	<b>40</b> 0.019	<b>42</b> 0.0070	<b>43</b> 0.0040
	3	<b>39</b> 0.181	<b>42</b> 0.088	<b>44</b> 0.050	<b>47</b> 0.018	<b>49</b> 0.0076	<b>50</b> 0.0048
19	2	<b>45</b> 0.174	<b>48</b> 0.091	<b>51</b> 0.042	<b>53</b> 0.023	<b>56</b> 0.0079	<b>58</b> 0.0033
	3	<b>50</b> 0.200	<b>54</b> 0.093	<b>57</b> 0.046	<b>60</b> 0.021	<b>63</b> 0.0079	<b>65</b> 0.0037
20	2	<b>56</b> 0.190	<b>60</b> 0.094	<b>64</b> 0.040	<b>66</b> 0.024	<b>70</b> 0.0078	<b>72</b> 0.0040
	3	<b>17</b> 0.182	<b>19</b> 0.061	<b>20</b> 0.030			
21	2	<b>24</b> 0.161	<b>26</b> 0.077	<b>27</b> 0.049	<b>29</b> 0.014	<b>30</b> 0.0070	
	3	<b>30</b> 0.188	<b>33</b> 0.076	<b>35</b> 0.036	<b>36</b> 0.024	<b>38</b> 0.0080	<b>39</b> 0.0040
22	2	<b>37</b> 0.165	<b>39</b> 0.099	<b>42</b> 0.040	<b>44</b> 0.019	<b>46</b> 0.0080	<b>47</b> 0.0047
	3	<b>43</b> 0.181	<b>46</b> 0.093	<b>49</b> 0.042	<b>51</b> 0.023	<b>54</b> 0.0075	<b>55</b> 0.0047
23	2	<b>49</b> 0.193	<b>53</b> 0.088	<b>56</b> 0.043	<b>58</b> 0.025	<b>61</b> 0.0097	<b>63</b> 0.0046
	3	<b>56</b> 0.173	<b>60</b> 0.083	<b>63</b> 0.043	<b>66</b> 0.021	<b>69</b> 0.0085	<b>71</b> 0.0044
24	2	<b>62</b> 0.182	<b>66</b> 0.095	<b>70</b> 0.044	<b>73</b> 0.022	<b>77</b> 0.0076	<b>79</b> 0.0041
	3	<b>68</b> 0.190	<b>73</b> 0.089	<b>77</b> 0.043	<b>80</b> 0.023	<b>84</b> 0.0089	<b>87</b> 0.0039

*Continued*

**TABLE 6 Critical values and P-values of  $U_s$  for the Wilcoxon-Mann-Whitney test (continued)**

$n$	$n'$	0.20	0.10	0.05	0.025	0.01	0.005
11	2	<b>19</b> 0.154	<b>21</b> 0.051	<b>22</b> 0.026			
	3	<b>26</b> 0.170	<b>28</b> 0.088	<b>30</b> 0.039	<b>31</b> 0.022	<b>33</b> 0.0055	
	4	<b>33</b> 0.177	<b>36</b> 0.078	<b>38</b> 0.040	<b>40</b> 0.018	<b>42</b> 0.0059	<b>43</b> 0.0029
	5	<b>40</b> 0.180	<b>43</b> 0.090	<b>46</b> 0.038	<b>48</b> 0.019	<b>50</b> 0.0087	<b>52</b> 0.0032
	6	<b>47</b> 0.180	<b>50</b> 0.098	<b>53</b> 0.048	<b>56</b> 0.020	<b>59</b> 0.0071	<b>60</b> 0.0048
	7	<b>54</b> 0.179	<b>58</b> 0.085	<b>61</b> 0.044	<b>64</b> 0.020	<b>67</b> 0.0083	<b>69</b> 0.0041
	8	<b>61</b> 0.177	<b>65</b> 0.091	<b>69</b> 0.041	<b>72</b> 0.020	<b>75</b> 0.0091	<b>77</b> 0.0050
	9	<b>68</b> 0.175	<b>72</b> 0.095	<b>76</b> 0.047	<b>80</b> 0.020	<b>83</b> 0.0097	<b>86</b> 0.0042
	10	<b>74</b> 0.197	<b>79</b> 0.099	<b>84</b> 0.043	<b>87</b> 0.024	<b>92</b> 0.0079	<b>94</b> 0.0048
	11	<b>81</b> 0.193	<b>87</b> 0.088	<b>91</b> 0.047	<b>95</b> 0.023	<b>100</b> 0.0083	<b>103</b> 0.0041
12	2	<b>20</b> 0.198	<b>22</b> 0.088	<b>23</b> 0.044	<b>24</b> 0.022		
	3	<b>28</b> 0.180	<b>31</b> 0.070	<b>32</b> 0.048	<b>34</b> 0.018	<b>35</b> 0.0088	<b>36</b> 0.0044
	4	<b>36</b> 0.170	<b>39</b> 0.078	<b>41</b> 0.042	<b>43</b> 0.020	<b>45</b> 0.0077	<b>46</b> 0.0044
	5	<b>43</b> 0.195	<b>47</b> 0.082	<b>49</b> 0.049	<b>52</b> 0.019	<b>54</b> 0.0094	<b>56</b> 0.0039
	6	<b>51</b> 0.180	<b>55</b> 0.083	<b>58</b> 0.042	<b>60</b> 0.025	<b>63</b> 0.0097	<b>65</b> 0.0047
	7	<b>58</b> 0.196	<b>63</b> 0.083	<b>66</b> 0.045	<b>69</b> 0.022	<b>72</b> 0.0098	<b>75</b> 0.0037
	8	<b>66</b> 0.181	<b>70</b> 0.098	<b>74</b> 0.047	<b>78</b> 0.020	<b>81</b> 0.0096	<b>84</b> 0.0041
	9	<b>73</b> 0.193	<b>78</b> 0.096	<b>82</b> 0.049	<b>86</b> 0.023	<b>90</b> 0.0093	<b>93</b> 0.0043
	10	<b>81</b> 0.180	<b>86</b> 0.093	<b>91</b> 0.043	<b>94</b> 0.025	<b>99</b> 0.0090	<b>102</b> 0.0044
	11	<b>88</b> 0.190	<b>94</b> 0.091	<b>99</b> 0.044	<b>103</b> 0.023	<b>108</b> 0.0086	<b>111</b> 0.0045
	12	<b>95</b> 0.198	<b>102</b> 0.089	<b>107</b> 0.045	<b>111</b> 0.024	<b>117</b> 0.0083	<b>120</b> 0.0045
13	2	<b>22</b> 0.171	<b>24</b> 0.076	<b>25</b> 0.038	<b>26</b> 0.019		
	3	<b>30</b> 0.189	<b>33</b> 0.082	<b>35</b> 0.039	<b>36</b> 0.025	<b>38</b> 0.0071	<b>39</b> 0.0036
	4	<b>39</b> 0.163	<b>42</b> 0.079	<b>44</b> 0.045	<b>46</b> 0.023	<b>49</b> 0.0059	<b>50</b> 0.0034
	5	<b>47</b> 0.173	<b>50</b> 0.095	<b>53</b> 0.046	<b>56</b> 0.019	<b>58</b> 0.0098	<b>60</b> 0.0044
	6	<b>55</b> 0.179	<b>59</b> 0.087	<b>62</b> 0.046	<b>65</b> 0.022	<b>68</b> 0.0092	<b>70</b> 0.0047
	7	<b>63</b> 0.183	<b>67</b> 0.097	<b>71</b> 0.046	<b>74</b> 0.024	<b>78</b> 0.0085	<b>80</b> 0.0047
	8	<b>71</b> 0.185	<b>76</b> 0.089	<b>80</b> 0.045	<b>83</b> 0.025	<b>87</b> 0.0099	<b>90</b> 0.0045
	9	<b>79</b> 0.186	<b>84</b> 0.096	<b>89</b> 0.043	<b>93</b> 0.021	<b>97</b> 0.0089	<b>100</b> 0.0043
	10	<b>87</b> 0.186	<b>93</b> 0.088	<b>97</b> 0.049	<b>102</b> 0.021	<b>106</b> 0.0099	<b>110</b> 0.0041
	11	<b>95</b> 0.186	<b>101</b> 0.093	<b>106</b> 0.047	<b>111</b> 0.022	<b>116</b> 0.0088	<b>119</b> 0.0048
	12	<b>103</b> 0.186	<b>109</b> 0.098	<b>115</b> 0.046	<b>120</b> 0.022	<b>125</b> 0.0096	<b>129</b> 0.0045
	13	<b>111</b> 0.186	<b>118</b> 0.091	<b>124</b> 0.044	<b>129</b> 0.022	<b>135</b> 0.0086	<b>139</b> 0.0042
14	2	<b>23</b> 0.200	<b>25</b> 0.100	<b>27</b> 0.033	<b>28</b> 0.017		
	3	<b>32</b> 0.197	<b>35</b> 0.091	<b>37</b> 0.047	<b>39</b> 0.021	<b>41</b> 0.0059	<b>42</b> 0.0029
	4	<b>41</b> 0.192	<b>45</b> 0.079	<b>47</b> 0.046	<b>49</b> 0.025	<b>52</b> 0.0078	<b>53</b> 0.0046
	5	<b>50</b> 0.186	<b>54</b> 0.087	<b>57</b> 0.044	<b>60</b> 0.019	<b>63</b> 0.0072	<b>64</b> 0.0050
	6	<b>59</b> 0.179	<b>63</b> 0.091	<b>67</b> 0.041	<b>70</b> 0.020	<b>73</b> 0.0087	<b>75</b> 0.0046
	7	<b>67</b> 0.197	<b>72</b> 0.094	<b>76</b> 0.046	<b>79</b> 0.025	<b>83</b> 0.0097	<b>86</b> 0.0042
	8	<b>76</b> 0.188	<b>81</b> 0.095	<b>86</b> 0.042	<b>89</b> 0.024	<b>94</b> 0.0081	<b>96</b> 0.0050
	9	<b>85</b> 0.179	<b>90</b> 0.096	<b>95</b> 0.046	<b>99</b> 0.023	<b>104</b> 0.0086	<b>107</b> 0.0043
	10	<b>93</b> 0.192	<b>99</b> 0.096	<b>104</b> 0.048	<b>109</b> 0.022	<b>114</b> 0.0089	<b>117</b> 0.0048
	11	<b>102</b> 0.183	<b>108</b> 0.095	<b>114</b> 0.044	<b>118</b> 0.025	<b>124</b> 0.0090	<b>128</b> 0.0042
	12	<b>110</b> 0.193	<b>117</b> 0.095	<b>123</b> 0.046	<b>128</b> 0.023	<b>134</b> 0.0091	<b>138</b> 0.0045
	13	<b>119</b> 0.185	<b>126</b> 0.095	<b>132</b> 0.048	<b>138</b> 0.022	<b>144</b> 0.0091	<b>148</b> 0.0047
	14	<b>127</b> 0.194	<b>135</b> 0.094	<b>141</b> 0.050	<b>147</b> 0.024	<b>154</b> 0.0091	<b>158</b> 0.0049
15	2	<b>25</b> 0.177	<b>27</b> 0.088	<b>29</b> 0.029	<b>30</b> 0.015		
	3	<b>35</b> 0.164	<b>35</b> 0.076	<b>40</b> 0.039	<b>42</b> 0.017	<b>43</b> 0.0098	<b>44</b> 0.0049
	4	<b>44</b> 0.185	<b>48</b> 0.080	<b>50</b> 0.049	<b>53</b> 0.020	<b>55</b> 0.0093	<b>57</b> 0.0036

**TABLE 6 Critical values and P-values of  $U_s$  for the Wilcoxon-Mann-Whitney test (continued)**

<i>n</i>	<i>n'</i>	0.20	0.10	0.05	0.025	0.01	0.005
5	5	53 0.197	57 0.098	61 0.042	64 0.019	67 0.0077	69 0.0037
	6	63 0.178	67 0.095	71 0.045	74 0.023	78 0.0084	80 0.0046
	7	72 0.185	77 0.091	81 0.047	85 0.021	89 0.0085	92 0.0038
	8	81 0.190	87 0.087	91 0.047	95 0.024	100 0.0085	103 0.0042
	9	90 0.194	96 0.096	101 0.048	106 0.021	111 0.0083	114 0.0044
	10	99 0.196	106 0.091	111 0.048	116 0.023	121 0.0096	125 0.0044
	11	108 0.198	115 0.097	121 0.047	126 0.024	132 0.0092	136 0.0045
	12	117 0.200	125 0.093	131 0.047	136 0.025	143 0.0087	147 0.0044
	13	127 0.185	134 0.098	141 0.046	147 0.022	153 0.0096	158 0.0044
	14	136 0.186	144 0.093	151 0.046	157 0.023	164 0.0091	169 0.0043
	15	145 0.187	153 0.098	161 0.045	167 0.024	174 0.0099	179 0.0049
	16	2 0.157	29 0.078	31 0.026	32 0.013		
	3	37 0.171	40 0.085	42 0.048	44 0.023	46 0.0083	47 0.0041
	4	47 0.178	50 0.100	53 0.050	56 0.022	59 0.0074	60 0.0050
	5	57 0.179	61 0.091	65 0.040	67 0.025	71 0.0082	73 0.0041
	6	67 0.178	71 0.098	75 0.049	79 0.021	83 0.0080	85 0.0045
	7	76 0.198	82 0.089	86 0.047	90 0.023	94 0.0096	97 0.0046
	8	86 0.192	92 0.093	97 0.045	101 0.023	106 0.0087	109 0.0045
	9	96 0.187	102 0.095	107 0.049	112 0.023	117 0.0096	121 0.0043
	10	106 0.182	112 0.097	118 0.047	123 0.023	129 0.0087	133 0.0041
	11	115 0.195	122 0.099	129 0.044	134 0.023	140 0.0093	144 0.0047
	12	125 0.189	132 0.100	139 0.047	145 0.023	151 0.0097	156 0.0044
	13	134 0.199	143 0.092	149 0.050	156 0.022	163 0.0087	167 0.0048
	14	144 0.193	153 0.093	160 0.047	166 0.025	174 0.0091	179 0.0045
	15	154 0.188	163 0.093	170 0.049	177 0.024	185 0.0093	190 0.0048
	16	163 0.196	173 0.094	181 0.047	188 0.023	196 0.0096	202 0.0045
17	2	28 0.187	31 0.070	32 0.047	33 0.023		
	3	39 0.179	42 0.093	45 0.040	47 0.019	49 0.0070	50 0.0035
	4	50 0.172	53 0.099	57 0.040	59 0.024	62 0.0090	64 0.0040
	5	60 0.189	65 0.085	68 0.048	71 0.025	75 0.0086	77 0.0046
	6	71 0.177	76 0.087	80 0.044	83 0.024	87 0.0099	90 0.0045
	7	81 0.187	86 0.100	91 0.047	95 0.024	100 0.0085	103 0.0042
	8	91 0.194	97 0.098	102 0.050	107 0.023	112 0.0090	115 0.0048
	9	101 0.200	108 0.095	114 0.045	118 0.025	124 0.0092	128 0.0043
	10	112 0.187	119 0.093	125 0.046	130 0.024	136 0.0093	140 0.0047
	11	122 0.191	130 0.091	136 0.047	142 0.022	136 0.0093	152 0.0049
	12	132 0.195	140 0.097	147 0.048	153 0.024	160 0.0093	165 0.0043
	13	142 0.198	151 0.095	158 0.048	164 0.025	172 0.0091	177 0.0045
	14	153 0.186	161 0.100	169 0.048	176 0.023	184 0.0090	189 0.0046
	15	163 0.189	172 0.097	180 0.049	187 0.024	195 0.0100	201 0.0047
	16	173 0.191	183 0.094	191 0.049	199 0.023	207 0.0097	213 0.0048
	17	183 0.193	193 0.099	202 0.049	210 0.024	219 0.0095	225 0.0048
18	2	30 0.168	32 0.095	34 0.042	35 0.021		
	3	41 0.185	45 0.080	47 0.047	49 0.024	52 0.0060	53 0.0030
	4	52 0.195	56 0.098	60 0.042	63 0.019	66 0.0074	67 0.0049
	5	63 0.200	68 0.094	72 0.046	75 0.024	79 0.0089	81 0.0049
	6	74 0.199	80 0.090	84 0.047	88 0.022	92 0.0094	95 0.0044
	7	85 0.198	91 0.097	96 0.047	100 0.025	105 0.0094	108 0.0049
	8	96 0.196	103 0.091	108 0.047	113 0.022	118 0.0092	122 0.0042

**TABLE 6 Critical values and P-values of  $U_s$  for the Wilcoxon-Mann-Whitney test (continued)**

<i>n</i>	<i>n'</i>	0.20	0.10	0.05	0.025	0.01	0.005
	9	<b>107</b> 0.194	<b>114</b> 0.095	<b>120</b> 0.046	<b>125</b> 0.023	<b>131</b> 0.0089	<b>135</b> 0.0043
	10	<b>118</b> 0.191	<b>125</b> 0.099	<b>132</b> 0.045	<b>137</b> 0.024	<b>143</b> 0.0100	<b>148</b> 0.0044
	11	<b>129</b> 0.188	<b>137</b> 0.092	<b>143</b> 0.049	<b>149</b> 0.024	<b>156</b> 0.0094	<b>161</b> 0.0043
	12	<b>139</b> 0.200	<b>148</b> 0.095	<b>155</b> 0.048	<b>161</b> 0.025	<b>169</b> 0.0089	<b>173</b> 0.0050
	13	<b>150</b> 0.196	<b>159</b> 0.097	<b>167</b> 0.046	<b>173</b> 0.025	<b>181</b> 0.0095	<b>186</b> 0.0049
	14	<b>161</b> 0.193	<b>170</b> 0.099	<b>178</b> 0.049	<b>185</b> 0.025	<b>194</b> 0.0089	<b>199</b> 0.0047
	15	<b>172</b> 0.190	<b>182</b> 0.093	<b>190</b> 0.048	<b>197</b> 0.025	<b>206</b> 0.0094	<b>212</b> 0.0046
	16	<b>182</b> 0.199	<b>193</b> 0.095	<b>202</b> 0.046	<b>209</b> 0.025	<b>218</b> 0.0099	<b>224</b> 0.0050
	17	<b>193</b> 0.195	<b>204</b> 0.096	<b>213</b> 0.049	<b>221</b> 0.025	<b>231</b> 0.0093	<b>237</b> 0.0048
	18	<b>204</b> 0.192	<b>215</b> 0.097	<b>225</b> 0.047	<b>233</b> 0.024	<b>243</b> 0.0096	<b>250</b> 0.0046
19	2	<b>31</b> 0.191	<b>34</b> 0.086	<b>36</b> 0.038	<b>37</b> 0.019	<b>38</b> 0.0095	
	3	<b>43</b> 0.191	<b>47</b> 0.087	<b>50</b> 0.040	<b>52</b> 0.021	<b>54</b> 0.0091	<b>56</b> 0.0026
	4	<b>55</b> 0.188	<b>59</b> 0.097	<b>63</b> 0.044	<b>66</b> 0.021	<b>69</b> 0.0086	<b>71</b> 0.0041
	5	<b>67</b> 0.183	<b>72</b> 0.088	<b>76</b> 0.044	<b>79</b> 0.024	<b>83</b> 0.0093	<b>86</b> 0.0039
	6	<b>78</b> 0.198	<b>84</b> 0.092	<b>89</b> 0.043	<b>93</b> 0.021	<b>97</b> 0.0090	<b>100</b> 0.0044
	7	<b>90</b> 0.188	<b>96</b> 0.094	<b>101</b> 0.048	<b>106</b> 0.022	<b>111</b> 0.0085	<b>114</b> 0.0045
	8	<b>101</b> 0.198	<b>108</b> 0.095	<b>114</b> 0.045	<b>119</b> 0.022	<b>124</b> 0.0094	<b>128</b> 0.0044
	9	<b>113</b> 0.188	<b>120</b> 0.095	<b>126</b> 0.048	<b>131</b> 0.025	<b>138</b> 0.0086	<b>142</b> 0.0043
	10	<b>124</b> 0.195	<b>132</b> 0.094	<b>138</b> 0.050	<b>144</b> 0.024	<b>151</b> 0.0091	<b>155</b> 0.0048
	11	<b>136</b> 0.185	<b>144</b> 0.094	<b>151</b> 0.047	<b>157</b> 0.023	<b>164</b> 0.0094	<b>169</b> 0.0045
	12	<b>147</b> 0.191	<b>156</b> 0.093	<b>163</b> 0.048	<b>170</b> 0.023	<b>177</b> 0.0097	<b>182</b> 0.0049
	13	<b>158</b> 0.195	<b>167</b> 0.100	<b>175</b> 0.049	<b>182</b> 0.025	<b>190</b> 0.0098	<b>196</b> 0.0045
	14	<b>169</b> 0.199	<b>179</b> 0.098	<b>188</b> 0.046	<b>195</b> 0.024	<b>203</b> 0.0099	<b>209</b> 0.0048
	15	<b>181</b> 0.190	<b>191</b> 0.096	<b>200</b> 0.047	<b>208</b> 0.023	<b>216</b> 0.0100	<b>223</b> 0.0045
	16	<b>192</b> 0.194	<b>203</b> 0.095	<b>212</b> 0.048	<b>220</b> 0.024	<b>230</b> 0.0090	<b>236</b> 0.0047
	17	<b>203</b> 0.196	<b>214</b> 0.100	<b>224</b> 0.049	<b>233</b> 0.023	<b>242</b> 0.0100	<b>249</b> 0.0048
	18	<b>214</b> 0.199	<b>226</b> 0.098	<b>236</b> 0.049	<b>245</b> 0.024	<b>255</b> 0.0100	<b>262</b> 0.0050
	19	<b>226</b> 0.191	<b>238</b> 0.096	<b>248</b> 0.050	<b>258</b> 0.023	<b>268</b> 0.0099	<b>276</b> 0.0046
20	2	<b>33</b> 0.173	<b>36</b> 0.078	<b>38</b> 0.035	<b>39</b> 0.017	<b>40</b> 0.0087	
	3	<b>45</b> 0.197	<b>49</b> 0.094	<b>52</b> 0.046	<b>55</b> 0.018	<b>57</b> 0.0079	<b>58</b> 0.0045
	4	<b>58</b> 0.183	<b>62</b> 0.097	<b>66</b> 0.045	<b>69</b> 0.023	<b>72</b> 0.0100	<b>75</b> 0.0034
	5	<b>70</b> 0.192	<b>75</b> 0.097	<b>80</b> 0.042	<b>83</b> 0.024	<b>87</b> 0.0096	<b>90</b> 0.0043
	6	<b>82</b> 0.196	<b>88</b> 0.095	<b>93</b> 0.046	<b>97</b> 0.023	<b>102</b> 0.0087	<b>105</b> 0.0043
	7	<b>94</b> 0.198	<b>101</b> 0.092	<b>106</b> 0.048	<b>111</b> 0.022	<b>116</b> 0.0093	<b>120</b> 0.0041
	8	<b>106</b> 0.199	<b>113</b> 0.099	<b>119</b> 0.049	<b>124</b> 0.025	<b>130</b> 0.0096	<b>134</b> 0.0047
	9	<b>118</b> 0.199	<b>126</b> 0.095	<b>132</b> 0.049	<b>138</b> 0.023	<b>144</b> 0.0097	<b>149</b> 0.0043
	10	<b>130</b> 0.198	<b>138</b> 0.100	<b>145</b> 0.049	<b>151</b> 0.024	<b>158</b> 0.0096	<b>163</b> 0.0045
	11	<b>142</b> 0.197	<b>151</b> 0.095	<b>158</b> 0.049	<b>165</b> 0.023	<b>172</b> 0.0095	<b>177</b> 0.0047
	12	<b>154</b> 0.195	<b>163</b> 0.099	<b>171</b> 0.048	<b>178</b> 0.024	<b>186</b> 0.0092	<b>191</b> 0.0048
	13	<b>166</b> 0.194	<b>176</b> 0.094	<b>184</b> 0.048	<b>191</b> 0.024	<b>200</b> 0.0090	<b>205</b> 0.0049
	14	<b>178</b> 0.192	<b>188</b> 0.097	<b>197</b> 0.047	<b>204</b> 0.025	<b>213</b> 0.0098	<b>219</b> 0.0049
	15	<b>190</b> 0.191	<b>200</b> 0.099	<b>210</b> 0.046	<b>218</b> 0.023	<b>227</b> 0.0095	<b>233</b> 0.0049
	16	<b>201</b> 0.200	<b>213</b> 0.095	<b>222</b> 0.049	<b>231</b> 0.024	<b>241</b> 0.0091	<b>247</b> 0.0049
	17	<b>213</b> 0.198	<b>225</b> 0.097	<b>235</b> 0.049	<b>244</b> 0.024	<b>254</b> 0.0097	<b>261</b> 0.0048
	18	<b>225</b> 0.196	<b>237</b> 0.099	<b>248</b> 0.048	<b>257</b> 0.024	<b>268</b> 0.0094	<b>275</b> 0.0048
	19	<b>237</b> 0.194	<b>250</b> 0.095	<b>261</b> 0.047	<b>270</b> 0.024	<b>281</b> 0.0099	<b>289</b> 0.0047
	20	<b>249</b> 0.192	<b>262</b> 0.097	<b>273</b> 0.049	<b>283</b> 0.025	<b>295</b> 0.0095	<b>303</b> 0.0047

**TABLE 7 Critical Values and P-Values of  $B_s$  for the Sign Test**

Note: Because the Sign test null distribution is discrete, this table provides selected values of the test statistic  $B_s$  in **bold type** and corresponding  $P$ -values for a non-directional alternative in *italics*. Directional  $P$ -values are found by dividing the numbers in italics in half.

$n_d$	0.20	0.10	0.05	0.02	0.01	0.002	0.001
1							
2							
3							
4							
5	<b>5</b> 0.063	<b>5</b> 0.063					
6	<b>6</b> 0.031	<b>6</b> 0.031	<b>6</b> 0.031				
7	<b>6</b> 0.125	<b>7</b> 0.016	<b>7</b> 0.016	<b>7</b> 0.016			
8	<b>7</b> 0.070	<b>7</b> 0.070	<b>8</b> 0.008	<b>8</b> 0.008	<b>8</b> 0.008		
9	<b>7</b> 0.180	<b>8</b> 0.039	<b>8</b> 0.039	<b>9</b> 0.004	<b>9</b> 0.004		
10	<b>8</b> 0.109	<b>9</b> 0.021	<b>9</b> 0.021	<b>10</b> 0.002	<b>10</b> 0.002	<b>10</b> 0.0020	
11	<b>9</b> 0.065	<b>9</b> 0.065	<b>10</b> 0.012	<b>10</b> 0.012	<b>11</b> 0.001	<b>11</b> 0.0010	<b>11</b> 0.0010
12	<b>9</b> 0.146	<b>10</b> 0.039	<b>10</b> 0.039	<b>11</b> 0.006	<b>11</b> 0.006	<b>12</b> 0.0005	<b>12</b> 0.0005
13	<b>10</b> 0.092	<b>10</b> 0.093	<b>11</b> 0.023	<b>12</b> 0.003	<b>12</b> 0.003	<b>13</b> 0.0002	<b>13</b> 0.0002
14	<b>10</b> 0.180	<b>11</b> 0.057	<b>12</b> 0.013	<b>12</b> 0.013	<b>13</b> 0.0018	<b>13</b> 0.0018	<b>14</b> 0.0001
15	<b>11</b> 0.118	<b>12</b> 0.035	<b>12</b> 0.035	<b>13</b> 0.007	<b>13</b> 0.007	<b>14</b> 0.0010	<b>14</b> 0.0010
16	<b>12</b> 0.077	<b>12</b> 0.077	<b>13</b> 0.021	<b>14</b> 0.004	<b>14</b> 0.004	<b>15</b> 0.0005	<b>15</b> 0.0005
17	<b>12</b> 0.143	<b>13</b> 0.049	<b>13</b> 0.049	<b>14</b> 0.013	<b>15</b> 0.002	<b>16</b> 0.0003	<b>16</b> 0.0003
18	<b>13</b> 0.096	<b>13</b> 0.096	<b>14</b> 0.031	<b>15</b> 0.008	<b>15</b> 0.008	<b>16</b> 0.0013	<b>17</b> 0.0001
19	<b>13</b> 0.167	<b>14</b> 0.064	<b>15</b> 0.019	<b>15</b> 0.019	<b>16</b> 0.004	<b>17</b> 0.0007	<b>17</b> 0.0007
20	<b>14</b> 0.115	<b>15</b> 0.041	<b>15</b> 0.041	<b>16</b> 0.012	<b>17</b> 0.003	<b>18</b> 0.0004	<b>18</b> 0.0004
21	<b>14</b> 0.189	<b>15</b> 0.078	<b>16</b> 0.027	<b>17</b> 0.007	<b>17</b> 0.007	<b>18</b> 0.0015	<b>19</b> 0.0002
22	<b>15</b> 0.134	<b>16</b> 0.052	<b>17</b> 0.017	<b>17</b> 0.017	<b>18</b> 0.004	<b>19</b> 0.0009	<b>19</b> 0.0009
23	<b>16</b> 0.093	<b>16</b> 0.093	<b>17</b> 0.037	<b>18</b> 0.011	<b>19</b> 0.003	<b>20</b> 0.0005	<b>20</b> 0.0005
24	<b>16</b> 0.152	<b>17</b> 0.064	<b>18</b> 0.023	<b>19</b> 0.007	<b>19</b> 0.007	<b>20</b> 0.0015	<b>21</b> 0.0003
25	<b>17</b> 0.108	<b>18</b> 0.043	<b>18</b> 0.043	<b>19</b> 0.015	<b>20</b> 0.004	<b>21</b> 0.0009	<b>21</b> 0.0009
26	<b>17</b> 0.168	<b>18</b> 0.076	<b>19</b> 0.029	<b>20</b> 0.009	<b>20</b> 0.009	<b>22</b> 0.0005	<b>22</b> 0.0005
27	<b>18</b> 0.122	<b>19</b> 0.052	<b>20</b> 0.019	<b>20</b> 0.019	<b>21</b> 0.006	<b>22</b> 0.0015	<b>23</b> 0.0003
28	<b>18</b> 0.185	<b>19</b> 0.087	<b>20</b> 0.036	<b>21</b> 0.013	<b>22</b> 0.004	<b>23</b> 0.0009	<b>23</b> 0.0009
29	<b>19</b> 0.136	<b>20</b> 0.061	<b>21</b> 0.024	<b>22</b> 0.008	<b>22</b> 0.008	<b>24</b> 0.0005	<b>24</b> 0.0005
30	<b>20</b> 0.099	<b>20</b> 0.099	<b>21</b> 0.043	<b>22</b> 0.016	<b>23</b> 0.005	<b>24</b> 0.0014	<b>25</b> 0.0003
31	<b>20</b> 0.152	<b>21</b> 0.071	<b>22</b> 0.029	<b>23</b> 0.011	<b>24</b> 0.003	<b>25</b> 0.0009	<b>25</b> 0.0009

**TABLE 8 Critical Values and P-Values of  $W_s$  for the Wilcoxon Signed-Rank Test**

Note: Because the Wilcoxon Signed-Rank test null distribution is discrete, this table provides selected values of the test statistic  $W_s$  in bold type and corresponding  $P$ -values for a non-directional alternative in *italics*. Directional  $P$ -values are found by dividing the numbers in italics in half.

<i>n</i>	0.20	0.10	0.05	0.02	0.01	0.002	0.001
1							
2							
3							
4	<b>10</b> 0.125						
5	<b>13</b> 0.188	<b>15</b> 0.063					
6	<b>18</b> 0.156	<b>19</b> 0.093	<b>21</b> 0.031				
7	<b>23</b> 0.156	<b>25</b> 0.078	<b>26</b> 0.047	<b>28</b> 0.016			
8	<b>28</b> 0.195	<b>31</b> 0.078	<b>33</b> 0.039	<b>35</b> 0.016	<b>36</b> 0.0078		
9	<b>35</b> 0.164	<b>37</b> 0.098	<b>40</b> 0.039	<b>42</b> 0.020	<b>44</b> 0.0078		
10	<b>41</b> 0.193	<b>45</b> 0.084	<b>47</b> 0.049	<b>50</b> 0.020	<b>52</b> 0.0098	<b>55</b> 0.0020	
11	<b>49</b> 0.175	<b>53</b> 0.083	<b>56</b> 0.042	<b>59</b> 0.019	<b>61</b> 0.0098	<b>65</b> 0.0020	<b>66</b> 0.0010
12	<b>57</b> 0.176	<b>61</b> 0.092	<b>65</b> 0.042	<b>69</b> 0.016	<b>71</b> 0.0093	<b>76</b> 0.0015	<b>77</b> 0.0010
13	<b>65</b> 0.191	<b>70</b> 0.094	<b>74</b> 0.048	<b>79</b> 0.017	<b>82</b> 0.0081	<b>87</b> 0.0017	<b>89</b> 0.0007
14	<b>74</b> 0.194	<b>80</b> 0.091	<b>84</b> 0.049	<b>90</b> 0.017	<b>93</b> 0.0085	<b>99</b> 0.0017	<b>101</b> 0.0009
15	<b>84</b> 0.188	<b>90</b> 0.095	<b>95</b> 0.048	<b>101</b> 0.018	<b>105</b> 0.0084	<b>112</b> 0.0015	<b>114</b> 0.0009
16	<b>94</b> 0.193	<b>101</b> 0.093	<b>107</b> 0.044	<b>113</b> 0.018	<b>117</b> 0.0092	<b>125</b> 0.0017	<b>128</b> 0.0008
17	<b>105</b> 0.190	<b>112</b> 0.098	<b>119</b> 0.045	<b>126</b> 0.017	<b>130</b> 0.0093	<b>139</b> 0.0017	<b>142</b> 0.0008
18	<b>116</b> 0.196	<b>124</b> 0.099	<b>131</b> 0.048	<b>139</b> 0.018	<b>144</b> 0.0090	<b>153</b> 0.0019	<b>157</b> 0.0008
19	<b>128</b> 0.196	<b>137</b> 0.096	<b>144</b> 0.049	<b>153</b> 0.018	<b>158</b> 0.0094	<b>169</b> 0.0017	<b>172</b> 0.0010
20	<b>141</b> 0.189	<b>150</b> 0.097	<b>158</b> 0.048	<b>167</b> 0.019	<b>173</b> 0.0094	<b>184</b> 0.0020	<b>189</b> 0.0009
21	<b>154</b> 0.191	<b>164</b> 0.096	<b>173</b> 0.046	<b>182</b> 0.019	<b>189</b> 0.0090	<b>201</b> 0.0019	<b>206</b> 0.0009
22	<b>167</b> 0.198	<b>178</b> 0.094	<b>188</b> 0.046	<b>198</b> 0.019	<b>205</b> 0.0093	<b>218</b> 0.0019	<b>223</b> 0.0009
23	<b>182</b> 0.190	<b>193</b> 0.098	<b>203</b> 0.048	<b>214</b> 0.020	<b>222</b> 0.0091	<b>236</b> 0.0019	<b>241</b> 0.0010
24	<b>196</b> 0.197	<b>209</b> 0.095	<b>219</b> 0.049	<b>231</b> 0.019	<b>239</b> 0.0096	<b>255</b> 0.0018	<b>260</b> 0.0010
25	<b>212</b> 0.191	<b>225</b> 0.096	<b>236</b> 0.048	<b>249</b> 0.019	<b>257</b> 0.0096	<b>274</b> 0.0018	<b>280</b> 0.0009
26	<b>227</b> 0.199	<b>241</b> 0.099	<b>253</b> 0.049	<b>267</b> 0.019	<b>276</b> 0.0094	<b>293</b> 0.0020	<b>300</b> 0.0009
27	<b>244</b> 0.194	<b>259</b> 0.095	<b>271</b> 0.049	<b>286</b> 0.019	<b>295</b> 0.0096	<b>314</b> 0.0019	<b>321</b> 0.0009
28	<b>261</b> 0.194	<b>276</b> 0.099	<b>290</b> 0.048	<b>305</b> 0.019	<b>315</b> 0.0095	<b>335</b> 0.0019	<b>342</b> 0.0010
29	<b>278</b> 0.198	<b>295</b> 0.096	<b>309</b> 0.048	<b>325</b> 0.019	<b>335</b> 0.0099	<b>356</b> 0.0020	<b>364</b> 0.0010
30	<b>296</b> 0.198	<b>314</b> 0.096	<b>328</b> 0.050	<b>345</b> 0.020	<b>356</b> 0.0099	<b>379</b> 0.0019	<b>387</b> 0.0010
31	<b>315</b> 0.195	<b>333</b> 0.098	<b>349</b> 0.048	<b>366</b> 0.020	<b>378</b> 0.0097	<b>402</b> 0.0019	<b>410</b> 0.0010

**TABLE 9 Critical Values of the Chi-Square Distribution**

Note: Column headings are non-directional (omni-directional)  $P$ -values. If  $H_A$  is directional (which is only possible when  $df = 1$ ), the directional  $P$ -values are found by dividing the column headings in half.

df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	1.64	2.71	3.84	5.41	6.63	10.83	15.14
2	3.22	4.61	5.99	7.82	9.21	13.82	18.42
3	4.64	6.25	7.81	9.84	11.34	16.27	21.11
4	5.99	7.78	9.49	11.67	13.28	18.47	23.51
5	7.29	9.24	11.07	13.39	15.09	20.51	25.74
6	8.56	10.64	12.59	15.03	16.81	22.46	27.86
7	9.80	12.02	14.07	16.62	18.48	24.32	29.88
8	11.03	13.36	15.51	18.17	20.09	26.12	31.83
9	12.24	14.68	16.92	19.68	21.67	27.88	33.72
10	13.44	15.99	18.31	21.16	23.21	29.59	35.56
11	14.63	17.28	19.68	22.62	24.72	31.26	37.37
12	15.81	18.55	21.03	24.05	26.22	32.91	39.13
13	16.98	19.81	22.36	25.47	27.69	34.53	40.87
14	18.15	21.06	23.68	26.87	29.14	36.12	42.58
15	19.31	22.31	25.00	28.26	30.58	37.70	44.26
16	20.47	23.54	26.30	29.63	32.00	39.25	45.92
17	21.61	24.77	27.59	31.00	33.41	40.79	47.57
18	22.76	25.99	28.87	32.35	34.81	42.31	49.19
19	23.90	27.20	30.14	33.69	36.19	43.82	50.80
20	25.04	28.41	31.41	35.02	37.57	45.31	52.39
21	26.17	29.62	32.67	36.34	38.93	46.80	53.96
22	27.30	30.81	33.92	37.66	40.29	48.27	55.52
23	28.43	32.01	35.17	38.97	41.64	49.73	57.08
24	29.55	33.20	36.42	40.27	42.98	51.18	58.61
25	30.68	34.38	37.65	41.57	44.31	52.62	60.14
26	31.79	35.56	38.89	42.86	45.64	54.05	61.66
27	32.91	36.74	40.11	44.14	46.96	55.48	63.16
28	34.03	37.92	41.34	45.42	48.28	56.89	64.66
29	35.14	39.09	42.56	46.69	49.59	58.30	66.15
30	36.25	40.26	43.77	47.96	50.89	59.70	67.63

**TABLE 10 Critical Values of the F Distribution**

Numerator df = 1

Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	9.47	39.86	161	101 <sup>1</sup>	405 <sup>1</sup>	406 <sup>3</sup>	405 <sup>5</sup>
2	3.56	8.53	18.51	48.51	98.50	998	100 <sup>2</sup>
3	2.68	5.54	10.13	20.62	34.12	167	784
4	2.35	4.54	7.71	14.04	21.20	74.14	242
5	2.18	4.06	6.61	11.32	16.26	47.18	125
6	2.07	3.78	5.99	9.88	13.75	35.51	82.49
7	2.00	3.59	5.59	8.99	12.25	29.25	62.17
8	1.95	3.46	5.32	8.39	11.26	25.41	50.69
9	1.91	3.36	5.12	7.96	10.56	22.86	43.48
10	1.88	3.29	4.96	7.64	10.04	21.04	38.58
11	1.86	3.23	4.84	7.39	9.65	19.69	35.06
12	1.84	3.18	4.75	7.19	9.33	18.64	32.43
13	1.82	3.14	4.67	7.02	9.07	17.82	30.39
14	1.81	3.10	4.60	6.89	8.86	17.14	28.77
15	1.80	3.07	4.54	6.77	8.68	16.59	27.45
16	1.79	3.05	4.49	6.67	8.53	16.12	26.36
17	1.78	3.03	4.45	6.59	8.40	15.72	25.44
18	1.77	3.01	4.41	6.51	8.29	15.38	24.66
19	1.76	2.99	4.38	6.45	8.18	15.08	23.99
20	1.76	2.97	4.35	6.39	8.10	14.82	23.40
21	1.75	2.96	4.32	6.34	8.02	14.59	22.89
22	1.75	2.95	4.30	6.29	7.95	14.38	22.43
23	1.74	2.94	4.28	6.25	7.88	14.20	22.03
24	1.74	2.93	4.26	6.21	7.82	14.03	21.66
25	1.73	2.92	4.24	6.18	7.77	13.88	21.34
26	1.73	2.91	4.23	6.14	7.72	13.74	21.04
27	1.73	2.90	4.21	6.11	7.68	13.61	20.77
28	1.72	2.89	4.20	6.09	7.64	13.50	20.53
29	1.72	2.89	4.18	6.06	7.60	13.39	20.30
30	1.72	2.88	4.17	6.04	7.56	13.29	20.09
40	1.70	2.84	4.08	5.87	7.31	12.61	18.67
60	1.68	2.79	4.00	5.71	7.08	11.97	17.38
100	1.66	2.76	3.94	5.59	6.90	11.50	16.43
140	1.66	2.74	3.91	5.54	6.82	11.30	16.05
$\infty$	1.64	2.71	3.84	5.41	6.63	10.83	15.14

Notation:  $406^3$  means  $406 \times 10^3$ .*Continued*

**TABLE 10 Critical Values of the F Distribution  
(continued)**

Denom. df	Numerator df = 2						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	12.00	49.50	200	125 <sup>1</sup>	500 <sup>1</sup>	500 <sup>3</sup>	500 <sup>5</sup>
2	4.00	9.00	19.00	49.00	99.00	999	100 <sup>2</sup>
3	2.89	5.46	9.55	18.86	30.82	149	695
4	2.47	4.32	6.94	12.14	18.00	61.25	198
5	2.26	3.78	5.79	9.45	13.27	37.12	97.03
6	2.13	3.46	5.14	8.05	10.92	27.00	61.63
7	2.04	3.26	4.74	7.20	9.55	21.69	45.13
8	1.98	3.11	4.46	6.64	8.65	18.49	36.00
9	1.93	3.01	4.26	6.23	8.02	16.39	30.34
10	1.90	2.92	4.10	5.93	7.56	14.91	26.55
11	1.87	2.86	3.98	5.70	7.21	13.81	23.85
12	1.85	2.81	3.89	5.52	6.93	12.97	21.85
13	1.83	2.76	3.81	5.37	6.70	12.31	20.31
14	1.81	2.73	3.74	5.24	6.51	11.78	19.09
15	1.80	2.70	3.68	5.14	6.36	11.34	18.11
16	1.78	2.67	3.63	5.05	6.23	10.97	17.30
17	1.77	2.64	3.59	4.97	6.11	10.66	16.62
18	1.76	2.62	3.55	4.90	6.01	10.39	16.04
19	1.75	2.61	3.52	4.84	5.93	10.16	15.55
20	1.75	2.59	3.49	4.79	5.85	9.95	15.12
21	1.74	2.57	3.47	4.74	5.78	9.77	14.74
22	1.73	2.56	3.44	4.70	5.72	9.61	14.41
23	1.73	2.55	3.42	4.66	5.66	9.47	14.12
24	1.72	2.54	3.40	4.63	5.61	9.34	13.85
25	1.72	2.53	3.39	4.59	5.57	9.22	13.62
26	1.71	2.52	3.37	4.56	5.53	9.12	13.40
27	1.71	2.51	3.35	4.54	5.49	9.02	13.21
28	1.71	2.50	3.34	4.51	5.45	8.93	13.03
29	1.70	2.50	3.33	4.49	5.42	8.85	12.87
30	1.70	2.49	3.32	4.47	5.39	8.77	12.72
40	1.68	2.44	3.23	4.32	5.18	8.25	11.70
60	1.65	2.39	3.15	4.18	4.98	7.77	10.78
100	1.64	2.36	3.09	4.07	4.82	7.41	10.11
140	1.63	2.34	3.06	4.02	4.76	7.26	9.84
$\infty$	1.61	2.30	3.00	3.91	4.61	6.91	9.21

**TABLE 10 Critical Values of the *F* Distribution  
(continued)**

Denom. df	Numerator df = 3						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	13.06	53.59	216	135 <sup>1</sup>	540 <sup>1</sup>	540 <sup>3</sup>	540 <sup>5</sup>
2	4.16	9.16	19.16	49.17	99.17	999	100 <sup>2</sup>
3	2.94	5.39	9.28	18.11	29.46	141	659
4	2.48	4.19	6.59	11.34	16.69	56.18	181
5	2.25	3.62	5.41	8.67	12.06	33.20	86.29
6	2.11	3.29	4.76	7.29	9.78	23.70	53.68
7	2.02	3.07	4.35	6.45	8.45	18.77	38.68
8	1.95	2.92	4.07	5.90	7.59	15.83	30.46
9	1.90	2.81	3.86	5.51	6.99	13.90	25.40
10	1.86	2.73	3.71	5.22	6.55	12.55	22.04
11	1.83	2.66	3.59	4.99	6.22	11.56	19.66
12	1.80	2.61	3.49	4.81	5.95	10.80	17.90
13	1.78	2.56	3.41	4.67	5.74	10.21	16.55
14	1.76	2.52	3.34	4.55	5.56	9.73	15.49
15	1.75	2.49	3.29	4.45	5.42	9.34	14.64
16	1.74	2.46	3.24	4.36	5.29	9.01	13.93
17	1.72	2.44	3.20	4.29	5.18	8.73	13.34
18	1.71	2.42	3.16	4.22	5.09	8.49	12.85
19	1.70	2.40	3.13	4.16	5.01	8.28	12.42
20	1.70	2.38	3.10	4.11	4.94	8.10	12.05
21	1.69	2.36	3.07	4.07	4.87	7.94	11.73
22	1.68	2.35	3.05	4.03	4.82	7.80	11.44
23	1.68	2.34	3.03	3.99	4.76	7.67	11.19
24	1.67	2.33	3.01	3.96	4.72	7.55	10.96
25	1.66	2.32	2.99	3.93	4.68	7.45	10.76
26	1.66	2.31	2.98	3.90	4.64	7.36	10.58
27	1.65	2.30	2.96	3.87	4.60	7.27	10.41
28	1.65	2.29	2.95	3.85	4.57	7.19	10.26
29	1.65	2.28	2.93	3.83	4.54	7.12	10.12
30	1.64	2.28	2.92	3.81	4.51	7.05	9.99
40	1.62	2.23	2.84	3.67	4.31	6.59	9.13
60	1.60	2.18	2.76	3.53	4.13	6.17	8.35
100	1.58	2.14	2.70	3.43	3.98	5.86	7.79
140	1.57	2.12	2.67	3.38	3.92	5.73	7.57
$\infty$	1.55	2.08	2.60	3.28	3.78	5.42	7.04

**TABLE 10 Critical Values of the F Distribution  
(continued)**

Numerator df = 4							
Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	13.64	55.83	225	141 <sup>1</sup>	562 <sup>1</sup>	562 <sup>3</sup>	562 <sup>5</sup>
2	4.24	9.24	19.25	49.25	99.25	999	100 <sup>2</sup>
3	2.96	5.34	9.12	17.69	28.71	137	640
4	2.48	4.11	6.39	10.90	15.98	53.44	172
5	2.24	3.52	5.19	8.23	11.39	31.09	80.53
6	2.09	3.18	4.53	6.86	9.15	21.92	49.42
7	1.99	2.96	4.12	6.03	7.85	17.20	35.22
8	1.92	2.81	3.84	5.49	7.01	14.39	27.49
9	1.87	2.69	3.63	5.10	6.42	12.56	22.77
10	1.83	2.61	3.48	4.82	5.99	11.28	19.63
11	1.80	2.54	3.36	4.59	5.67	10.35	17.42
12	1.77	2.48	3.26	4.42	5.41	9.63	15.79
13	1.75	2.43	3.18	4.28	5.21	9.07	14.55
14	1.73	2.39	3.11	4.16	5.04	8.62	13.57
15	1.71	2.36	3.06	4.06	4.89	8.25	12.78
16	1.70	2.33	3.01	3.97	4.77	7.94	12.14
17	1.68	2.31	2.96	3.90	4.67	7.68	11.60
18	1.67	2.29	2.93	3.84	4.58	7.46	11.14
19	1.66	2.27	2.90	3.78	4.50	7.27	10.75
20	1.65	2.25	2.87	3.73	4.43	7.10	10.41
21	1.65	2.23	2.84	3.69	4.37	6.95	10.12
22	1.64	2.22	2.82	3.65	4.31	6.81	9.86
23	1.63	2.21	2.80	3.61	4.26	6.70	9.63
24	1.63	2.19	2.78	3.58	4.22	6.59	9.42
25	1.62	2.18	2.76	3.55	4.18	6.49	9.24
26	1.62	2.17	2.74	3.52	4.14	6.41	9.07
27	1.61	2.17	2.73	3.50	4.11	6.33	8.92
28	1.61	2.16	2.71	3.47	4.07	6.25	8.79
29	1.60	2.15	2.70	3.45	4.04	6.19	8.66
30	1.60	2.14	2.69	3.43	4.02	6.12	8.54
40	1.57	2.09	2.61	3.30	3.83	5.70	7.76
60	1.55	2.04	2.53	3.16	3.65	5.31	7.06
100	1.53	2.00	2.46	3.06	3.51	5.02	6.55
140	1.52	1.99	2.44	3.02	3.46	4.90	6.35
$\infty$	1.50	1.94	2.37	2.92	3.32	4.62	5.88

**TABLE 10 Critical Values of the F Distribution  
(continued)**

Numerator df = 5

Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	14.01	57.24	230	144 <sup>1</sup>	576 <sup>1</sup>	576 <sup>3</sup>	576 <sup>5</sup>
2	4.28	9.29	19.30	49.30	99.30	999	100 <sup>2</sup>
3	2.97	5.31	9.01	17.43	28.24	135	628
4	2.48	4.05	6.26	10.62	15.52	51.71	166
5	2.23	3.45	5.05	7.95	10.97	29.75	76.91
6	2.08	3.11	4.39	6.58	8.75	20.80	46.75
7	1.97	2.88	3.97	5.76	7.46	16.21	33.06
8	1.90	2.73	3.69	5.22	6.63	13.48	25.63
9	1.85	2.61	3.48	4.84	6.06	11.71	21.11
10	1.80	2.52	3.33	4.55	5.64	10.48	18.12
11	1.77	2.45	3.20	4.34	5.32	9.58	16.02
12	1.74	2.39	3.11	4.16	5.06	8.89	14.47
13	1.72	2.35	3.03	4.02	4.86	8.35	13.29
14	1.70	2.31	2.96	3.90	4.69	7.92	12.37
15	1.68	2.27	2.90	3.81	4.56	7.57	11.62
16	1.67	2.24	2.85	3.72	4.44	7.27	11.01
17	1.65	2.22	2.81	3.65	4.34	7.02	10.50
18	1.64	2.20	2.77	3.59	4.25	6.81	10.07
19	1.63	2.18	2.74	3.53	4.17	6.62	9.71
20	1.62	2.16	2.71	3.48	4.10	6.46	9.39
21	1.61	2.14	2.68	3.44	4.04	6.32	9.11
22	1.61	2.13	2.66	3.40	3.99	6.19	8.87
23	1.60	2.11	2.64	3.36	3.94	6.08	8.65
24	1.59	2.10	2.62	3.33	3.90	5.98	8.46
25	1.59	2.09	2.60	3.30	3.85	5.89	8.28
26	1.58	2.08	2.59	3.28	3.82	5.80	8.13
27	1.58	2.07	2.57	3.25	3.78	5.73	7.99
28	1.57	2.06	2.56	3.23	3.75	5.66	7.86
29	1.57	2.06	2.55	3.21	3.73	5.59	7.74
30	1.57	2.05	2.53	3.19	3.70	5.53	7.63
40	1.54	2.00	2.45	3.05	3.51	5.13	6.90
60	1.51	1.95	2.37	2.92	3.34	4.76	6.25
100	1.49	1.91	2.31	2.82	3.21	4.48	5.78
140	1.48	1.89	2.28	2.78	3.15	4.37	5.59
$\infty$	1.46	1.85	2.21	2.68	3.02	4.10	5.15

**TABLE 10 Critical Values of the F Distribution  
(continued)**

Numerator df = 6							
Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	14.26	58.20	234	146 <sup>1</sup>	586 <sup>1</sup>	586 <sup>3</sup>	586 <sup>5</sup>
2	4.32	9.33	19.33	49.33	99.33	999	100 <sup>2</sup>
3	2.97	5.28	8.94	17.25	27.91	133	620
4	2.47	4.01	6.16	10.42	15.21	50.53	162
5	2.22	3.40	4.95	7.76	10.67	28.83	74.43
6	2.06	3.05	4.28	6.39	8.47	20.03	44.91
7	1.96	2.83	3.87	5.58	7.19	15.52	31.57
8	1.88	2.67	3.58	5.04	6.37	12.86	24.36
9	1.83	2.55	3.37	4.65	5.80	11.13	19.97
10	1.78	2.46	3.22	4.37	5.39	9.93	17.08
11	1.75	2.39	3.09	4.15	5.07	9.05	15.05
12	1.72	2.33	3.00	3.98	4.82	8.38	13.56
13	1.69	2.28	2.92	3.84	4.62	7.86	12.42
14	1.67	2.24	2.85	3.72	4.46	7.44	11.53
15	1.66	2.21	2.79	3.63	4.32	7.09	10.82
16	1.64	2.18	2.74	3.54	4.20	6.80	10.23
17	1.63	2.15	2.70	3.47	4.10	6.56	9.75
18	1.62	2.13	2.66	3.41	4.01	6.35	9.33
19	1.61	2.11	2.63	3.35	3.94	6.18	8.98
20	1.60	2.09	2.60	3.30	3.87	6.02	8.68
21	1.59	2.08	2.57	3.26	3.81	5.88	8.41
22	1.58	2.06	2.55	3.22	3.76	5.76	8.18
23	1.57	2.05	2.53	3.19	3.71	5.65	7.97
24	1.57	2.04	2.51	3.15	3.67	5.55	7.79
25	1.56	2.02	2.49	3.13	3.63	5.46	7.62
26	1.56	2.01	2.47	3.10	3.59	5.38	7.48
27	1.55	2.00	2.46	3.07	3.56	5.31	7.34
28	1.55	2.00	2.45	3.05	3.53	5.24	7.22
29	1.54	1.99	2.43	3.03	3.50	5.18	7.10
30	1.54	1.98	2.42	3.01	3.47	5.12	7.00
40	1.51	1.93	2.34	2.88	3.29	4.73	6.30
60	1.48	1.87	2.25	2.75	3.12	4.37	5.68
100	1.46	1.83	2.19	2.65	2.99	4.11	5.24
140	1.45	1.82	2.16	2.61	2.93	4.00	5.06
$\infty$	1.43	1.77	2.10	2.51	2.80	3.74	4.64

**TABLE 10 Critical Values of the F Distribution  
(continued)**

Numerator df = 7

Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	14.44	58.91	237	148 <sup>1</sup>	593 <sup>1</sup>	593 <sup>3</sup>	593 <sup>5</sup>
2	4.34	9.35	19.35	49.36	99.36	999	100 <sup>2</sup>
3	2.97	5.27	8.89	17.11	27.67	132	614
4	2.47	3.98	6.09	10.27	14.98	49.66	159
5	2.21	3.37	4.88	7.61	10.46	28.16	72.61
6	2.05	3.01	4.21	6.25	8.26	19.46	43.57
7	1.94	2.78	3.79	5.44	6.99	15.02	30.48
8	1.87	2.62	3.50	4.90	6.18	12.40	23.42
9	1.81	2.51	3.29	4.52	5.61	10.70	19.14
10	1.77	2.41	3.14	4.23	5.20	9.52	16.32
11	1.73	2.34	3.01	4.02	4.89	8.66	14.34
12	1.70	2.28	2.91	3.85	4.64	8.00	12.89
13	1.68	2.23	2.83	3.71	4.44	7.49	11.79
14	1.65	2.19	2.76	3.59	4.28	7.08	10.92
15	1.64	2.16	2.71	3.49	4.14	6.74	10.23
16	1.62	2.13	2.66	3.41	4.03	6.46	9.66
17	1.61	2.10	2.61	3.34	3.93	6.22	9.19
18	1.60	2.08	2.58	3.27	3.84	6.02	8.79
19	1.58	2.06	2.54	3.22	3.77	5.85	8.45
20	1.58	2.04	2.51	3.17	3.70	5.69	8.16
21	1.57	2.02	2.49	3.13	3.64	5.56	7.90
22	1.56	2.01	2.46	3.09	3.59	5.44	7.68
23	1.55	1.99	2.44	3.05	3.54	5.33	7.48
24	1.55	1.98	2.42	3.02	3.50	5.23	7.30
25	1.54	1.97	2.40	2.99	3.46	5.15	7.14
26	1.53	1.96	2.39	2.97	3.42	5.07	6.99
27	1.53	1.95	2.37	2.94	3.39	5.00	6.86
28	1.52	1.94	2.36	2.92	3.36	4.93	6.75
29	1.52	1.93	2.35	2.90	3.33	4.87	6.64
30	1.52	1.93	2.33	2.88	3.30	4.82	6.54
40	1.49	1.87	2.25	2.74	3.12	4.44	5.86
60	1.46	1.82	2.17	2.62	2.95	4.09	5.27
100	1.43	1.78	2.10	2.52	2.82	3.83	4.84
140	1.42	1.76	2.08	2.48	2.77	3.72	4.67
$\infty$	1.40	1.72	2.01	2.37	2.64	3.47	4.27

**TABLE 10 Critical Values of the F Distribution  
(continued)**

Numerator df = 8							
Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	14.58	59.44	239	149 <sup>1</sup>	598 <sup>1</sup>	598 <sup>3</sup>	598 <sup>5</sup>
2	4.36	9.37	19.37	49.37	99.37	999	100 <sup>2</sup>
3	2.98	5.25	8.85	17.01	27.49	131	609
4	2.47	3.95	6.04	10.16	14.80	49.00	157
5	2.20	3.34	4.82	7.50	10.29	27.65	71.23
6	2.04	2.98	4.15	6.14	8.10	19.03	42.54
7	1.93	2.75	3.73	5.33	6.84	14.63	29.64
8	1.86	2.59	3.44	4.79	6.03	12.05	22.71
9	1.80	2.47	3.23	4.41	5.47	10.37	18.50
10	1.75	2.38	3.07	4.13	5.06	9.20	15.74
11	1.72	2.30	2.95	3.91	4.74	8.35	13.80
12	1.69	2.24	2.85	3.74	4.50	7.71	12.38
13	1.66	2.20	2.77	3.60	4.30	7.21	11.30
14	1.64	2.15	2.70	3.48	4.14	6.80	10.46
15	1.62	2.12	2.64	3.39	4.00	6.47	9.78
16	1.61	2.09	2.59	3.30	3.89	6.19	9.23
17	1.59	2.06	2.55	3.23	3.79	5.96	8.76
18	1.58	2.04	2.51	3.17	3.71	5.76	8.38
19	1.57	2.02	2.48	3.12	3.63	5.59	8.04
20	1.56	2.00	2.45	3.07	3.56	5.44	7.76
21	1.55	1.98	2.42	3.02	3.51	5.31	7.51
22	1.54	1.97	2.40	2.99	3.45	5.19	7.29
23	1.53	1.95	2.37	2.95	3.41	5.09	7.09
24	1.53	1.94	2.36	2.92	3.36	4.99	6.92
25	1.52	1.93	2.34	2.89	3.32	4.91	6.76
26	1.52	1.92	2.32	2.86	3.29	4.83	6.62
27	1.51	1.91	2.31	2.84	3.26	4.76	6.50
28	1.51	1.90	2.29	2.82	3.23	4.69	6.38
29	1.50	1.89	2.28	2.80	3.20	4.64	6.28
30	1.50	1.88	2.27	2.78	3.17	4.58	6.18
40	1.47	1.83	2.18	2.64	2.99	4.21	5.53
60	1.44	1.77	2.10	2.51	2.82	3.86	4.95
100	1.41	1.73	2.03	2.41	2.69	3.61	4.53
140	1.40	1.71	2.01	2.37	2.64	3.51	4.36
$\infty$	1.38	1.67	1.94	2.27	2.51	3.27	3.98

**TABLE 10 Critical Values of the F Distribution  
(continued)**

Numerator df = 9							
Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	14.68	59.86	241	151 <sup>1</sup>	602 <sup>1</sup>	602 <sup>3</sup>	602 <sup>5</sup>
2	4.37	9.38	19.38	49.39	99.39	999	100 <sup>2</sup>
3	2.98	5.24	8.81	16.93	27.35	130	606
4	2.46	3.94	6.00	10.07	14.66	48.47	155
5	2.20	3.32	4.77	7.42	10.16	27.24	70.13
6	2.03	2.96	4.10	6.05	7.98	18.69	41.73
7	1.93	2.72	3.68	5.24	6.72	14.33	28.99
8	1.85	2.56	3.39	4.70	5.91	11.77	22.14
9	1.79	2.44	3.18	4.33	5.35	10.11	18.00
10	1.74	2.35	3.02	4.04	4.94	8.96	15.27
11	1.70	2.27	2.90	3.83	4.63	8.12	13.37
12	1.67	2.21	2.80	3.66	4.39	7.48	11.98
13	1.65	2.16	2.71	3.52	4.19	6.98	10.92
14	1.63	2.12	2.65	3.40	4.03	6.58	10.09
15	1.61	2.09	2.59	3.30	3.89	6.26	9.42
16	1.59	2.06	2.54	3.22	3.78	5.98	8.88
17	1.58	2.03	2.49	3.15	3.68	5.75	8.43
18	1.56	2.00	2.46	3.09	3.60	5.56	8.05
19	1.55	1.98	2.42	3.03	3.52	5.39	7.72
20	1.54	1.96	2.39	2.98	3.46	5.24	7.44
21	1.53	1.95	2.37	2.94	3.40	5.11	7.19
22	1.53	1.93	2.34	2.90	3.35	4.99	6.98
23	1.52	1.92	2.32	2.87	3.30	4.89	6.79
24	1.51	1.91	2.30	2.83	3.26	4.80	6.62
25	1.51	1.89	2.28	2.81	3.22	4.71	6.47
26	1.50	1.88	2.27	2.78	3.18	4.64	6.33
27	1.49	1.87	2.25	2.76	3.15	4.57	6.21
28	1.49	1.87	2.24	2.73	3.12	4.50	6.09
29	1.49	1.86	2.22	2.71	3.09	4.45	5.99
30	1.48	1.85	2.21	2.69	3.07	4.39	5.90
40	1.45	1.79	2.12	2.56	2.89	4.02	5.26
60	1.42	1.74	2.04	2.43	2.72	3.69	4.69
100	1.40	1.69	1.97	2.33	2.59	3.44	4.29
140	1.39	1.68	1.95	2.29	2.54	3.34	4.12
$\infty$	1.36	1.63	1.88	2.19	2.41	3.10	3.75

**TABLE 10 Critical Values of the *F* Distribution  
(continued)**

Numerator df = 10							
Denom. df	TAIL PROBABILITY						
	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	14.77	60.19	242	151 <sup>1</sup>	606 <sup>1</sup>	606 <sup>3</sup>	606 <sup>5</sup>
2	4.38	9.39	19.40	49.40	99.40	999	100 <sup>2</sup>
3	2.98	5.23	8.79	16.86	27.23	129	603
4	2.46	3.92	5.96	10.00	14.55	48.05	154
5	2.19	3.30	4.74	7.34	10.05	26.92	69.25
6	2.03	2.94	4.06	5.98	7.87	18.41	41.08
7	1.92	2.70	3.64	5.17	6.62	14.08	28.45
8	1.84	2.54	3.35	4.63	5.81	11.54	21.68
9	1.78	2.42	3.14	4.26	5.26	9.89	17.59
10	1.73	2.32	2.98	3.97	4.85	8.75	14.90
11	1.69	2.25	2.85	3.76	4.54	7.92	13.02
12	1.66	2.19	2.75	3.59	4.30	7.29	11.65
13	1.64	2.14	2.67	3.45	4.10	6.80	10.60
14	1.62	2.10	2.60	3.33	3.94	6.40	9.79
15	1.60	2.06	2.54	3.23	3.80	6.08	9.13
16	1.58	2.03	2.49	3.15	3.69	5.81	8.60
17	1.57	2.00	2.45	3.08	3.59	5.58	8.15
18	1.55	1.98	2.41	3.02	3.51	5.39	7.78
19	1.54	1.96	2.38	2.96	3.43	5.22	7.46
20	1.53	1.94	2.35	2.91	3.37	5.08	7.18
21	1.52	1.92	2.32	2.87	3.31	4.95	6.94
22	1.51	1.90	2.30	2.83	3.26	4.83	6.73
23	1.51	1.89	2.27	2.80	3.21	4.73	6.54
24	1.50	1.88	2.25	2.77	3.17	4.64	6.37
25	1.49	1.87	2.24	2.74	3.13	4.56	6.23
26	1.49	1.86	2.22	2.71	3.09	4.48	6.09
27	1.48	1.85	2.20	2.69	3.06	4.41	5.97
28	1.48	1.84	2.19	2.66	3.03	4.35	5.86
29	1.47	1.83	2.18	2.64	3.00	4.29	5.76
30	1.47	1.82	2.16	2.62	2.98	4.24	5.66
40	1.44	1.76	2.08	2.49	2.80	3.87	5.04
60	1.41	1.71	1.99	2.36	2.63	3.54	4.48
100	1.38	1.66	1.93	2.26	2.50	3.30	4.08
140	1.37	1.64	1.90	2.22	2.45	3.20	3.93
$\infty$	1.34	1.60	1.83	2.12	2.32	2.96	3.56

**TABLE 11 Bonferroni Multipliers for 95% Confidence Intervals**

The values given in the table are  $t_{df,0.025/k}$  where  $k$  is the number of tests.

df	NUMBER OF TESTS									
	1	2	3	4	5	6	8	10	15	20
1	12.706	25.452	38.185	50.923	63.657	76.384	101.856	127.321	190.946	254.647
2	4.303	6.205	7.648	8.860	9.925	10.885	12.590	14.089	17.275	19.963
3	3.182	4.177	4.857	5.392	5.841	6.231	6.895	7.453	8.575	9.465
4	2.776	3.495	3.961	4.315	4.604	4.851	5.261	5.598	6.254	6.758
5	2.571	3.163	3.534	3.810	4.032	4.219	4.526	4.773	5.247	5.604
6	2.447	2.969	3.287	3.521	3.707	3.863	4.115	4.317	4.698	4.981
7	2.365	2.841	3.128	3.335	3.499	3.636	3.855	4.029	4.355	4.595
8	2.306	2.752	3.016	3.206	3.355	3.479	3.677	3.833	4.122	4.334
9	2.262	2.685	2.933	3.111	3.250	3.364	3.547	3.690	3.954	4.146
10	2.228	2.634	2.870	3.038	3.169	3.277	3.448	3.581	3.827	4.005
11	2.201	2.593	2.820	2.981	3.106	3.208	3.370	3.497	3.728	3.895
12	2.179	2.560	2.779	2.934	3.055	3.153	3.308	3.428	3.649	3.807
13	2.160	2.533	2.746	2.896	3.012	3.107	3.256	3.372	3.584	3.735
14	2.145	2.510	2.718	2.864	2.977	3.069	3.214	3.326	3.529	3.675
15	2.131	2.490	2.694	2.837	2.947	3.036	3.177	3.286	3.484	3.624
16	2.120	2.473	2.673	2.813	2.921	3.008	3.146	3.252	3.444	3.581
17	2.110	2.458	2.655	2.793	2.898	2.984	3.119	3.222	3.410	3.543
18	2.101	2.445	2.639	2.775	2.878	2.963	3.095	3.197	3.380	3.510
19	2.093	2.433	2.625	2.759	2.861	2.944	3.074	3.174	3.354	3.481
20	2.086	2.423	2.613	2.744	2.845	2.927	3.055	3.153	3.331	3.455
25	2.060	2.385	2.566	2.692	2.787	2.865	2.986	3.078	3.244	3.361
30	2.042	2.360	2.536	2.657	2.750	2.825	2.941	3.030	3.189	3.300
40	2.021	2.329	2.499	2.616	2.704	2.776	2.887	2.971	3.122	3.227
50	2.009	2.311	2.477	2.591	2.678	2.747	2.855	2.937	3.083	3.184
60	2.000	2.299	2.463	2.575	2.660	2.729	2.834	2.915	3.057	3.156
70	1.994	2.291	2.453	2.564	2.648	2.715	2.820	2.899	3.039	3.137
80	1.990	2.284	2.445	2.555	2.639	2.705	2.809	2.887	3.026	3.122
100	1.984	2.276	2.435	2.544	2.626	2.692	2.793	2.871	3.007	3.102
140	1.977	2.266	2.423	2.530	2.611	2.676	2.776	2.852	2.986	3.079
1000	1.962	2.245	2.398	2.502	2.581	2.643	2.740	2.813	2.942	3.031
$\infty$	1.960	2.241	2.394	2.498	2.576	2.638	2.734	2.807	2.935	3.023

# ANSWERS TO SELECTED EXERCISES

## Chapter 1

**1.2.3** The acupuncturist expects acupuncture to work better than aspirin, so she or he is apt to “see” more improvement in someone given acupuncture than in someone given aspirin—even if the two groups are truly equivalent to each other in their response to treatment.

**1.3.1 (a)** Cluster sampling. The three clinics are the three clusters.

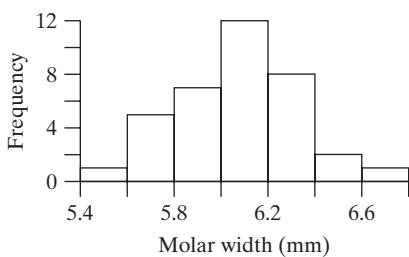
**1.3.2 (a)** The sample is nonrandom and likely nonrepresentative of the general population because it consists of (1) volunteers from (2) nightclubs. (i) The social anxiety level of people who attend nightclubs is likely lower than the social anxiety level of the general public. (ii) A better sampling strategy would be to recruit subjects from across the population.

## Chapter 2

**2.1.2 (a)** (i) Height and weight (ii) Continuous variables (iii) A child (iv) 37   **(b)** (i) Blood type and cholesterol level (ii) Blood type is categorical, cholesterol level is continuous (iii) A person (iv) 129

**2.2.1 (a)** There is no single correct answer. One possibility is

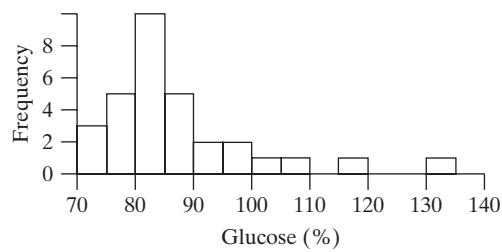
Molar width	Frequency (no. specimens)
[5.4, 5.6)	1
[5.6, 5.8)	5
[5.8, 6.0)	7
[6.0, 6.2)	12
[6.2, 6.4)	8
[6.4, 6.6)	2
[6.6, 6.8)	1
Total	36



**(b)** The distribution is fairly symmetric.

**2.2.7** There is no single correct answer. One possibility is

Glucose (%)	Frequency (no. of dogs)
[70, 75)	3
[75, 80)	5
[80, 85)	10
[85, 90)	5
[90, 95)	2
[95, 100)	2
[100, 105)	1
[105, 110)	1
[110, 115)	0
[115, 120)	1
[120, 125)	0
[125, 130)	0
[130, 135)	1
Total	31



**2.3.1** Any sample with  $\sum y_i = 100$  would be a correct answer. For example, 18, 19, 20, 21, 22.

**2.3.5**  $\bar{y} = 293.8 \text{ mg/dl}$ ; median = 283 mg/dl

**2.3.6**  $\bar{y} = 309 \text{ mg/dl}$ ; median = 292 mg/dl

**2.3.11** Median = 10.5 piglets

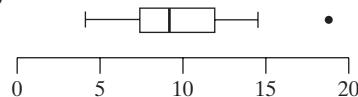
**2.3.13** Mean  $\approx$  median  $\approx$  50

**2.3.14** 15%

**2.4.2 (a)** Median = 9.2,  $Q_1 = 7.4$ ,  $Q_3 = 11.9$

**(b)** IQR = 4.5   **(c)** Upper fence = 18.65

**(d)**



**2.6.1 (a)**  $s = 2.45$    **(b)**  $s = 3.32$

**2.6.4 (a)**  $\bar{y} = 33.10 \text{ lb}$ ;  $s = 3.444 \text{ lb}$

**(b)** Coeff. of var. = 10.4%

**2.6.9** (a)  $32.23 \pm 8.07$  contains  $10/15$  or  $67\%$  of the observations. (b)  $16.09$  to  $48.37$  contains  $15/15$  or  $100\%$  of the observations.

**2.6.14** 4%

**2.6.15**  $\bar{y} = 45; s = 12$

**2.7.1** Mean =  $37.3$ ; SD =  $12.9$

**2.S.13** (a) Median =  $38$  (b)  $Q_1 = 36, Q_3 = 41$  (d)  $66.4\%$

## Chapter 3

**3.2.1** (a)  $0.51$  (b)  $0.94$  (c)  $0.46$  (d)  $0.54$

**3.2.5** (a)  $0.107$

**3.2.6** (a)  $0.916$

**3.3.1** (a)  $0.185$  (b)  $0.117$

(c) No;  $\text{Pr}\{\text{Smoke}\} \neq \text{Pr}\{\text{Smoke}|\text{High income}\}$

**3.4.3** (a)  $0.62$

**3.5.5**  $0.9$

**3.5.6**  $0.794$

**3.6.6** (a)  $0.3746$  (b)  $0.0688$  (c)  $0.1254$

**3.6.9** (a)  $0.75^6 = 0.1780$  (b)  $1 - 0.1780 = 0.8220$

**3.7.1** expected frequencies:  $939.5; 5,982.5; 15,873.1; 22,461.8; 17,879.3; 7590.2; 1,342.6$

**3.S.3**  $0.3369$

**3.S.7** (a)  $1 - 0.99^{100} = 0.6340$

(b)  $1 - 0.99^n \geq 0.95$ , so  $n \geq \log(0.05)/\log(0.99)$ , so  $n \geq 299$

**3.S.10** (a)  $0.0209$

## Chapter 4

**4.3.3** (a)  $84.13\%$  (b)  $61.47\%$  (c)  $77.34\%$  (d)  $22.66\%$

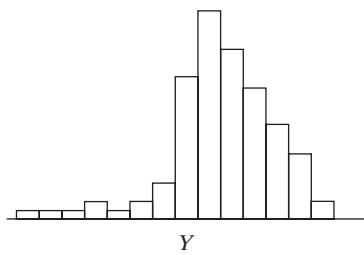
(e)  $20.38\%$  (f)  $20.38\%$

**4.3.4** (a)  $22.66\%$  (b)  $20.38\%$

**4.3.8** (a)  $90.7$  lb (b)  $85.3$  lb

**4.3.12** (a)  $98.76\%$  (b)  $98.76\%$  (c)  $1.24\%$

**4.4.3** (b)



**(c)** Independence of the observations would be questionable, because birthweights of the members of a twin pair might be dependent.

## Chapter 7

*Remark concerning tests of hypotheses* The answer to a hypothesis testing exercise includes verbal statements of the hypotheses and a verbal statement of the conclusion from the test in the context of the problem. In phrasing these statements, we have tried to capture the essence of the biological question being addressed; nevertheless the statements are necessarily oversimplified and they gloss over many issues that in reality might be quite important. For instance, the hypotheses and conclusion may refer to a causal connection between treatment and response; in reality the validity of such a causal interpretation usually depends on a number of factors related to the design of the investigation (such as unbiased allocation of animals to treatment groups) and to the specific experimental procedures (such as the accuracy of assays or measurement techniques). In short, the student should be aware that the verbal statements are intended to clarify the *statistical* concepts; their *biological* content may be open to question.

### 7.1.2 (b) 9

**7.2.1 (a)**  $t_s = -3.13$  so  $0.02 < P\text{-value} < 0.04$

**(b)**  $t_s = 1.25$  so  $0.20 < P < 0.40$

**(c)**  $t_s = 4.62$  so  $P < 0.001$

**7.2.3 (a)** yes **(b)** no **(c)** yes **(d)** no

**7.2.7 (a)**  $H_0$ : Mean serotonin concentration is the same in heart patients and in controls ( $\mu_1 = \mu_2$ );  $H_A$ : Mean serotonin concentration is not the same in heart patients and in controls ( $\mu_1 \neq \mu_2$ ).  $t_s = -1.38$ .  $H_0$  is not rejected. There is insufficient evidence ( $0.10 < P < 0.20$ ) to conclude that serotonin levels are different in heart patients than in controls.

**7.2.11**  $H_0$ : Flooding has no effect on ATP ( $\mu_1 = \mu_2$ );  $H_A$ : Flooding has some effect on ATP ( $\mu_1 \neq \mu_2$ ).  $t_s = -3.92$ .  $H_0$  is rejected.

### 7.3.4 Type II

**7.3.6** Yes; because zero is outside of the confidence interval, we know that the  $P$ -value is less than 0.05, so we reject the hypothesis that  $\mu_1 - \mu_2 = 0$ .

**7.4.1** People with respiratory problems move to Arizona (because the dry air is good for them).

**7.4.4 (a)** Coffee consumption rate

**(b)** Coronary heart disease (present or absent)

**(c)** Subjects (i.e., the 1,040 persons)

**7.5.1 (a)**  $0.10 < P < 0.20$  **(b)**  $0.03 < P < 0.04$

**7.5.3 (a)** yes **(b)** yes **(c)** yes **(d)** no

**7.5.9**  $H_0$ : Wounding the plant has no effect on larval growth ( $\mu_1 = \mu_2$ );  $H_A$ : Wounding the plant tends to diminish larval growth ( $\mu_1 < \mu_2$ ), where 1 denotes wounded and 2 denotes control.  $t_s = -2.69$ .  $H_0$  is rejected. There is sufficient evidence ( $0.005 < P < 0.01$ ) to conclude that wounding the plant tends to diminish larval growth.

**7.5.10 (a)**  $H_0$ : The drug has no effect on pain ( $\mu_1 = \mu_2$ );  $H_A$ : The drug increases pain relief ( $\mu_1 > \mu_2$ ).

$t_s = 1.81$ .  $H_0$  is rejected. There is sufficient evidence ( $0.03 < P < 0.04$ ) to conclude that the drug is effective.

**(b)** The  $P$ -value would be between 0.06 and 0.08. At  $\alpha = 0.05$  we would not reject  $H_0$ .

**7.6.4** No, according to the confidence interval the data do not indicate whether the true difference is “important.”

**7.6.6** 0.33

**7.7.1 (a)** 23 **(b)** 11

**7.7.4 (a)** 71 **(b)** 101 **(c)** 58

**7.7.6** 0.5

**7.10.1 (a)**  $P > 0.149$  **(b)**  $P = 0.048$

**(c)**  $P = 0.0025$

**7.10.3 (a)**  $H_0$ : Toluene has no effect on dopamine in rat striatum;  $H_A$ : Toluene has some effect on dopamine in rat striatum.  $U_s = 32$ .  $H_0$  is rejected. There is sufficient evidence ( $0.015 < P < 0.041$ ) to conclude that toluene increases dopamine in rat striatum.

**7.5.2**  $H_0$ : Mean platelet calcium is the same in people with high blood pressure as in people with normal blood pressure ( $\mu_1 = \mu_2$ );  $H_A$ : Mean platelet calcium is different in people with high blood pressure than in people with normal blood pressure ( $\mu_1 \neq \mu_2$ ).

$t_s = 11.2$ .  $H_0$  is rejected. There is sufficient evidence ( $P < 0.0001$ ) to conclude that platelet calcium is higher in people with high blood pressure.

**7.5.4** No; the  $t$  test is valid because the sample sizes are rather large.

**7.5.8**  $H_0$ : Stress has no effect on growth;  $H_A$ : Stress tends to retard growth.  $U_s = 148.5$ .  $H_0$  is rejected. There is sufficient evidence ( $P < 0.0021$ ) to conclude that stress tends to retard growth.

**7.5.21** False: Zero is in the confidence interval.

## Chapter 8

### 8.2.1 (a) 0.34

**8.2.3**  $H_0$ : Progesterone has no effect on cAMP ( $\mu_1 = \mu_2$ );  $H_A$ : Progesterone has some effect on cAMP ( $\mu_1 \neq \mu_2$ ).  $t_s = 3.4$ .  $H_0$  is rejected. There is sufficient evidence ( $0.04 < P\text{-value} < 0.05$ ) to conclude that progesterone decreases cAMP under these conditions.

**8.2.6 (a)**  $-0.50 < \mu_1 - \mu_2 < 0.74^\circ\text{C}$ , where 1 denotes treated and 2 denotes control.

**8.4.1 (a)**  $P > 0.20$  **(b)**  $P = 0.180$

**(c)**  $P = 0.039$  **(d)**  $P = 0.004$

**8.4.4**  $H_0$ : Weight of the cerebral cortex is not affected by environment ( $p = 0.5$ );  $H_A$ : Environmental enrichment increases cortex weight ( $p > 0.5$ ).  $B_s = 10$ .  $H_0$  is rejected. There is sufficient evidence ( $P = 0.0195$ ) to conclude that environmental enrichment increases cortex weight.

**8.4.8** 0.000061

**8.4.11**  $n = 6$ ;  $P$ -value = 0.03125

**8.5.1 (a)**  $P > 0.20$    **(b)**  $P = 0.078$ **(c)**  $P = 0.047$    **(d)**  $P = 0.016$ 

**8.5.3**  $H_0$ : Hunger rating is not affected by treatment (mCPP vs. placebo);  $H_A$ : Treatment does affect hunger rating.  $W_s = 27$  and  $n_D = 8$ .  $H_0$  is not rejected. There is insufficient evidence ( $P > 0.20$ ) to conclude that treatment has an effect.

**8.6.4** No. “Accurate” prediction would mean that the individual differences ( $d$ 's) are small. To judge whether this is the case, one would need the individual values of the  $d$ 's; using these, one could see whether most of the magnitudes ( $|d|$ 's) are small.

**8.8.8**  $H_0$ : The average number of species is the same in pools as in riffles ( $\mu_1 = \mu_2$ );  $H_A$ : The average numbers of species in pools and in riffles differ ( $\mu_1 \neq \mu_2$ ).  $t_s = 4.58$ .  $H_0$  is rejected. There is sufficient evidence ( $P < 0.001$ ) to conclude that the average number of species in pools is greater than in riffles.

**8.8.12**  $H_0$ : Caffeine has no effect on RER ( $\mu_1 = \mu_2$ );  $H_A$ : Caffeine has some effect on RER ( $\mu_1 \neq \mu_2$ ).  $t_s = 3.94$ .  $H_0$  is rejected. There is sufficient evidence ( $0.001 < P < 0.01$ ) to conclude that caffeine tends to decrease RER under these conditions.

## Chapter 9

**9.1.2 (a)** 0.250   **(b)** 0.441

**(c)** No; the fewest mutants possible is zero, in which case  $\tilde{p}$  is  $2/7$ .

**9.1.4 (a)** 0.2501   **(b)** 0.0352

**9.1.5 (a)** (i) 0.3164, (ii) 0.4219, (iii) 0.2109, (iv) 0.0469, (v) 0.0039

**9.1.9** 0.5259**9.2.2 (a)** 0.040   **(b)** 0.020**9.2.3 (a)** (0.134, 0.290)   **(b)** (0.164, 0.242)**9.2.5 (a)** (0.164, 0.250)

**(b)** We are 95% confident that the probability of adverse reaction in infants who receive their first injection of vaccine is between 0.164 and 0.250.

**9.2.7**  $n \geq 146$ **9.3.4** (0.646, 0.838)

**9.4.1**  $H_0$ : The population ratio is 12:3:1 ( $\Pr[\text{white}] = 0.75$ ,  $\Pr[\text{yellow}] = 0.1875$ ,  $\Pr[\text{green}] = 0.0625$ );  $H_A$ : The ratio is not 12:3:1.  $\chi^2_s = 0.69$ .  $H_0$  is not rejected. There is little or no evidence ( $P > 0.20$ ) that the model is not correct; the data are consistent with the model.

**9.4.2**  $H_0$  and  $H_A$  as in Exercise 9.4.1.  $\chi^2_s = 6.9$ .  $H_0$  is rejected. There is sufficient evidence ( $0.02 < P < 0.05$ ) to conclude that the model is incorrect; the data are not consistent with the model.

**9.4.8**  $H_0$ : The drug does not cause tumors ( $\Pr[T] = \frac{1}{3}$ );  $H_A$ : The drug causes tumors ( $\Pr[T] > \frac{1}{3}$ ), where  $T$  denotes the event that a tumor occurs first in the treated rat.  $\chi^2_s = 6.4$ .

$H_0$  is rejected. There is sufficient evidence ( $0.005 < P < 0.01$ ) to conclude that the drug does cause tumors.

**9.8.2 (a)** 0.2111   **(b)** 0.5700**9.8.3** (0.707, 0.853)

**9.8.14 (a)**  $H_0$ : Directional choice is random ( $\Pr[\text{toward}] = 0.25$ ,  $\Pr[\text{away}] = 0.25$ ,  $\Pr[\text{right}] = 0.25$ ,  $\Pr[\text{left}] = 0.25$ );  $H_A$ : Directional choice is not random.  $\chi^2_s = 4.88$ .  $H_0$  is not rejected. There is insufficient evidence ( $0.10 < P < 0.20$ ) to conclude that the directional choice is not random.

**9.8.16**  $H_0$ : The probability of an egg being on a particular type of bean is 0.25 for all four types of beans;  $H_A$ :  $H_0$  is false.  $\chi^2_s = 2.23$ .  $H_0$  is not rejected. There is insufficient evidence ( $P > 0.20$ ) to conclude that cowpea weevils prefer one type of bean over the others.

## Chapter 10

**10.2.3 (a)**

5	20
10	40

**(b)**  $\hat{p}_1 = 5/15 = 1/3$  and  $\hat{p}_2 = 20/60 = 1/3$ ; yes

**10.2.5**  $H_0$ : Mites do not induce resistance to wilt ( $p_1 = p_2$ );  $H_A$ : Mites do induce resistance to wilt ( $p_1 < p_2$ ), where  $p$  denotes the probability of wilt and 1 denotes mites and 2 denotes no mites.  $\chi^2_S = 7.21$ .  $H_0$  is rejected. There is sufficient evidence ( $0.0005 < P\text{-value} < 0.005$ ) to conclude that mites do induce resistance to wilt.

**10.2.10**  $H_0$ : The two timings are equally effective ( $p_1 = p_2$ );  $H_A$ : The two timings are not equally effective ( $p_1 \neq p_2$ ).  $\chi^2_S = 4.48$ .  $H_0$  is rejected. There is sufficient evidence ( $0.02 < P\text{-value} < 0.05$ ) to conclude that the simultaneous timing is superior to the sequential timing.

**10.2.13**  $H_0$ : Ancrod and placebo are equally effective ( $p_1 = p_2$ );  $H_A$ : Ancrod and placebo are not equally effective ( $p_1 \neq p_2$ ).  $\chi^2_S = 3.82$ . We do not reject  $H_0$ ; there is insufficient evidence ( $0.05 < P\text{-value} < 0.10$ ) to conclude that the treatments differ.

**10.3.3 (a)**  $\hat{\Pr}[D|S] = 0.239$ ,  $\hat{\Pr}[D|WW] = 0.305$ ,  $\hat{\Pr}[S|D] = 0.439$ ,  $\hat{\Pr}[S|A] = 0.522$ .

**(b)**  $H_0$ : There is no association between treatment and survival ( $\Pr[D|S] = \Pr[D|WW]$ );  $H_A$ : There is some association between treatment method (surgery versus watchful waiting) and survival ( $\Pr[D|S] \neq \Pr[D|WW]$ ).  $H_0$  is rejected. There is insufficient evidence ( $0.05 < P\text{-value} < 0.10$ ) to conclude that the survival rates differ for the two treatments.

**10.3.4 (a)**  $\hat{\Pr}[RF|RH] = 0.934$    **(b)**  $\hat{\Pr}[RF|LH] = 0.511$

**(c)**  $\chi^2_S = 398$    **(d)**  $\chi^2_S = 1,623$

**10.4.1**

5	1	6	0
9	15	8	16

**10.5.3 (a)**  $H_0$ : The blood type distributions are the same for ulcer patients and controls ( $\Pr\{O|UP\} = \Pr\{O|C\}$ ,  $\Pr\{A|UP\} = \Pr\{A|C\}$ ,  $\Pr\{B|UP\} = \Pr\{B|C\}$ ,  $\Pr\{B|UP\} = \Pr\{AB|C\}$ );  $H_A$ : The blood type distributions are not the same.  $H_0$  is rejected. There is sufficient evidence ( $P\text{-value} < 0.0001$ ,  $df = 3$ ) to conclude that the blood type distribution of ulcer patients is different from that of controls.

**10.5.5 (a)**  $H_0$ : Change in ADAS-Cog score is independent of treatment;  $H_A$ : Change in ADAS-Cog score is related to treatment.  $\chi^2_S = 10.26$ ,  $df = 4$ .  $H_0$  is rejected. There is sufficient evidence ( $0.02 < P\text{-value} < 0.05$ ) to conclude that Egb and placebo are not equally effective.

**10.6.2** This analysis is not appropriate because the observational units (mice) are nested within the units (litters) that were randomly allocated to treatments. This hierarchical structure casts doubt on the condition that the observations on the 224 mice are independent, especially in light of the investigator's comment that the response varied considerably from litter to litter.

**10.7.3**  $0.001 < p_1 - p_2 < 0.230$ . No; the confidence interval suggests that bed rest may actually be harmful.

**10.7.5 (a)**  $0.067 < p_1 - p_2 < 0.118$

**(b)** We are 95% confident that the proportion of persons with type O blood among ulcer patients is higher than the proportion of persons with type O blood among healthy individuals by between 0.067 and 0.118. That is, we are 95% confident that  $p_1$  exceeds  $p_2$  by between 0.067 and 0.118.

**10.8.1**  $H_0$ : There is no association between oral contraceptive use and stroke ( $p = 0.5$ );  $H_A$ : There is an association between oral contraceptive use and stroke ( $p \neq 0.5$ ), where  $p$  denotes the probability that a discordant pair will be Yes(case)/No(control).  $\chi^2_S = 6.72$ .  $H_0$  is rejected. There is sufficient evidence ( $0.001 < P\text{-value} < 0.01$ ) to conclude that stroke victims are more likely to be oral contraceptive users ( $p > 0.5$ ).

**10.9.1 (a) (i)** 1.339 **(ii)** 1.356 **(b) (i)** 1.314 **(ii)** 1.355

**10.9.7 (a)** 1.241 **(b)** (1.036, 1.488)

**(c)** We are 95% confident that taking heparin increases the odds of a negative response by a factor of between 1.036 and 1.488 when compared to taking enoxaparin.

**10.8.3 (a)**  $H_0$ : Sex ratio is 1:1 in warm environment ( $p_1 = 0.5$ );  $H_A$ : Sex ratio is not 1:1 in warm environment ( $p_1 \neq 0.5$ ), where  $p_1$  denotes the probability of a female in the warm environment.  $\chi^2_S = 0.18$ .  $H_0$  is not rejected. There is insufficient evidence ( $P\text{-value} > 0.20$ ) to conclude that the sex ratio is not 1:1 in warm environment.

**(c)**  $H_0$ : Sex ratio is the same in the two environments ( $p_1 = p_2$ );  $H_A$ : Sex ratio is not the same in the two environ-

ments ( $p_1 \neq p_2$ ), where  $p$  denotes the probability of a female and 1 and 2 denote the warm and cold environments.  $\chi^2_S = 4.20$ .  $H_0$  is rejected. There is sufficient evidence ( $0.02 < P\text{-value} < 0.05$ ) to conclude that the probability of a female is higher in the cold than the warm environment.

**10.S.12**  $H_0$ : Site of capture and site of recapture are independent ( $\Pr\{RI|CI\} = \Pr\{RI|CII\}$ );  $H_A$ : Flies preferentially return to their site of capture ( $\Pr\{RI|CI\} > \Pr\{RI|CII\}$ ), where  $C$  and  $R$  denote capture and recapture and I and II denote the sites.  $H_0$  is rejected. There is sufficient evidence ( $0.0005 < P\text{-value} < 0.005$ ) to conclude that flies preferentially return to their site of capture.

**10.S.14 (a)** 1.709 **(b)**  $1.55 < \theta < 1.89$

**(c)** The odds ratio gives the (estimated) odds of survival for men compared to women. This ratio (of 1.709) is a good approximation to the relative risk of death for women compared to men (which is 1.658), because death is fairly rare.

## Chapter 11

**11.2.1 (a)** SS(between) = 228, SS(within) = 120

**(b)** SS(total) = 348

**(c)** MS(between) = 114, MS(within) = 15,  $s_{\text{pooled}} = 3.87$

**11.2.4 (a)**

Source	df	SS	MS
Between groups	3	135	45
Within groups	12	337	28.08
Total	15	472	

**(b)** 4 **(c)** 16

**11.4.2 (a)**  $H_0$ : The stress conditions all produce the same mean lymphocyte concentration ( $\mu_1 = \mu_2 = \mu_3 = \mu_4$ );  $H_A$ : Some of the stress conditions produce different mean lymphocyte concentrations (the  $\mu$ 's are not all equal).  $F_s = 3.84$ .  $H_0$  is rejected. There is sufficient evidence ( $0.01 < P\text{-value} < 0.02$ ) to conclude that some of the stress conditions produce different mean lymphocyte concentrations.

**(b)**  $s_{\text{pooled}} = 2.78 \text{ cells/ml} \times 10^{-6}$

**11.4.3 (a)**  $H_0$ : Mean HBE is the same in all three populations ( $\mu_1 = \mu_2 = \mu_3$ );  $H_A$ : Mean HBE is not the same in all three populations (the  $\mu$ 's are not all equal).  $F_s = 0.58$ .  $H_0$  is not rejected. There is insufficient evidence ( $P\text{-value} > 0.20$ ) to conclude that mean HBE is not the same in all three populations.

**(d)**  $s_{\text{pooled}} = 14.4 \text{ pg/ml}$

**11.6.2** There is no single correct answer. One possibility is as follows:

Treatment	Piglet				
	Litter 1	Litter 2	Litter 3	Litter 4	Litter 5
1	2	5	2	4	5
2	1	4	1	1	2
3	4	2	5	2	4
4	5	3	3	3	3
5	3	1	4	5	1

**11.6.5** Plan II is better. We want units within a block to be similar to each other; plan II achieves this. Under plan I the effect of rain could be confounded with the effect of a variety.

**11.7.2 (a)**

Source	df	SS	MS
Between species	1	2.19781	2.19781
Between flooding levels	1	2.25751	2.25751
Interaction	1	0.097656	0.097656
Within groups	12	0.47438	0.03953
Total	15	5.027356	

**(b)**  $F_s = 0.097656/0.03953 = 2.47$ . With  $df = 1$  and 12, Table 10 gives  $F_{.20} = 1.84$  and  $F_{.10} = 3.18$ . Thus,  $0.10 < P\text{-value} < 0.20$  and we do not reject  $H_0$ . There is insufficient evidence ( $P\text{-value} > 0.10$ ) to conclude that there is an interaction present.

**(c)**  $F_s = 2.19781/0.03953 = 55.60$ . With  $df = 1$  and 12, Table 10 gives  $F_{.0001} = 32.43$ . Thus,  $P\text{-value} < 0.0001$  and  $H_0$  is rejected. There is strong evidence ( $P\text{-value} < 0.0001$ ) to conclude that species affects ATP concentration.

**(d)**  $s_{\text{pooled}} = \sqrt{0.03953} = 0.199$ .

$$\mathbf{11.7.4 (a)} F_s = \frac{31.33/1}{30648.81/(223 - 4)} = 31.33/139.95 = 0.22.$$

With  $df = 1$  and 140, Table 10 gives  $F_{.20} = 1.66$ . Thus,  $P\text{-value} > 0.20$  and we do not reject  $H_0$ . There is insufficient evidence ( $P\text{-value} > 0.20$ ) to conclude that there is an interaction present.

**11.8.2 (a)** 123 mm Hg    **(b)** 123.2 mm Hg

**(d)** 0.851 mm Hg

**11.8.7**  $0.67 < \mu_E - \mu_S < 1.48$  gm, where  $\mu_E = \frac{1}{2}(\mu_{E,\text{Low}} + \mu_{E,\text{High}})$  and  $\mu_S = \frac{1}{2}(\mu_{S,\text{Low}} + \mu_{S,\text{High}})$ .

**11.8.8 (b)**  $L = 3.685 \text{ nmol}/10^8 \text{ platelets/hour}$ ;

$\text{SE}_L = 1.048 \text{ nmol}/10^8 \text{ platelets/hour}$

**11.9.1** The following hypotheses are rejected:  $H_0: \mu_C = \mu_D$ ;  $H_0: \mu_A = \mu_D$ ;  $H_0: \mu_B = \mu_D$ ;  $H_0: \mu_C = \mu_E$ ;  $H_0: \mu_A = \mu_E$ ;  $H_0: \mu_B = \mu_E$ ;  $H_0: \mu_B = \mu_C$ ;  $H_0: \mu_A = \mu_C$ . The following hypotheses are not rejected:  $H_0: \mu_A = \mu_B$ ;  $H_0: \mu_D = \mu_E$ . Summary:

C A B E D

There is sufficient evidence to conclude that treatments D and E give the largest means, treatments A and B the next largest, and treatment C the smallest. There is insufficient evidence to conclude that treatments A and B give different means or that treatments D and E give different means.

**11.9.2** The following hypotheses are not rejected:  $H_0: \mu_A = \mu_B$ ;  $H_0: \mu_B = \mu_D$ ;  $H_0: \mu_B = \mu_E$ ;  $H_0: \mu_D = \mu_E$ . Summary:

C A B E D

**11.9.4 (a)** Yes, each of diets B, C, and D differs from A, as none of the intervals includes zero.

**11.S.1**  $H_0$ : The three classes produce the same mean change in fat-free mass ( $\mu_1 = \mu_2 = \mu_3$ );  $H_A$ : At least one class produces a different mean (the  $\mu$ 's are not all equal).  $F_s = 0.64$ . We do not reject  $H_0$ . There is insufficient evidence ( $P\text{-value} > 0.20$ ) to conclude that the population means differ.

**11.S.3**  $H_0$ : The mean refractive error is the same in the four populations ( $\mu_1 = \mu_2 = \mu_3 = \mu_4$ );  $H_A$ : Some of the populations have different mean refractive errors (the  $\mu$ 's are not all equal).  $F_s = 3.56$ .  $H_0$  is rejected. There is sufficient evidence ( $0.01 < P\text{-value} < 0.02$ ) to conclude that some of the populations have different mean refractive errors.

**11.S.13** Let 1, 2, 3, and 4 denote placebo; probucol; multivitamins; and probucol and multivitamins.

**(a)**  $\bar{y}_2 - \bar{y}_1 = 1.79 - 1.43 = 0.36$

**(b)**  $\bar{y}_4 - \bar{y}_3 = 1.54 - 1.40 = 0.14$

**(c)** The contrast that measures the interaction between probucol and multivitamins is “the difference in differences” from parts (a) and (b):

$$(\bar{y}_2 - \bar{y}_1) - (\bar{y}_4 - \bar{y}_3) = 0.36 - 0.14 = 0.22$$

(Note: This is not the only correct answer; reversing the signs in (a) and (b), or in (c), is also correct.)

## Chapter 12

**12.2.1 (d), (a), (b), (c), (e)** (The correlations are  $-0.97$ ,  $-0.63$ ,  $0.10$ ,  $0.58$ , and  $0.93$ .)

**12.2.2 (b)**  $r = 0.439$

**12.2.3**  $H_0$ : There is no correlation between blood urea and uric acid concentration ( $\rho = 0$ );  $H_A$ : Blood urea and uric acid concentration are positively correlated ( $\rho > 0$ ).  $t_s = 3.952$ .  $H_0$  is rejected. There is strong evidence ( $P\text{-value} < 0.0005$ ) to conclude that blood urea and uric acid concentration are positively correlated.

**12.2.5 (a)**  $H_0$ : There is no correlation between plant density and mean cob weight ( $\rho = 0$ );  $H_A$ : Plant density and mean cob weight are correlated ( $\rho \neq 0$ ).  $t_s = -11.9$ .  $H_0$  is rejected. There is strong evidence ( $P\text{-value} < 0.001$ ) to conclude that plant density and mean cob weight are negatively correlated.

**(b)** Observational study

**(c)** No; this is an observational study in which plant density was observed but not manipulated. The study suggests that density manipulation is worth exploring in a follow-up experiment.

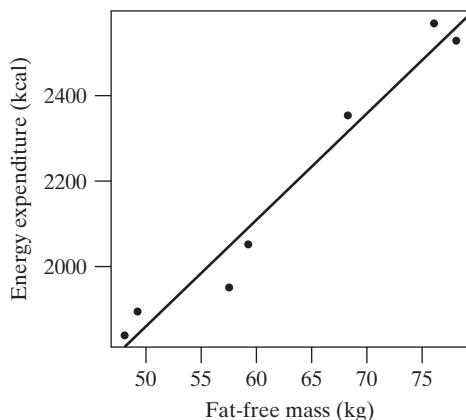
**12.3.1 (b)** Leucine =  $-0.05 + 0.02928 \times \text{Time}$ ; the slope is  $0.02928 \text{ ng/min}$

**(d)**  $s_e = 0.0839$ .

**12.3.2 (c)**  $\hat{y} = -0.592 + 7.641x$ ;  $s_e = 0.881^\circ\text{C}$

**12.3.5 (a)**  $\hat{y} = 607.7 + 25.01x$

**(b)**



**(c)** As fat-free mass goes up by 1 kg, energy expenditure goes up by 25.01 kcal, on average.

**(d)**  $s_e = 64.85$  kcal

**12.3.8 (b)**  $r^2 = 0.107 = 10.7\%$    **(f)**  $12/17 = 71\%$

**12.4.5** Estimated mean = 21.1 mm; estimated SD = 1.3 mm

**12.4.9** Estimated mean = 658.1 l/min; estimated SD = 115.16 l/min

**12.5.1 (a)**  $0.0252 < \beta_1 < 0.0334$  ng/min

**(b)** We are 95% confident that the rate at which leucine is incorporated into protein in the population of all *Xenopus* oocytes is between 0.0252 ng/min and 0.0334 ng/min.

**12.5.5 (a)**  $19.4 < \beta_1 < 30.6$  kcal/kg

**(b)**  $20.6 < \beta_1 < 29.4$  kcal/kg

**12.5.7**  $H_0$ : There is no linear relationship between respiration rate and altitude of origin ( $\beta_1 = 0$ );  $H_A$ : Trees from

higher altitudes tend to have higher respiration rates ( $\beta_1 > 0$ ).  $t_s = 6.06$ .  $H_0$  is rejected. There is sufficient evidence ( $P\text{-value} < 0.0005$ ) to conclude that trees from higher altitudes tend to have higher respiration rates.

**12.6.6 (a) – (iii)**   **(b) – (i)**   **(c) – (ii)**

**12.7.1 (a)** The dashed lines, which tell us where the true (population) regression line lies.

**12.S.1** 0.24 gm

**12.S.3 (a)** Estimated mean = 0.85 kg; estimated SD = 0.17 kg

**12.S.6 (a)**  $s_e = 0.137$  cm

**(b)**  $H_0: \rho = 0$  or  $H_0: \beta_1 = 0$ .  $t_s = 3.01$ .  $H_0$  is rejected. There is sufficient evidence ( $0.02 < P\text{-value} < 0.04$ ) to conclude that there is a positive correlation between diameter of forage branch and wing length.

## Chapter 13

**13.2.1** A chi-square test of independence would be appropriate. The null hypothesis of interest is  $H_0: p_1 = p_2$ , where  $p_1 = \Pr\{\text{clinically important improvement if given clozapine}\}$  and  $p_2 = \Pr\{\text{clinically important improvement if given haloperidol}\}$ . A confidence interval for  $p_1 - p_2$  would also be relevant.

**13.2.10** A two-sample comparison is called for here, but the data do not support the condition of normality. Thus, the Wilcoxon-Mann-Whitney test is appropriate.

**13.2.12** It would be natural to consider correlation and regression with these data. For example, we could regress  $Y = \text{forearm length}$  on  $X = \text{height}$ ; we could also find the correlation between forearm length and height and test the null hypothesis that the population correlation is zero.

*This page intentionally left blank*

# INDEX

## A

Addition rules, 95–97  
Additive factors, 450  
Additive transformation, 70–71  
“Age-adjusted” mean, 457  
Alanine aminotransferase (ALT), 38  
Alternative hypothesis, 223, 224, 278  
  directional, 251, 382  
Analysis of covariance, 536–38  
Analysis of variance (ANOVA), 415  
  applicability of methods, 433–36  
  basic one-way, 418  
  “between-groups,” 425  
  conditions verification, 433  
  factorial, 449–55  
  fundamental relationship, 423  
  global *F* test, 468  
  graphical perspective, 417–18  
  group effect, 428–29  
  model, 427  
  notation, 421–22  
  null hypothesis, 427  
  one-way, 418–19  
  pooled standard deviation, 420–21  
  population SDs equality, 434–35  
  quantities with formulas, 426  
  standard conditions, 433  
  table, 425  
  two-way, 449–55  
  within-groups, 425  
    variation measure, 420  
Anecdote, 7, 181  
ANOVA, *See* Analysis of variance (ANOVA)  
Anterior commissure (AC), 8  
Arithmetic mean, *See* Mean

## B

Bar chart, 28

distributions visual

  impression, 386

  stacked, 54

Bayesian view, 281

Bias, 20

  nonresponse, 22

  panel, 13

  sampling, 20

  selection, 75

Biased sample, 16

Bimodality, 35

Binomial coefficient, 110–11, 567–68

Binomial distribution, 107–8, 338, 566–67

  application to sampling, 114

  binomial coefficient, 110–11

  fitting to data, 116–18

  formula, 110

  illustration, 108–10

  independent-trials model, 108

  mean and standard

    deviation, 114, 569

  normal approximation, 162, 163

Binomial random variable, 109

Bivariate frequency table, 52

Bivariate random sampling model, 485–86, 520

Blinding, 11

Blocking, 440, *See also*

  Randomized blocks

  design, one-way

  agricultural field study, 440, 441

  randomization procedure, 440

Bonferroni method, 470–71

  advantage, 473

  Bonferroni adjustment, 470

Boxplots, 45, 47, 55

  IQR, 46–47

  modified, 50–51

  quartiles, 45–46

## C

Categorical data:

  chi-square goodness-of-fit

    test, 348–50, 352

  chi-square statistic, 350–51

  chi-square distribution, 352–54

  compound hypotheses, 354

  confidence interval:

    one-sided, 344

    planning study, 345–46

  for population

    proportion, 341–42, 343

  confidence levels, 347

  dichotomous variables:

    directional alternative, 356–57

    directional conclusion, 355

  inference methods

    summary, 359

  univariate summaries, 52

Wilson-adjusted sample

  proportion, 336–37

  dependence on sample

    size, 339–40

  relationship to statistical

    inference, 339

  sampling distribution, 337–39

  standard error (SE), 342

Categorical variable, 26

Central Limit Theorem, 151, 153, 159, 343

  and normal approximation

    to binomial

    distribution, 572

Chance error due to

  sampling, 20

Chance operation, 85, 108, 147

  coin tossing, 85, 86–87

  tossing die, 102

Chi-square ( $\chi^2$ ) distribution, 352

Chi-square goodness-of-fit

  test, 348, 352

  chi-square statistic, 350,

  351

  compound null hypothesis, 354

bar charts, 350

dichotomous variables, 355

  directional alternative, 356–57

  directional conclusion, 355

Chi-square test, 350, 365

  features, 354

  Fisher’s exact test, 381

$r \times k$  contingency table, 387

$2 \times 2$  contingency table, 365–66, 368

Classes, 32

Cluster sample, 18

Coding, 69

Coefficient of

  determination, 501–2

Coefficient of variation, 63

Comparisonwise Type I

  error rate, 465

Compound null hypothesis, 354

Concordant pairs, 398

Conditional distributions, 505, 507

Conditional populations, 505

Confidence interval, 181, 302–3, 459–60, 578–79

  one-sided, 185

population means, 177

  conditions for validity, 194–96

  condition verification, 196–97

  critical value

    determination, 179

    invisible man analogy, 177–78

  student’s *t* distributions, 178–79

    student’s *t* method

      condition, 196

  population means difference, 206

  conditions for validity, 210

  confidence interval construction, 206–10

  degrees of freedom calculation, 206

  population proportions, 341–46

- Confidence interval (*cont.*)  
 95% confidence interval  
     for  $p$ , 342–44  
 other confidence levels,  
     344–45  
 planning a study to  
     estimate  $p$ , 345–46  
 standard error of  $\tilde{p}$ , 342  
 and randomness, 181  
 relationship, 184  
 Wilson interval, 343
- Confounding, 246–47
- Conjugated equine estrogen (CEE), 183
- Contagion, 114–15
- Contingency tables, 364
- Continuity correction,  
 164–66
- Continuous variable, 27
- Contrasts, 457  
 interaction assessment,  
     461–62
- Control groups, 12–13
- Conventional medical therapy (CMT), 381
- Correlation analysis, 480
- Correlation coefficient, 482,  
 484, 542  
 bivariate random sampling  
     model, 485–86  
 confidence interval:  
     for population  
         correlation, 487  
 correlation and causation,  
     488  
 alga reproduction,  
     488–89  
 formula, 484  
 inference, 486  
 interpretation, 485–86  
 population correlation,  
     485  
 sample correlation, 485  
 linear association strength  
     measurement, 482  
 null hypothesis, 486  
 significant, 489
- Creatine phosphokinase (CK), 32
- Curvilinear regression, 535
- D**
- Data analysis, 552, *See also* Exploratory data analysis
- Degrees of freedom (df), 62,  
 178, 181
- denominator, 429  
 numerator, 429  
 within groups (df(within)),  
     422  
 between groups  
     (df(between)), 423
- Density curves, 100  
 continuum paradox, 101  
 interpretation, 100  
 probabilities, 101
- Density function, 124
- Density scale, 100
- Descriptive statistics, 40  
 mean, 41  
 median, 40  
 robustance, 42
- df, *See* Degrees of freedom  
 (df)  
 df(between), 423  
 df(total), 425  
 df(within), 422
- Dichotomous variables, 355
- Directional alternative  
 hypothesis, 251, 356–57,  
     368
- chi-square goodness-of-fit  
 test, 356
- nondirectional alternatives  
 versus, 254–56  
 rules, 257  
 in sign test, 318  
 in Wilcoxon signed-rank  
 test, 323
- in Wilcoxon-Mann-Whitney test, 285–86
- Discordant pairs, 398
- Discrete variable, 27
- Dispersion measures, 59  
 comparison, 66  
 range, 59  
 standard deviation, 60  
 variation coefficient, 63  
 visualization, 63
- Distributions shapes, 35, 36  
 bimodality, 35  
 unimodal, 35
- Distribution-free test, 282
- Dotplots, 30
- Double replacement, 147
- Double-blind experiment,  
 11
- E**
- Effect size, 262–63
- Empirical rule, 65–66
- Error probabilities  
 interpretation, 280
- medical testing analysis,  
 280  
 hypothetical results, 281  
 probability tree, 281
- Expected frequencies, 351  
 in chi-square test, 387  
 in contingency table, 367
- Experiment, 9, 242, 243
- Experimentwise Type I error  
 rate, 465
- Explanatory variable, 242
- Exploratory data analysis,  
 552
- Extracorporeal membrane  
 oxygenation (ECMO),  
 381
- Extrapolation, 509
- F**
- F* distributions, 429  
 parameters, 429  
 shapes, 35–37
- F* test, global, 429  
*F* distributions, 429  
*F* statistic, 430  
 and *t* test, 431
- Factors, 449
- Fences, 49
- Finite population correction  
 factor, 151
- Fisher transformation,  
 487–88
- Fisher's exact test, 381  
 comparison to chi-square  
 test, 383  
 alternative hypothesis,  
     382–83  
 binomial coefficient, 382  
 nondirectional alternatives  
     and, 383–84
- Fisher's LSD, 465  
 experimentwise Type I  
     error rate, 468  
 formula for computation,  
     467–68  
 intermediate  
     computations, 467
- Fitted regression line, 482,  
 492, 542
- determination coefficient,  
 501–2
- equation, 496
- least-squares criterion, 499
- least-squares formulas,  
 580–81
- least-squares regression  
 line, 496
- line of averages, 496
- residual standard  
 deviation, 500
- residual sum of squares,  
 498
- SD Line, 493
- Fitted value, 435
- Five number summary, 47
- Food and Drug Administration (FDA),  
 227
- Forced vital capacity (FVC),  
 456, 459
- Frequency, 28
- Frequency distributions, 28  
 grouped, 32, 33  
 infant mortality, 30  
 linear transformation  
     effect, 69–70  
     tails of, 33
- Frequency interpretation of  
 probability, 86–88
- Frequentist view, 281
- G**
- Gibberellic acid (GA), 552
- Goodness-of-fit test, 350,  
 352
- Grand mean, 419, 420  
 drawback, 458
- Graph of averages, 497
- Grouped frequency  
 distributions, 32
- H**
- Heat shock protein (HSP),  
 557
- Hierarchical structure, 190
- High-level residential care  
 (HLRC), 394
- Histogram, 30, *See also* Bar chart  
 areas interpretation, 34  
 CK distribution, 35  
 relative frequency, 99  
 SD estimation, 65–66
- Historical controls, 13
- Honest Significant Difference (HSD), 472
- Hypothesis:  
 alternative, 223  
 null, 223  
 statistical test, 224  
 testing, 223
- error occurrence situations, 239

- randomization test, 219–21  
 $t$  test, 221, 223  
 Type I error, 238–39, 240  
 Type II error, 239, 240
- I**  
 Incomplete blocks design, 438  
 Indefinitely extended regions area, 570–71  
 Independent samples mean comparison, 414  
 ANOVA, 415  
   two-way, 449  
 experimental designs, 475  
 global approach  
   advantages, 475  
 global  $F$  test, 429  
 linear combinations, 456  
 multiple comparisons, 464  
 nonparametric approaches, 475  
 organic methods treatment efficiency, 414  
 randomized blocks design, 437, 441, 444  
 ranking and selection, 476  
 $t$  test limitations, 416  
   multiple comparisons problem, 416–17  
   standard deviation estimation, 417  
   structure in groups, 417  
 Independent-trials model, 108  
 Indicator variable, 532  
 Inference, 543  
   conditions, 519–20  
   correlation, 486  
   for proportions, 348  
   statistical, 73  
 Inference methods, 550  
   flowchart, 551, 552  
 Influential point, 518  
   effect in correlation coefficient, 519  
 Interaction, 451, 462  
 Interaction graph, 451  
 Interpolation, 509  
 Interpretation of density, 100  
 Interpretation of the definition of  $s$ , 61–63  
 Interquartile range (IQR), 46, 59, 63, 66  
 Intersection, 95
- IQR, *See* Interquartile range (IQR)
- J**  
 Jowett, Geoff, 177
- L**  
 Lactate dehydrogenase (LD), 261  
 Least significant difference (LSD), 465  
 Least-squares, *See* Fitted regression line  
 Least-squares criterion, 499, 535  
 Least-squares formula, 580–81  
 Least-squares regression line, 496  
 Levels, factor, 453  
 Leverage points, 518  
 Linear combinations, 457  
   for adjustment, 457  
   “age-adjusted” mean, 458  
 confidence intervals, 459  
 contrasts, 458  
   to assess interaction, 461  
   chromosomal aberrations, 462  
   standard error (SE), 458  
 $t$  tests, 460  
 Linear model, 506, 532  
   estimation, 508  
   interpolation in, 509  
   prediction and, 510  
 Linear regression and  
   correlation analysis, 480–549  
   analysis of covariance, 536–38  
   bivariate random sampling model, 485  
   coefficient of determination, 501–2  
   correlation coefficient, 482–89  
   confidence interval for  $\rho$ , 487–88  
   defined, 482  
   formulas, 542  
   *significant*, use of term, 489  
   statistical inference concerning correlation, 511–15  
 examples of, 482, 485–87, 488–89  
 fitted regression line, 492–502  
 equation of the regression line, 496  
 formulas, 542  
 least-squares criterion, 499  
 least-squares line, 499  
 regression line, 496–97  
 residual standard deviation, 500–01  
 residual sum of squares, 498–99  
 inference formulas, 543  
 interpretation guidelines, 516–25  
 conditions for inference, 519–20  
 correlation and causation, 488  
 design conditions, 513, 519  
 inadequate descriptions of data set, 516–19  
 linear model and normality condition, 522  
 parameter conditions, 520  
 population distribution conditions, 520  
 residual plots, 522–23  
 sampling conditions, 519–21  
 transformations, use of, 524–25  
 linear model, 505–10  
 conditional distributions, 505  
 conditional populations, 505  
 constancy of standard deviation, 506  
 defined, 506–8  
 estimation of, 508  
 graph of averages, 496–97  
 interpolation in, 509–10  
 linearity, 506  
   and prediction, 510  
 random subsampling model, 508  
 logistic regression, 538–42  
 nonparametric and robust regression and correlation, 536
- regression and the  $t$  test, 531–35  
 statistical inference  
   concerning  $\beta_1$ , 511–15  
   confidence interval for  $\beta_1$ , 513  
   standard error of  $b_1$ , 511–13  
   testing the hypothesis, 513–15  
   summary of formulas, 542–43
- Linear transformations, 68  
   coding, 69  
   effect, 70  
   frequency distribution, 69–70
- Logistic regression, 538, 539  
 Logistic response function, 540
- Low-level residential care (LLRC), 394
- M**  
 Main effect, 451  
 Mann-Whitney test,  
   *See* Wilcoxon-Mann-Whitney test  
 Matched-pair designs, 310  
 m-chlorophenylpiperazine (mCPP), 303  
 McNemar’s test, 399  
   chi-square distribution, 399  
   HIV transmission to children analysis, 399–400  
 Mean, 41, 103–4  
   deviations, 42  
   median versus, 43  
 Mean comparisons, 199  
   notation, 199–200  
   observational studies, 246–48  
 pooled standard error, 203  
   standard deviation (SD), 204  
   vital capacity calculation, 203–4  
 standard error (SE):  
   tonsillectomy experiment, 202–3  
   of two sample means difference, 200
- Mean square between groups (MS(between)), 422

- Mean square within groups (MS(within)), 422
- Mean squares for blocks (MS(blocks)), 443, 444
- Measurement error, 123
- Measures of dispersion, 59–67
- coefficient of variation, 63
  - comparison of, 66–67
  - interpretation of the definition of  $s$ , 61–63
  - range, 59–60
  - standard deviation (SD), 60–61
  - estimating from a histogram, 65–66
  - visualizing, 63–65
- Median, 40, 42
- distribution, 45–46
  - mean versus, 43–44
  - sample, 78
  - visualization, 43
- Meta-study, 146
- sampling distribution visualization, 150
  - for  $t$  test, 237
- Missing data, 23
- Mode, 33
- Modified boxplot, 50–51
- Monoamine oxidase (MAO), 4, 174, 431
- MS(between), 422
- MS(blocks), 443, 444
- MS(within), 422
- Multiple comparisons, 464, 475
- Bonferroni method, 470
- conditions for validity, 472–73
  - experimentwise versus comparisonwise error, 465
- Fisher's LSD, 465
- problem, 416
  - Tukey's HSD, 472
- Multiple regression and correlation, 535
- Multiplication rules, 97–98
- Multiplicative transformation, 69
- Myocardial blood flow (MBF), 299
- N**
- Nondirectional alternative, 250
- Nonlinear transformations, 71–72
- Nonnormal data, transformations for, 552
- Nonparametric methods, 552
- Nonparametric test, 282
- Nonresponse bias, 22
- Nonsampling error, 22, *See also* Sampling error nonresponse bias, 22
- Nonsimple random sampling methods, 18
- stratified random sample, 18
- Normal approximation to the binomial distribution, 162–66
- Normal curve, 121, 124
- areas, 125
  - determination, 127–29
  - standardized scale, 125–27
  - density function, 124
  - inverse reading, 129–31
  - with mean and SD, 124
- Normal distribution, 121
- measurement error, 123
  - normal probability plot, 134–36, 137
- Normality assessment, 132
- decision making, 136–38
  - normal probability plots, 134
  - functionality, 134–35
  - transformation for nonnormal data, 138–39
- Null distribution, 277, 278
- of chi-square distribution, 352
  - for sign test, 318
  - test statistic, 316
  - Wilcoxon-Mann-Whitney, 287
- Null hypothesis, 223, 369, *See also* Alternative hypothesis global, 417–18
- Numeric variable, 27
- O**
- Observational studies, 8, 242, 243, 310
- confounding, 246–47
  - difficulties, 244–45
  - experimental studies versus, 243–44
- spurious association, 247–48
- Observational units, 27
- nested, 190–91
  - notation, 27
- Observed frequency, 351
- Odds ratio, 402, 403
- advantage, 403–5
  - case-control design, 405–6
  - confidence interval, 406–7
  - standard error (SE), 407, 408
- One-sided confidence intervals, 344
- One-tailed  $t$  tests, 250, 256
- directional alternative hypotheses, 251
  - nondirectional alternatives versus, 254–56
  - rule, 256–57
  - test procedure, 251–52
  - $P$ -value, 252
- Ordinal variable, 26
- Outliers, 48, 518
- lower fence, 49
  - radish growth in light, 49–50
  - upper fence, 49
- P**
- Paired design, 299, 310
- data analysis, 314
  - examples of, 310–12
  - experiments with unit pairs, 310
  - limitations, 326–30
  - purposes of pairing, 312–13
  - randomized, completely randomized design versus, 313–14
  - repeated measurements, 311
- Paired samples comparisons: analyzing differences, 300–301
- confidence interval, 302–3
- dotplot of differences, 304
  - parallel dotplots, 305
  - standard error (SE), 304, 305
- ignoring pairing result, 303
- student's  $t$  analysis:
- conditions for validity, 306
  - formulas, 307
- Panel bias, 13
- Parameter, 40*fn*, 76, 79
- Placebo, 10
- Pooled standard deviation, 420–21
- df(within), 422
  - MS(between), 422
  - MS(within), 422
  - SS(within), 422
- Pooled variance, 203, 421–22, 532
- Population, 15, 75
- categorical variable, 76–77
  - correlation, 485
  - description, 76
  - mean, 78
  - parameter, 76
  - SD, 78
  - tobacco leaves, 78
- Population distributions, 195
- conditional, 505
  - conditions, 196
  - of differences, 306
  - distributed variable, 152
  - mean, 155
  - standard deviation, 155
- Population mean estimation, 187
- standard error (SE), 188
- Positron emission tomography (PET), 299*fn*
- Power, 240–41, 267
- calculation, 574–75
  - dependence on  $(\mu_1 - \mu_2)$ , 268
  - normal distributions, 269
  - planning study, 269
  - dependence on  $n$ , 268
  - dependence on  $\alpha$ , 267
  - dependence on  $\sigma$ , 267–68
- Precision in prediction, 527
- confidence and prediction intervals, 528–29
  - intervals computation, 529
- Prediction, 543
- Probability, 84
- chance operation, 85
  - coin tossing experiment, 85
  - combination, 90
  - conditional, 365–66
  - frequency interpretation, 86
  - rules, 94
  - addition, 95–97
  - basic, 94–95
  - multiplication, 97–98

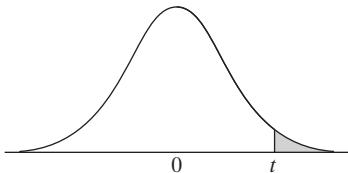
- Probability tree, 88  
*P*-value, 226, 227
- Q**  
Quartiles, 45
- R**  
 $r \times k$  contingency table, 385  
chi-square test, 387  
conditions for validity, 391  
expected frequencies, 387  
power considerations, 394  
design conditions verification, 392  
contexts, 388–89  
Random cluster sample, 18  
Random sample, 20, 145  
selection procedure, 17  
simple, 16  
stratified, 19  
Random sampling, 15  
biased sample, 16  
employing randomness, 17  
nonsimple methods, 18  
population, 15  
practical concerns, 18  
random sample selection, 17  
samples, 15, 16  
sampling bias, 20  
sampling error, 20  
simple random sample, 16  
Random subsampling model, 486, 508, 520  
Random variable, 102  
binomial, 107–8, 109  
distribution formula, 100  
mean, 114  
standard deviation (SD), 114  
continuous, 103  
discrete, 103  
mean, 103–4  
variance, 104–5  
rules, 105–6  
Randomization distribution, 248–49  
Randomization test, 218–21, 289  
Randomized blocks  
ANOVA model, 441  
visualizing block effects, 442  
Randomized blocks design, one-way, 437  
randomized complete block *F* test, 444  
within-subject blocking, 439  
df(blocks), 445  
mean squares between blocks, 444  
SS(blocks), 445  
Range, 59–60  
Regression and correlation: analysis of covariance, 536  
curvilinear relationship with  $X$ , 517  
inadequate description causes, 516–17  
inference conditions, 519–20  
interpretation, 516  
least squares extensions, 535  
linear model and normality condition, 522  
logistic regression, 538  
nonparametric and robust, 536  
residual plots, 522  
sampling conditions guidelines, 520  
*t* test, 531  
transformations use, 524  
 $X$  and  $Y$  labeling, 522  
Regression line, 57  
Regression parametric interpretation:  
conditional distributions, 505  
conditional populations, 505  
linear model, 506  
interpolation in, 509  
prediction and, 510  
random subsampling model, 508  
Relative frequency, 31  
cumulative, 87  
histogram, 99  
stacked, 54  
Relative risk, 401–08  
Research hypothesis, *See* Alternative hypothesis  
Residual, 434, 436, 442  
plots, 522, 523, 526, 527  
standard deviation, 500  
Residual sum of squares (SS(resid)), 498  
Response variable, 242, 437*fn*  
Robustance, 42–43
- S**  
Sample correlation, 485  
Sample mean, 41, 149  
sampling distribution, 149–50, 151, 156  
Central Limit Theorem, 153  
dependence on sample size, 153–54  
shape, 151  
standard deviation, 150, 151, 155  
Sample space, 95  
Samples, 15  
Sampling bias, 20  
Sampling distribution, 145, 147, *See also* Sampling variability  
and data analysis, 212–13  
relationship to statistical inference, 148, 149  
sample mean, 149–50, 152  
sample proportion, 337–38  
Sampling error, 20, 145, *See also* Nonsampling error  
magnitude, 396  
Sampling frame, 17  
Sampling variability, 145, 147, *See also* Probability aspects, 156  
meta-study, 146  
Satterthwaite's method, 206*fn*  
Scatterplot, 56  
Score confidence interval, 578  
SD, *See* Standard deviation (SD)  
SD line, 493–95  
SE, *See* Standard error (SE)  
Shape characteristics, 35  
Shapiro–Wilk test, 139–40  
Side-by-side boxplots, 55  
Sign test, 315, 325, *See also* *t* test  
applicability, 319–20  
bracketing *P*-value, 318  
critical value calculation, 319  
directional alternative, 318  
null distribution, 318–19  
critical values, 317  
finding *P*-value, 316, 317  
survival times, 316  
treatment of zeros, 318  
Significance level, 227  
Significant difference, 261, 489  
Significant digits, 573  
Simple random sample, 16  
Skewed to the right, 33  
Skewness:  
effect, 43  
moderate, 274  
Soil respiration, 282–83, 284–85  
Spurious association:  
SS(between), 423  
SS(resid), 498  
SS(total), 424  
SS(within), 422  
Stacked bar charts, 54  
Stacked relative frequency, 54  
Standard deviation (SD), 60, 172  
empirical rule, 65  
estimation from histogram, 65  
interpretation, 61  
visualization, 64  
Standard error (SE), 171–72  
linear combination, 458–59  
groups of people, 176  
regression parameter, 511  
structure, 512  
standard deviation (SD) versus, 172  
Wilson-adjusted sample proportion, 342  
Standard error of the mean, 172  
Standard normal, 125  
Standardized scale, 125  
Statistic(s), 1, 26, 40, 76, 79  
chi-square, 350–51, 366–67  
computer, 7  
descriptive, 40–44  
*t* statistic, 224–25  
Statistical estimation, 170  
mean, 170, 171  
notation, 171  
standard deviation, 170, 171  
Statistical inference, 73, 170  
concerning  $\beta_1$ , 511  
confidence interval, 513  
implications for design, 513  
null hypothesis formulation, 513–14  
standard error (SE), 511  
population, 75

- Statistical significance  
interpretation:  
confidence intervals, 263–65  
effect size, 262–63  
significant difference  
versus important  
difference, 260–62
- Strata, 19  
Stratified random sample, 18  
Student's *t* distribution,  
178–79  
conditions, 273  
conditions verification,  
273–74  
inappropriate use  
consequences, 274  
*t* test mechanics summary,  
276–77  
Studentized range  
distribution, 472  
Sum of squares between  
groups (SS(between)),  
423  
Sum of squares within  
groups (SS(within)),  
422
- T**  
*t* test, 221, 223, 460, *See also*  
Sign test  
alternative hypothesis, 223,  
278  
conditions, 273–274  
meta-study for, 237  
null hypothesis, 223–24,  
278  
power, 240–41  
*P*-value, 226, 277–78, 279  
conservative, 229–30  
determination, 229  
drawing conclusions,  
227–29  
interpretation, 236–38,  
280–81  
significance level versus,  
238  
two-tailed, 226  
reporting results, 230–31  
*t* statistic, 224–26  
test and confidence  
interval relationship,  
234–35
- Test for association, *See* Chi-square test  
Test of hypothesis, 224  
Test of independence, *See*  
Chi-square test  
Test statistic, 224–25  
Therapeutic touch (TT),  
558–59  
Total degrees of freedom  
(df(total)), 425  
Total sum of squares  
(SS(total)), 424  
Transformations, 524  
effect of, 68–72  
linear, 68–71  
coding, 69  
effect on frequency  
distribution, 69–71  
multiplicative, 69  
nonlinear, 71–72  
Tukey's Honest Significant  
Difference (HSD), 472  
 $2 \times 2$  contingency tables,  
364, *See also*  $r \times k$   
contingency table  
chi-square statistic, 366  
expected frequencies, 367  
observed frequencies, 366  
chi-square test, 365  
conditional probability,  
365  
confidence interval,  
395–97  
null hypothesis, 365–66  
relationship to test, 397  
computational notes, 369  
contexts, 373  
facts about rows and  
columns, 376–77  
independence and  
association, 373  
odds ratio, 402, 403–8  
paired data, 398  
HIV transmission,  
398–99  
McNemar's test, 399–400  
relative risk, 401, 402  
test procedure, 367–68  
verbal description of  
association, 377
- Two sample *t*-test, 554  
Two-tailed *t* test, 250
- Type I error, 239, 416, 475  
consequences analysis,  
239–40  
risk, 281  
Type I error rate,  
experimentwise, 465,  
472  
Type II error, 239  
consequences analysis,  
239–40  
probability, 240  
risk, 281
- U**  
Unimodality, 35  
Univariate summary, 52
- V**  
Variable(s), 26  
categorical, 26  
continuous, 27  
dichotomous, 355  
discrete, 27  
notation, 27  
numeric, 27  
ordinal, 26  
random, 102–3  
relationships:  
categorical–categorical  
relationships, 52  
numeric–categorical  
relationships, 55  
numeric–numeric  
relationships, 56  
transformation effect, 68  
additive transformation,  
70–71  
linear transformations, 68  
multiplicative  
transformation, 69
- Variance:  
model analysis, 427–28  
one-way analysis, 418–19  
random variable, 104
- Variation sources, 38  
serum ALT, 38  
Venn diagram, 95
- W**  
Wald confidence interval,  
578  
Welch's method, 206/*fn*
- Wilcoxon signed-rank test,  
321–22, *See also*  
Wilcoxon-Mann-Whitney test  
applicability, 324–25  
bracketing *P*-value, 323  
directional alternative, 323  
absolute value  
calculation, 322  
critical values, 323  
signed ranks, 323  
treatment of ties, 324  
treatment of zeros, 324
- Wilcoxon-Mann-Whitney  
test, 274, 282, 576–77,  
*See also* Wilcoxon  
signed-rank test  
applicability, 283–84  
conditions, 288  
data arrays, 286  
directional alternative,  
285–86  
directionality, 285  
null distributions, 287  
*P*-values, 288  
randomization test versus,  
289  
rationale, 286  
statement of  $H_0$  and  $H_A$ ,  
282–83  
statistic calculations, 284–85  
*t* test versus, 288–89
- Wilson-adjusted sample  
proportion, 336–37  
confidence interval, 341–43  
one-sided, 344  
confidence levels, 347  
dependence on sample  
size, 339–40  
planning study:  
to estimate *p*, 345  
in ignorance, 345–46  
relationship to statistical  
inference, 339  
sampling distribution,  
337–39  
standard error (SE) for, 342
- Wilson confidence interval,  
578–579
- X**  
 $X^2$  distribution, 352

# INDEX OF EXAMPLES

- A**борт funding, 22  
Acne, treatment of, 329  
Adenosine triphosphate (ATP), and flooding, 3  
Agricultural field study, 439, 440, 441  
Alanine aminotransferase (ALT), 38  
Albinism, 108, 109  
Alcohol and MOPEG, 75–76  
Alfalfa and acid rain, 437–38, 441, 445, 446  
Alga, reproduction of, 488–89  
Amphetamine and food consumption, 480–81, 497, 502, 505–6, 507, 510  
Anthrax, vaccine for, 2  
Arsenic in Rice, 481–82, 493–94, 495, 496, 499, 500, 501, 502, 509, 529  
Aspirin, and heart attacks, 408  
Asthma, bronchial, 10  
Autism, 10
- B**acteria and cancer, 2  
Bacterial growth, 147  
Beef steers growth, 520  
Biofeedback and blood pressure, 326–27  
Birthweight and smoking, 246–47  
Blocking by litter, 438  
Blocking in an agricultural field study, 439, 440, 441  
Blood flow, 299, 301, 302–3  
Blood glucose, 99, 100, 101  
Blood pressure, 46, 59 and biofeedback, 326–27 and platelet calcium, 486–87, 488, 514–15, 534–35 and serum cholesterol, 536 Blood type, 74, 75, 94–95, 97, 113, 114  
Body size and energy expenditure, 6  
Body temperature, 69, 70  
Body weight, 261–62, 263, 264  
Bone mineral density, 183
- Brain weight, 37  
Breast cancer, 343–44  
Bronchial asthma, 10  
Butterfly wings, 170–71, 172, 179, 180, 188  
Butterfly thorax weight, 208–10
- C**cancer:  
and bacteria, 2–3  
breast, 343–44  
esophageal, 538–41  
and hair dye, 251  
lung, 77  
and smoking, 310, 401, 402, 403–4, 405, 406, 407  
Canine anatomy, 190–91  
Caterpillar head size, 536–37  
Cattle, daily gain, 64  
Cats, mutants, 108, 111  
Cell firing times, 37  
Chemotherapy and THC, 320  
Chickenpox, 114–115  
Chromosomal aberrations, 462  
Chromosome puffs, 557–58  
Cigarette Smoking, 243–44  
Chrysanthemum growth, 60–61, 62, 63  
Clofibrate, 12  
Coin tossing, 85, 86, 89–90, 97  
Color:  
of hair and eye, 95, 96, 97, 373, 374, 377–78, 388–89  
of poinsettias, 28, 31–32  
Common cold, 12–13  
Contaminated soda, 336, 337, 338, 339–40  
Coronary artery disease, 13–14  
Crabs, sand, 19–20  
Crawfish length, 235–36  
Creatine phosphokinase (CK), 32–33, 35  
Crickets, singing times, 43, 72
- D**aily gain of cattle, 64  
Damselflies, 562  
Deafness and lightning, 7
- Deer habitat and fire, 348, 349–50, 351, 353, 354, 355  
Dice, 102, 104, 105  
Dogs, toxicity in, 9
- E**CMO, 344, 381, 382, 383  
E. Coli watershed contamination, 53  
Eggplant fertilizer, 310, 313–14  
Eggshell thickness, 122, 181  
Energy expenditure and body size 6  
Esophageal cancer, 538–41  
Estrogen and steroids, 561–62  
Exercise and serum triglycerides, 311  
Eye color and hair color, 95, 96, 97, 373, 374, 377–78, 388–89  
Eye facets, 159
- F**alse positives, 93  
Family size, 103  
Fast plants, 206–8, 227–28, 229–30  
Feet to inches, 106  
Fertilizers for eggplants, 310, 313–14  
Fire and deer habitat, 348, 349–50, 351, 353, 354, 355  
Fish, lengths of, 20, 127–28, 130–31  
Fish vertebrae, 103, 104  
Flax seeds, 353  
Flexibility, 218, 219–20  
Flooding and ATP, 3  
Flower pollination, 393  
Flu shots, 384  
Food choice by insect larvae, 5–6, 392  
Forced vital capacity (FVC), 456–57, 459  
Fruitflies, sampling, 85, 87–88, 91, 112  
Fungus resistance in corn, 21
- G**ermination of spores, 191–93
- H**air color and eye color, 95, 96, 97, 373, 374, 377–78, 388–89  
Hair dye and cancer, 251  
Hand size, 98  
Harvest Moon Festival, 356–57  
Headache pain, 248  
migraine, 363–64, 365, 366, 367, 368, 369, 396–97  
Heart attacks and aspirin, 408  
Health and marriage, 9  
Height and weight of girls, 63  
of young men, 506, 507–8, 522  
Heights:  
of men, 103  
of people, 268, 270  
of students, 33–34  
of women, 135–136  
Hematocrit in males and females, 242  
HIV testing, 22, 364, 370  
HIV transmission to children, 398–399, 400  
Hunger rating, 303–5  
Hyperactivity and sugar, 23
- I**mmunotherapy, 240  
Infant mortality, 30  
Insect larvae, food choice by, 5–6, 392  
Interspike times in nerve cells, 122  
Iron supplements, 452, 453

- K**nee replacement, 147–48
- L**amb birthweights, 172, 173–74
- La Graciosa thistle, 19
- Leaf area, 221–22
- Left-handedness, 345–46, 347
- Length and weight of snakes, 482, 485, 508, 512, 513
- Lengths of fish, 20, 127–28, 130–31
- Lentil growth, 138–39, 140
- Lightning and deafness, 7
- Litter size of sows, 30–31
- Lung cancer, 77 and smoking, 310, 401, 402, 403–4, 405, 406, 407
- M**ammary artery ligation, 10–11
- Mao and schizophrenia, 4, 174–75
- Marijuana and intelligence, 190
- Marijuana and the pituitary, 239
- Marriage and health, 9
- Mass, 106
- Measurement error, 123
- Medical testing, 92, 93, 280
- Medications, 103
- Microfossils, 36
- Migraine headache, 363–64, 365, 366, 367, 368, 369, 396–97
- Moisture content, 133
- Monoamine oxidase (MAO) and schizophrenia, 4, 174–75
- MOPEG and alcohol, 75–76
- Music and marigolds, 237–38, 256–57
- Mutant cats, 108, 111
- N**erve cells: density, 322–23 interspike times in, 121 sizes of, 20
- Neck pain and school bags, 29
- Niacin supplementation, 251, 252–53, 254–55
- Nitric oxide, 91–92
- Nitrite metabolism, 21
- O**at plants, 76
- Ocean temperature, 492–93
- Oysters and seagrass, 465–67, 469, 471
- P**argyline and sucrose consumption, 242–43
- Physiotherapy, 394
- Plant height and disease resistance, 375–76, 377
- Platelet calcium and blood pressure 486–87, 488, 514–15, 534–35
- Plover nesting, 385–86, 387–88
- Poinsettias, color of, 28, 31–32
- Pollination of flowers, 393
- Postpartum weight loss, 270–71
- Pregnancy, smoking during, 342
- Pulse, 46, 49 after exercise, 66
- R**ace and brain size, 245
- Radish growth, 48, 55 in light, 49
- Rat blood pressure, 147
- Reaction time, 160, 161–62
- Reproduction of alga, 488–89
- S**ampling fruitflies, 85, 87–88, 91, 112
- Sand crabs, 19–20
- Schizophrenia and MAO, 4, 174–75
- School bags and neck pain, 29
- Seagrass and oysters, 465–67, 469, 471
- Seastars, 560–61
- Sediment yield, 197
- Seeds per fruit, 183–84, 185
- Serum ALT, 38
- Serum cholesterol, 121, 133, 150–51 and blood pressure, 536 measuring, 328 and serum glucose, 520
- Serum CK, 32–33, 35
- Serum LD, 261, 262–63, 264
- Serum triglycerides and exercise, 311
- Sexes of children, 116–18
- Sexual orientation, 8
- Skin grafts, 315–16, 324–25
- Smoking: and birthweight, 246–47 and lung cancer, 310, 401, 402, 403–4, 405, 406, 407 during pregnancy, 342
- Snakes, length and weight of, 482, 485, 508, 512, 513
- Soda, contaminated, 336, 337, 338, 339–40
- Soil respiration, 282–83, 284–85
- Soil samples, 561
- Sows, litter size of, 30–31
- Soybean growth, 449, 450, 454, 458, 459, 461, 524–25
- Squirrels, 306–7
- Sucrose in beet roots, 21
- Sugar and hyperactivity, 23
- Sweet corn, 414–15, 434–35, 436
- T**amoxifen, 557
- Temperature, 105
- THC and chemotherapy, 320
- Therapeutic touch, 558–60
- Thistle, La Graciosa, 19
- Thorax weight, butterfly, 208–10
- Tissue inflammation, 274–75
- Toads, 454–55
- Tobacco leaves, 78
- Tobacco use prevention, 562
- Toluene and the brain, 223–24, 225, 226, 227, 531–34
- Tonsillectomy, 202–3
- Toxicity in dogs, 9
- Treatment of acne, 329
- Tree diameters, 101
- Twins, 561
- U**lcerative colitis, treatment of, 21–22
- Ultrasound, 247–48
- V**accinations, 561
- Vaccine for anthrax, 2
- Virus growth, 311, 317
- Vital capacity, 200, 202, 203 Forced (FVC), 456–57, 459
- W**atershed contamination, 53
- Weight, 69
- Weight gain of lambs, 40, 41–42, 419–420, 421, 422, 423, 425–426, 428, 430, 434
- Weight of seeds, 38, 152, 154, 155, 156
- Whale Selenium, 56
- Whale swimming speed, 555–56
- Y**ield of tomatoes, 264, 265



## Critical Values of Student's *t* Distribution

df	Upper Tail Probability									
	0.20	0.10	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.888	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850
21	0.859	1.323	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.819
22	0.858	1.321	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.792
23	0.858	1.319	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.768
24	0.857	1.318	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.745
25	0.856	1.316	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.725
26	0.856	1.315	1.706	1.822	1.967	2.056	2.162	2.479	2.779	3.707
27	0.855	1.314	1.703	1.819	1.963	2.052	2.158	2.473	2.771	3.690
28	0.855	1.313	1.701	1.817	1.960	2.048	2.154	2.467	2.763	3.674
29	0.854	1.311	1.699	1.814	1.957	2.045	2.150	2.462	2.756	3.659
30	0.854	1.310	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.646
40	0.851	1.303	1.684	1.796	1.936	2.021	2.123	2.423	2.704	3.551
50	0.849	1.299	1.676	1.787	1.924	2.009	2.109	2.403	2.678	3.496
60	0.848	1.296	1.671	1.781	1.917	2.000	2.099	2.390	2.660	3.460
70	0.847	1.294	1.667	1.776	1.912	1.994	2.093	2.381	2.648	3.435
80	0.846	1.292	1.664	1.773	1.908	1.990	2.088	2.374	2.639	3.416
100	0.845	1.290	1.660	1.769	1.902	1.984	2.081	2.364	2.626	3.390
140	0.844	1.288	1.656	1.763	1.896	1.977	2.073	2.353	2.611	3.361
1000	0.842	1.282	1.646	1.752	1.883	1.962	2.056	2.330	2.581	3.300
$\infty$	0.842	1.282	1.645	1.751	1.881	1.960	2.054	2.326	2.576	3.291